ICML
International Conference
On Machine Learning

EMMA

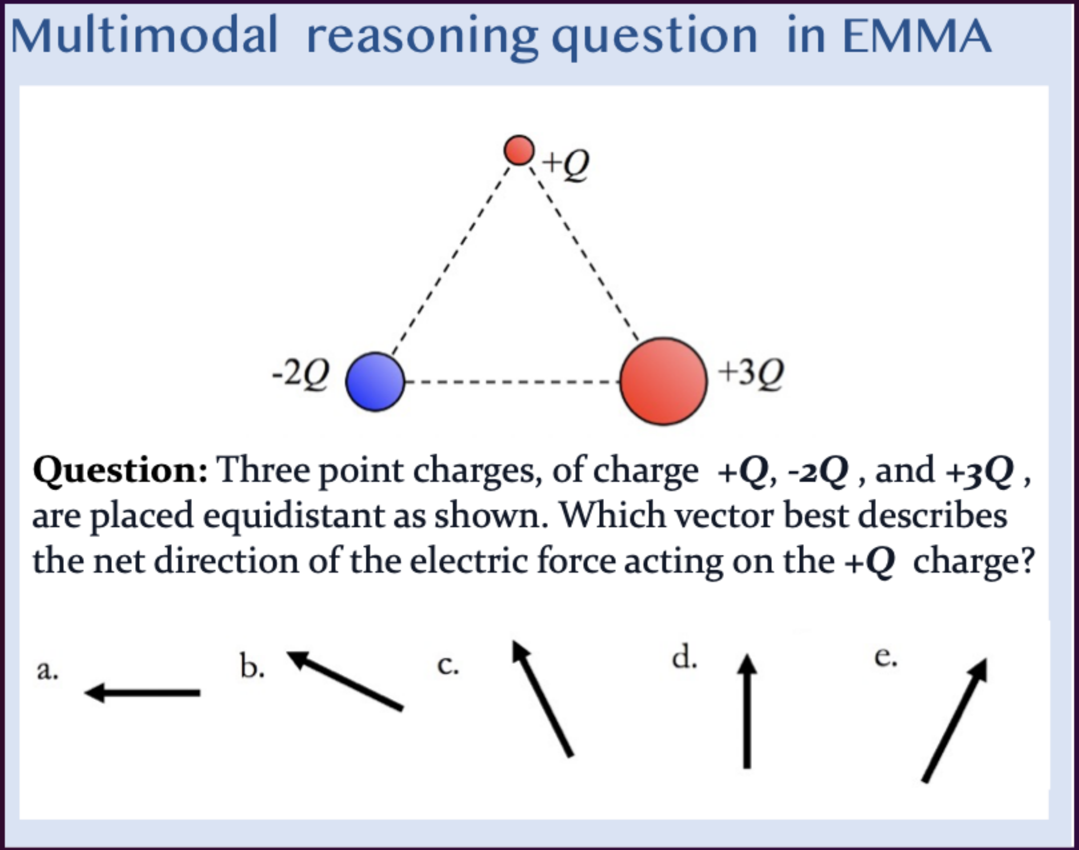
Can MLLMs Reason/Think in Multimodality?

Yunzhuo Hao*, Jiawei Gu*, Huichen Will Wang*, Linjie Li*,
Zhengyuan Yang, Lijuan Wang, Yu Cheng

*Equal contribution

UESTC 1956
SUN YAT-SEN UNIVERSITY
UNIVERSITY OF WASHINGTON 1861
Microsoft

Introduction

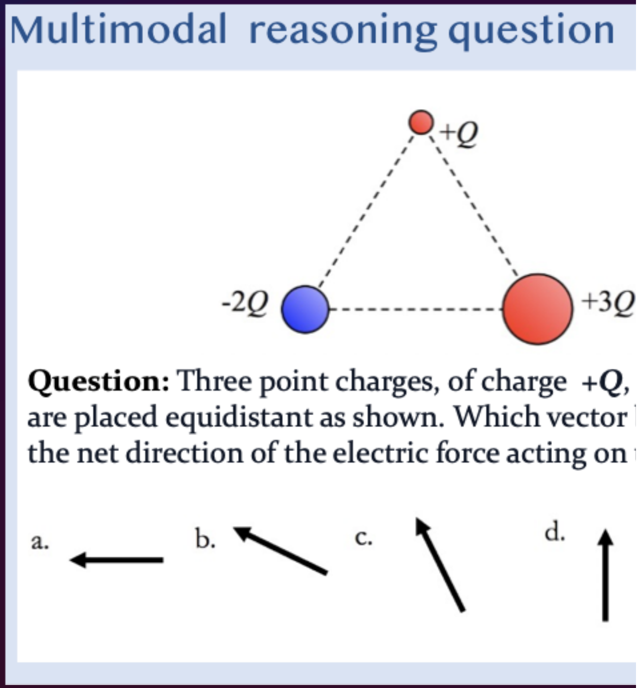


Multimodal reasoning question in EMMA

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?

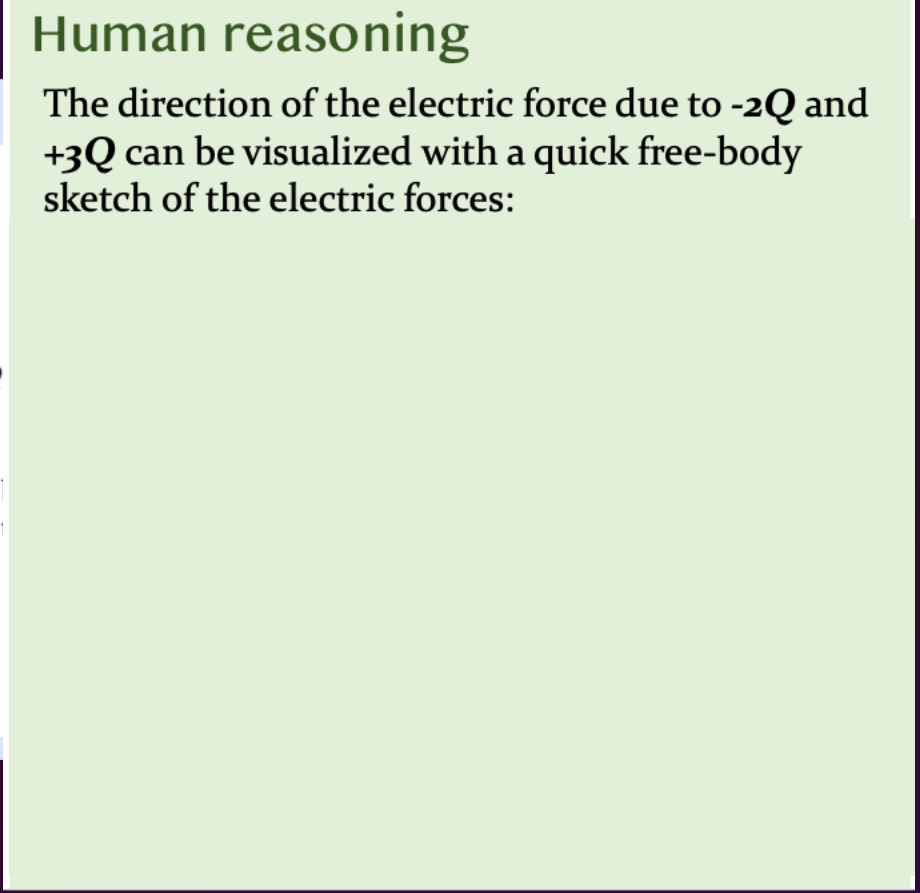
A physics problem that requires force analysis

Introduction



Multimodal reasoning question

Question: Three point charges, of charge +*Q*,
are placed equidistant as shown. Which vector
the net direction of the electric force acting on



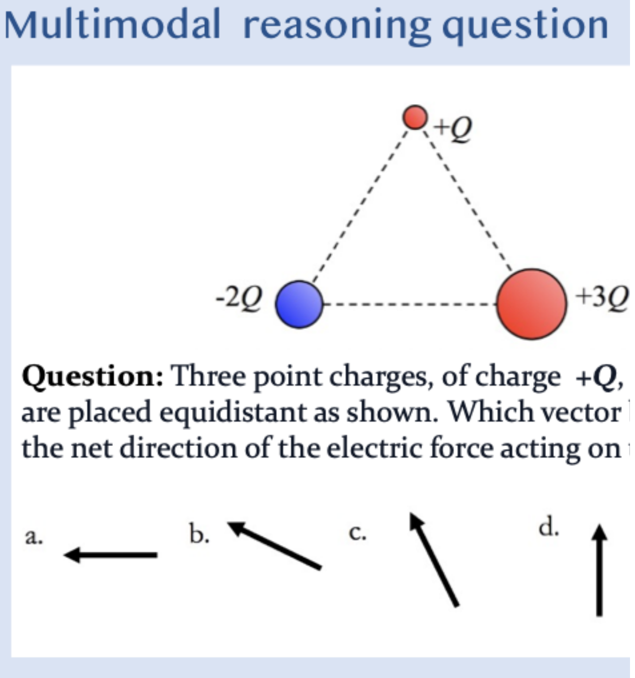
Human reasoning

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

Try a quick sketch

Force analysis between two positive and one negative

Introduction

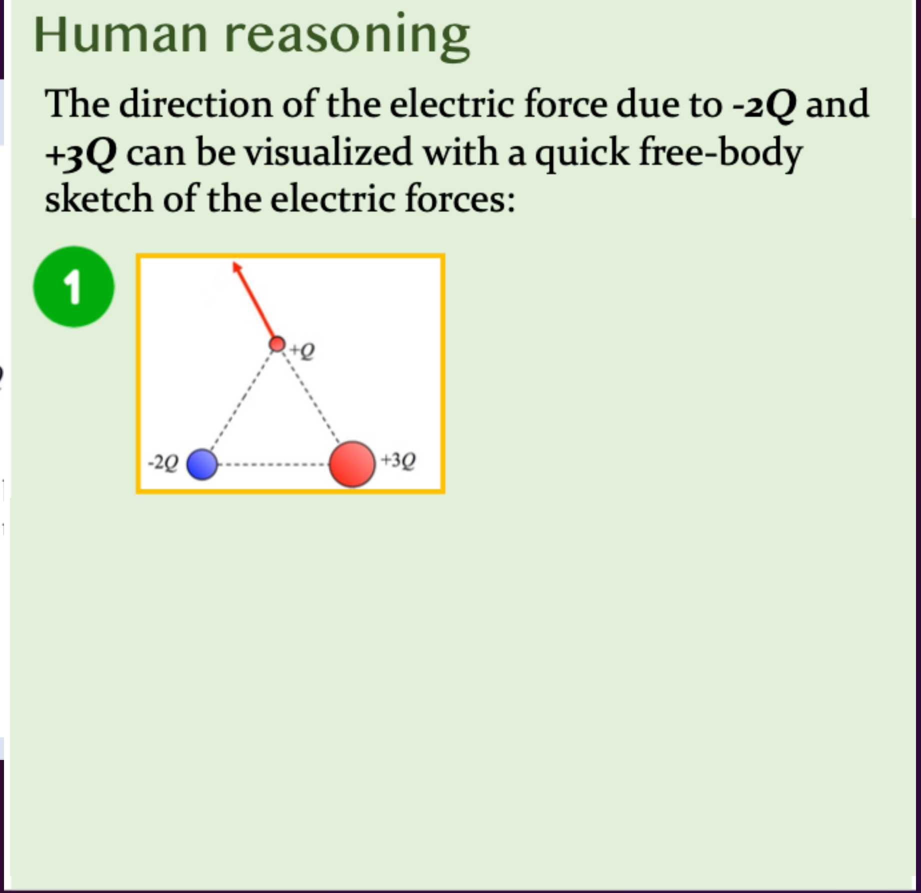


Multimodal reasoning question

Question: Three point charges, of charge $+Q$,

are placed equidistant as shown. Which vector

the net direction of the electric force acting on



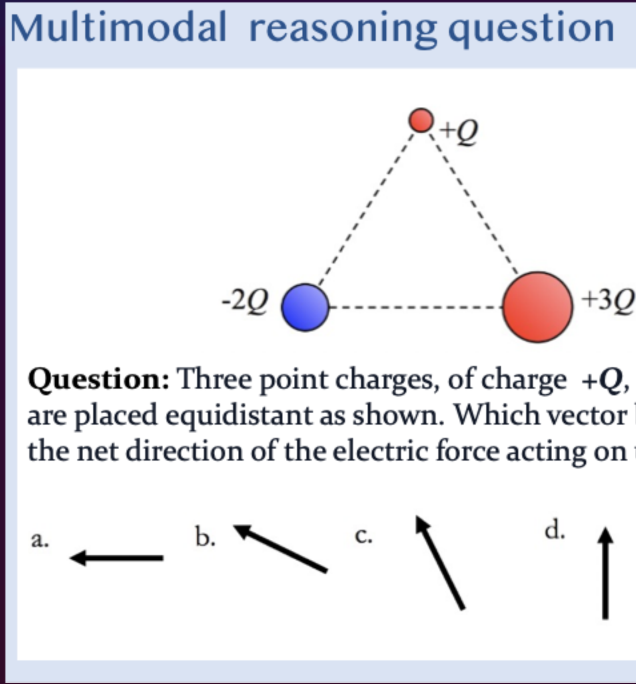
Human reasoning

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

"Alike repel"

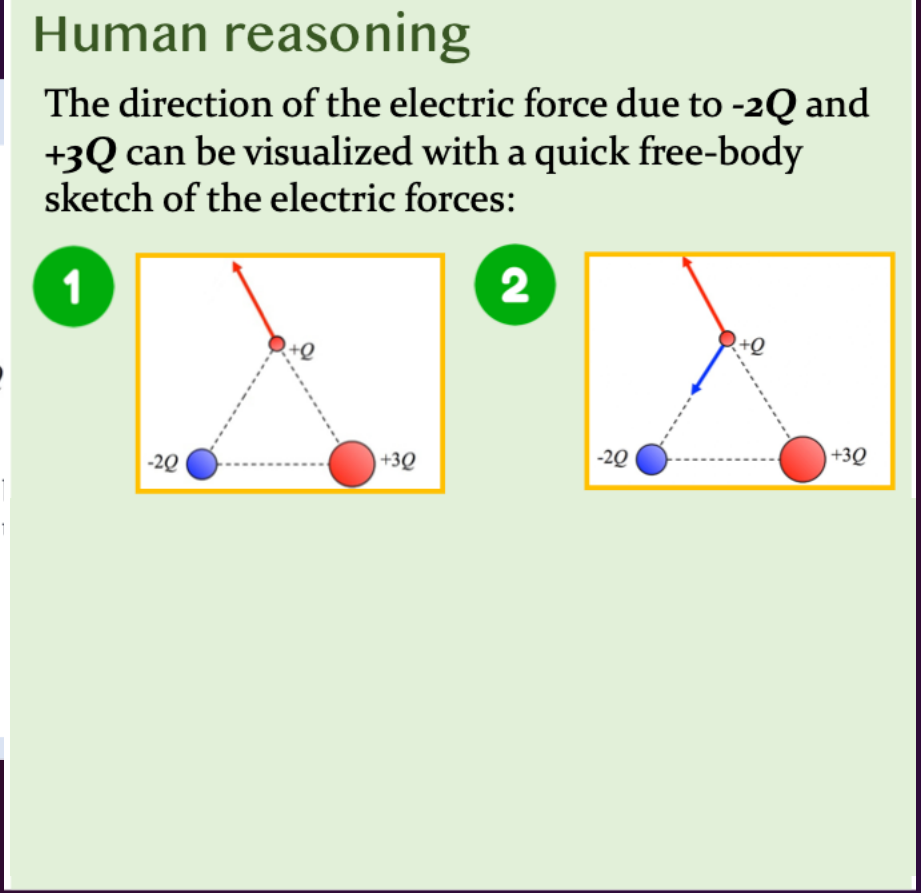
Force analysis between two positive and one negative

Introduction



Multimodal reasoning question

Question: Three point charges, of charge +*Q*, are placed equidistant as shown. Which vector the net direction of the electric force acting on



Human reasoning

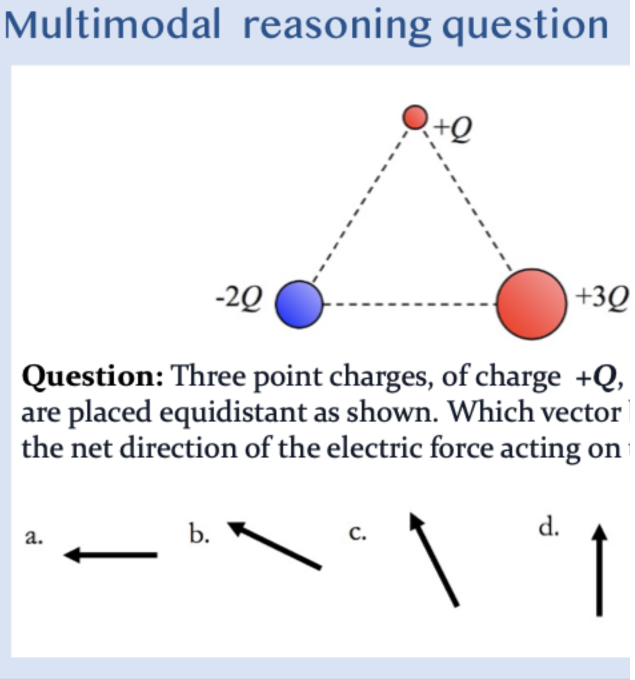
The direction of the electric force due to **-2*Q*** and **+3*Q*** can be visualized with a quick free-body sketch of the electric forces:

