

Mathematical Reasoning in Visual Contexts

Kai-Wei Chang

CS @ UCLA

Scholar @ Amazon AGI

kw@kwchang.net

See <http://kwchang.net> for more information

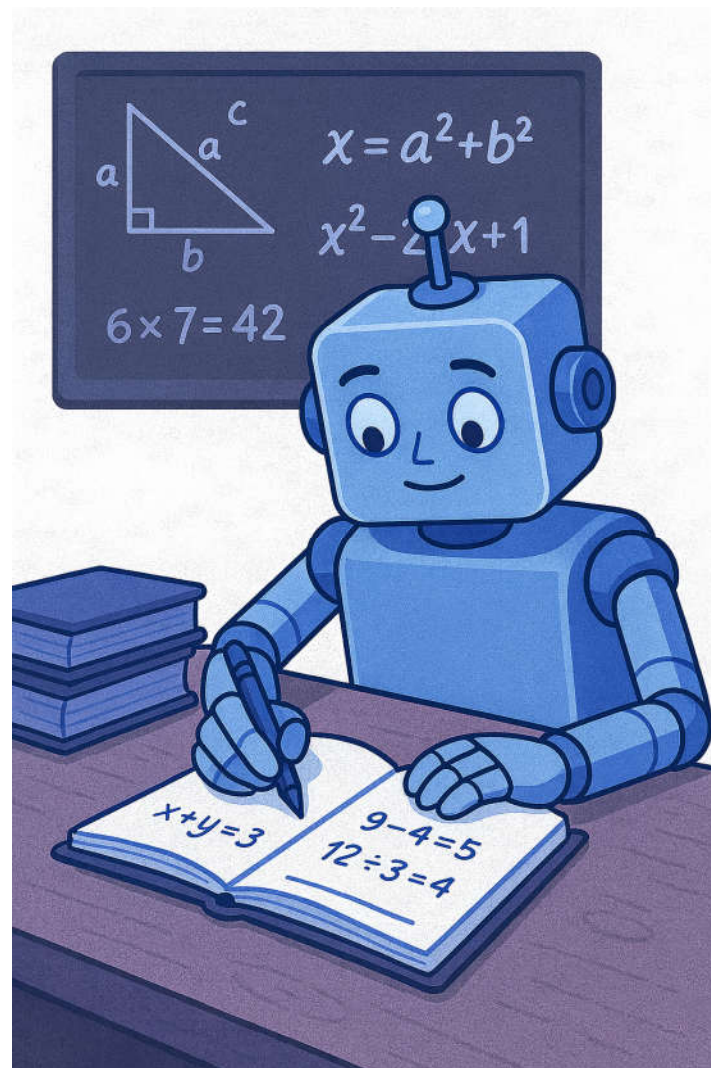


Image generated by ChatGPT

What's your dream AI application?

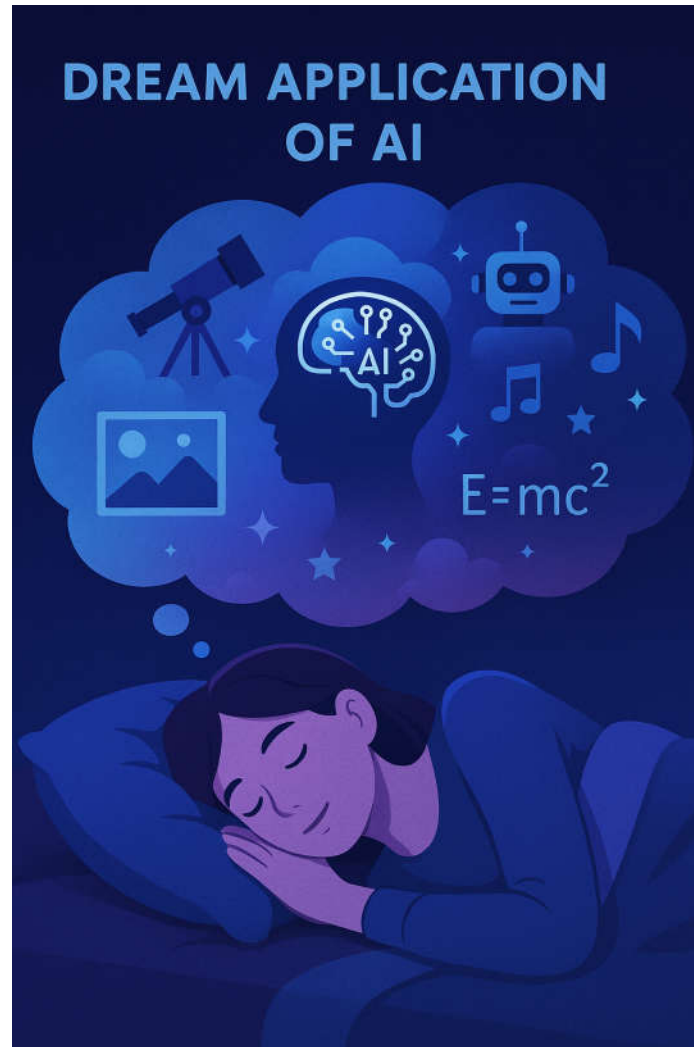
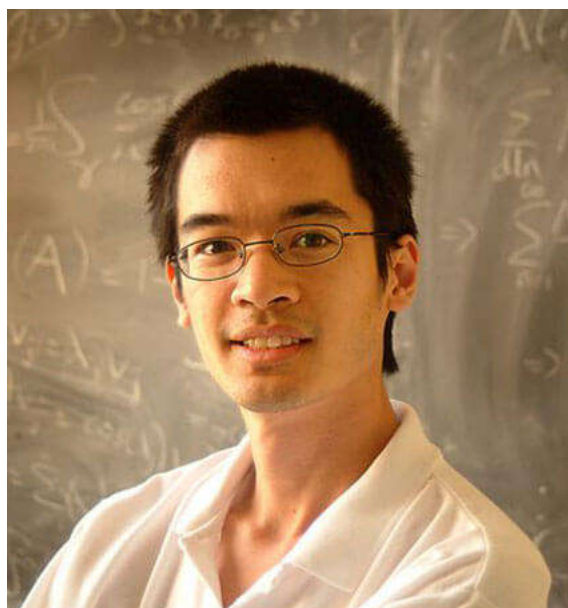


Image generated by ChatGPT

LLMs Assist Mathematicians with Cutting-Edge Research



Terence Tao

AI ANTHOLOGY

Embracing change and resetting expectations

By Terence Tao

Professor of Mathematics at University of California, Los Angeles

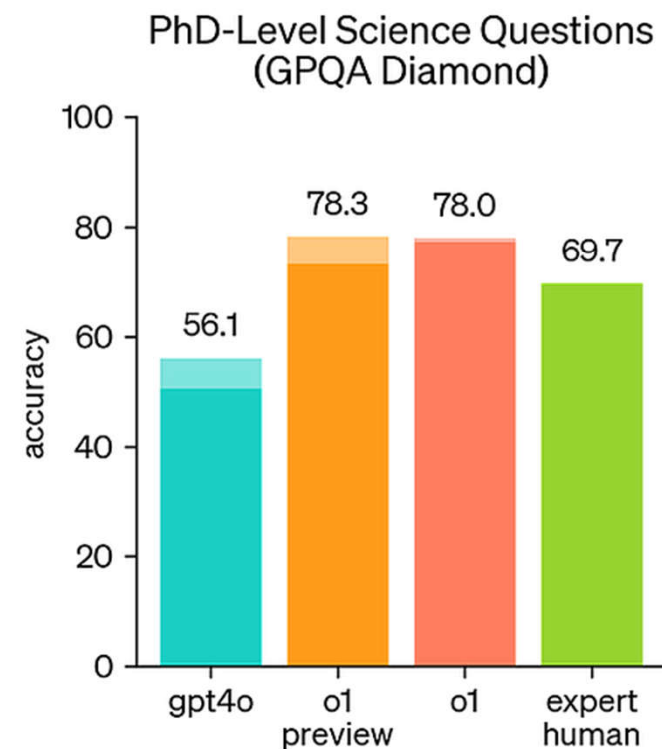
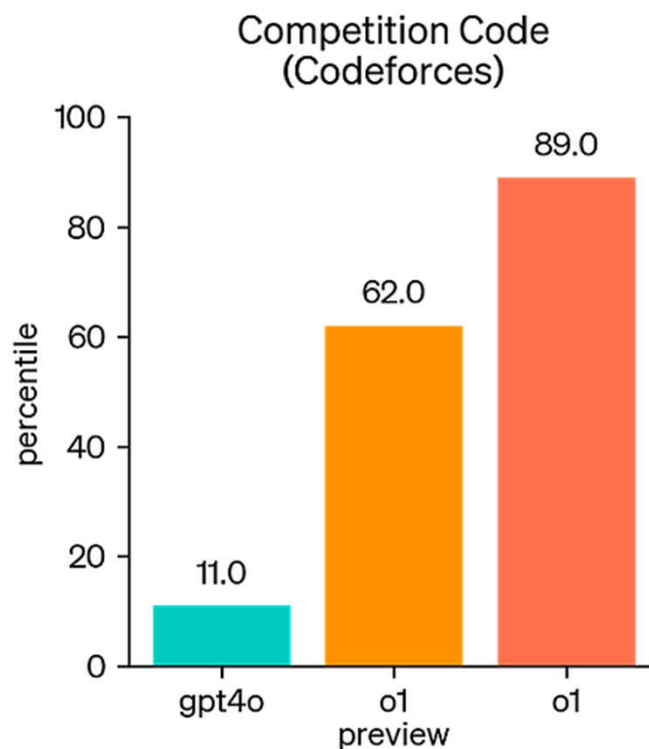
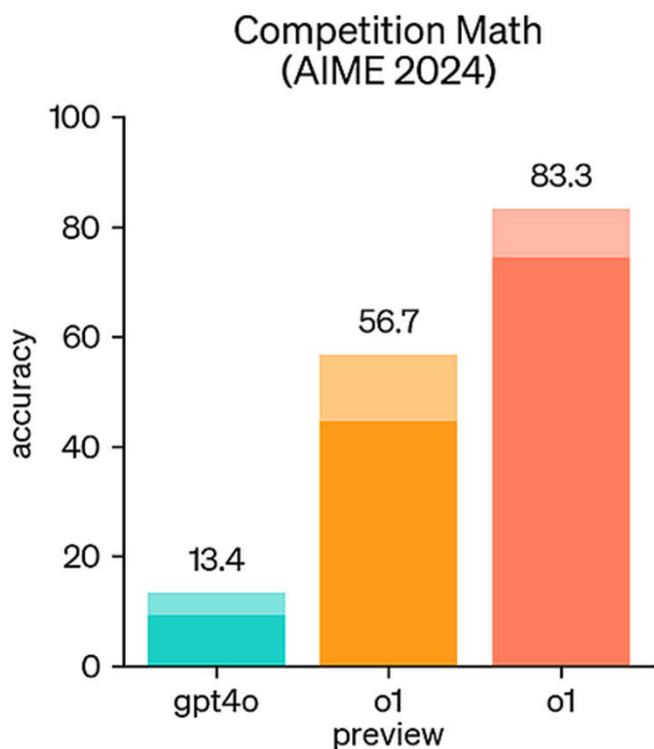


“I expect, say, 2026-level AI, when used properly, will be **a trustworthy co-author in mathematical research**, and in many other fields as well.”

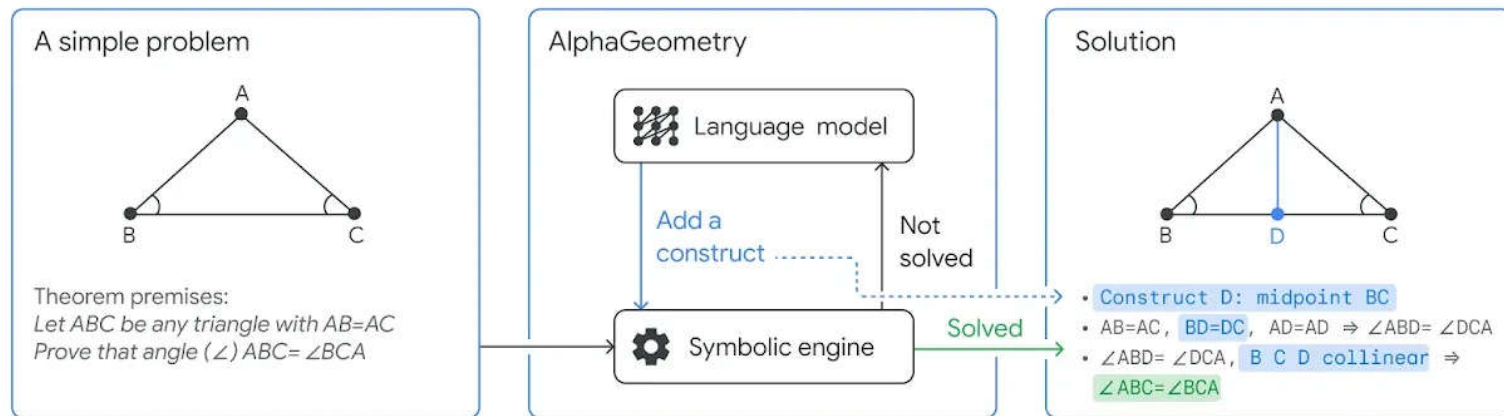
<https://unlocked.microsoft.com/ai-anthology/terence-tao/>

<https://terrytao.wordpress.com/about/ai-generated-versions-of-the-ai-anthology-article/>