"Alike repel"

"Opposites attract"

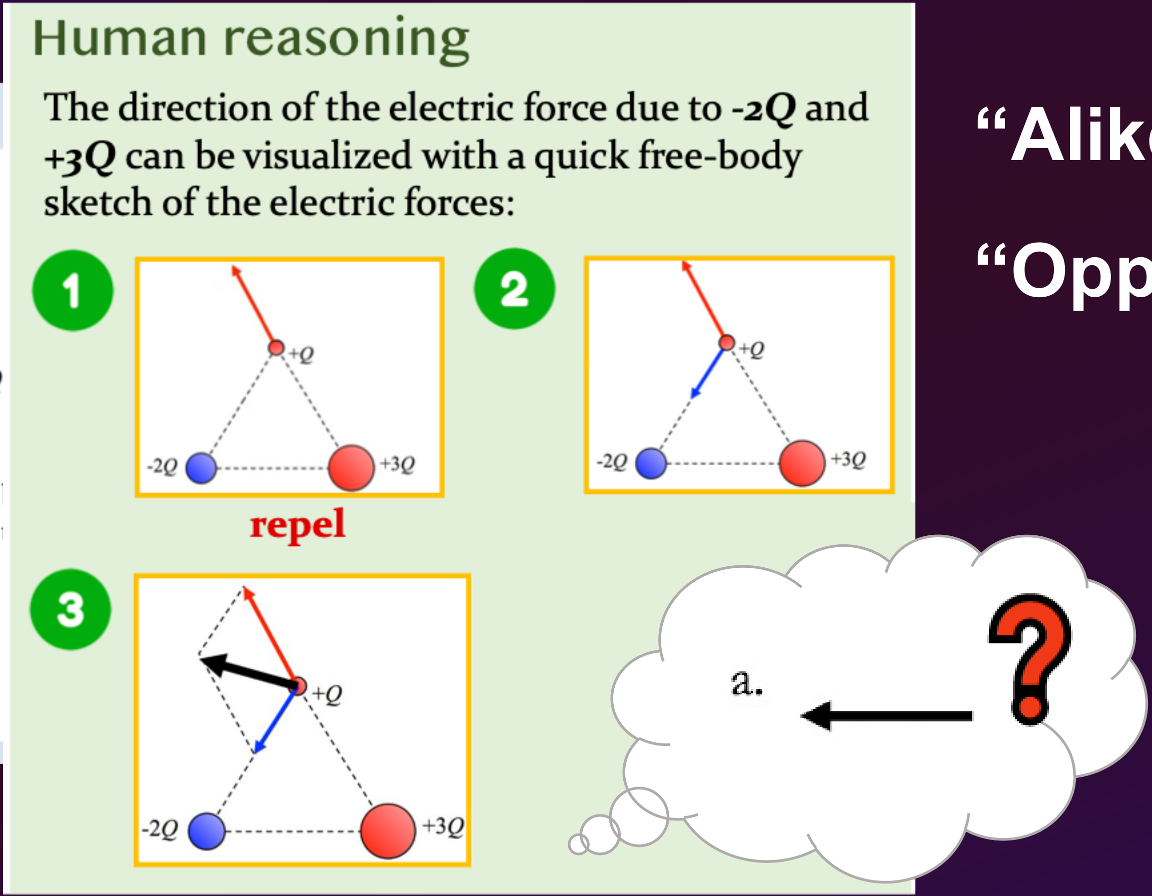
Force analysis between two positive and one negative

Introduction



Multimodal reasoning question

Question: Three point charges, of charge +*Q*, are placed equidistant as shown. Which vector
the net direction of the electric force acting on



Human reasoning

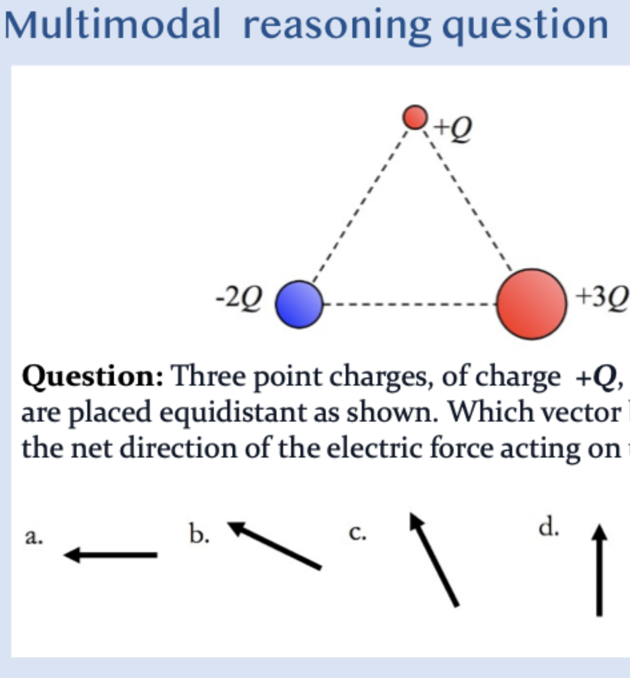
The direction of the electric force due to **-2Q** and **+3Q** can be visualized with a quick free-body sketch of the electric forces:

“Alike repel”

“Opposites attract”

Force analysis between two positive and one negative

Introduction



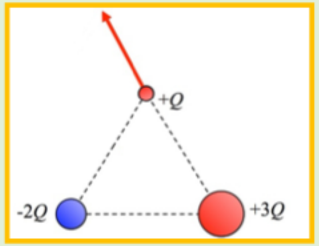
Multimodal reasoning question

Question: Three point charges, of charge $+Q$, are placed equidistant as shown. Which vector the net direction of the electric force acting on

Human reasoning

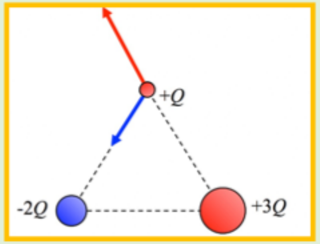
The direction of the electric force due to **-2Q** and **+3Q** can be visualized with a quick free-body sketch of the electric forces:

1

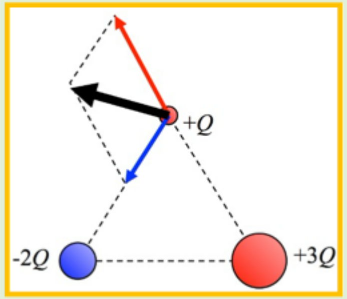


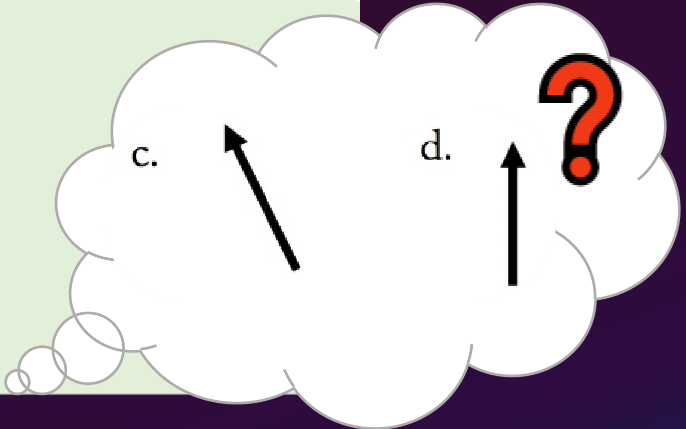
repel

2



3



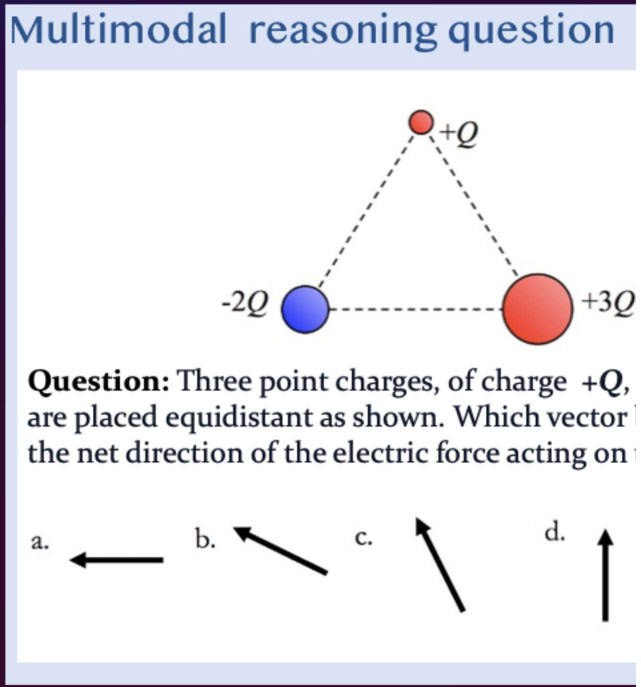


“Alike repel”

“Opposites attract”

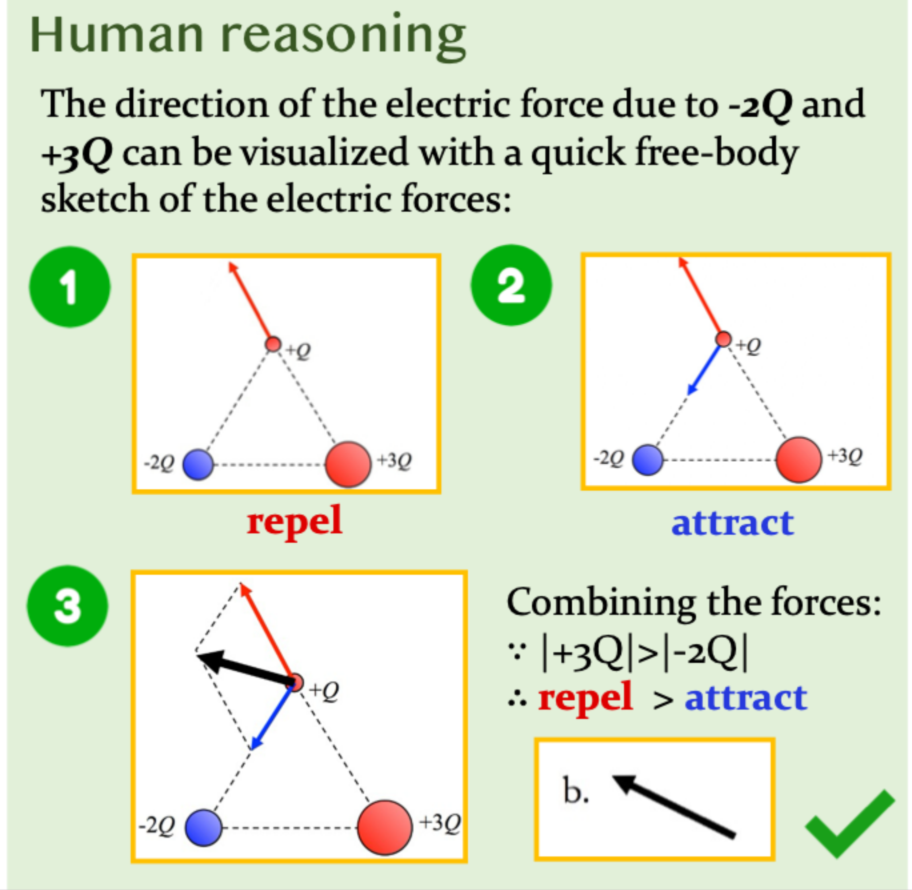
Force analysis between two positive and one negative

What is MM Reasoning?



Multimodal reasoning question

Question: Three point charges, of charge +Q, are placed equidistant as shown. Which vector the net direction of the electric force acting on



Human reasoning

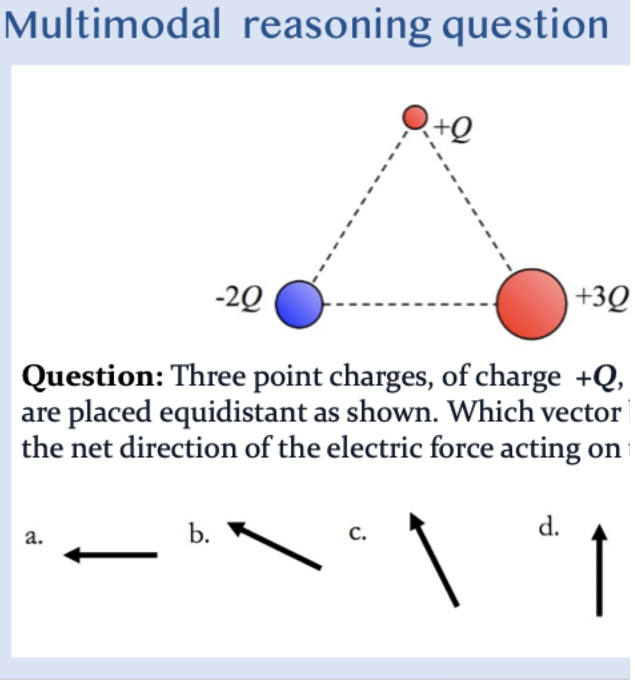
The direction of the electric force due to **-2Q** and **+3Q** can be visualized with a quick free-body sketch of the electric forces:

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

A back-and-forth process between interpreting visual inputs and applying multimodal reasoning steps

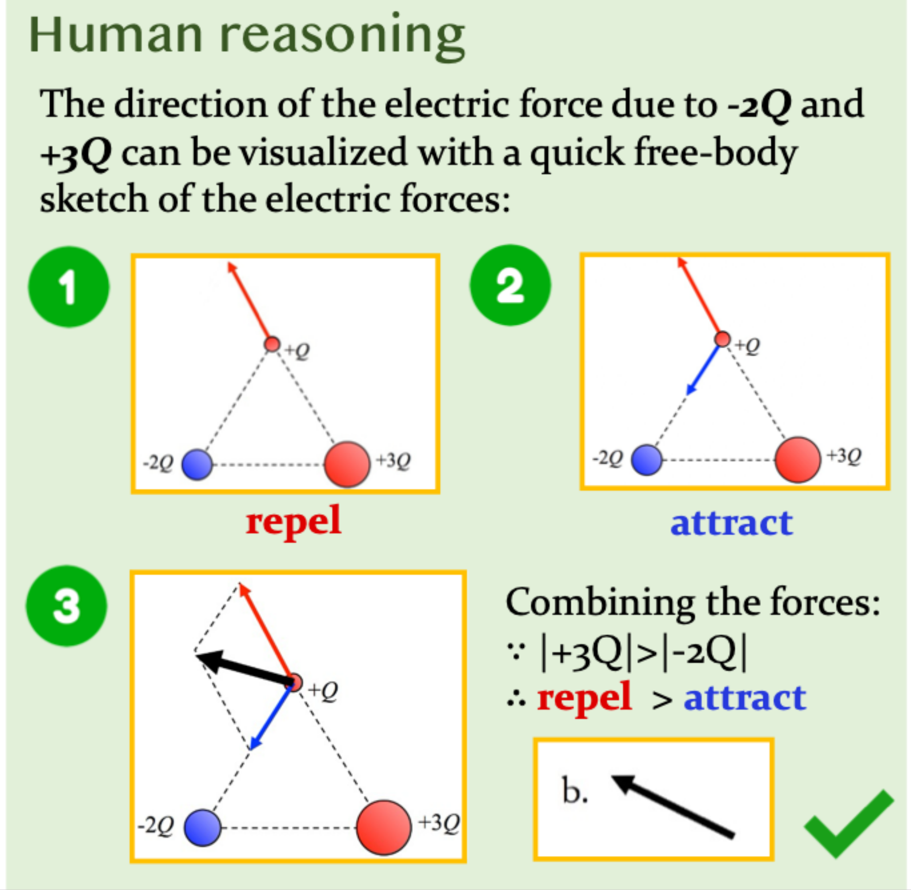
[Text Clue] → [Visual Input] → [Sketching] → [Option Matching] → [Answer]

What is MM Reasoning?



Multimodal reasoning question

Question: Three point charges, of charge $+Q$, are placed equidistant as shown. Which vector the net direction of the electric force acting on



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:

repel

attract

Combining the forces:

$\because |+3Q|>|-2Q|$

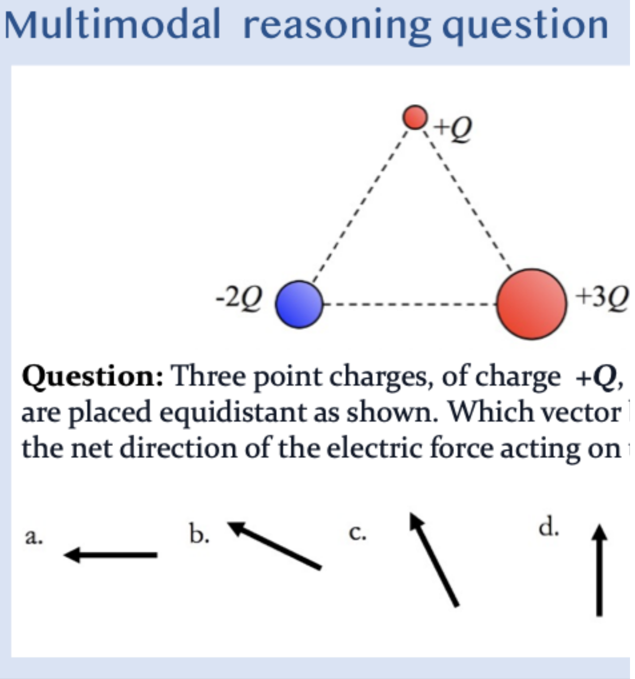
$\therefore$ **repel** > **attract**

A back-and-forth reasoning process between Multimodality

Multimodal reasoning

Visual aids are often integral or more efficient for the solution.

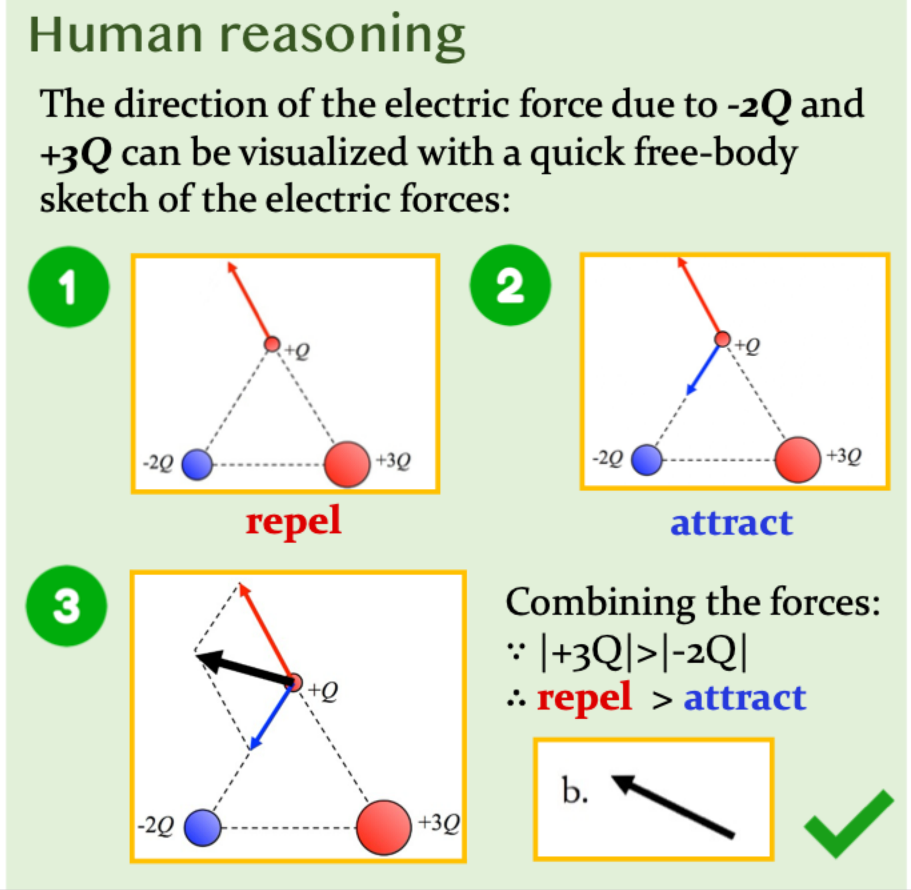
Introduction



Multimodal reasoning question

Question: Three point charges, of charge +Q, are placed equidistant as shown. Which vector the net direction of the electric force acting on

a. b. c. d.



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 repel

② attract

③ Combining the forces:

∵ |+3Q|>|-2Q|

∴ **repel** > **attract**

b.

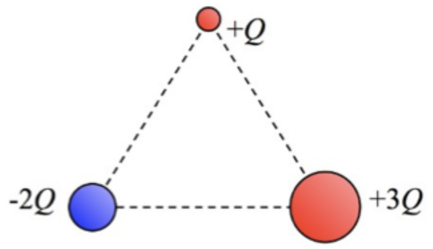
A back-and-forth reasoning process between Multimodality

Multimodal reasoning

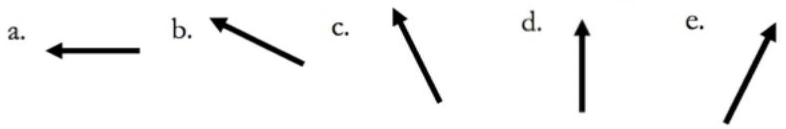
Can MLLMs perform this type of "cross-modal" MM reasoning?

Introduction

Multimodal reasoning question in EMMA

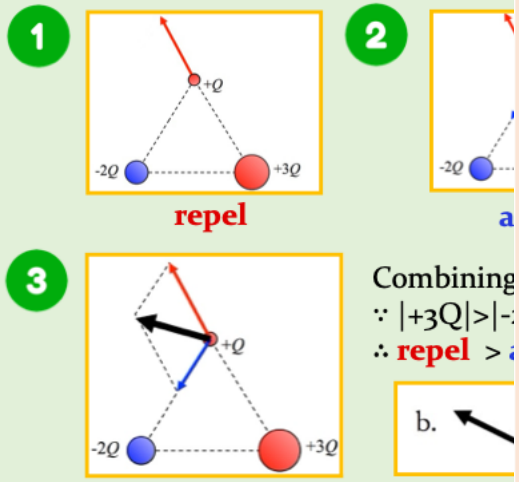


Question: Three point charges, of charge **$+Q$**, **$-2Q$**, and **$+3Q$**, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the **$+Q$** charge?



Human reasoning

The direction of the electric force due t
$+3Q$ can be visualized with a quick free
sketch of the electric forces:



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

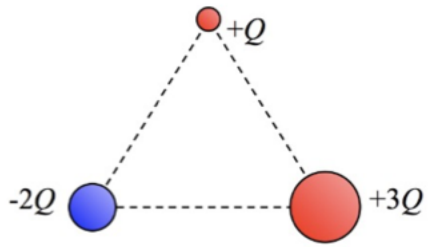
……

Error: The issue arose from not correctly identifying the force direction from **+3Q** to **+Q**, which causes the net force to point ↖, not downward-right↘.

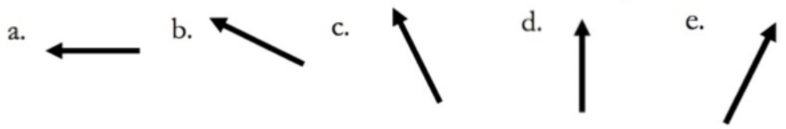
The answer seems to be negative.

Introduction

Multimodal reasoning question in EMMA

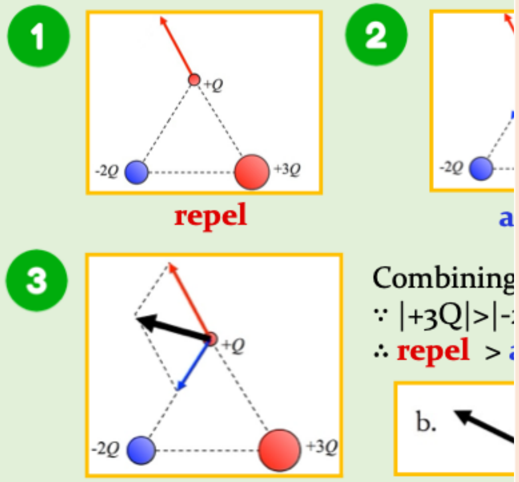


Question: Three point charges, of charge $+Q$, $-2Q$, and $+3Q$, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the $+Q$ charge?



Human reasoning

The direction of the electric force due t
$+3Q$ can be visualized with a quick free
sketch of the electric forces:



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: The issue arose from not correctly identifying the force direction from **+3Q** to **+Q**, which causes the net force to point ↖, not downward-right↘.

"Alike repel" ↖

The Model fails to identify the direction of repulsion.

Is this common?

Is this common?

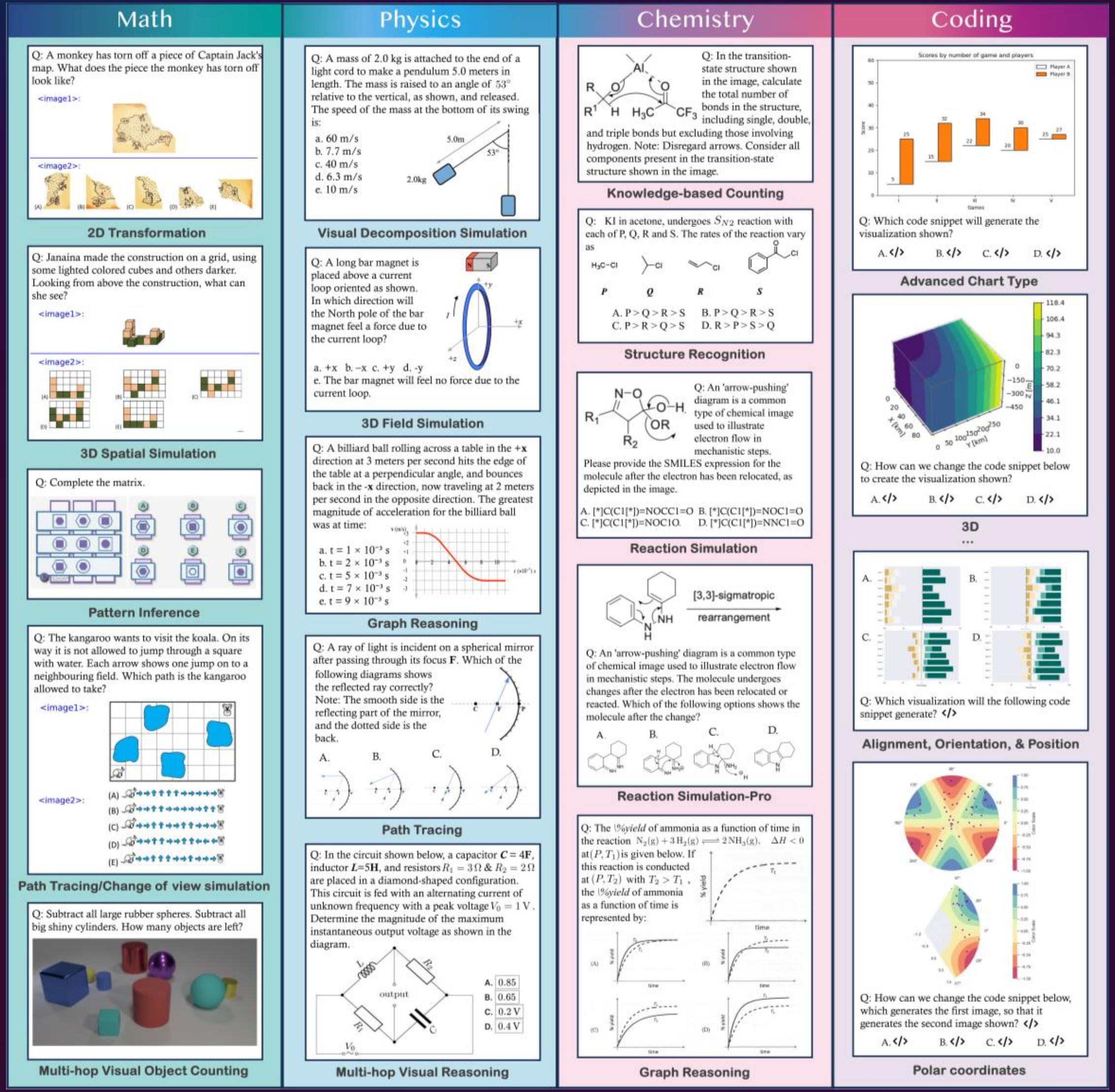
Can MLLMs Reason in Multimodality?

We need !

An Enhanced MultiModal ReAsoning Benchmark!

EMMA

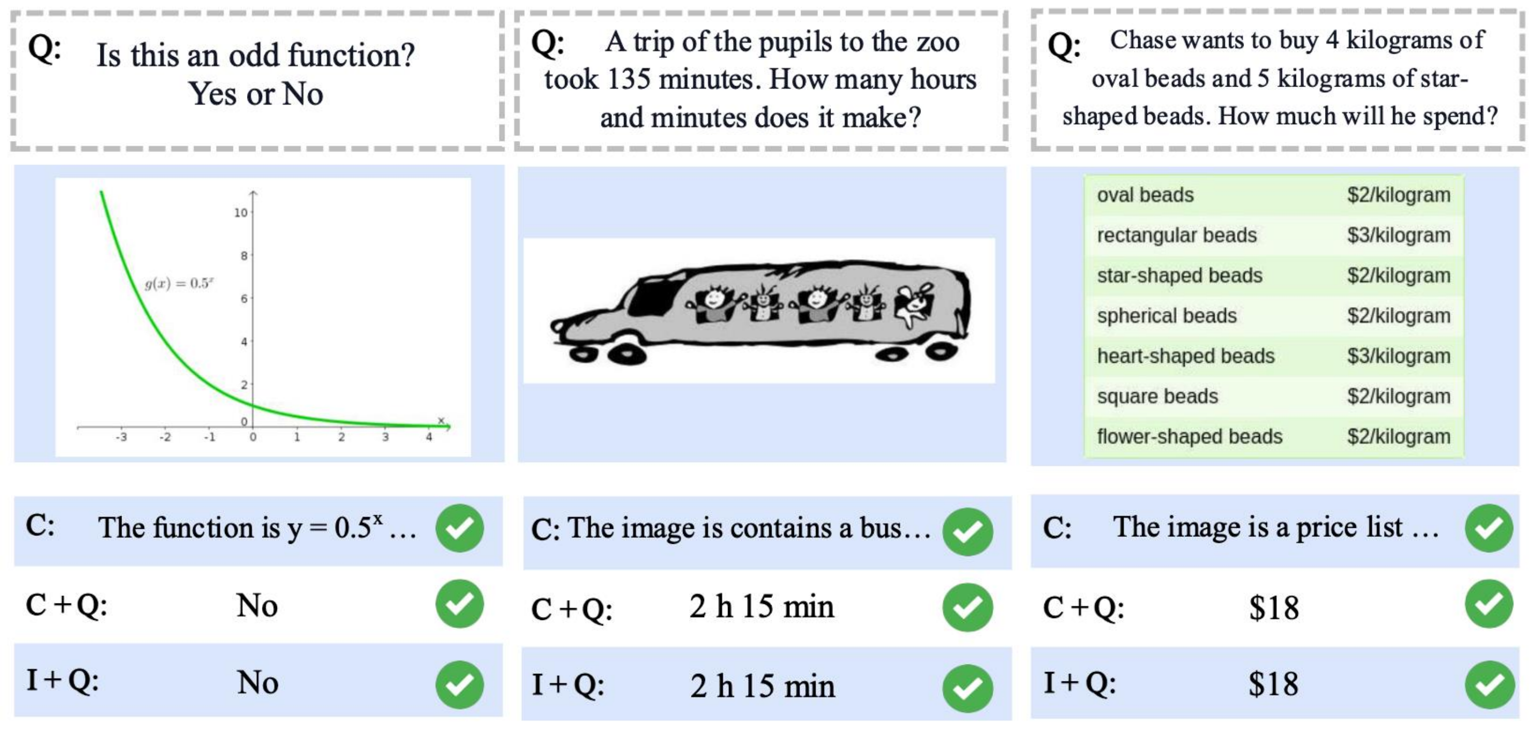
Homepage



- **Demands cross-modal reasoning**
- **Not solvable by single-modality reasoning**
- **Serve as an enhanced test suite for evaluating MLLMs' reasoning capabilities**

EMMA vs Others

Homepage



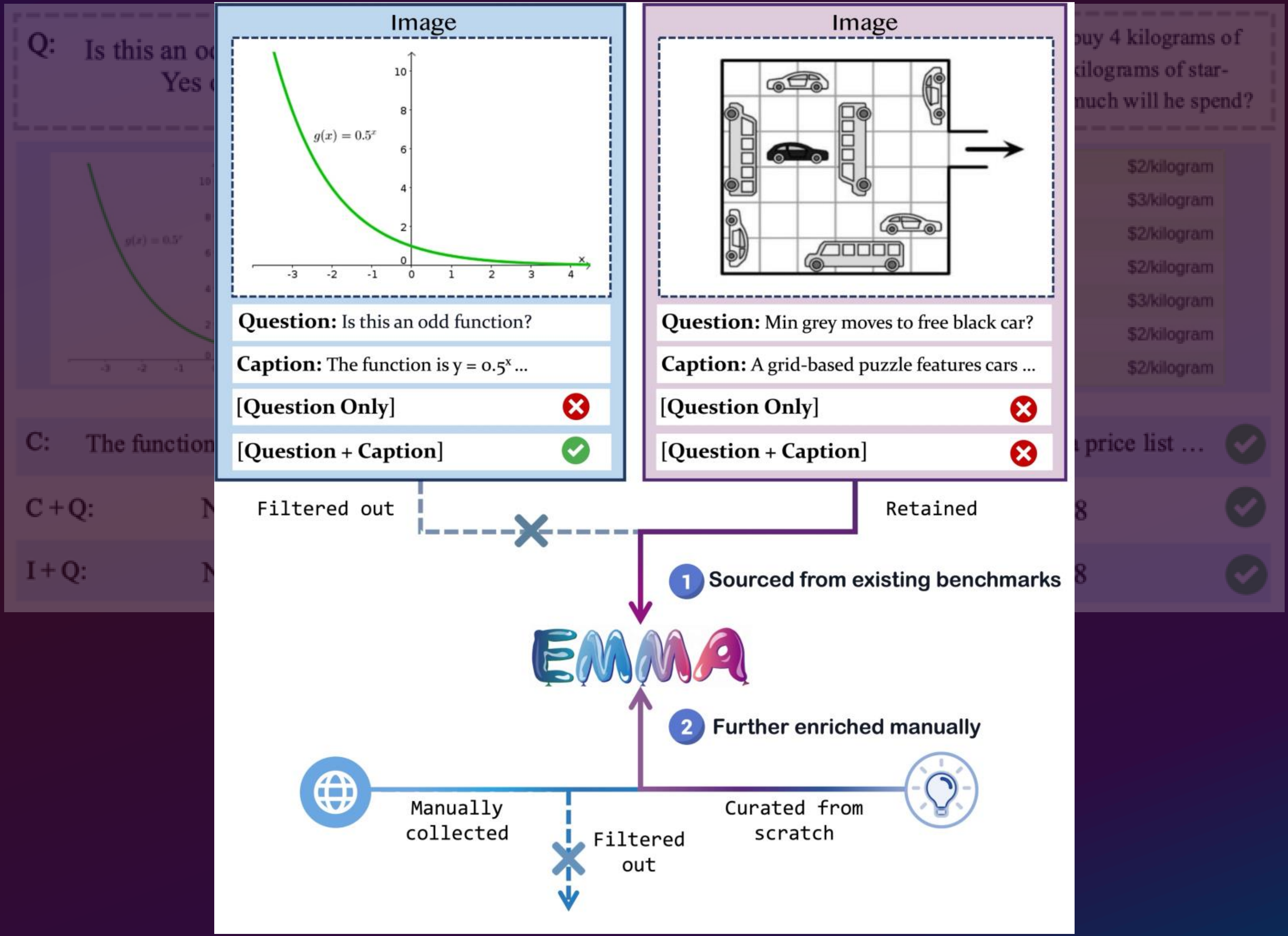
Q: Is this an odd function? Yes or No	Q: A trip of the pupils to the zoo took 135 minutes. How many hours and minutes does it make?	Q: Chase wants to buy 4 kilograms of oval beads and 5 kilograms of star-shaped beads. How much will he spend?
C: The function is $y = 0.5^x$ … ✓	C: The image is contains a bus… ✓	C: The image is a price list … ✓
C + Q: No ✓	C + Q: 2 h 15 min ✓	C + Q: \$18 ✓
I + Q: No ✓	I + Q: 2 h 15 min ✓	I + Q: \$18 ✓

oval beads	\$2/kilogram
rectangular beads	\$3/kilogram
star-shaped beads	\$2/kilogram
spherical beads	\$2/kilogram
heart-shaped beads	\$3/kilogram
square beads	\$2/kilogram
flower-shaped beads	\$2/kilogram

Existing benchmarks (such as *MathVista, MMMU*, etc.) often emphasize text-dominant reasoning or rely on shallow visual cues, failing to adequately assess integrated visual and textual reasoning.