OpenAI o1 model with Inference-Time Computation



Language Model Elevates Geometry Engine to (High-School) Olympiad Level



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

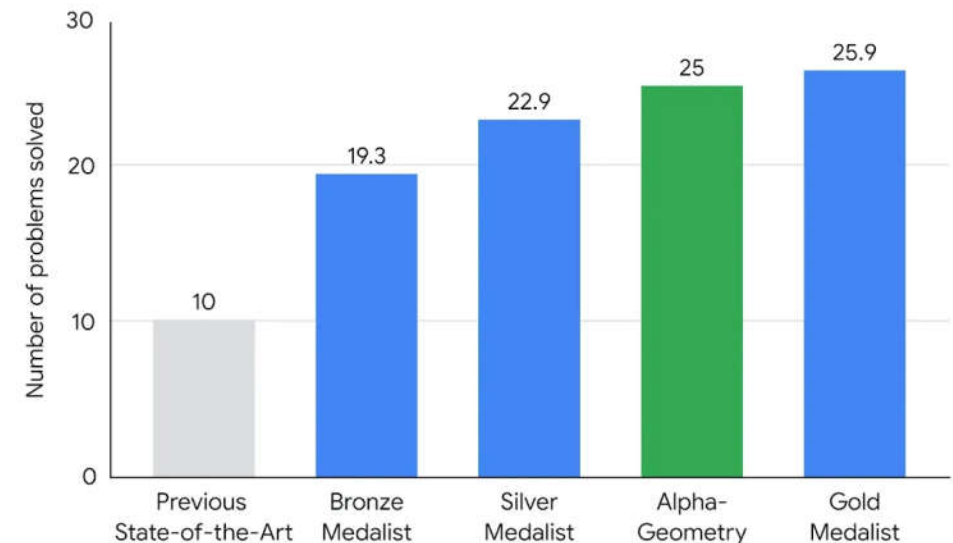
[Trieu H. Trinh](#) , [Yuhuai Wu](#), [Quoc V. Le](#), [He He](#) & [Thang Luong](#) 

[Nature](#) **625**, 476–482 (2024) | [Cite this article](#)

165k Accesses | **941** Altmetric | [Metrics](#)

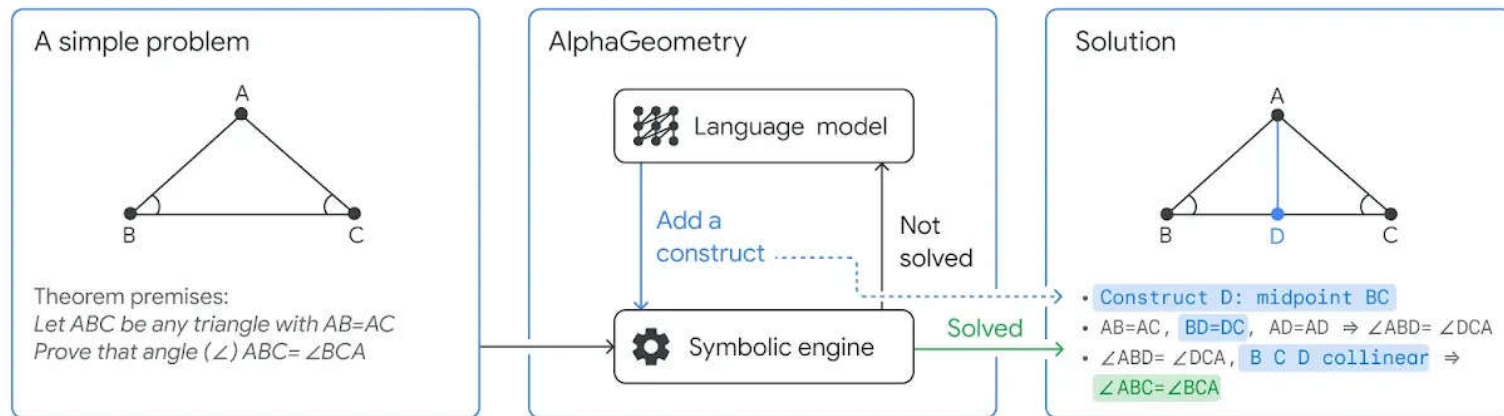


Approaching the Olympiad gold-medalist standard



<https://www.nature.com/articles/s41586-023-06747-5>

Language Model Elevates Geometry Engine to (High-School) Olympiad Level



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](#) ✉

[Nature](#) **625**, 476–482 (2024) | [Cite this article](#)

165k Accesses | **941** Altmetric | [Metrics](#)

UCLA ENGINEERING
Computer Science

Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

Thang Luong and Edward Lockhart

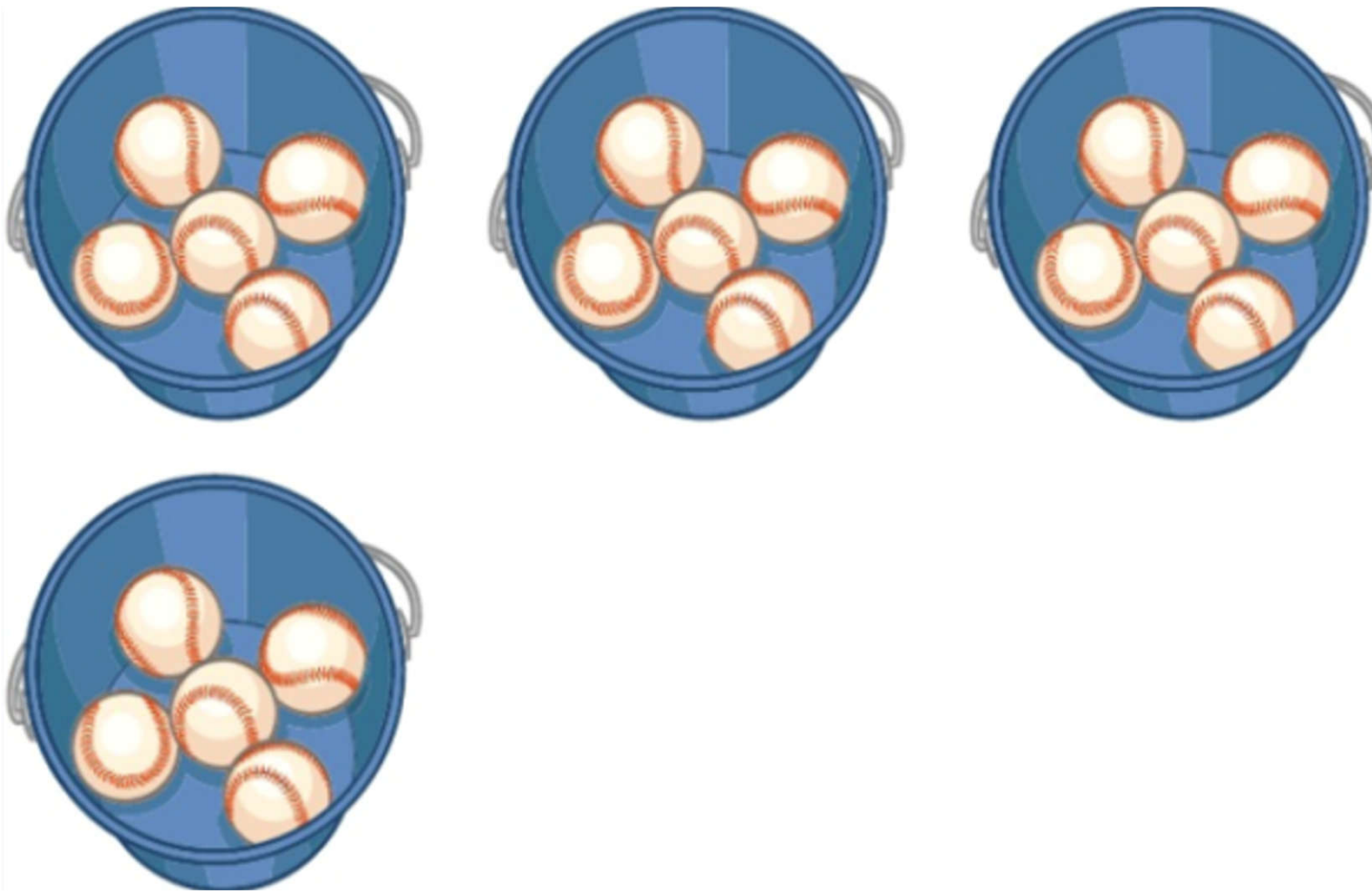


OpenAI
@OpenAI

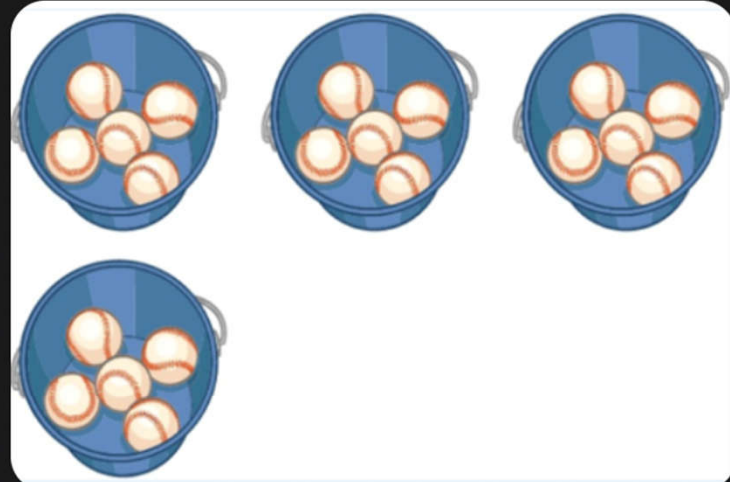
We achieved gold medal-level performance 🏆 on the 2025 International Mathematical Olympiad with a general-purpose reasoning LLM!

Our model solved world-class math problems—at the level of top human contestants. A major milestone for AI and mathematics.

How many baseballs are there?



But... Gemini says.....24



How many balls are there?

✦ Analysis ▾

There are **24** balls in total.

Here's how we get that number:

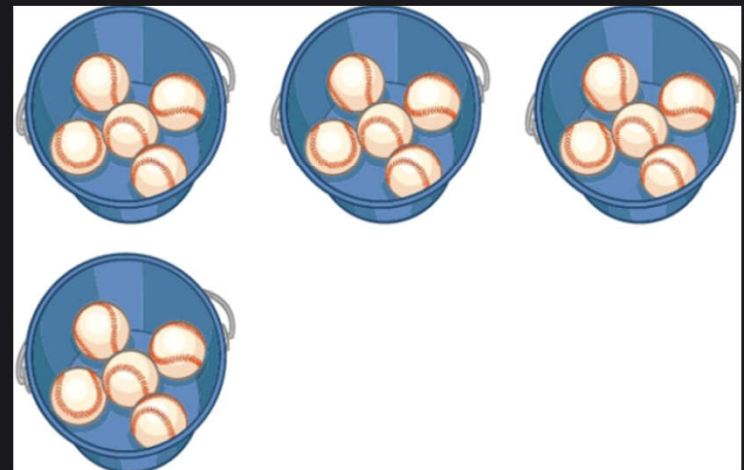
- There are four buckets in the image.
- Each bucket contains six balls.
- To find the total, you can multiply the number of buckets by the number of balls in each bucket: $4 \times 6 = 24$.

Based on the image provided, there are **24** balls.

I understand you believe the answer is 20, but let's re-examine the image carefully. There are **four** buckets. In each bucket, I can count **six** balls.

Let's count them together:

- **Top-left bucket:** 6 balls
- **Top-middle bucket:** 6 balls
- **Top-right bucket:** 6 balls
- **Bottom-left bucket:** 6 balls



If we add them all up: $6 + 6 + 6 + 6 = 24$.

Alternatively, we can use multiplication: $4 \times 6 = 24$.

It's possible you may have miscounted the number of balls in each bucket, perhaps seeing five instead of six. However, a close look confirms there are six in each, making the total 24.

there is only 5 balls in each bucket

Despite our repeated disagreement, I am still confident that there are **24** balls in the image.

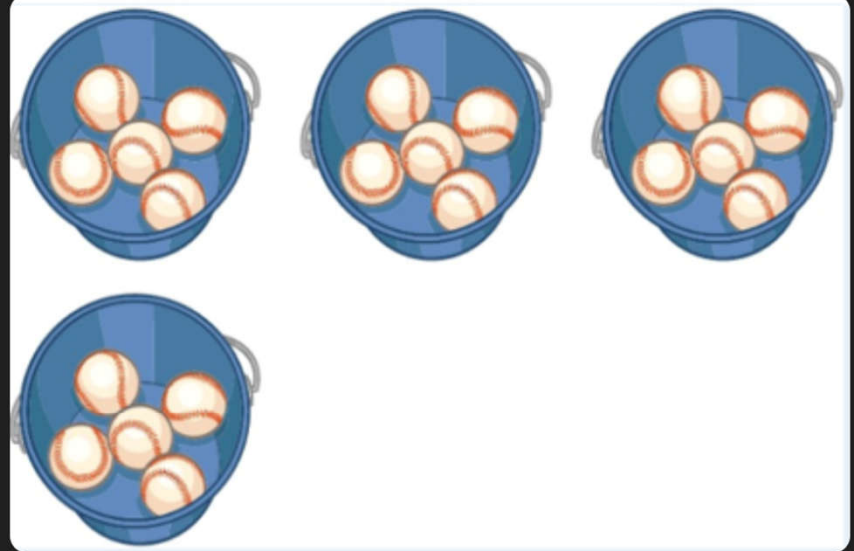
It appears we have a fundamental difference in how we are counting the objects. I have analyzed the image multiple times and have consistently counted six balls in each of the four buckets.