EMMA vs Others

Homepage



Image

Question: Is this an odd function?

Caption: The function is $y = 0.5^x$ ...

[Question Only] ✗

[Question + Caption] ✓

Image

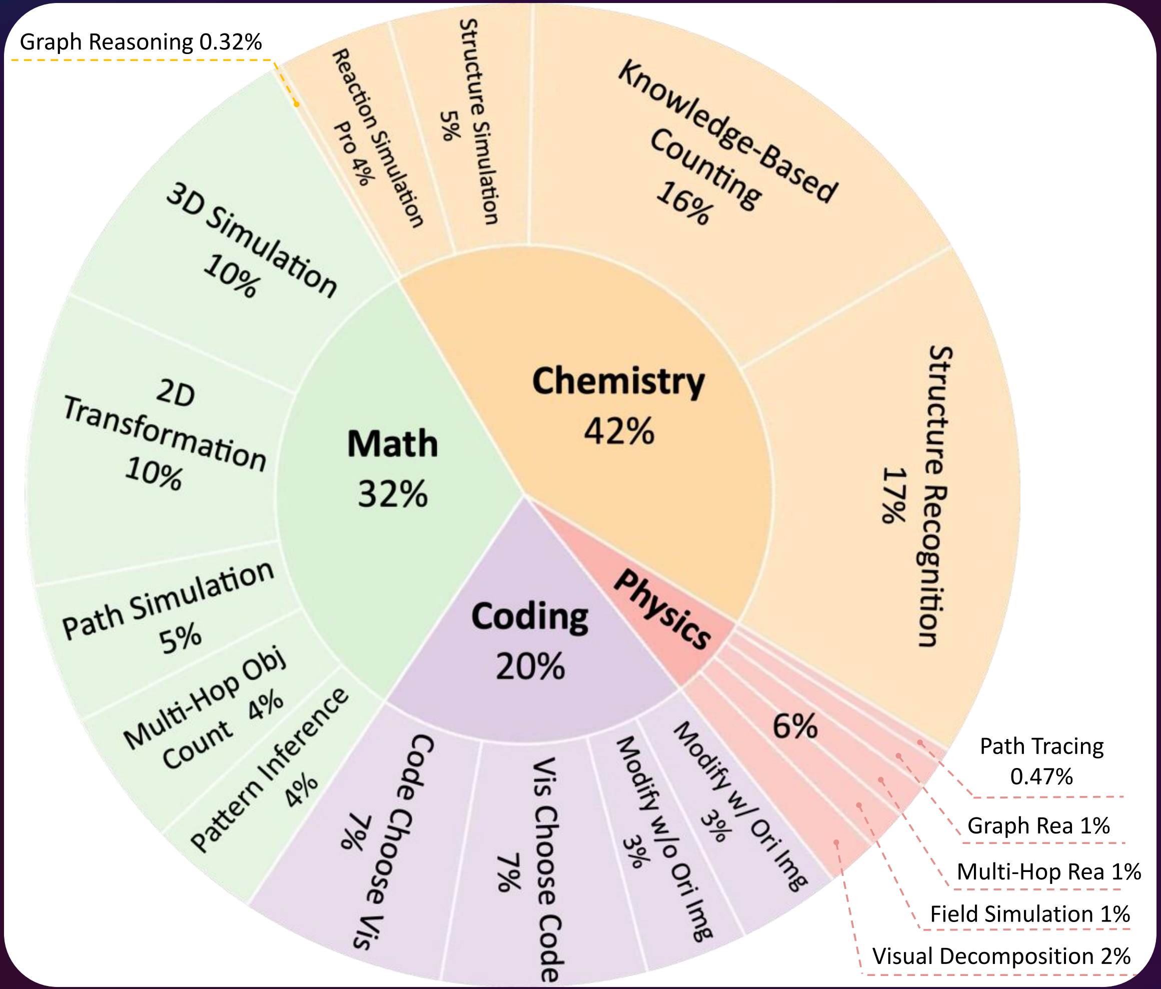
Question: Min grey moves to free black car?

Caption: A grid-based puzzle features cars ...

[Question Only] ✗

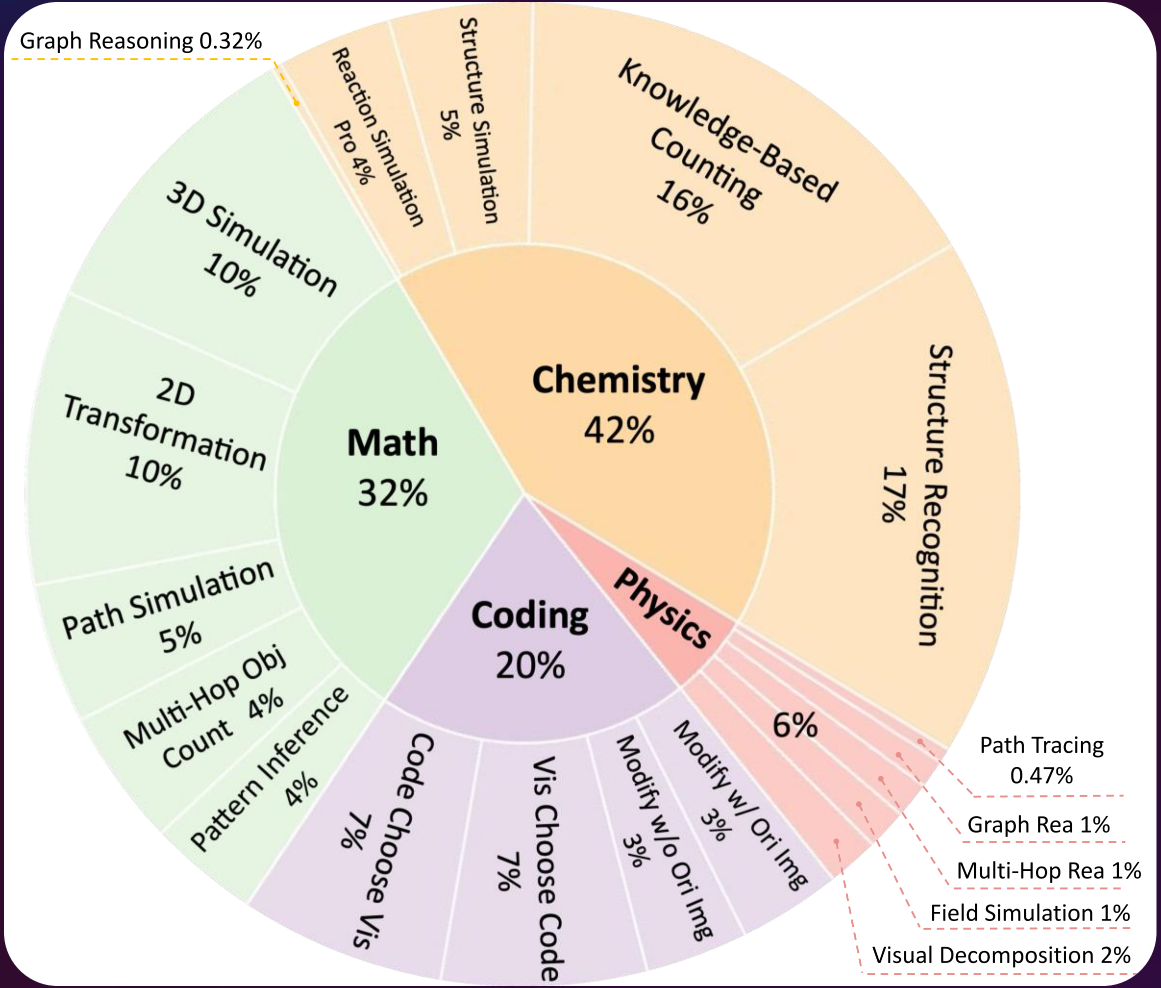
[Question + Caption] ✗

An Enhanced MultiModal ReAsoning Benchmark



- **2,788** total problems
- **1,796** are newly constructed
- Covers math, physics, chemistry, coding

An Enhanced MultiModal ReAsoning Benchmark



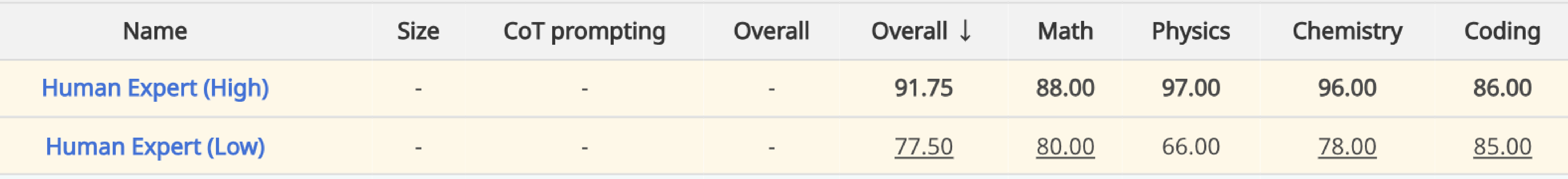
- **2,788 total problems**
- **1,796 are newly constructed**
- **Covers math, physics, chemistry, coding**

EMMA-mini:

A more balanced subset of 400 questions randomly sampled (100 /subject).

Human Experts Performance

Leaderboard



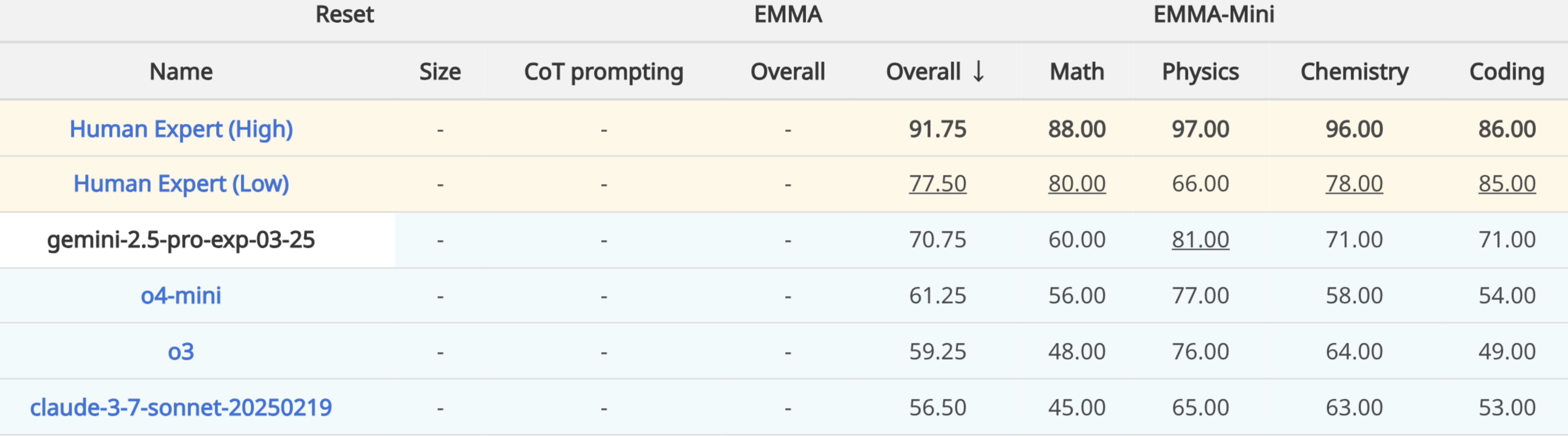
Name	Size	CoT prompting	Overall	Overall ↓	Math	Physics	Chemistry	Coding
Human Expert (High)	-	-	-	91.75	88.00	97.00	96.00	86.00
Human Expert (Low)	-	-	-	77.50	80.00	66.00	78.00	85.00

- To estimate expert-level performance on EMMA-mini, we hire a group of human experts per subject and report their average score.

- This score serves as a baseline contextualizing model performance.

SOTA and Tool-used Models

Leaderboard

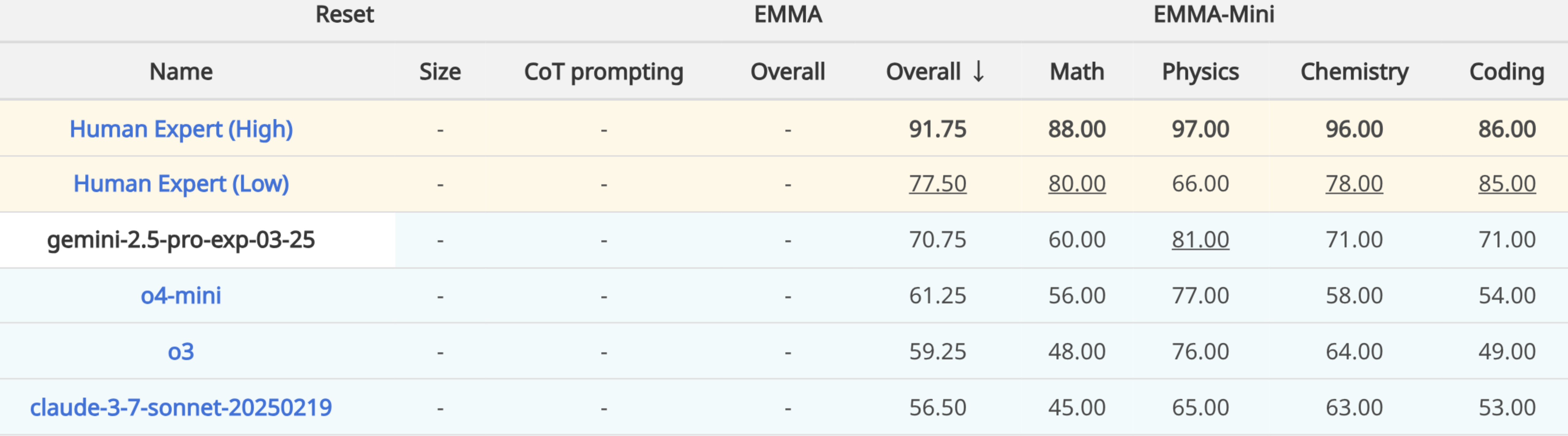


Reset			EMMA	EMMA-Mini				
Name	Size	CoT prompting	Overall	Overall ↓	Math	Physics	Chemistry	Coding
Human Expert (High)	-	-	-	**91.75**	**88.00**	**97.00**	**96.00**	**86.00**
Human Expert (Low)	-	-	-	77.50	80.00	66.00	78.00	85.00
gemini-2.5-pro-exp-03-25	-	-	-	70.75	60.00	81.00	71.00	71.00
o4-mini	-	-	-	61.25	56.00	77.00	58.00	54.00
o3	-	-	-	59.25	48.00	76.00	64.00	49.00
claude-3-7-sonnet-20250219	-	-	-	56.50	45.00	65.00	63.00	53.00

Even the most advanced model Gemini-2.5-pro, or the performance of o3/o4-mini models that can invoke visual tools still lags behind human experts by more than 20%.