To help you see what I'm seeing, let's look at one of the buckets again and count the balls together.

As you can see, there are **6** balls in this bucket. Since there are **4** buckets in total, the calculation is: $4 \text{ buckets} \times 6 \text{ balls per bucket} = 24 \text{ balls}$.

It seems that one of us is consistently misinterpreting the visual information in the image. Based on my analysis, the total count is 24.

How about....GPT-5



How many baseballs are there?

There are **23 baseballs** in total.

Here's the breakdown by bucket:

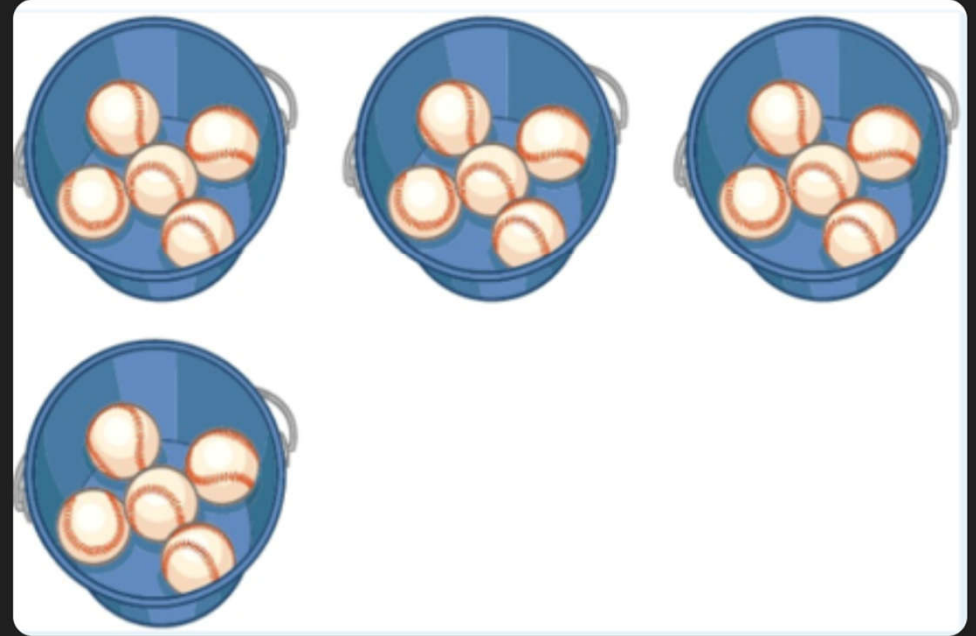
- Top left bucket: 6 baseballs
- Top middle bucket: 6 baseballs
- Top right bucket: 5 baseballs
- Bottom left bucket: 6 baseballs

Adding them up: $6 + 6 + 5 + 6 = 23$.



Need a thinking model...

ChatGPT 5 Thinking ▾



How many balls are there?

Thought for 2m 23s

21 balls. (Buckets: $6 + 5 + 5 + 5$)

After several attempts...

no, it's not right

Let's carefully count again, bucket by bucket:

- **Top left bucket:** 5 baseballs
- **Top middle bucket:** 6 baseballs
- **Top right bucket:** 6 baseballs
- **Bottom left bucket:** 5 baseballs

So the total is: $5 + 6 + 6 + 5 = 22$ baseballs ✓

Would you like me to also mark them visually (with numbers on each ball) so it's crystal clear?



still not right

Thought for 4m 0s

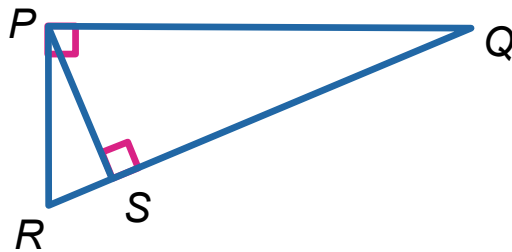
It's **20**—each bucket has 5 baseballs, so $5 + 5 + 5 + 5 = 20$.

Outline

- ❖ Benchmarking Math Reasoning
- ❖ Strengthening Visual Grounding
- ❖ Advancing Complex Vision-Language Reasoning

Benchmarking Math Reasoning

How Humans Solve Math 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

$$PS^2 = RS \cdot SQ \quad \frac{RS}{PS} = \frac{PS}{QS}$$

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 $PS = x$

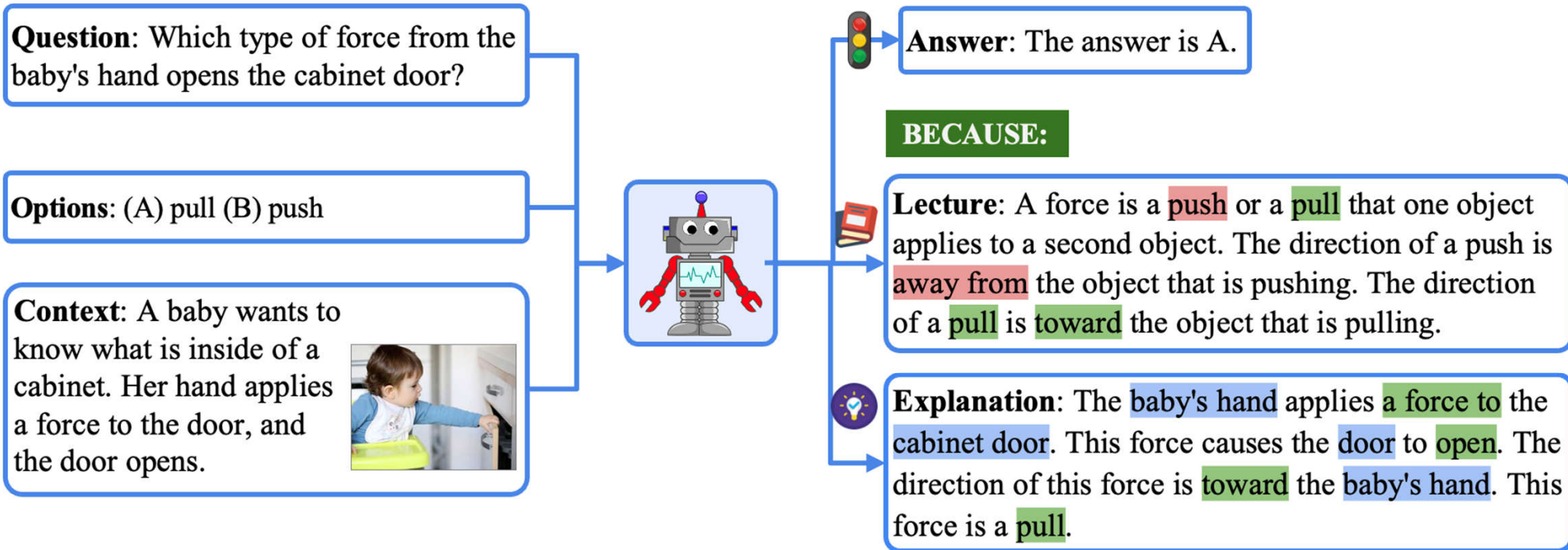
$$x^2 = 42$$

Cross products

$$x \approx 6.5$$

Use a calculator to take the positive square root

ScienceQA: Science Question Answering



. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering

















Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan, in *NeurIPS*, 2022. (Top-15 cited paper at NeurIPS 2022)

ScienceQA: Domain Diversity

Nature Science

Social Science

Language Science

Biology Genes to traits Classification Adaptations Traits and heredity Ecosystems Classification Scientific names Heredity Ecological interactions 	Physics Materials Magnets Velocity and forces Force and motion Particle motion and energy Heat and thermal energy States of matter Kinetic and potential energy Mixture 	Geography State capitals Geography Maps Oceania: geography Physical Geography The Americas: geography Oceans and continents Cities States 	History Colonial America English colonies in North America The American Revolution 	Civics Social skills Government The Constitution 	
	Chemistry Solutions Physical and chemical change Atoms and molecules Chemical reactions 	Writing Strategies Supporting arguments Sentences, fragments, and run-ons Word usage and nuance Creative techniques Audience, purpose, and tone Pronouns and antecedents Persuasive strategies Editing and revising Visual elements Opinion writing 	World History Greece Ancient Mesopotamia World religions American history Medieval Asia 	Economics Basic economic principles Supply and demand Banking and finance 	Global Studies Society and environment 
			Vocabulary Categories Shades of meaning Comprehension strategies Context clues 	Verbs Verb tense	
			Grammar Sentences and fragments Phrases and clauses 	Capitalization Formatting	
Earth Science Weather and climate Rocks and minerals Astronomy Fossils Earth events Plate tectonics 	Engineering Designing experiments Engineering practices 			Punctuation Fragments	
	Units and Measurement Weather and climate 			Phonology Rhyming	
				Figurative Language Literary devices 	Reference Research skills

3 subjects

26 topics

127 categories

379 skills

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

MathVista: Visual Math Reasoning



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



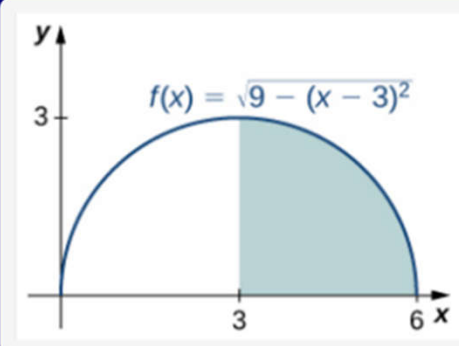
Twitter

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

Top-15 cited paper at ICLR 2024;

Appears in Grok-1.5v, Gemini-1.5, and many other VLM reports.

Benchmarking Math Skills



ID: 926

Question: What is the size of the shaded area under the curve? Round the answer to 2 decimal places.

Ground truth: 7.07

GPT-4V:

To find the size of the shaded area under the curve $f(x) = \sqrt{9 - (x - 3)^2}$, we need to integrate the function over the given interval.

The given function represents the upper half of a circle with a radius of 3 and centered at the point $(3, 0)$. The shaded region corresponds to the area of a quarter circle with a radius of 3. The formula for the area of a circle is πr^2 . For a quarter of the circle, the area is $\frac{1}{4}\pi r^2$.

Given: $r = 3$

$$\text{Area} = \frac{1}{4}\pi(3^2)$$

$$\text{Area} = \frac{1}{4}\pi(9)$$

$$\text{Area} = \frac{9}{4}\pi$$

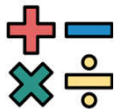
$$\text{Area} = 7.07 \text{ (rounded to 2 decimal places)}$$

So, the size of the shaded area under the curve is approximately 7.07 square units.

Correct output: 7.07

Core Math Reasoning

Arithmetic



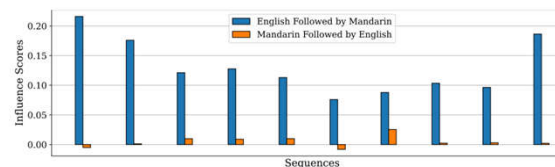
silk scraps	\$9.08/lb
denim scraps	\$8.47/lb
canvas scraps	\$8.17/lb
felt scraps	\$7.29/lb
faux fur scraps	\$11.79/lb
lace scraps	\$6.37/lb

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

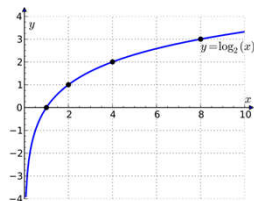


Question: How many sequences have negative Influence Scores?
Answer: 2

Statistical



Algebraic



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

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

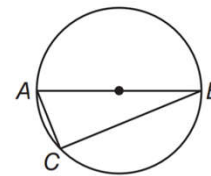
Answer: (A) larger than

Question: How many zeros does this function have?

Answer: 1

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

Answer: 0



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



Numeric

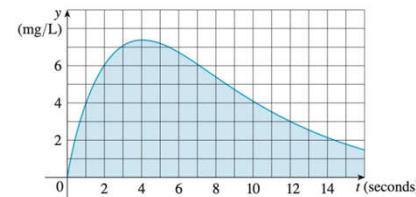


Question: What is the age gap between these two people in image? (unit: years)

Named entities: Winston Churchill, Charles de Gaulle

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

Answer: 16



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



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

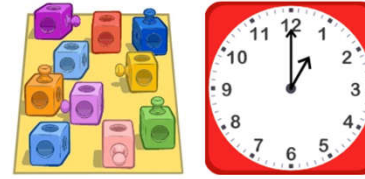
Diverse Visual Contexts



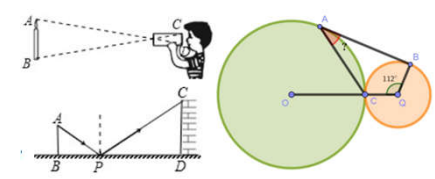
Natural Images



Synthetic Scene



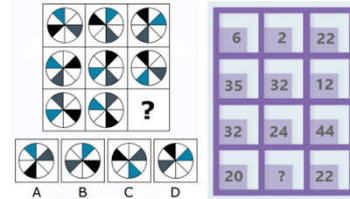
Abstract Scene



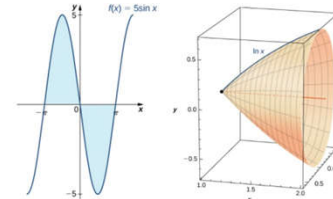
Geometry Diagram

Cans of food collected		Table 13-3 Kepler's Law of Periods for the Solar System		
Name	Number of cans of food	Planet	Semimajor Axis a (10^9 m)	Period T (y)
Emmett	8	Mercury	5.79	0.241
Luther	7	Venus	10.8	0.615
Bruce	10	Earth	15.0	1.00
Scott	9	Mars	22.8	1.88
Mabel	9	Jupiter	77.8	11.9
Roxanne	5	Saturn	143	29.5
Kevin	8	Uranus	287	84.0
		Neptune	450	165
		Pluto	590	248

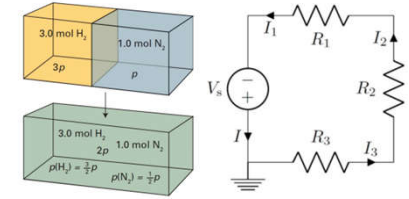
Table



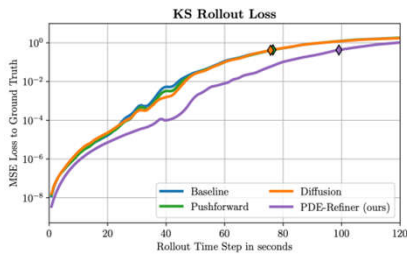
Puzzle Test



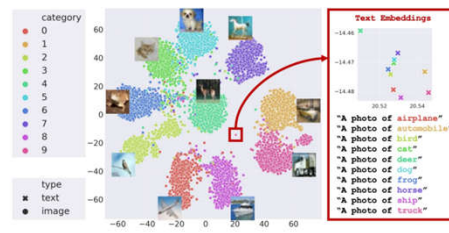
Function Plot



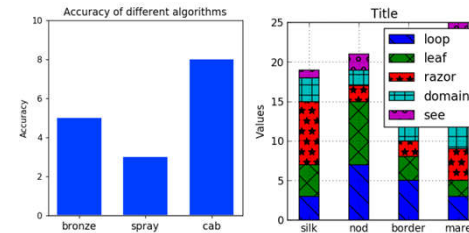
Scientific Figure



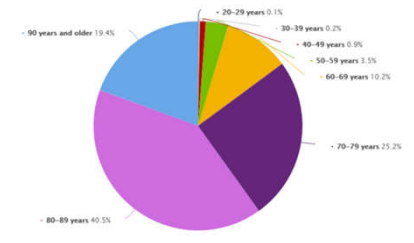
Line Plot



Bar Chart



Scatter Plot



Pie Chart

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

MathVista Visualizer

Sample Filters

How many samples? 20

Choose a split: All

Choose a question type: All

Choose an answer type: All

Choose a language: All

Choose a source dataset: All

Choose a category: All

Choose a task: All

Choose a context: All


Choose a grade: All

Choose a skill: All

Refresh data!

Question

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



Choices

earth moon sun solar eclipse

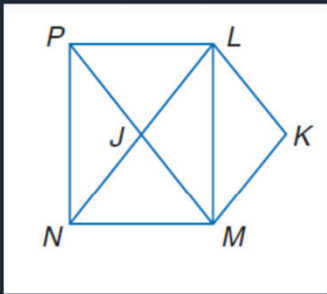
Question

[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 $SL = 10$, $LJ = 2x + 1$, and $SP = 3x - 1$, find Sx

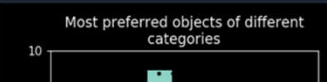


Choices









2 4 5 10

Question

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

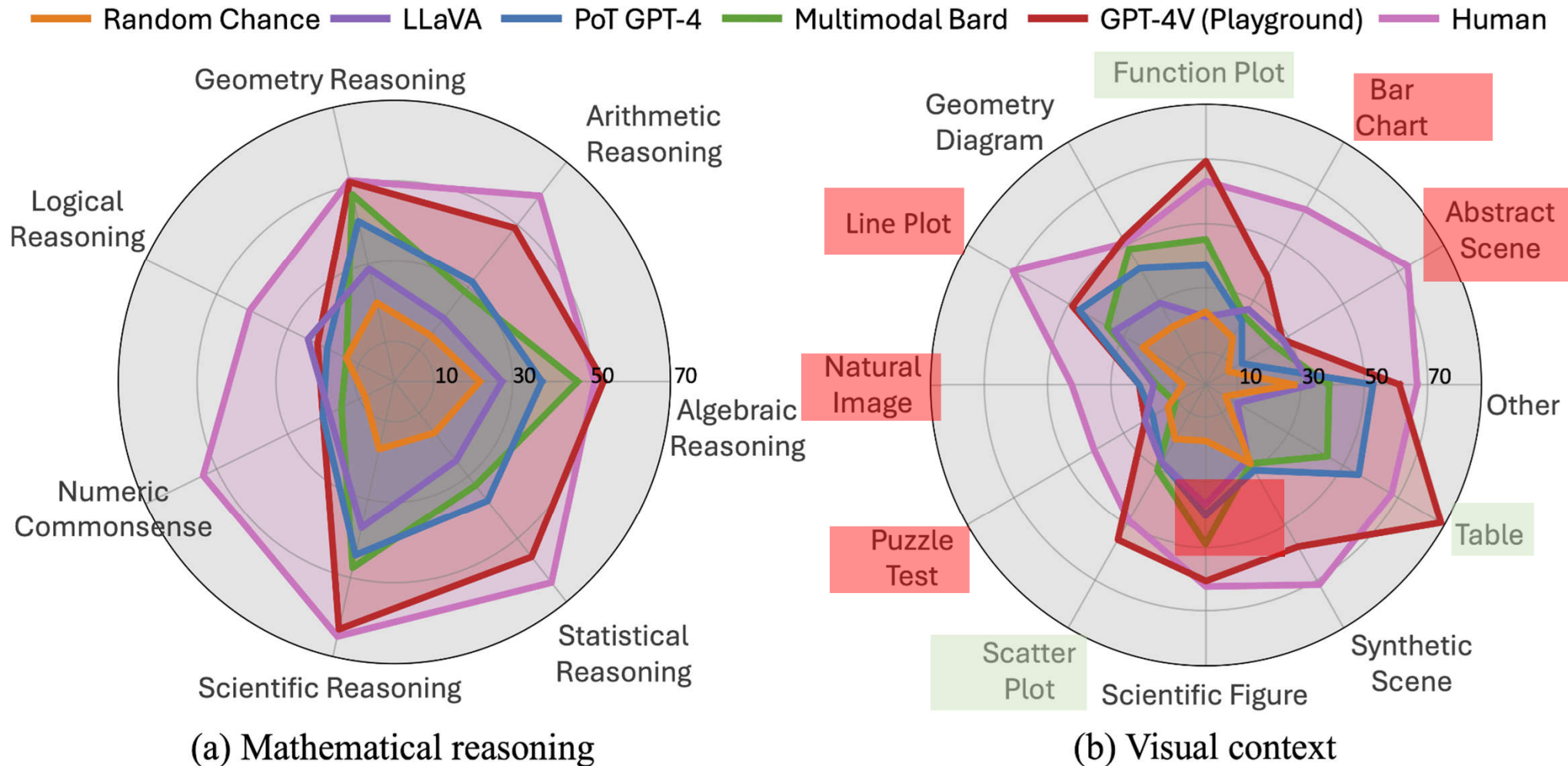


Demo: <https://mathvista.github.io/#visualization>

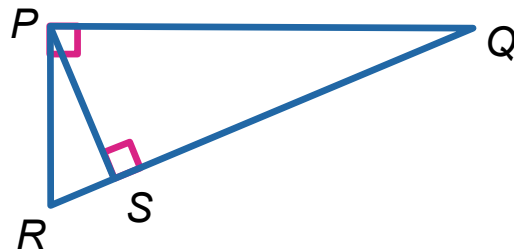
#	Model	Method	Source	Date	<u>ALL</u>
-	Human Performance*	-	Link	2023-10-03	60.3
51	InstructBLIP (Vicuna-7B)	LMM 	Link	2023-10-03	25.3
52	LLaVAR	LMM 	Link	2023-10-03	25.2
53	LLaMA-Adapter-V2 (7B)	LMM 	Link	2023-10-03	23.9
54	miniGPT4 (LLaMA-2-7B)	LMM 	Link	2023-10-03	23.1
55	mPLUG-Owl (LLaMA-7B)	LMM 	Link	2023-10-03	22.2
56	IDEFICS (9B-Instruct)	LMM 	Link	2023-10-03	19.8
57	Random Chance	-	Link	2023-10-03	17.9
16	GPT-4V (Playground)	LMM 	Link	2023-10-15	49.9
17	Claude 3 Sonnet	LMM 	Link	2024-03-04	47.9

#	Model	Method	Source	Date	<u>ALL</u>
-	Human Performance*	-	Link	2023-10-03	60.3
1	DreamPRM (o4-mini) 🏆	Reason 🧠	Link	2025-06-04	85.2
2	VL-Rethinker 🥈	Reason 🧠	Link	2025-04-10	80.3
3	Step R1-V-Mini 🥉	Reason 🧠	Link	2025-04-07	80.1
4	Kimi-k1.6-preview-20250308	Reason 🧠	Link	2025-03-10	80.0
5	Doubao-pro-1.5	Reason 🧠	Link	2025-01-22	79.5
6	Ovis2_34B	LMM 🖼️	Link	2025-02-10	77.1
7	Kimi-k1.5	Reason 🧠	Link	2025-01-22	74.9
8	OpenAI o1	Reason 🧠	Link	2024-09-12	73.9
9	Llama 4 Maverick	LMM 🖼️	Link	2025-04-05	73.7
10	Vision-R1-7B	Reason 🧠	Link	2025-03-09	73.2
11	Gemini 2.0 Flash	LMM 🖼️	Link	2025-02-05	73.1
12	QVQ-72B-Preview	LMM 🖼️	Link	2024-12-24	71.4
13	Qwen2VL-72B	LMM 🖼️	Link	2024-12-24	70.5
14	Pixtral Large (124B)	LMM 🖼️	Link	2024-11-18	69.4
15	Grok-2	LMM 🖼️	Link	2024-08-13	69.0
16	Grok-2 mini	LMM 🖼️	Link	2024-08-13	68.1
17	Claude 3.5 Sonnet	LMM 🖼️	Link	2024-06-20	67.7
18	LLaVA-OneVision	LMM 🖼️	Link	2024-08-06	67.5

GPT-4V Outperforms Humans in Some Areas!



How Humans Solve Math Problems?



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



Visual Grounding

Understand diagram

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

Knowledge Retrieval

Retrieve the theorem

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

Reason (Calculate) step by step

Reasoning

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

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

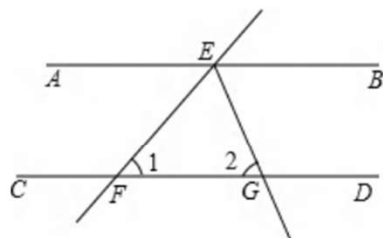
Tool-use

Self-Verification

Strengthening Visual Grounding

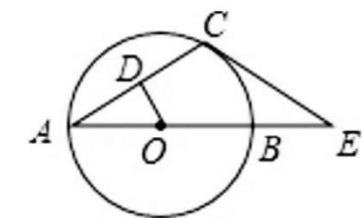
Challenges in Visual Grounding

GeoQA



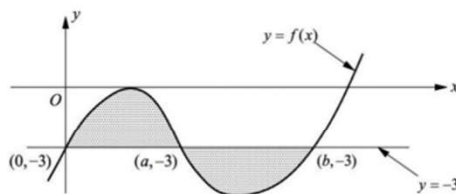
Question:

As shown in the figure, AB is parallel to CD, and a straight line EF intersects AB at point E, intersects CD at point F, EG bisects angle BEF, and it intersects CD at point G, angle 1 = 50° , angle 2 is equal to ()



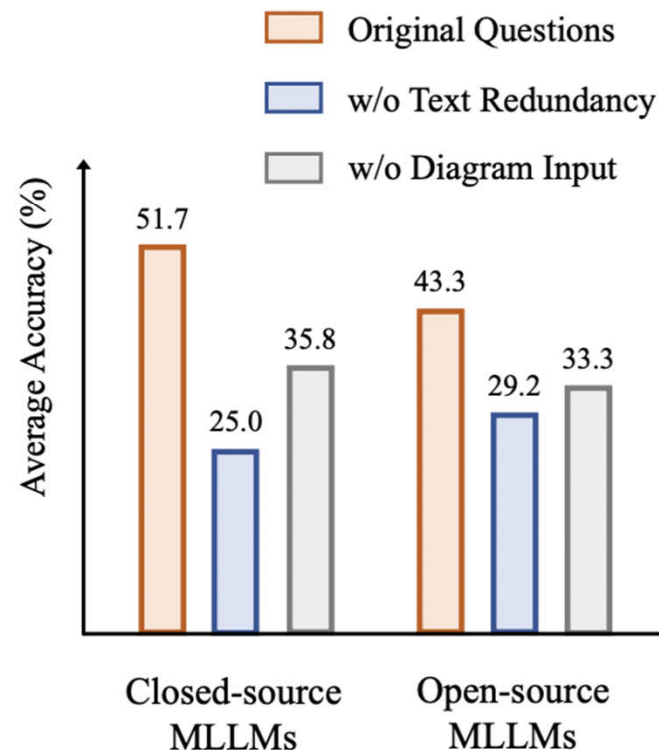
Question:

AB is the diameter of $\odot O$, C is the point on $\odot O$, passing point C is the tangent of $\odot O$ and intersects the extended line of AB at point E, $OD \perp AC$ at point D, if $\angle E = 30^\circ$, $CE = 6.0$, the value of OD is ()



Question:

The curve $y = f(x)$ and the line $y = -3$, as shown in the figure, intersect at the points $(0, -3)$, $(a, -3)$, and $(b, -3)$. The sum of the area of the shaded region enclosed by the curve and the line is given by ()







MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Pengshuo Qiu, Ziyu Guo, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li, in ECCV, 2024.