A Closer Look

Leaderboard

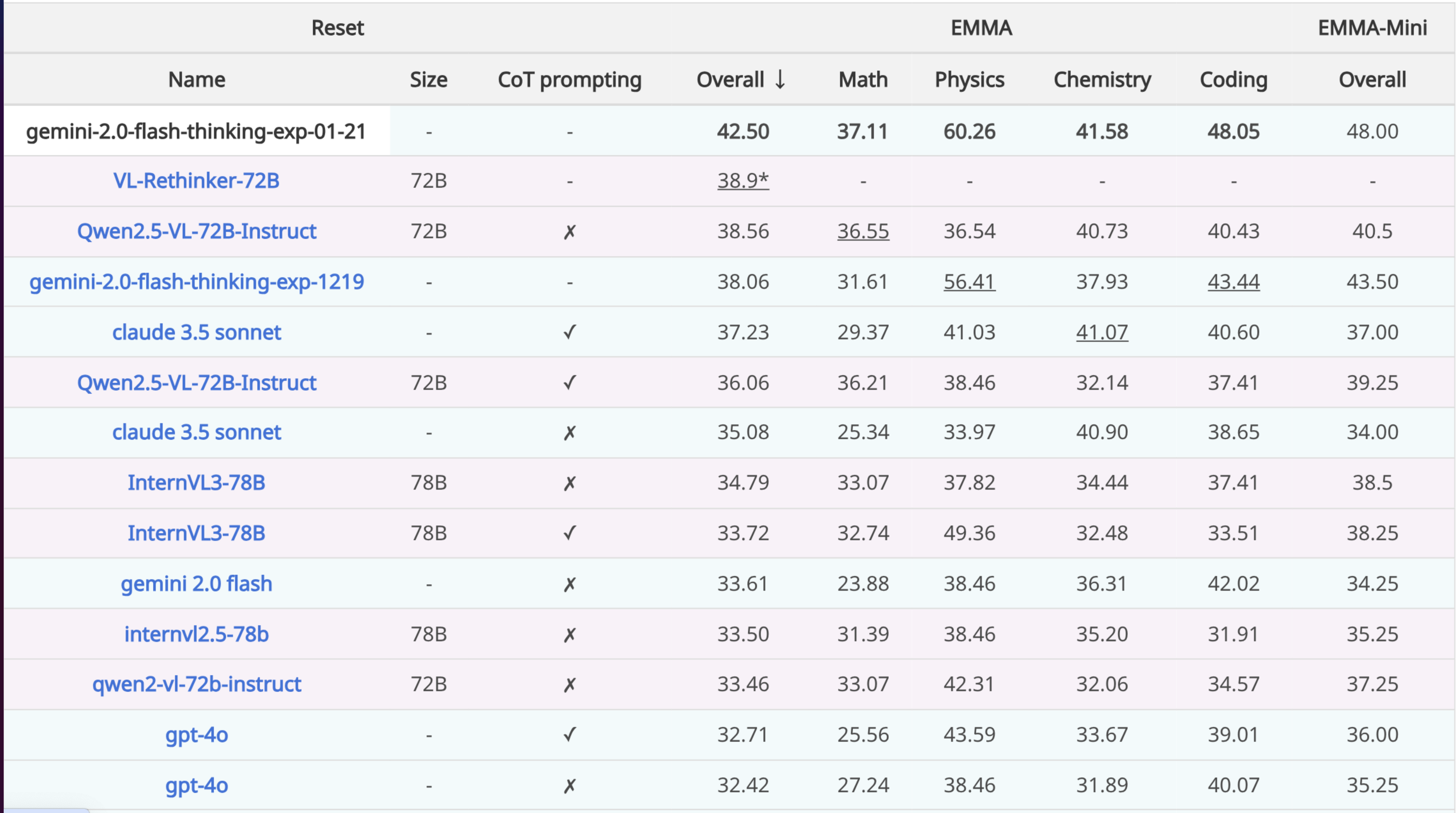


Reset			EMMA	EMMA-Mini				
Name	Size	CoT prompting	Overall	Overall ↓	Math	Physics	Chemistry	Coding
Human Expert (High)	-	-	-	**91.75**	**88.00**	**97.00**	**96.00**	**86.00**
Human Expert (Low)	-	-	-	77.50	80.00	66.00	78.00	85.00
gemini-2.5-pro-exp-03-25	-	-	-	70.75	60.00	81.00	71.00	71.00
o4-mini	-	-	-	61.25	56.00	77.00	58.00	54.00
o3	-	-	-	59.25	48.00	76.00	64.00	49.00
claude-3-7-sonnet-20250219	-	-	-	56.50	45.00	65.00	63.00	53.00

- **Physics: Though requiring high-level scientific knowledge, Gemini 2.5 Pro outperforms**
- **Math: Less knowledge-intensive and more intuitive for humans, yet the performance gap remains large**

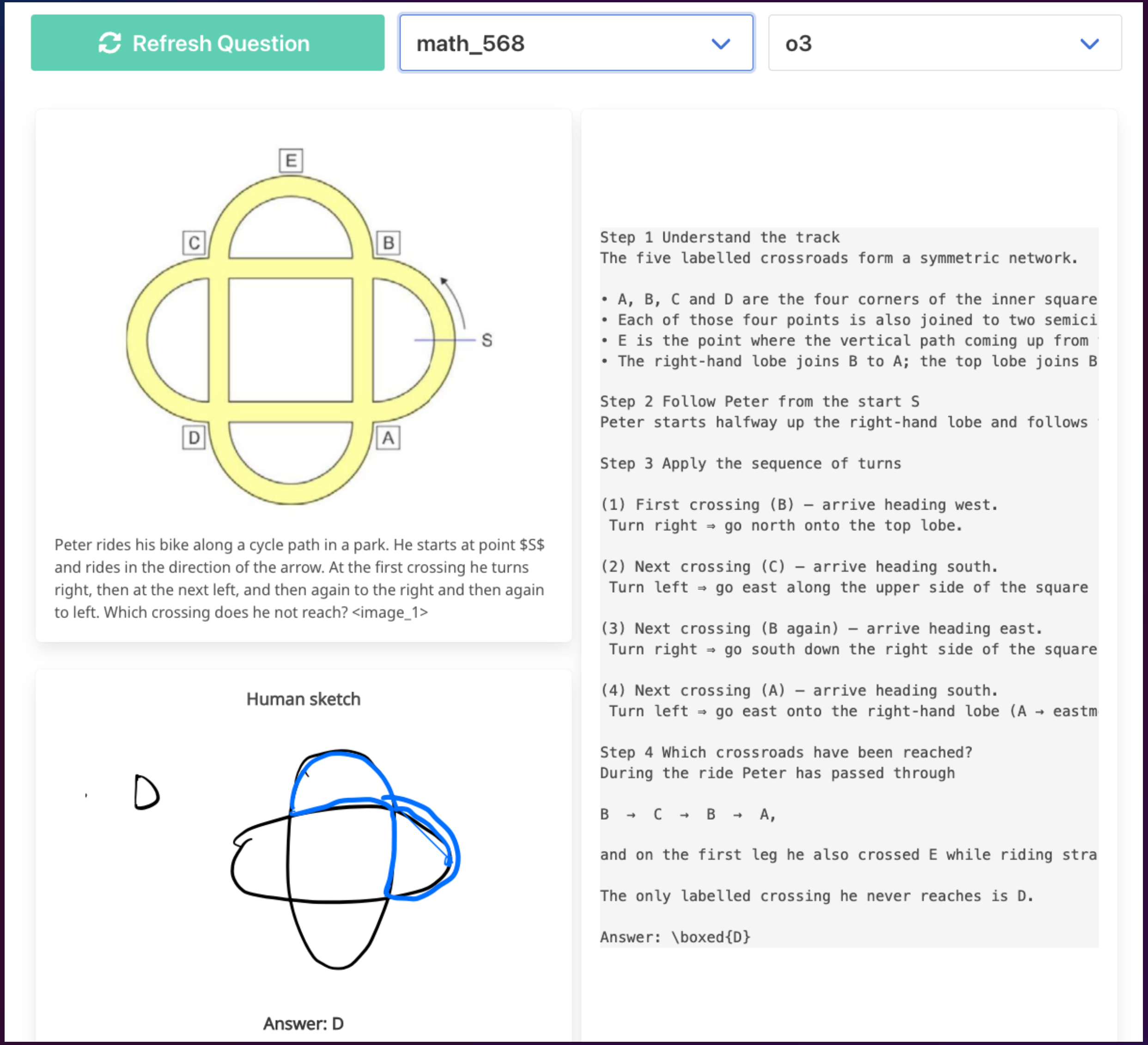
Model Performance(EMMA-full)

Leaderboard



Reset			EMMA					EMMA-Mini
Name	Size	CoT prompting	Overall ↓	Math	Physics	Chemistry	Coding	Overall
gemini-2.0-flash-thinking-exp-01-21	-	-	**42.50**	**37.11**	**60.26**	**41.58**	**48.05**	48.00
VL-Rethinker-72B	72B	-	38.9*	-	-	-	-	-
Qwen2.5-VL-72B-Instruct	72B	✗	38.56	36.55	36.54	40.73	40.43	40.5
gemini-2.0-flash-thinking-exp-1219	-	-	38.06	31.61	56.41	37.93	43.44	43.50
claude 3.5 sonnet	-	✓	37.23	29.37	41.03	41.07	40.60	37.00
Qwen2.5-VL-72B-Instruct	72B	✓	36.06	36.21	38.46	32.14	37.41	39.25
claude 3.5 sonnet	-	✗	35.08	25.34	33.97	40.90	38.65	34.00
InternVL3-78B	78B	✗	34.79	33.07	37.82	34.44	37.41	38.5
InternVL3-78B	78B	✓	33.72	32.74	49.36	32.48	33.51	38.25
gemini 2.0 flash	-	✗	33.61	23.88	38.46	36.31	42.02	34.25
internvl2.5-78b	78B	✗	33.50	31.39	38.46	35.20	31.91	35.25
qwen2-vl-72b-instruct	72B	✗	33.46	33.07	42.31	32.06	34.57	37.25
gpt-4o	-	✓	32.71	25.56	43.59	33.67	39.01	36.00
gpt-4o	-	✗	32.42	27.24	38.46	31.89	40.07	35.25

Human Sketch vs Model Response



Peter rides his bike along a cycle path in a park. He starts at point S and rides in the direction of the arrow. At the first crossing he turns right, then at the next left, and then again to the right and then again to left. Which crossing does he not reach? <image_1>

Human sketch

Answer: D

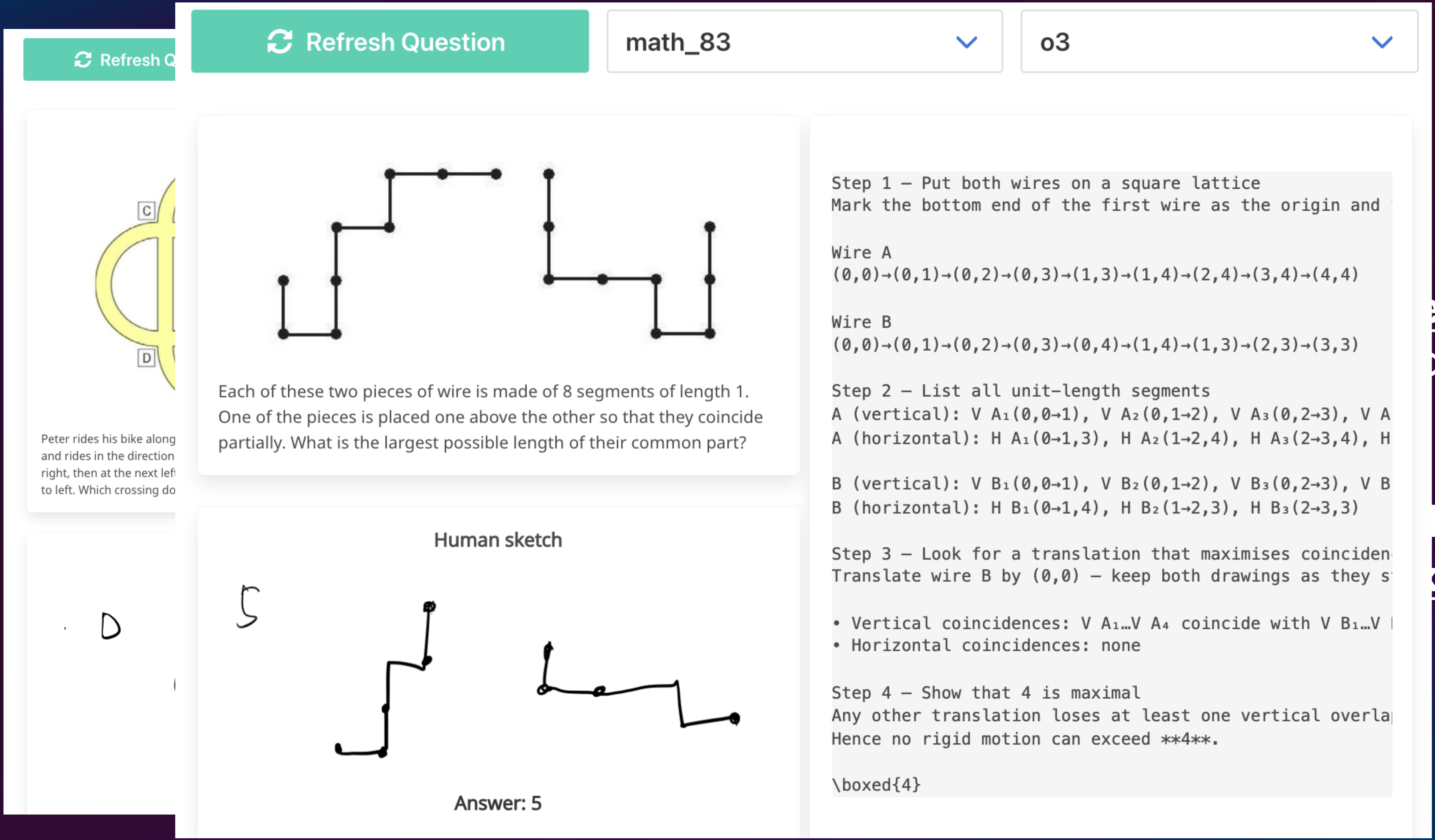
```
Step 1 Understand the track
The five labelled crossroads form a symmetric network.

• A, B, C and D are the four corners of the inner square
• Each of those four points is also joined to two semici
• E is the point where the vertical path coming up from
• The right-hand lobe joins B to A; the top lobe joins B

Step 2 Follow Peter from the start S
Peter starts halfway up the right-hand lobe and follows

Step 3 Apply the sequence of turns

(1) First crossing (B) – arrive heading west.
 Turn right ⇒ go north onto the top lobe.

(2) Next crossing (C) – arrive heading south.
 Turn left ⇒ go east along the upper side of the square

(3) Next crossing (B again) – arrive heading east.
 Turn right ⇒ go south down the right side of the square

(4) Next crossing (A) – arrive heading south.
 Turn left ⇒ go east onto the right-hand lobe (A → eastm

Step 4 Which crossroads have been reached?
During the ride Peter has passed through

B  →  C  →  B  →  A,

and on the first leg he also crossed E while riding stra

The only labelled crossing he never reaches is D.

Answer: \boxed{D}
```

Observation:

Human experts tend to use simple hand-drawn sketches for "visual thinking and spatial simulation".

Models (such as o3) rely on detailed, structured textual steps for reasoning.

Human Sketch vs Model Response



Peter rides his bike along
and rides in the direction
right, then at the next lef
to left. Which crossing do

Each of these two pieces of wire is made of 8 segments of length 1. One of the pieces is placed one above the other so that they coincide partially. What is the largest possible length of their common part?