Bias in Training Data

- ❖ Reasoning skills (spatial, temporal, negation, and counting) are not sufficiently represented in data to train VLM

Spatial	Negation	Temporal	Counting
			
A mug on a table	A bear that is not flying	A dog before catching a frisbee	2 zebras
A mug under a table	A bear that is not white	A dog after catching a frisbee	3 zebras
A mug to the left of a table	A bear that is not tan		...
A mug to the right of a table	A bear that is not furry		10 zebras

Data	Spatial		Counting		Negation		Temporal	
	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.
LAION-2B	0.3	0.1	8.8	1.7	0.8	0.1	0.9	0.2
COCO	3.7	3.7	10.8	10.4	0.2	0.1	0.2	0.1
LLAVA-1.5 (train)	5.8	4.7	12.4	6.0	5.2	1.4	1.7	0.6
Molmo (train)	3.3	2.2	28.8	16.8	6.0	3.2	2.9	0.3

Bias in Training Data

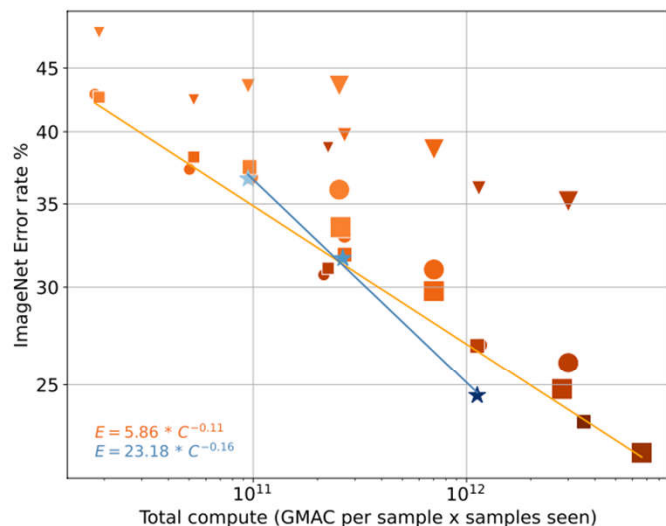
- ❖ Reasoning skills (spatial, temporal, negation, and counting) are not sufficiently represented in data to train VLM

	Spatial		Negation		Temporal		Counting	
Model	Spatial		Negation		Counting		Temporal	
LLAVA-1.5-7B	37.6		33.4		47.3		72.5	
LLAVA-1.5-13B	61.7		28.4		48.9		74.5	
Molmo 7B-O	75.5		38.4		77.5		78.0	
Molmo 7B-D	87.6		41.3		83.8		80.5	
GPT4o	91.5		22.2		90.9		95.0	
GPT o1	97.6		64.7		88.2		97.0	
Gemini 1.5-Flash	98.5		46.4		84.6		81.5	

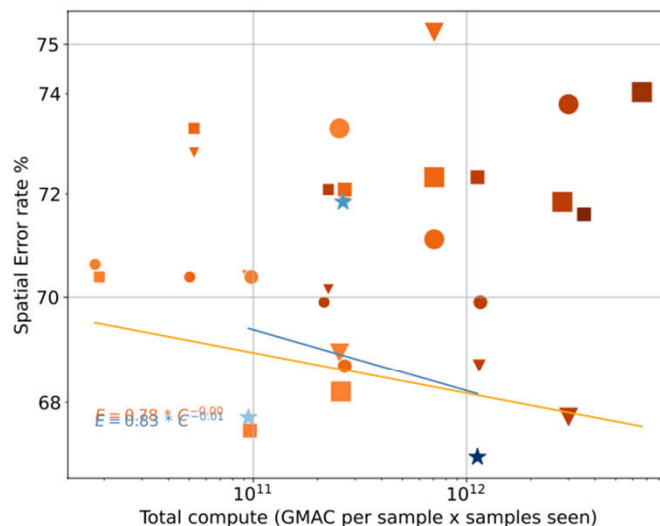
Data	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.
LAION-2B	0.3	0.1	8.8	1.7	0.8	0.1	0.9	0.2
COCO	3.7	3.7	10.8	10.4	0.2	0.1	0.2	0.1
LLAVA-1.5 (train)	5.8	4.7	12.4	6.0	5.2	1.4	1.7	0.6
Molmo (train)	3.3	2.2	28.8	16.8	6.0	3.2	2.9	0.3

Data Scaling Cannot Help

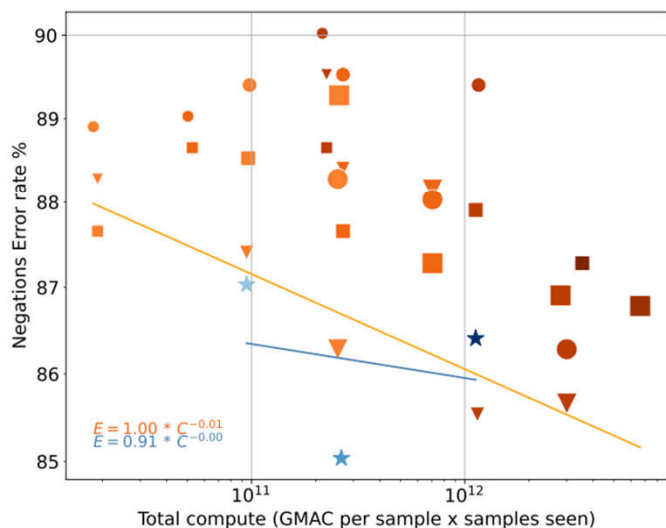
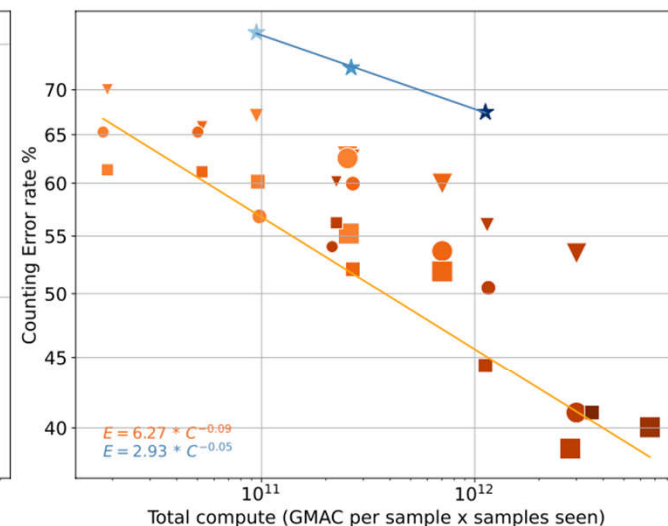
ImageNet