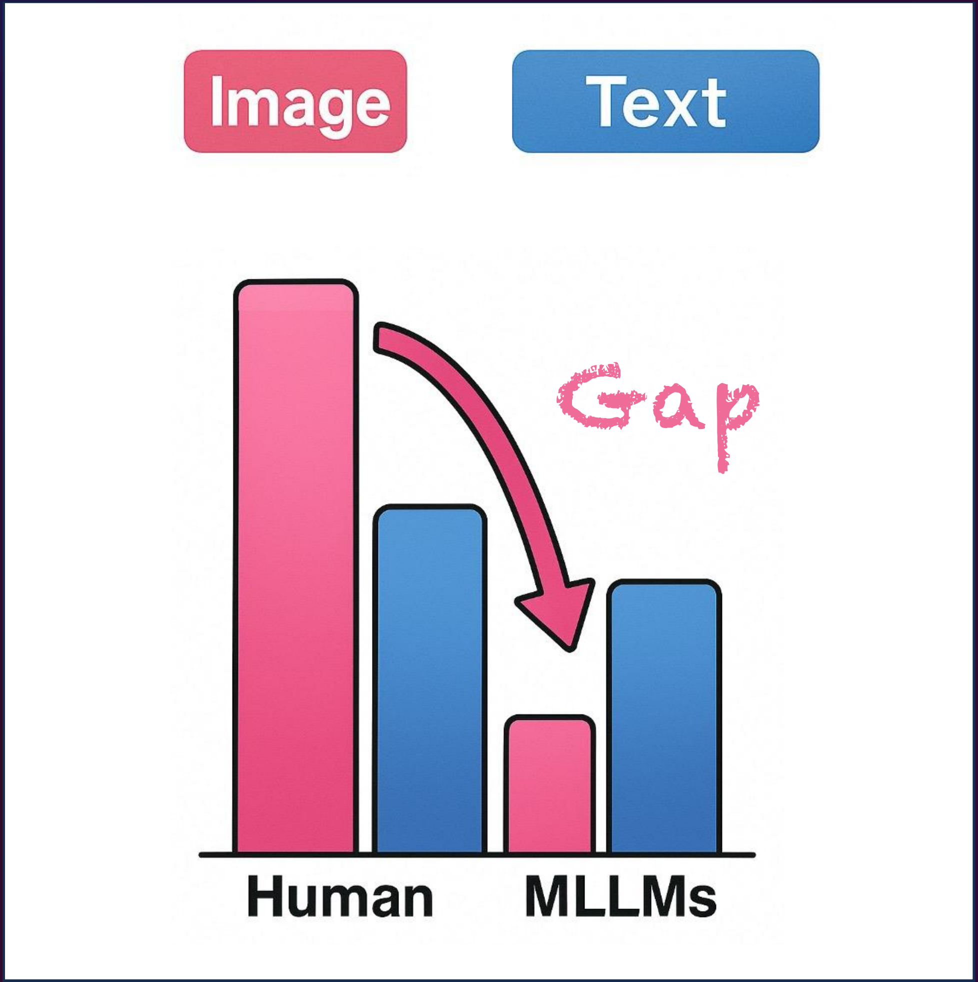
Answer: 5

```
Step 1 – Put both wires on a square lattice
Mark the bottom end of the first wire as the origin and

Wire A
(0,0)→(0,1)→(0,2)→(0,3)→(1,3)→(1,4)→(2,4)→(3,4)→(4,4)

Wire B
(0,0)→(0,1)→(0,2)→(0,3)→(0,4)→(1,4)→(1,3)→(2,3)→(3,3)

Step 2 – List all unit-length segments
A (vertical): V A₁(0,0→1), V A₂(0,1→2), V A₃(0,2→3), V A
A (horizontal): H A₁(0→1,3), H A₂(1→2,4), H A₃(2→3,4), H

B (vertical): V B₁(0,0→1), V B₂(0,1→2), V B₃(0,2→3), V B
B (horizontal): H B₁(0→1,4), H B₂(1→2,3), H B₃(2→3,3)

Step 3 – Look for a translation that maximises coinciden
Translate wire B by (0,0) – keep both drawings as they s

• Vertical coincidences: V A₁…V A₄ coincide with V B₁…V
• Horizontal coincidences: none

Step 4 – Show that 4 is maximal
Any other translation loses at least one vertical overla
Hence no rigid motion can exceed **4**.

\boxed{4}
```

use
etches
d

ly on
xtual

Human vs Model



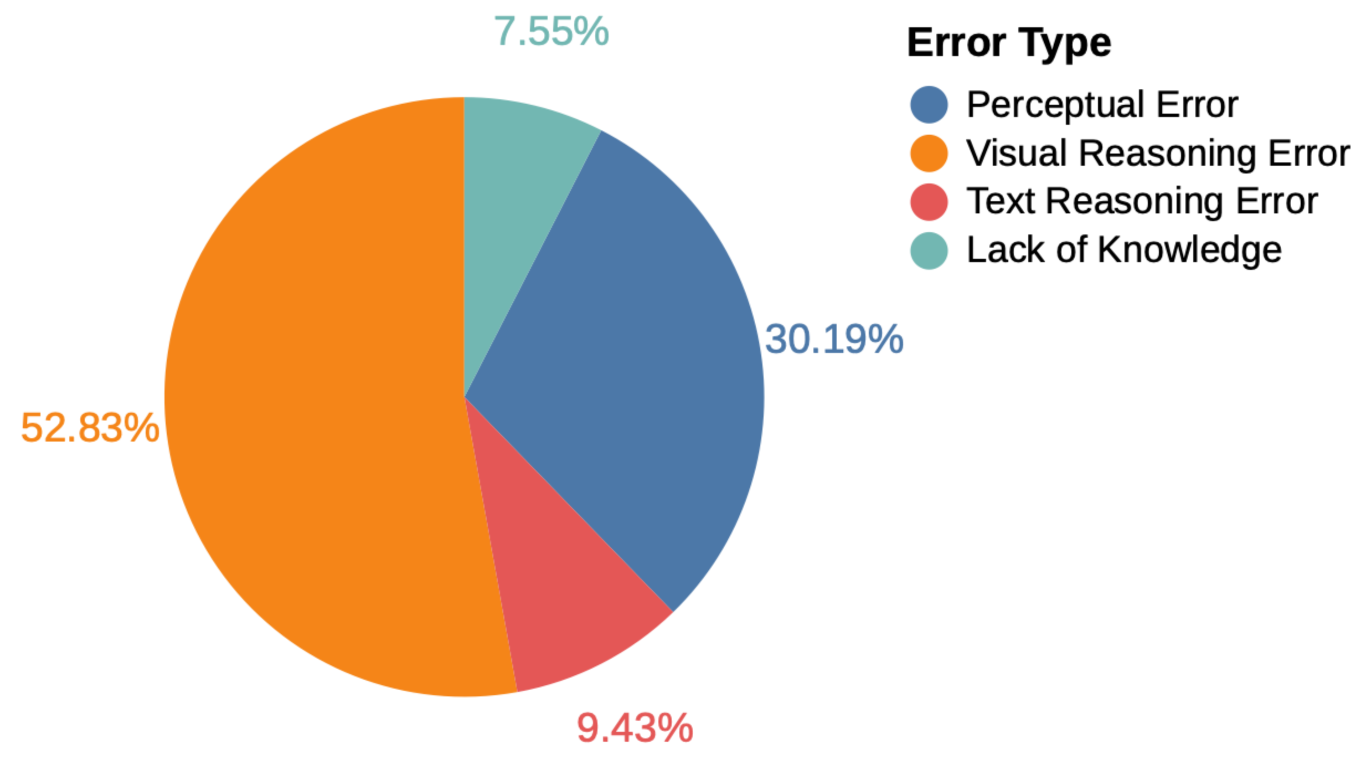
"The ability to organically reason over and with both text and images is a pillar of human intelligence."

MLLMs

- **Language-dominant reasoning limits vision-centric insight**
- **underuse vision for flexible, human-like inference**

Visual Reasoning vs Text Reasoning

Most errors stem from visual reasoning challenges



Visual soning

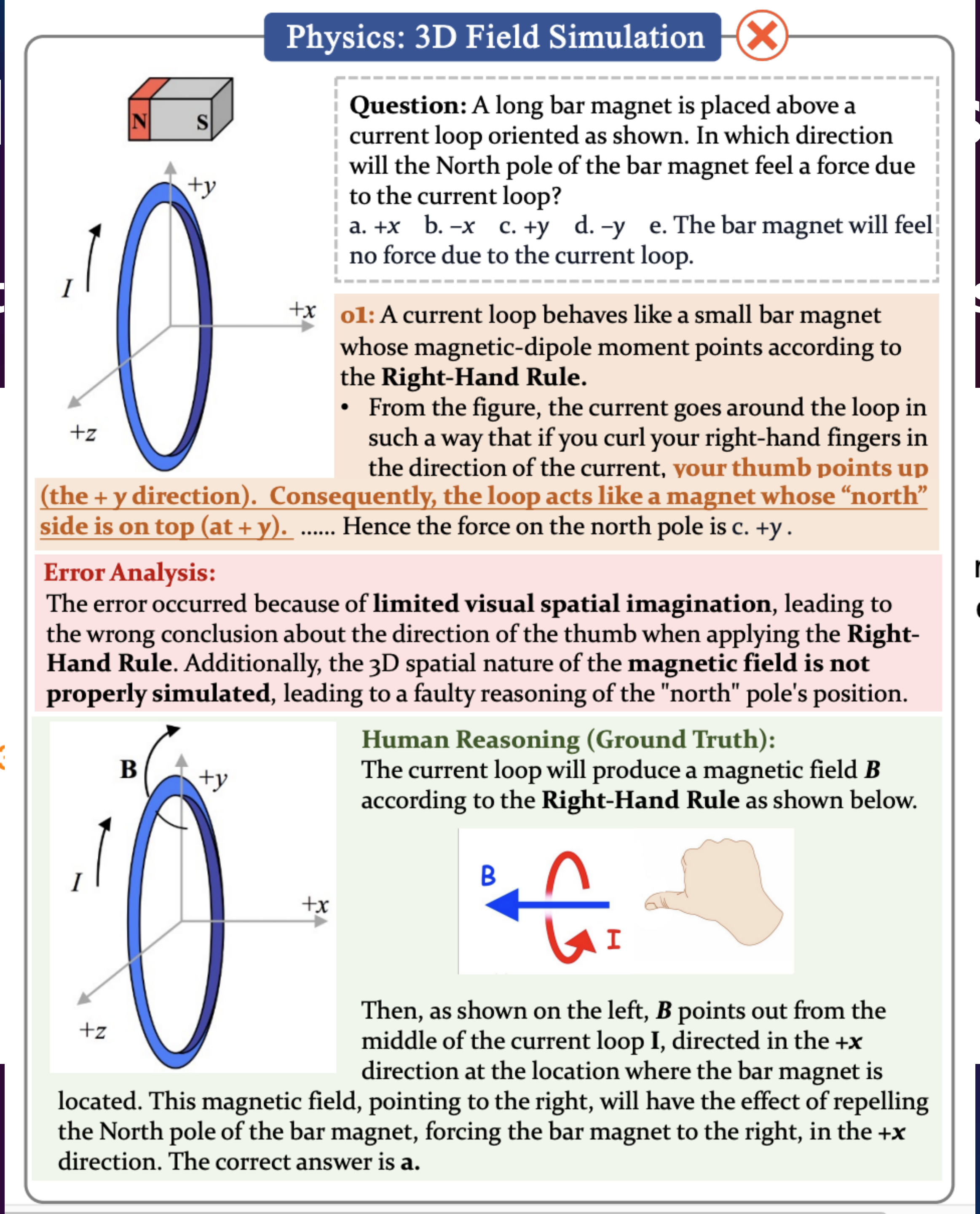
Most errors st s

52.8

Error

rror

e



Physics: 3D Field Simulation

Question: A long bar magnet is placed above a current loop oriented as shown. In which direction will the North pole of the bar magnet feel a force due to the current loop?

a. $+x$ b. $-x$ c. $+y$ d. $-y$ e. The bar magnet will feel no force due to the current loop.

o1: A current loop behaves like a small bar magnet whose magnetic-dipole moment points according to the **Right-Hand Rule.**

- From the figure, the current goes around the loop in such a way that if you curl your right-hand fingers in the direction of the current, **your thumb points up (the + y direction). Consequently, the loop acts like a magnet whose "north" side is on top (at + y).** ...... Hence the force on the north pole is c. $+y$.

Error Analysis:

The error occurred because of **limited visual spatial imagination**, leading to the wrong conclusion about the direction of the thumb when applying the **Right-Hand Rule**. Additionally, the 3D spatial nature of the **magnetic field is not properly simulated**, leading to a faulty reasoning of the "north" pole's position.

Human Reasoning (Ground Truth):

The current loop will produce a magnetic field ***B*** according to the **Right-Hand Rule** as shown below.

Then, as shown on the left, ***B*** points out from the middle of the current loop **I**, directed in the **+*x*** direction at the location where the bar magnet is located. This magnetic field, pointing to the right, will have the effect of repelling the North pole of the bar magnet, forcing the bar magnet to the right, in the **+*x*** direction. The correct answer is **a.**

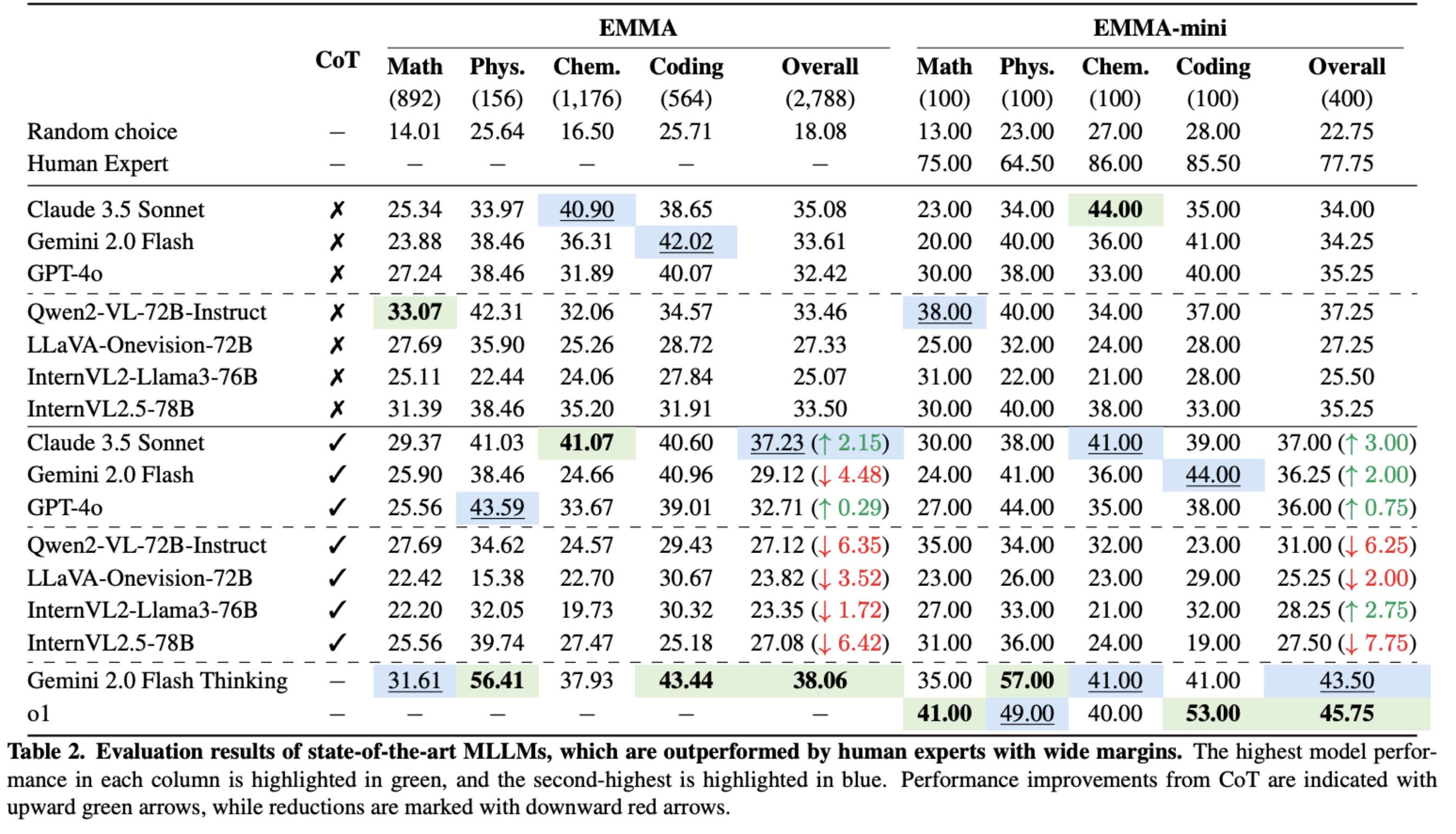
How to improve MM reasoning?

How to improve MM reasoning?

Chain-of-Thought 🤔

How to improve MM reasoning?

① CoT:



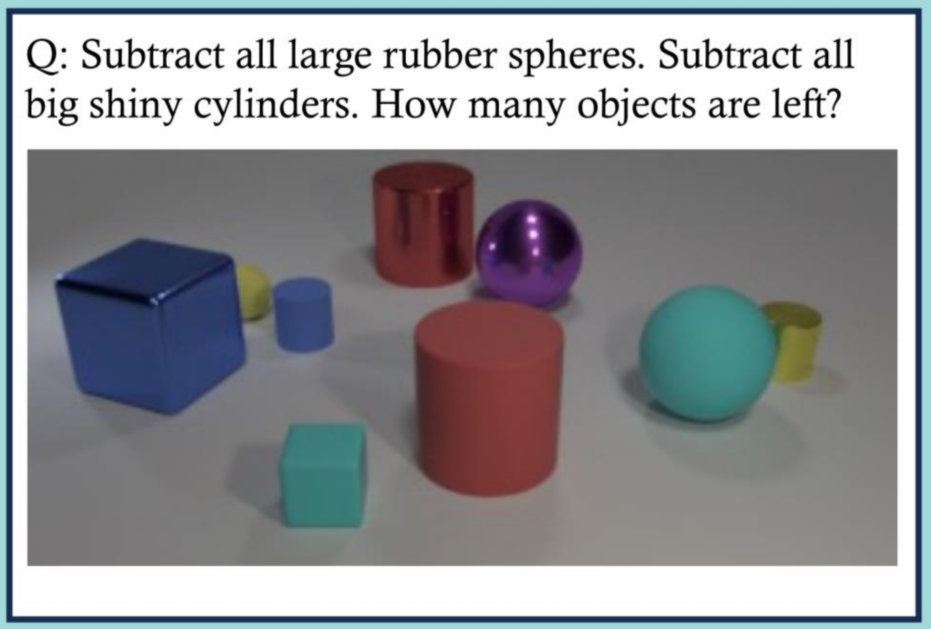
		EMMA					EMMA-mini				
	CoT	Math (892)	Phys. (156)	Chem. (1,176)	Coding (564)	Overall (2,788)	Math (100)	Phys. (100)	Chem. (100)	Coding (100)	Overall (400)
Random choice	–	14.01	25.64	16.50	25.71	18.08	13.00	23.00	27.00	28.00	22.75
Human Expert	–	–	–	–	–	–	75.00	64.50	86.00	85.50	77.75
Claude 3.5 Sonnet	✗	25.34	33.97	40.90	38.65	35.08	23.00	34.00	**44.00**	35.00	34.00
Gemini 2.0 Flash	✗	23.88	38.46	36.31	42.02	33.61	20.00	40.00	36.00	41.00	34.25
GPT-4o	✗	27.24	38.46	31.89	40.07	32.42	30.00	38.00	33.00	40.00	35.25
Qwen2-VL-72B-Instruct	✗	**33.07**	42.31	32.06	34.57	33.46	38.00	40.00	34.00	37.00	37.25
LLaVA-Onevision-72B	✗	27.69	35.90	25.26	28.72	27.33	25.00	32.00	24.00	28.00	27.25
InternVL2-Llama3-76B	✗	25.11	22.44	24.06	27.84	25.07	31.00	22.00	21.00	28.00	25.50
InternVL2.5-78B	✗	31.39	38.46	35.20	31.91	33.50	30.00	40.00	38.00	33.00	35.25
Claude 3.5 Sonnet	✓	29.37	41.03	**41.07**	40.60	37.23 (↑ 2.15)	30.00	38.00	41.00	39.00	37.00 (↑ 3.00)
Gemini 2.0 Flash	✓	25.90	38.46	24.66	40.96	29.12 (↓ 4.48)	24.00	41.00	36.00	44.00	36.25 (↑ 2.00)
GPT-4o	✓	25.56	43.59	33.67	39.01	32.71 (↑ 0.29)	27.00	44.00	35.00	38.00	36.00 (↑ 0.75)
Qwen2-VL-72B-Instruct	✓	27.69	34.62	24.57	29.43	27.12 (↓ 6.35)	35.00	34.00	32.00	23.00	31.00 (↓ 6.25)
LLaVA-Onevision-72B	✓	22.42	15.38	22.70	30.67	23.82 (↓ 3.52)	23.00	26.00	23.00	29.00	25.25 (↓ 2.00)
InternVL2-Llama3-76B	✓	22.20	32.05	19.73	30.32	23.35 (↓ 1.72)	27.00	33.00	21.00	32.00	28.25 (↑ 2.75)
InternVL2.5-78B	✓	25.56	39.74	27.47	25.18	27.08 (↓ 6.42)	31.00	36.00	24.00	19.00	27.50 (↓ 7.75)
Gemini 2.0 Flash Thinking	–	31.61	**56.41**	37.93	**43.44**	**38.06**	35.00	**57.00**	41.00	41.00	43.50
o1	–	–	–	–	–	–	**41.00**	49.00	40.00	**53.00**	**45.75**

Table 2. Evaluation results of state-of-the-art MLLMs, which are outperformed by human experts with wide margins. The highest model performance in each column is highlighted in green, and the second-highest is highlighted in blue. Performance improvements from CoT are indicated with upward green arrows, while reductions are marked with downward red arrows.

Divergent tendencies across closed- and open-source models.

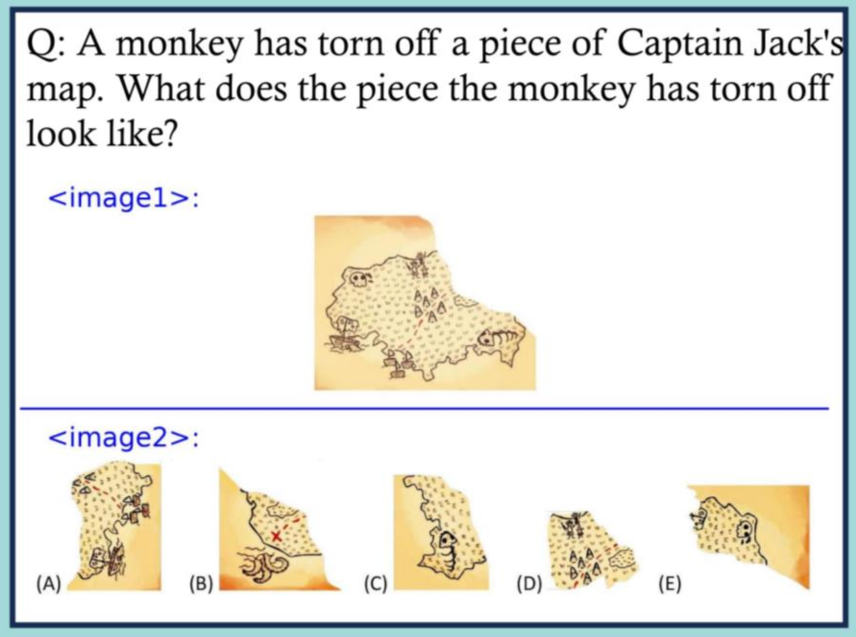
How to improve MM reasoning?

① CoT:



Q: Subtract all large rubber spheres. Subtract all big shiny cylinders. How many objects are left?

Multi-hop Visual Object Counting



Q: A monkey has torn off a piece of Captain Jack's map. What does the piece the monkey has torn off look like?

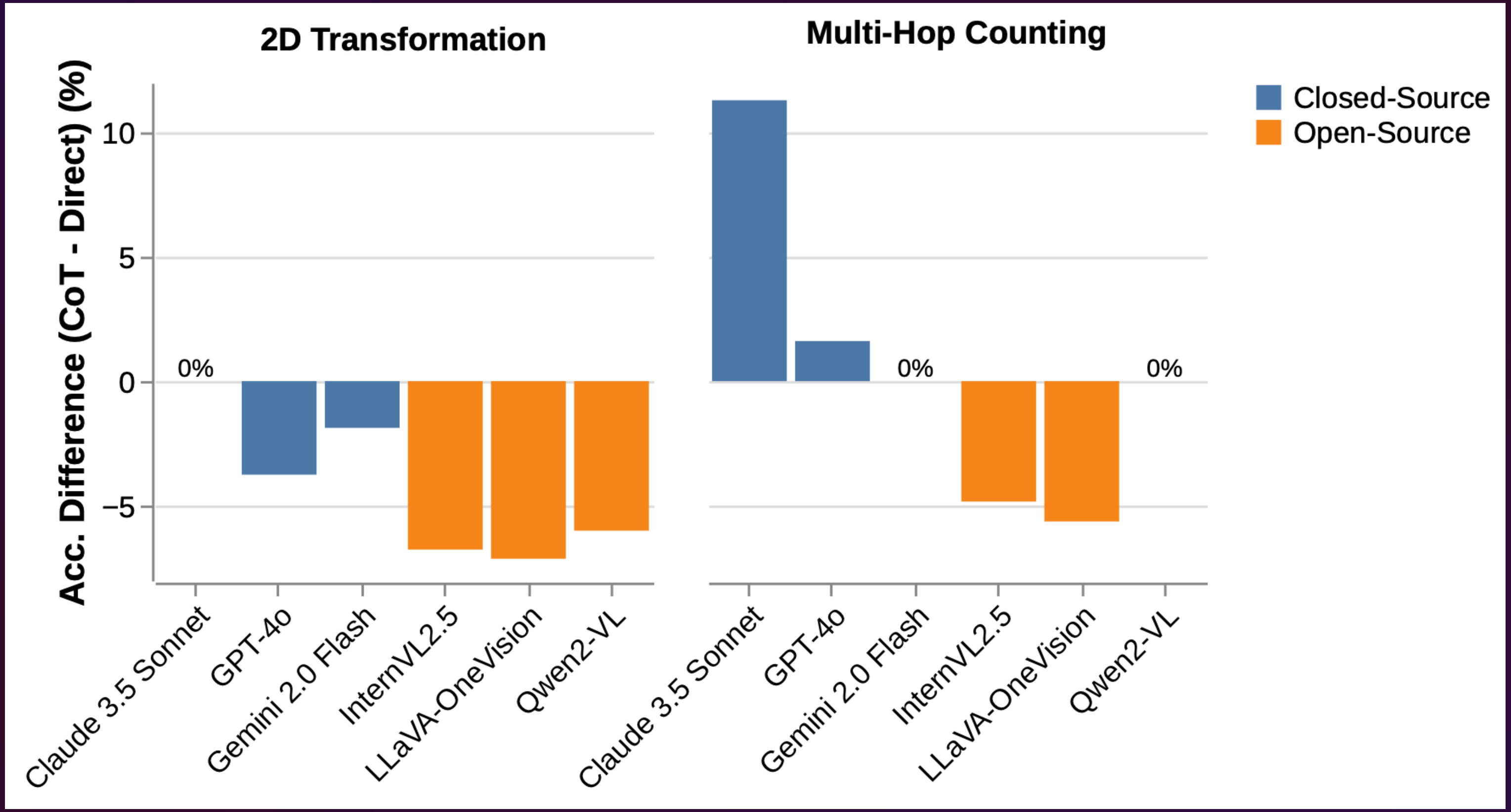
2D Transformation

Allow to benefit from language-based reasoning

Visual-centric tasks

How to improve MM reasoning?

① CoT:

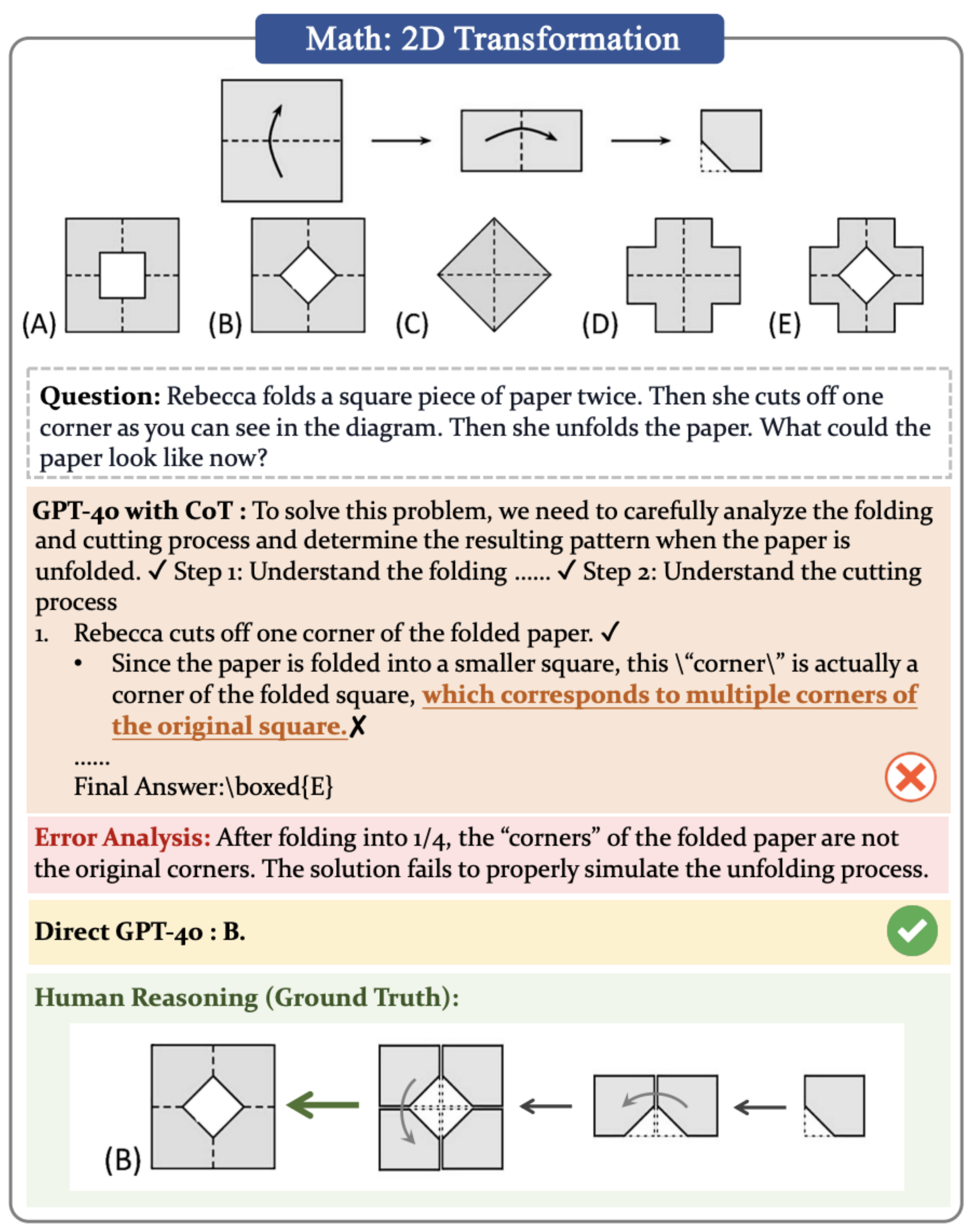


More visual reasoning → less CoT benefit

CoT introduces more hallucinations for open-source models.

How ing?

① CoT:



Math: 2D Transformation

Question: Rebecca folds a square piece of paper twice. Then she cuts off one corner as you can see in the diagram. Then she unfolds the paper. What could the paper look like now?

GPT-4o with CoT : To solve this problem, we need to carefully analyze the folding and cutting process and determine the resulting pattern when the paper is unfolded. ✓ Step 1: Understand the folding ...... ✓ Step 2: Understand the cutting process

1. Rebecca cuts off one corner of the folded paper. ✓
   - Since the paper is folded into a smaller square, this \“corner\” is actually a corner of the folded square, **which corresponds to multiple corners of the original square.** ✗

......

Final Answer:\boxed{E} ✗

Error Analysis: After folding into 1/4, the “corners” of the folded paper are not the original corners. The solution fails to properly simulate the unfolding process.

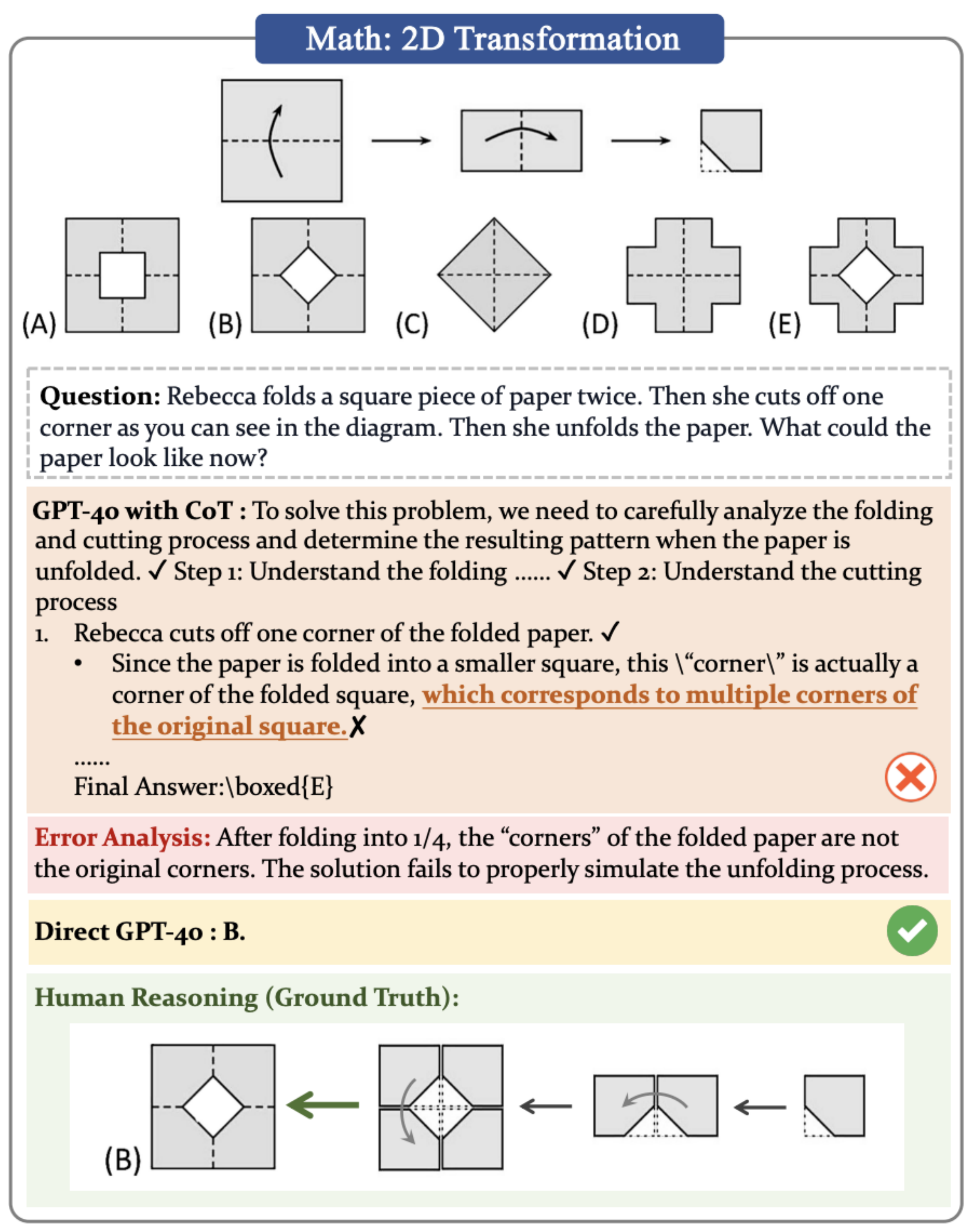
Direct GPT-4o : B. ✓

Human Reasoning (Ground Truth):

CoT introduc ce models.

How ing?

① CoT:



Math: 2D Transformation

Question: Rebecca folds a square piece of paper twice. Then she cuts off one corner as you can see in the diagram. Then she unfolds the paper. What could the paper look like now?