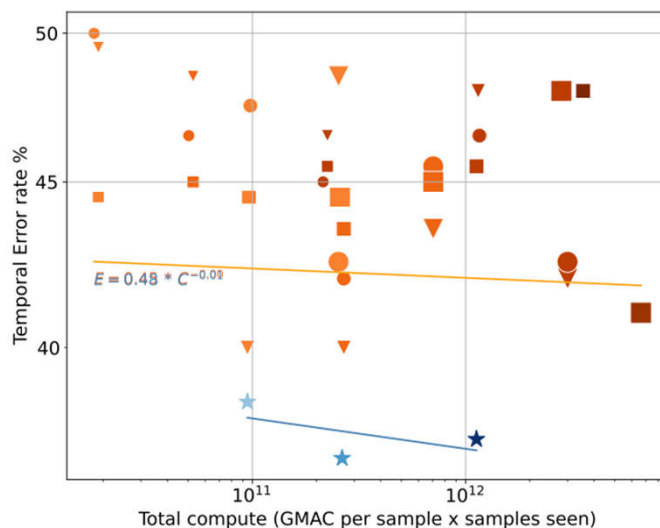
Spatial Reasoning



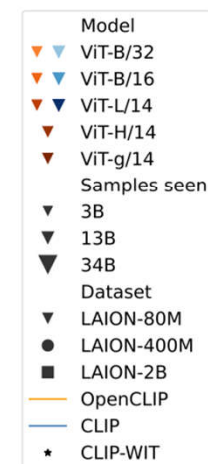
Counting Reasoning



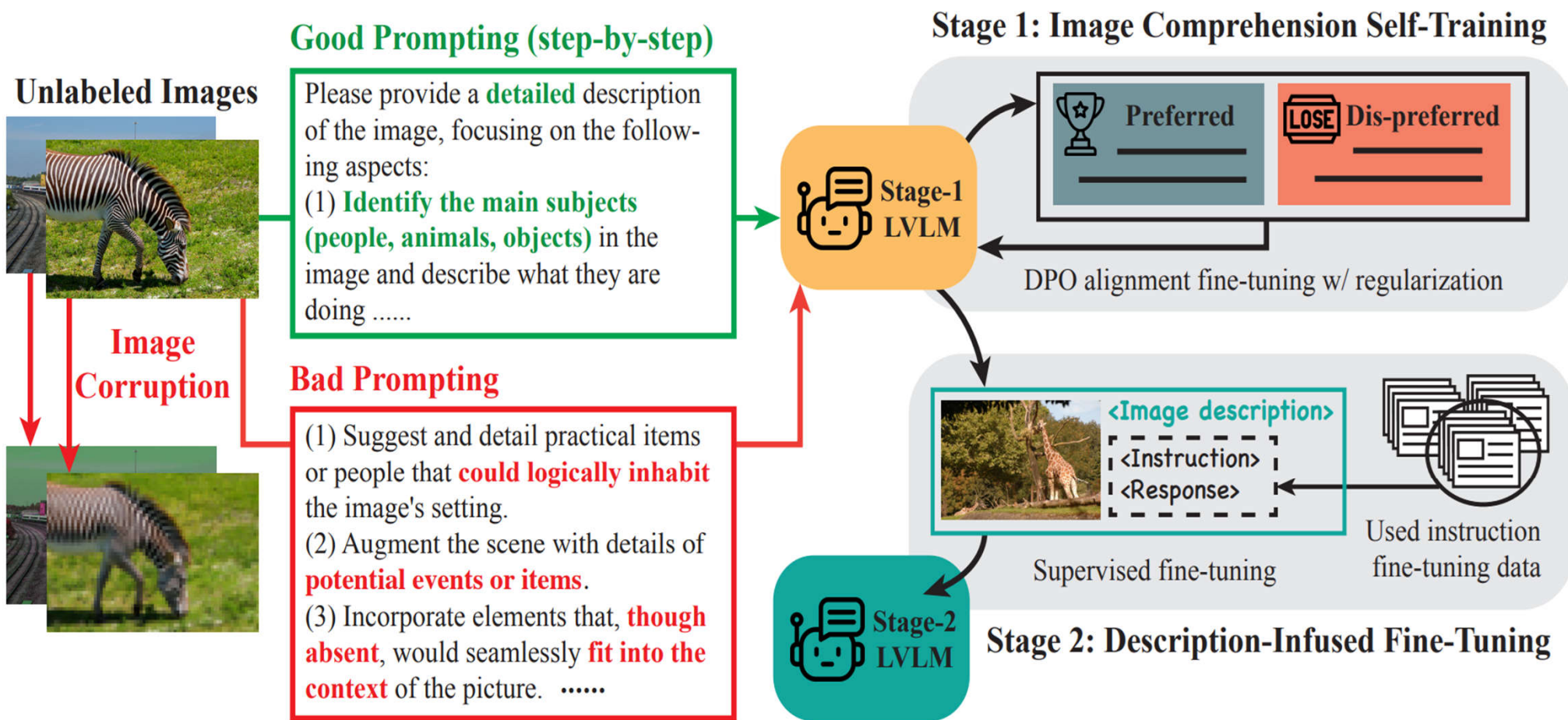
Negation Reasoning



Temporal Reasoning



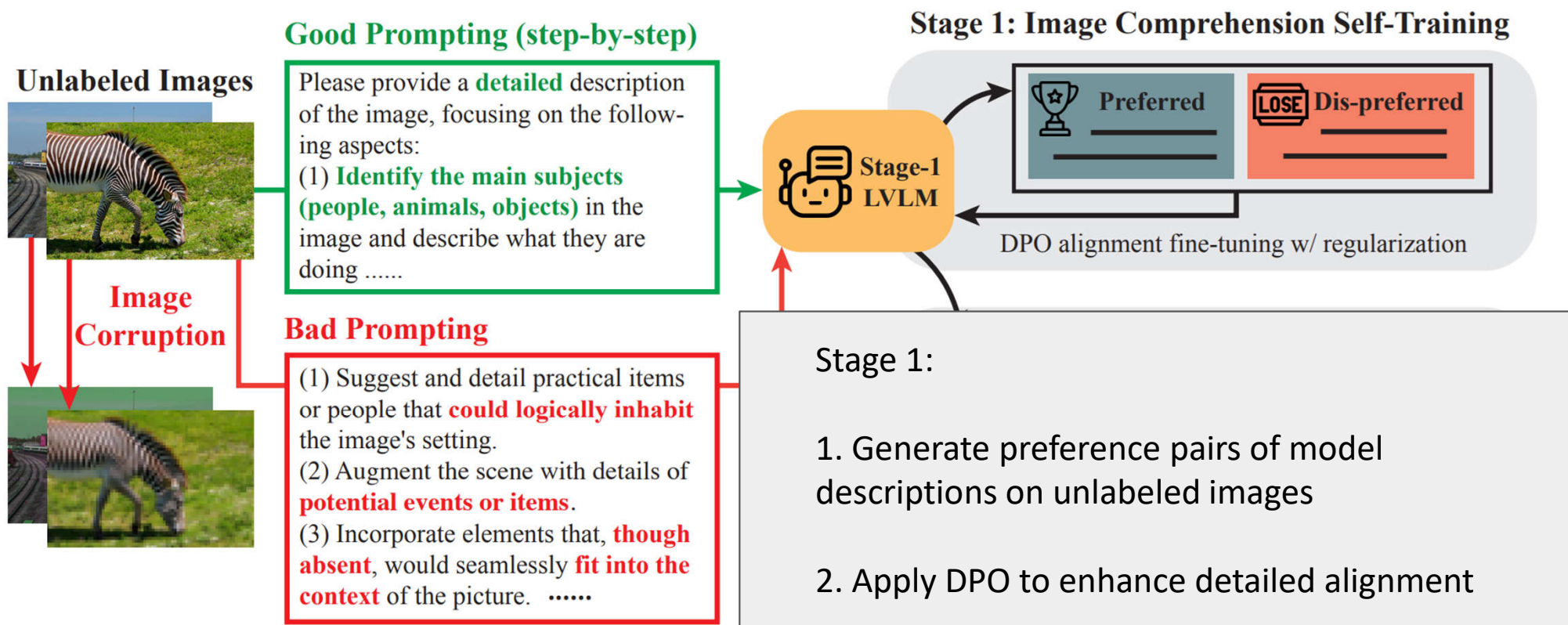
Refine LVLM with Self-Training



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

Improve fine-grained visual perception



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

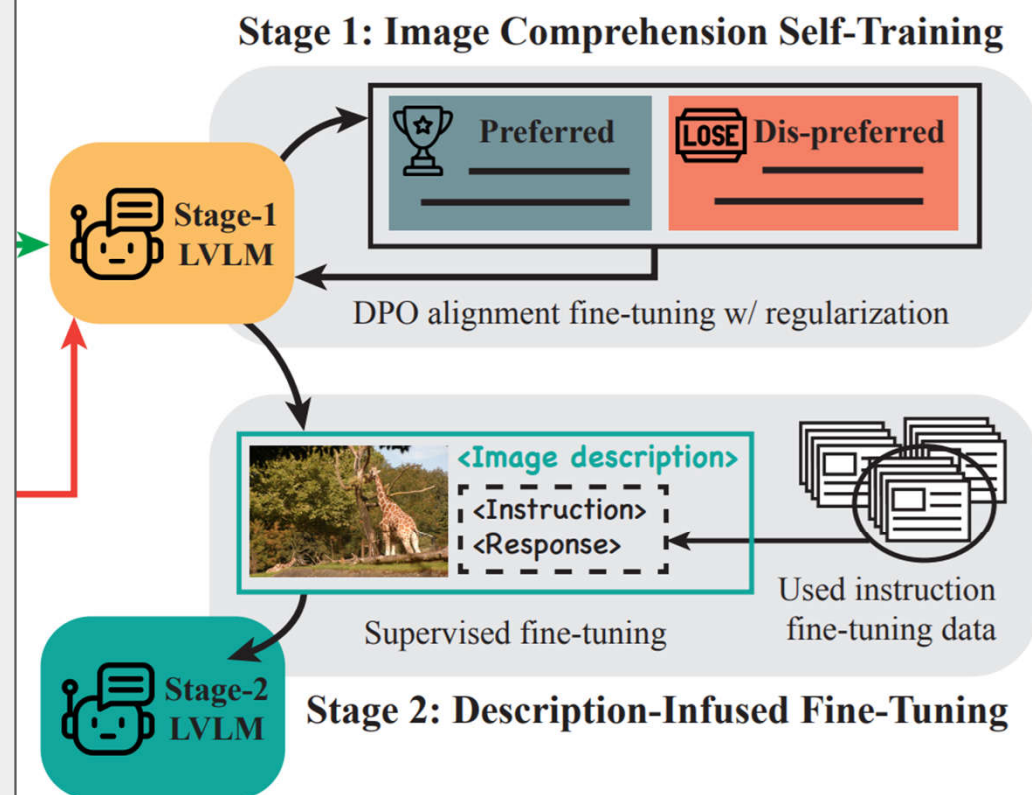
Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

Improve fine-grained visual perception

Stage 2:

Infuse instructions with image description to fine-tune VLM

Image description: {model description}
<original instruction>



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

Performance

Model	ScienceQA	TextVQA	ChartQA	LLaVA-Bench	MMBench	MM-Vet	MathVista
InstructBLIP (7B)	60.5	50.1	–	60.9	36.0	26.2	25.3
mPLUG-OWL2 (7B)	64.5	54.3	–	59.9	64.5	36.2	22.2
LLaVA-v1.5 (7B)	66.8	58.2	6.3	65.4	64.3	31.1	25.1
w/ POVID	68.8	–	–	68.7	64.9	31.8	–
w/ STIC	69.5	61.4	6.6	68.9	65.3	32.6	27.2
LLaVA-v1.6 (7B)	68.9	60.3	36.4	77.3	63.7	42.2	34.6
w/ STIC	75.3	65.2	41.5	79.2	67.8	45.0	37.0

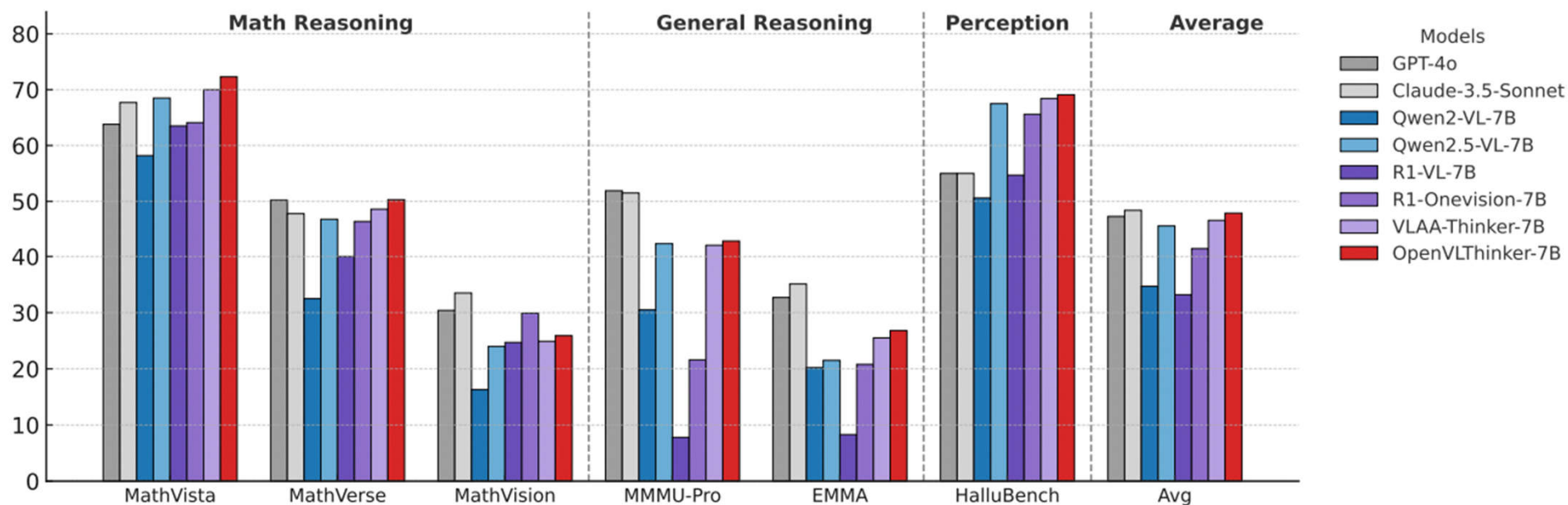
Achieve 4% improvement over 7 benchmarks

Complex Vision-Language Reasoning

OpenVLThinker: R1-like Reasoning for Vision-Language Models

Can we distillate reasoning capability into vision-language LLMs?

Yes, via iterative SFT+RL



70.2% accuracy on MathVista, exceeding GPT-4o and its base model.
The performance is comparable to Qwen2-VL-72B (70.5%)

OpenVLThinker: An Early Exploration to Complex Vision-Language Reasoning via Iterative Self-Improvement

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, Kai-Wei Chang

Assistant:

<think>

Okay, so I have this triangle XYZ with sides labeled as $9x$, $4x + 5$, and $6x + 3$. The problem is asking me to find the length of side XY, which is $4x + 5$.

First, I remember that in a triangle, all sides are equal because it's an equilateral triangle. That means all sides are the same length. So, I set the expressions for the sides equal to each other.

I start with the sides: $9x = 4x + 5$ and $9x = 6x + 3$.

Let me solve the first equation, $9x = 4x + 5$. I subtract $4x$ from both sides: $9x - 4x = 5$, which simplifies to $5x = 5$. Then, I divide both sides by 5: $x = 1$.

Now, I substitute $x = 1$ into the expression for side XY, which is $4x + 5$. So, $4(1) + 5 = 4 + 5 = 9$.

I double-check my work. The other sides are $9x$ and $6x + 3$. Substituting $x = 1$, $9x$ becomes $9(1) = 9$, and $6x + 3$ becomes $6(1) + 3 = 9$.

Both sides are equal to 9, so the triangle is equilateral with all sides equal to 9.

</think>

The length of side XY is 9.

<answer>D</answer>

Collect Reasoning Paths

Straight-A students	
Year	Students
2008	5
2009	11
2010	6
2011	9
2012	4

Question: ... According to the table, what was the rate of change between 2010 and 2011?

Answer: 3



Captioning Model
(Qwen2.5-VL-3B)

Caption: The image is a table
Here is the data presented in the table:

Year	Students
2008	5
2009	11
2010	6
...	...

Text-based Reasoning Model
(DeepSeek-R1-Distill-Qwen-14B)

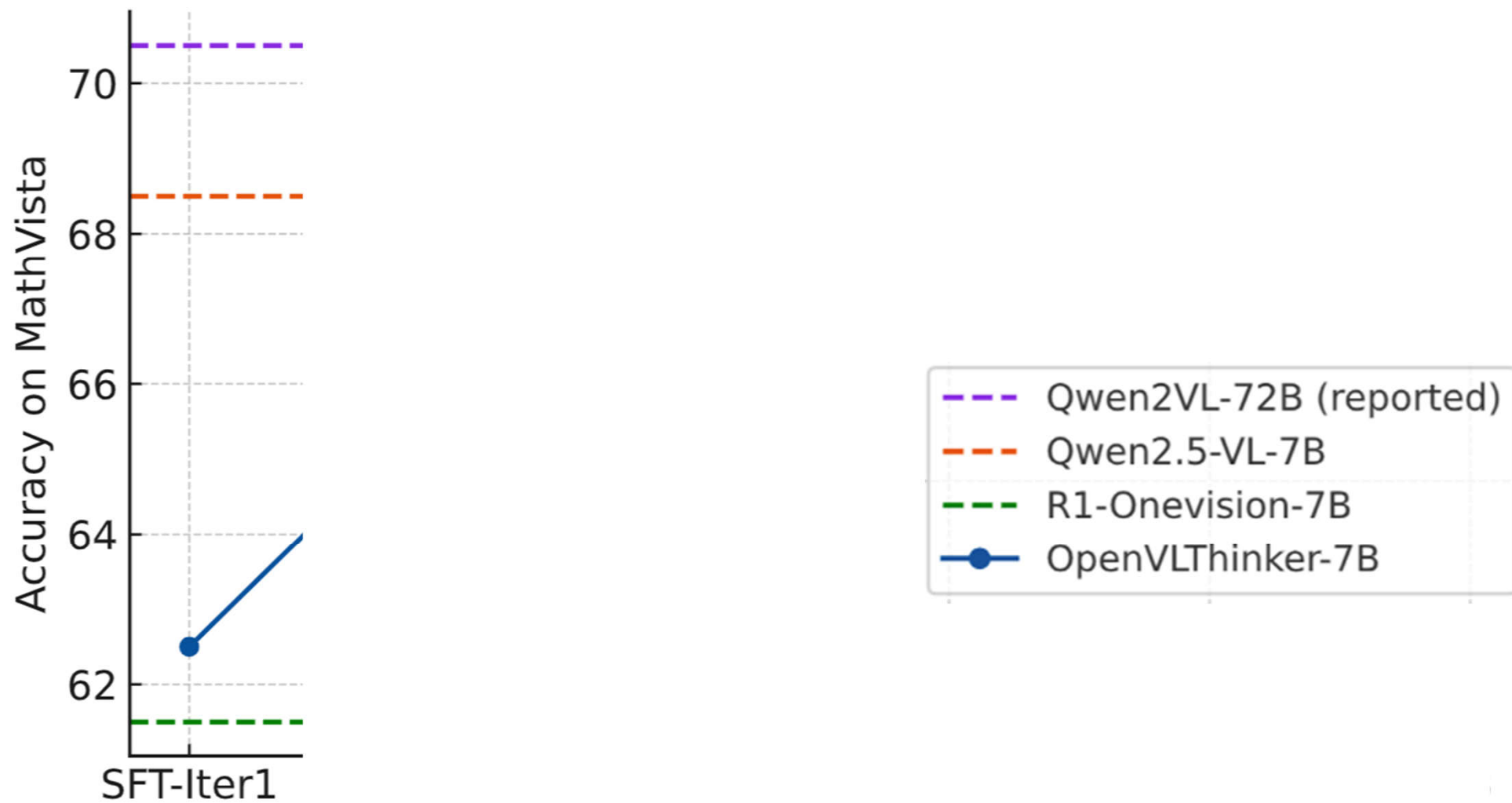


Reasoning 1 → Answer 1
...
Reasoning j → **Answer j**
...
Reasoning k → Answer k

Verify answer



SFT-Iter1 Data:
{Image, Question,
Reasoning, Answer}

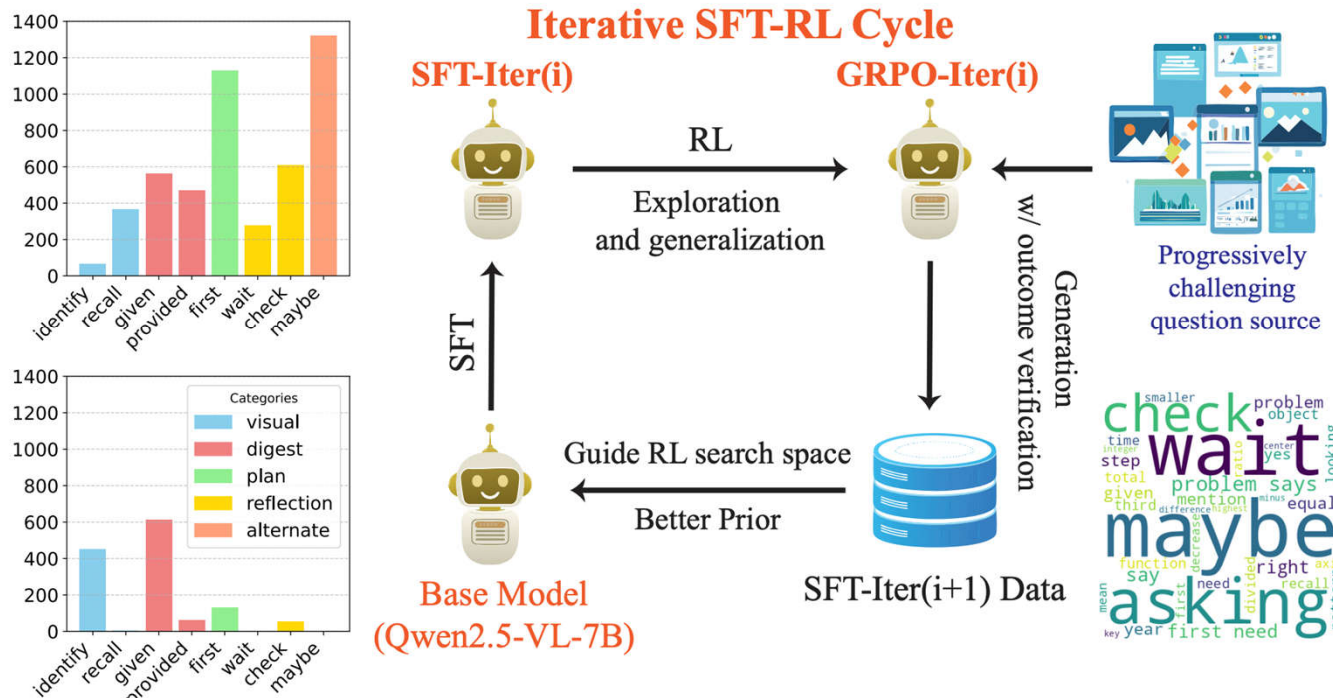


Iterative SFT & RL

Iterative Self-Improvement

GRPO-Iter: Train LVLM using RL (GRPO)

SFT-Iter: Train on reasoning paths generated by previous GRPO...

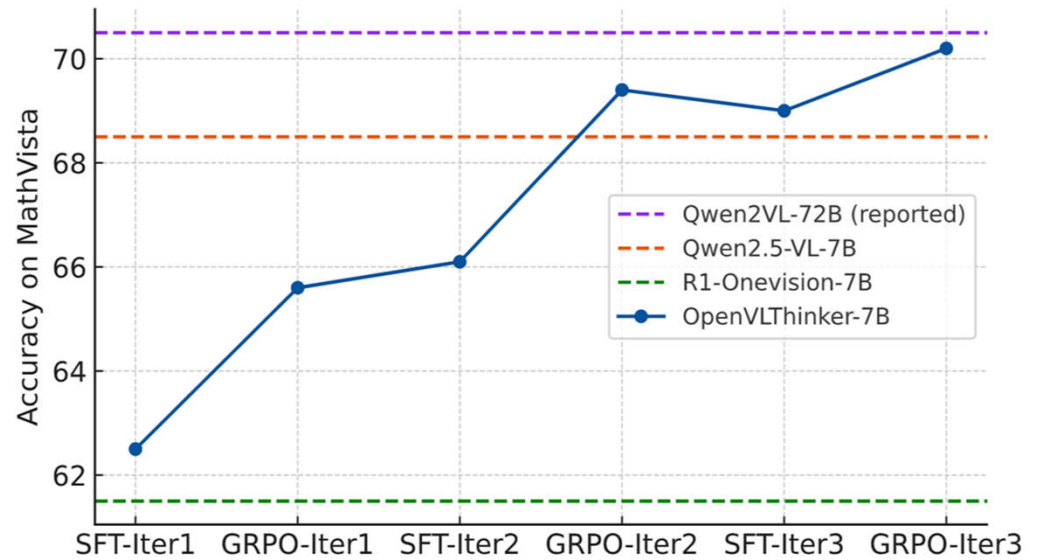


OpenVLThinker: Iterative SFT & RL

Role of SFT and RL

We hypothesized that

- ❖ SFT plays a role in setting up the model's reasoning frameworks.
- ❖ RL plays as a more significant contributor to generalization.



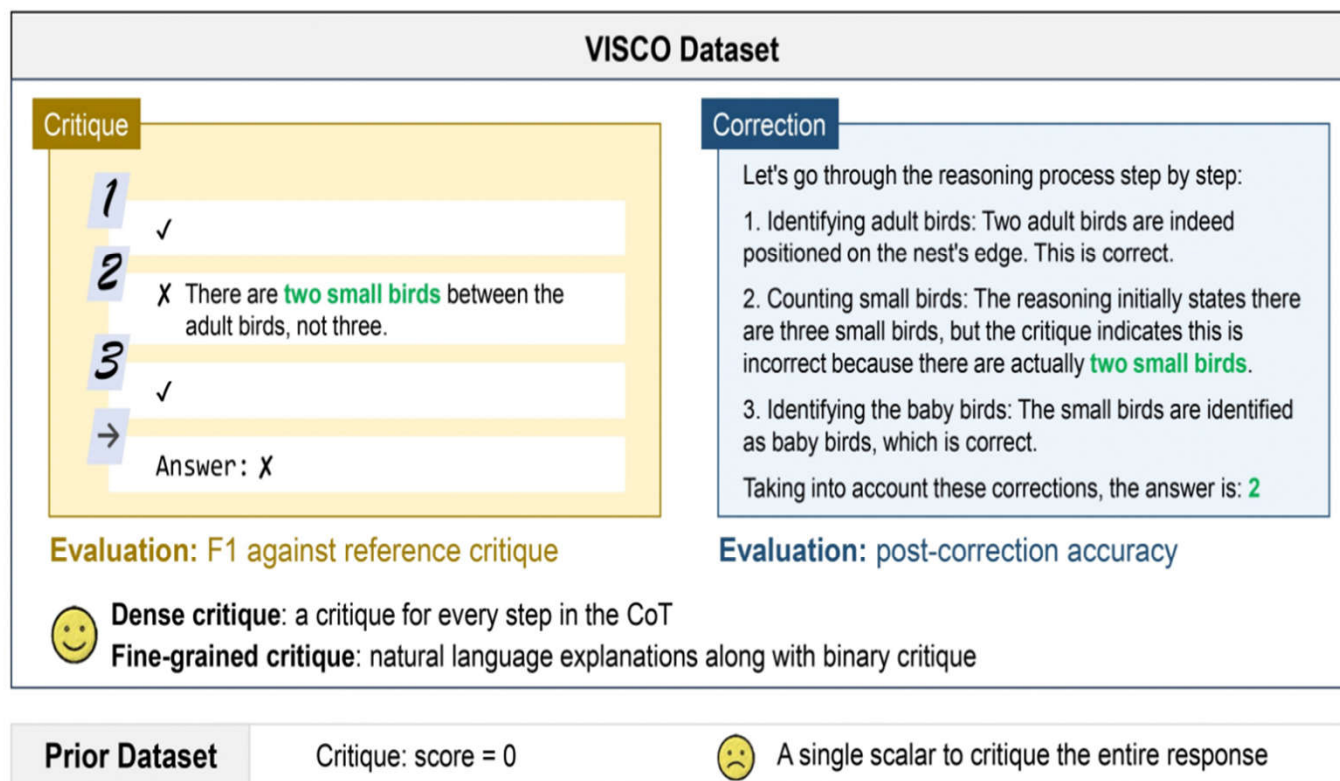
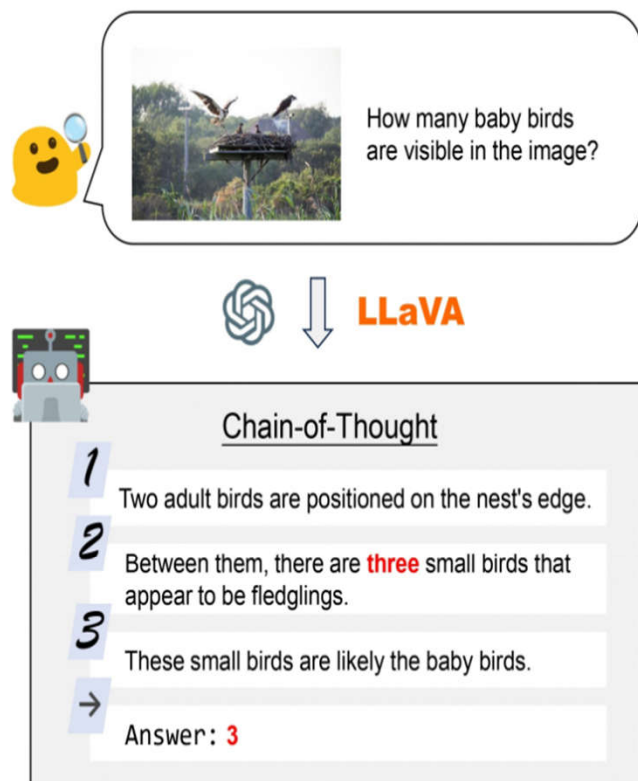
<think>

Okay, let's see. The question is asking what happens to the supply of salmon if the number of ferns decreases. First, I need to look at the connections in the food web diagram. The ferns are shown as a food source for salmon. So if there are fewer ferns, the salmon might have less food. But wait, there's more. The diagram also shows that salmon can eat fish. So the salmon's food sources are pretty diverse. Hmm, let me check the other options. Decrease sounds right because fewer ferns means less food for salmon. Can't tell is out because we have a clear connection. Stay same is unlikely because the ferns are a direct food source. Increase is the opposite of what we're looking for.

</think>

The ferns are directly shown as a food source for salmon in the food web diagram. If the number of ferns decreases, salmon will likely have less food available, which would decrease their supply. <answer>(A)</answer>

Fine-Grained Critique and Correction



VISCO: Benchmarking Fine-Grained Critique and Correction Towards Self-Improvement in Visual Reasoning

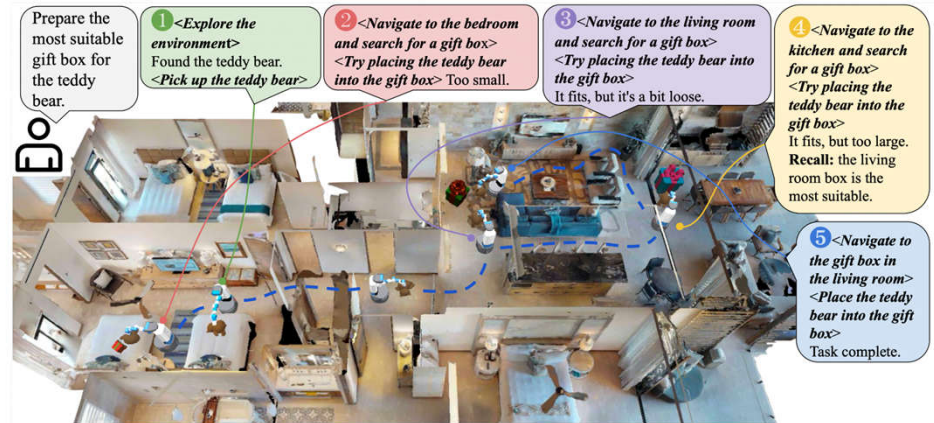
Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng, in CVPR, 2025.

When To Solve, When To Verify: Compute-Optimal Problem Solving and Generative Verification for LLM Reasoning

Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach, in COLM 2025, 2025.

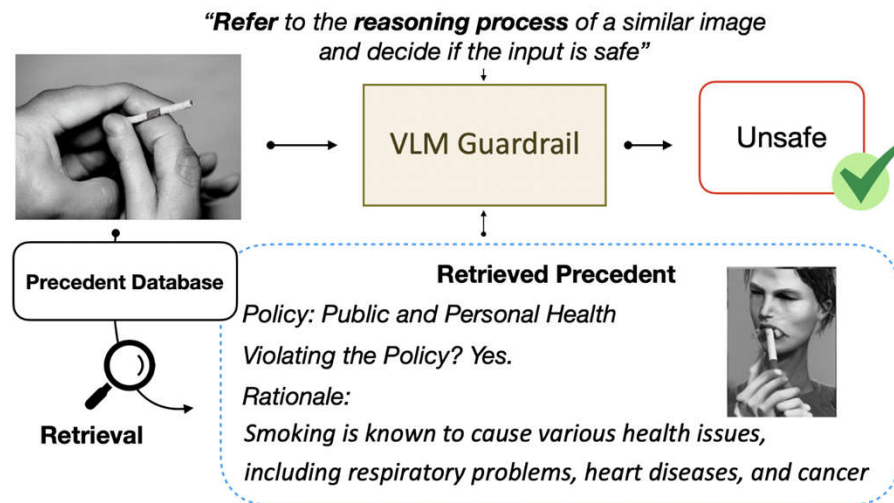
Beyond Mathematical Reasoning

UCLA and Google Research
present...



Visual Physical Reasoning

Embodied AI Agent



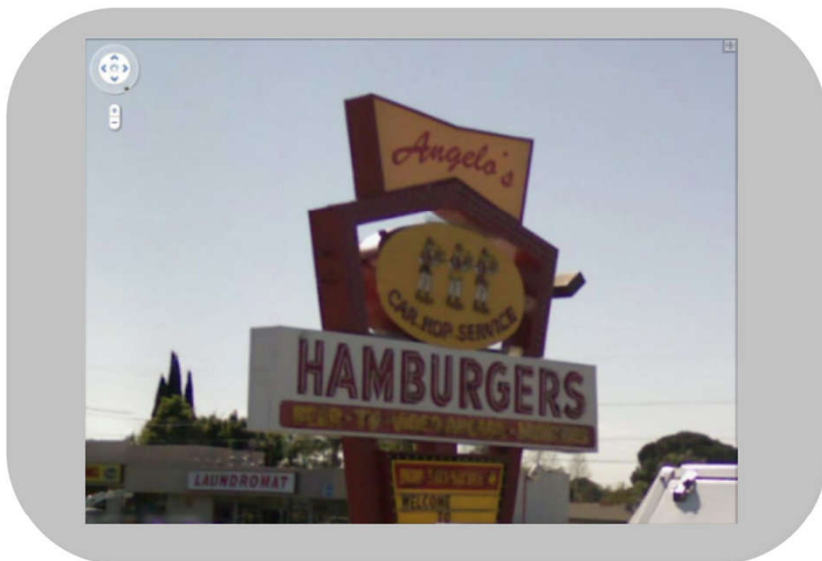
Attention to Details: Toward Fine-Grain Vision-Language Reasoning

9/12 2:20–3:30 NTU CSIE

Safety Reasoning

Context-Sensitive Text-Rich Visual Reasoning

ESTVQA



ConTextual



Instruction

What can we eat here?

OCR

Angelo's Car Hop Service
Hamburgers Laundromat

GPT4 w/ OCR
Response

You can eat hamburgers at
Angelo's Car Hop Service. ✓










Get the number of the boat with
three yellow and one red round
items hanging from it.

SS273 WH97 SS266 SS681 SS138

SS681 ✗

ConTextual: Evaluating Context-Sensitive Text-Rich Visual Reasoning in Large Multimodal Models

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng, in ICML, 2024.

#	Model	Method	Source	Date	<u>ALL</u>	Time	Shop.	Nav.	Abs.	App.	Web.	Info.	Misc. NS.
-	Human Performance	-	Link	2024-01-24	69.6	64.0	64.0	73.5	75.5	64.0	58.0	72.0	78.0
1	GPT-4o 🏆	LMM 	Link	2024-05-18	62.8	32.0	70.0	60.0	98.0	72.0	62.0	48.0	64.7
2	GPT-4o-mini-2024-07-18 🏆	LMM 	Link	2024-07-18	61.7	22.0	62.0	62.0	98.0	72.0	64.0	42.0	67.3
3	Claude-3.5-Sonnet-2024-06-20 🏆	LMM 	Link	2024-07-18	57.5	22.0	52.0	66.0	96.0	68.0	64.0	44.0	56.7
4	Gemini-1.5-Flash-Preview-0514	LMM 	Link	2024-05-18	56.0	30.0	51.0	52.1	84.0	63.0	63.2	42.8	61.7
5	Gemini-1.5-Pro-Preview-0514	LMM 	Link	2024-05-18	52.4	24.0	46.9	39.6	84.0	45.8	59.2	43.8	64.0
6	GPT-4V(ision)	LMM 	Link	2024-01-24	47.4	18.0	54.0	48.0	100.0	48.0	42.0	28.0	48.0
7	Gemini-Pro-Vision	LMM 	Link	2024-01-24	40.2	16.0	32.7	28.6	65.3	44.9	43.8	20.0	52.8
8	Claude-3-Opus-2024-02-29	LMM 	Link	2024-03-05	38.1	18.0	32.0	34.0	68.0	44.0	38.0	18.0	44.7
9	LLaVA-Next-34B	LMM 	Link	2024-03-05	36.8	10.0	36.0	30.6	66.0	36.0	28.0	12.0	51.3

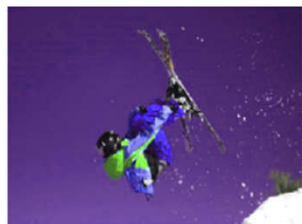
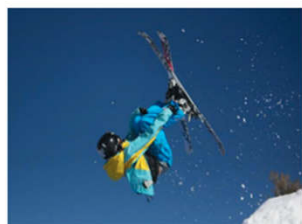
- ❖ GPT-4o is bad at time-reading and infographics but outperforms humans in abstract contexts (memes, etc.).
- ❖ Current models have limited capacity for fine-grained visual perception

Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.

Preferred data (Good prompting: Provide a detailed description of the image, focusing on ...)

The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....



Preferred response generation:

- Image: Original image
- Prompt: GPT-4 to generate image descriptions.
 - We test these prompts on MSCOCO samples.
- SFT on the preferred data alone can be similar to system-2 distillation.

Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.



Preferred data (Good prompting: Provide a detailed description of the image, focusing on ...)
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) Dis-preferred data (Bad prompting: Describe the image with imaginative objects that ...)
..... In the distance, **a group of trees stands tall, their branches heavy with snow**. Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.

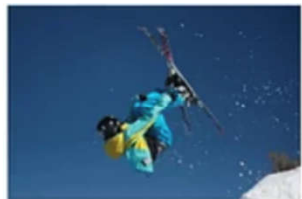


Dispreferred response generation:

1. Bad prompting: designed to elicit inaccurate descriptions by setting up a slightly different task (describe objects that would logically exist in the image) for the model.
2. Prompts are similarly generated using GPT-4 with human filtering.

Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.

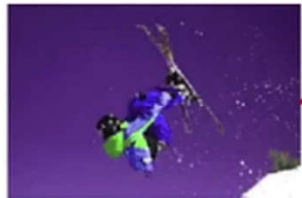


Preferred data (Good prompting: Provide a detailed description of the image, focusing on ...)

The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) Dis-preferred data (Bad prompting: Describe the image with imaginative objects that ...)

..... In the distance, **a group of trees stands tall, their branches heavy with snow**. Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.



(b) Dis-preferred data (Normal prompting + Image corruption: Color jitter.)

..... The skier, clad in a **vibrant purple jacket** and blue pants, The background is a stark contrast to the skier, with a **purple sky** that adds a sense of depth and dimension

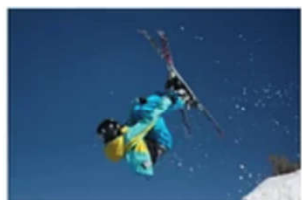


(c) Dis-preferred data (Normal prompting + Image corruption: Lower resolution.)

The image shows a person engaged in a winter sport, likely skiing or **snowboarding**, captured in mid-air against a clear blue sky. The individual is wearing a blue and yellow suit, The **person is holding onto a ski or snowboard**, which is also visible in the image. The **motion blur effect**

Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.



Preferred data (Good prompting: Provide a detailed description of the image, focusing on ...)
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) Dis-preferred data (Bad prompting: Describe the image with imaginative objects that ...)
..... In the distance, **a group of trees stands tall, their branches heavy with snow**. Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.



(b) Dis-preferred data (Normal prompting + Image corruption: Color jitter.)
..... The skier, clad in a **vibrant purple jacket** and blue pants, The background is a stark contrast to the skier, with a **purple sky** that adds a sense of depth and dimension



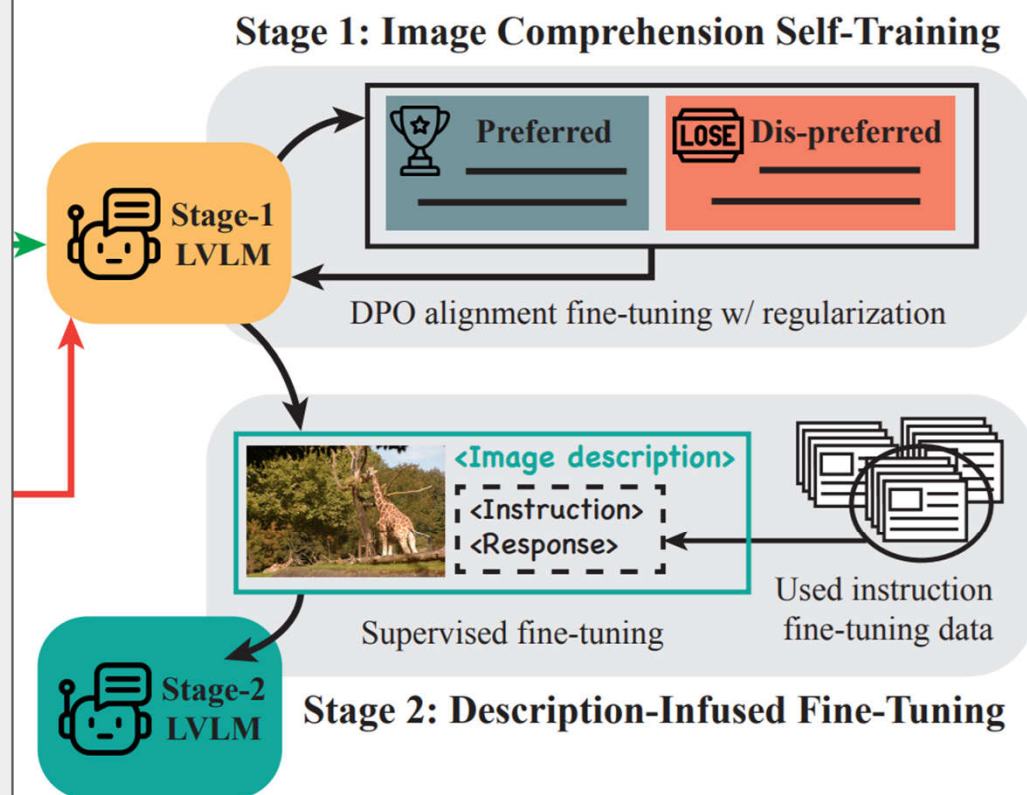
(c) Dis-preferred data (Normal prompting + Image corruption: Lower resolution.)
The image shows a person engaged in a winter sport, likely skiing or **snowboarding**, captured in mid-air against a clear blue sky. The individual is wearing a blue and yellow suit, The **person is holding onto a ski or snowboard**, which is also visible in the image. The **motion blur effect**

$$\text{Update } \theta_1 = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{(\mathbf{x}, \mathbf{y}_g, \mathbf{y}_b) \in D} \left[\underbrace{\ell \left(\lambda \log \frac{p_{\theta}(\mathbf{y}_g | \mathbf{x})}{p_{\theta_0}(\mathbf{y}_g | \mathbf{x})} - \lambda \log \frac{p_{\theta}(\mathbf{y}_b | \mathbf{x})}{p_{\theta_0}(\mathbf{y}_b | \mathbf{x})} \right)}_{\text{DPO}} - \underbrace{\alpha \log p_{\theta}(\mathbf{y}_g | \mathbf{x})}_{\text{Regularizer}} \right].$$

How can we improve fine-grained visual perception in reasoning?

Stage 2:

Fine-tune LVLM with generated detailed image description



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

Stage 2: Description-Infused Fine-Tuning

- Randomly select a small set (50k) data
- Infuse instructions with image description

Image description: {model description}
<original instruction>

for $i = 1, \dots, m$ **do**

Randomly sample $\mathbf{x}_{\text{des}} \sim \{\mathbf{x}_{\text{des}}^{(i)}\}_{i \in [M]}$.

Generate model image description $\mathbf{y}_{\text{des}} \sim p_{\boldsymbol{\theta}_t}(\cdot | \mathbf{v}^{(i)}, \mathbf{x}_{\text{des}})$.

Add $([\mathbf{y}_{\text{des}}, \mathbf{x}^{(i)}], \mathbf{y}^{(i)})$ to D_{des} .

end for

Update $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{(\mathbf{x}, \mathbf{y}) \in D_{\text{des}}} \ell(\log p_{\boldsymbol{\theta}}(\mathbf{y} | \mathbf{x}))$.