GPT-4o with CoT : To solve this problem, we need to carefully analyze the folding and cutting process and determine the resulting pattern when the paper is unfolded. ✓ Step 1: Understand the folding …… ✓ Step 2: Understand the cutting process

1. Rebecca cuts off one corner of the folded paper. ✓
   - Since the paper is folded into a smaller square, this \“corner\” is actually a corner of the folded square, which corresponds to multiple corners of the original square. ✗

……

Final Answer:\boxed{E}

Error Analysis: After folding into 1/4, the “corners” of the folded paper are not the original corners. The solution fails to properly simulate the unfolding process.

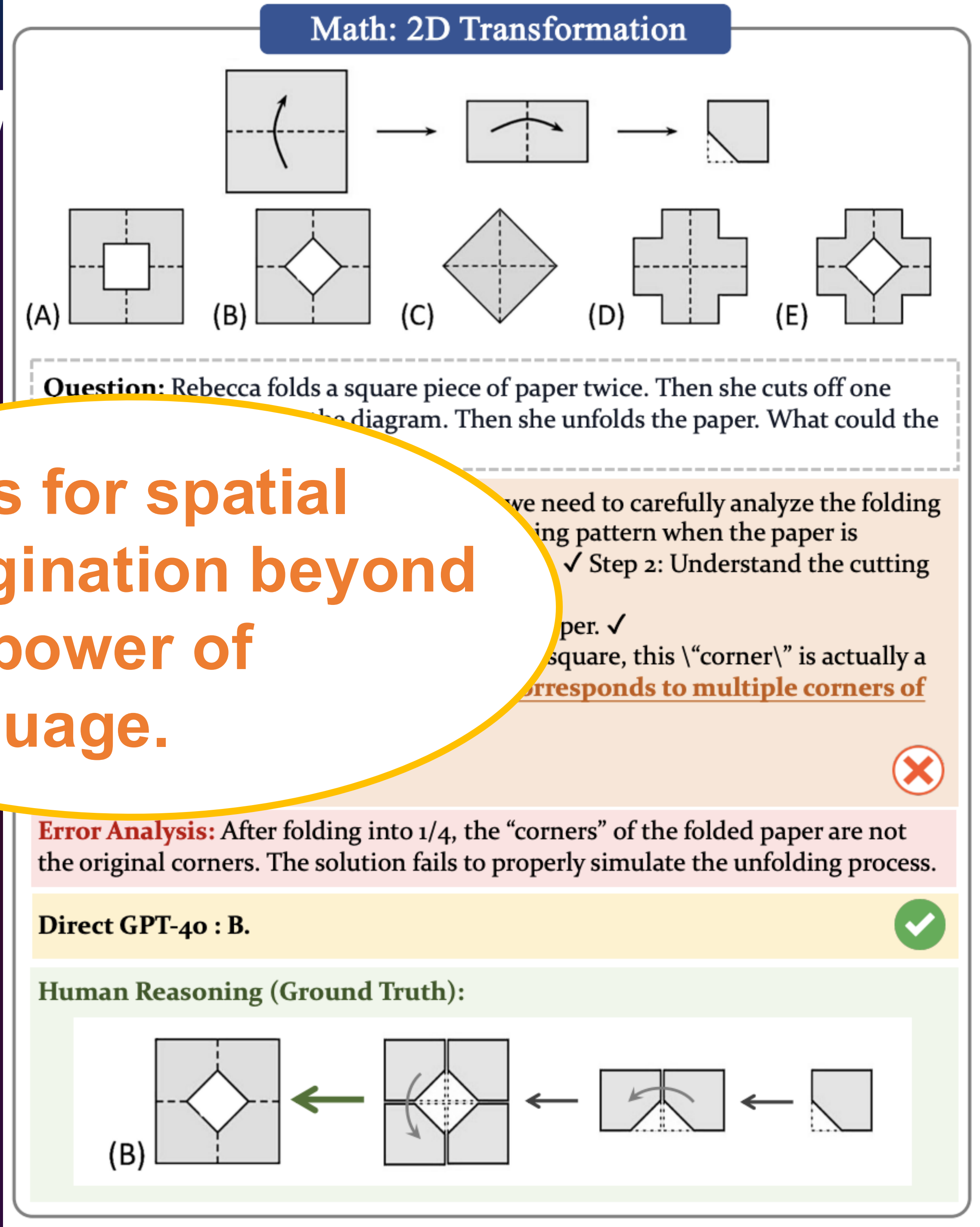
Direct GPT-4o : B.

Human Reasoning (Ground Truth):

Thoughts show superficial association with spatial manipulations and relations, rather than genuine visual reasoning.

How ing?

① CoT:



Math: 2D Transformation

Question: Rebecca folds a square piece of paper twice. Then she cuts off one
diagram. Then she unfolds the paper. What could the

we need to carefully analyze the folding
ing pattern when the paper is
✓ Step 2: Understand the cutting

per. ✓
square, this \“corner\” is actually a
orresponds to multiple corners of

Error Analysis: After folding into 1/4, the “corners” of the folded paper are not the original corners. The solution fails to properly simulate the unfolding process.

Direct GPT-4o : B.

Human Reasoning (Ground Truth):

Calls for spatial imagination beyond the power of language.

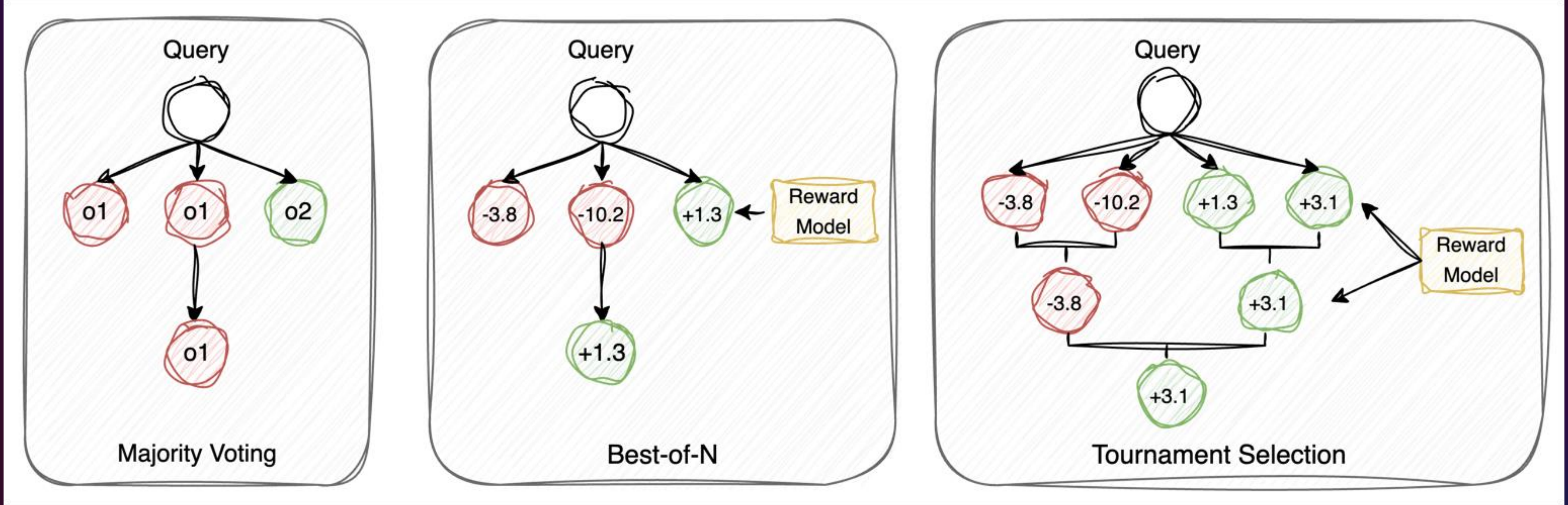
Thoughts show superficial association with spatial manipulations and relations, rather than genuine visual reasoning.

How to improve MM reasoning?

Test Time Compute Scaling 🤔

How to improve MM reasoning?

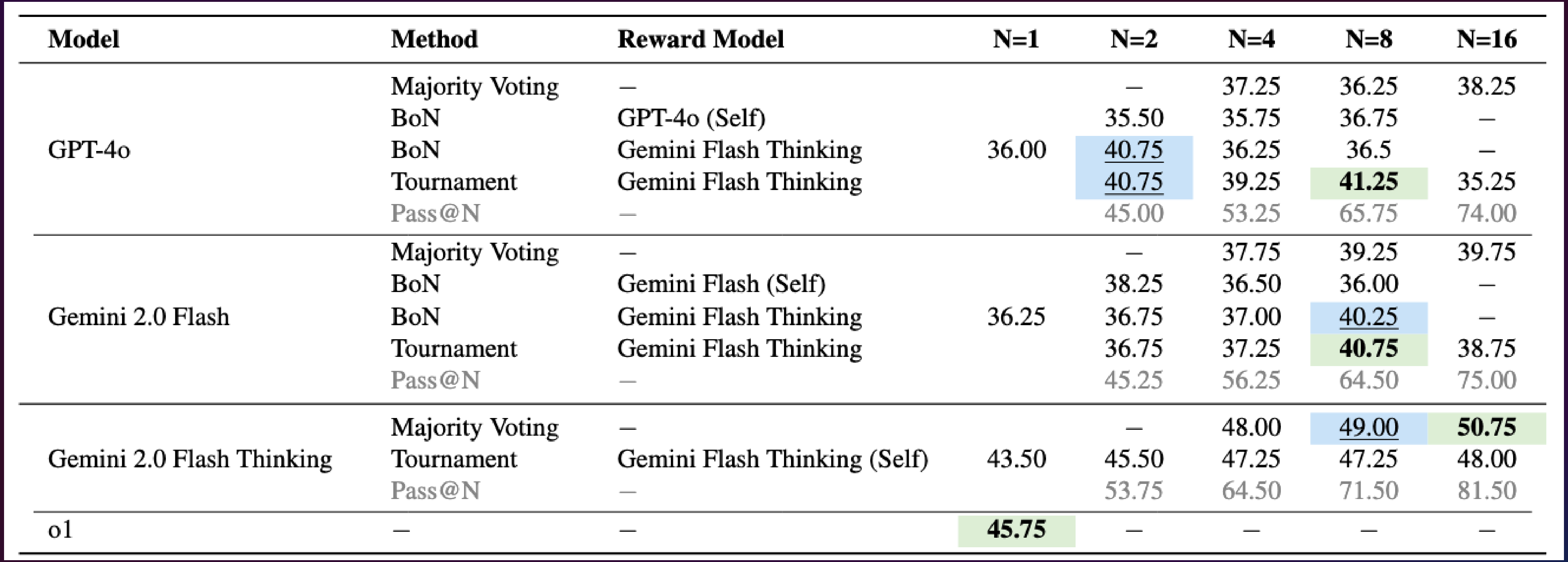
② **Try various test-time compute scaling strategies**



e.g., majority voting, best-of-N, and tournament (up to 16 times)

How to improve MM reasoning?

More textual CoTs candidate ≠ Better visual reasoning



Model	Method	Reward Model	N=1	N=2	N=4	N=8	N=16
GPT-4o	Majority Voting	–	36.00	–	37.25	36.25	38.25
	BoN	GPT-4o (Self)		35.50	35.75	36.75	–
	BoN	Gemini Flash Thinking		40.75	36.25	36.5	–
	Tournament	Gemini Flash Thinking		40.75	39.25	**41.25**	35.25
	Pass@N	–		45.00	53.25	65.75	74.00
Gemini 2.0 Flash	Majority Voting	–	36.25	–	37.75	39.25	39.75
	BoN	Gemini Flash (Self)		38.25	36.50	36.00	–
	BoN	Gemini Flash Thinking		36.75	37.00	40.25	–
	Tournament	Gemini Flash Thinking		36.75	37.25	**40.75**	38.75
	Pass@N	–		45.25	56.25	64.50	75.00
Gemini 2.0 Flash Thinking	Majority Voting	–	43.50	–	48.00	49.00	**50.75**
	Tournament	Gemini Flash Thinking (Self)		45.50	47.25	47.25	48.00
	Pass@N	–		53.75	64.50	71.50	81.50
o1	–	–	**45.75**	–	–	–	–

Particularly for tasks requiring fine-grained spatial understanding or multi-step reasoning

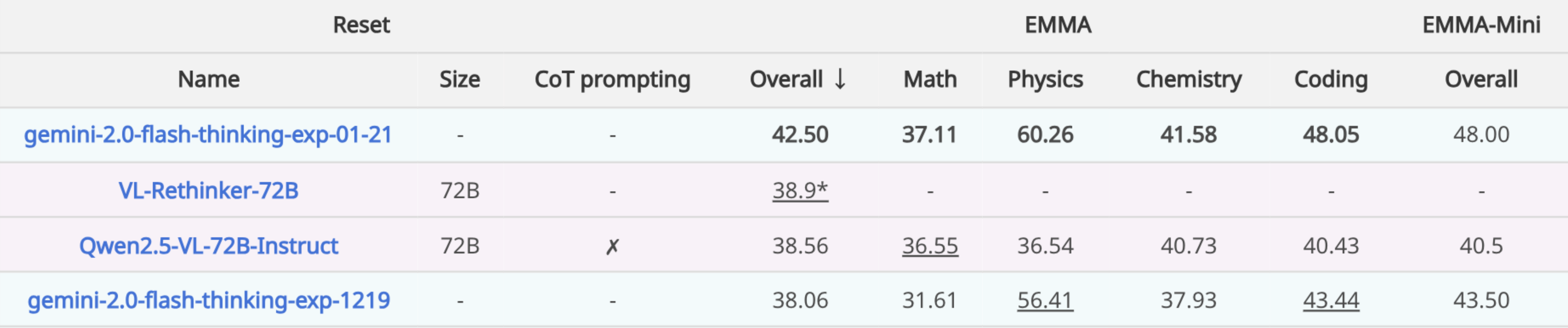
How to improve MM reasoning?

Reinforcement Learning Scaling

🤔

How to improve MM reasoning?

③ Try RL training to scale

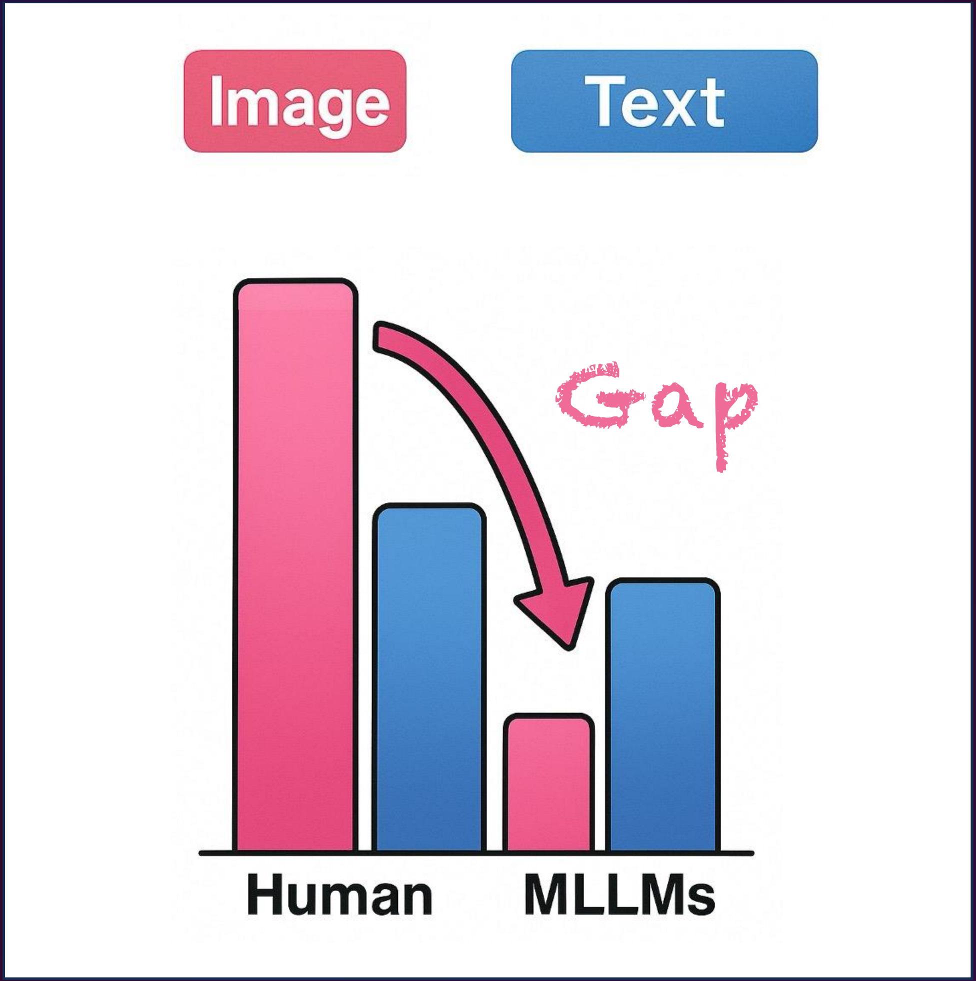


Reset			EMMA					EMMA-Mini
Name	Size	CoT prompting	Overall ↓	Math	Physics	Chemistry	Coding	Overall
gemini-2.0-flash-thinking-exp-01-21	-	-	**42.50**	**37.11**	**60.26**	**41.58**	**48.05**	48.00
VL-Rethinker-72B	72B	-	38.9*	-	-	-	-	-
Qwen2.5-VL-72B-Instruct	72B	✗	38.56	36.55	36.54	40.73	40.43	40.5
gemini-2.0-flash-thinking-exp-1219	-	-	38.06	31.61	56.41	37.93	43.44	43.50

Maybe promising

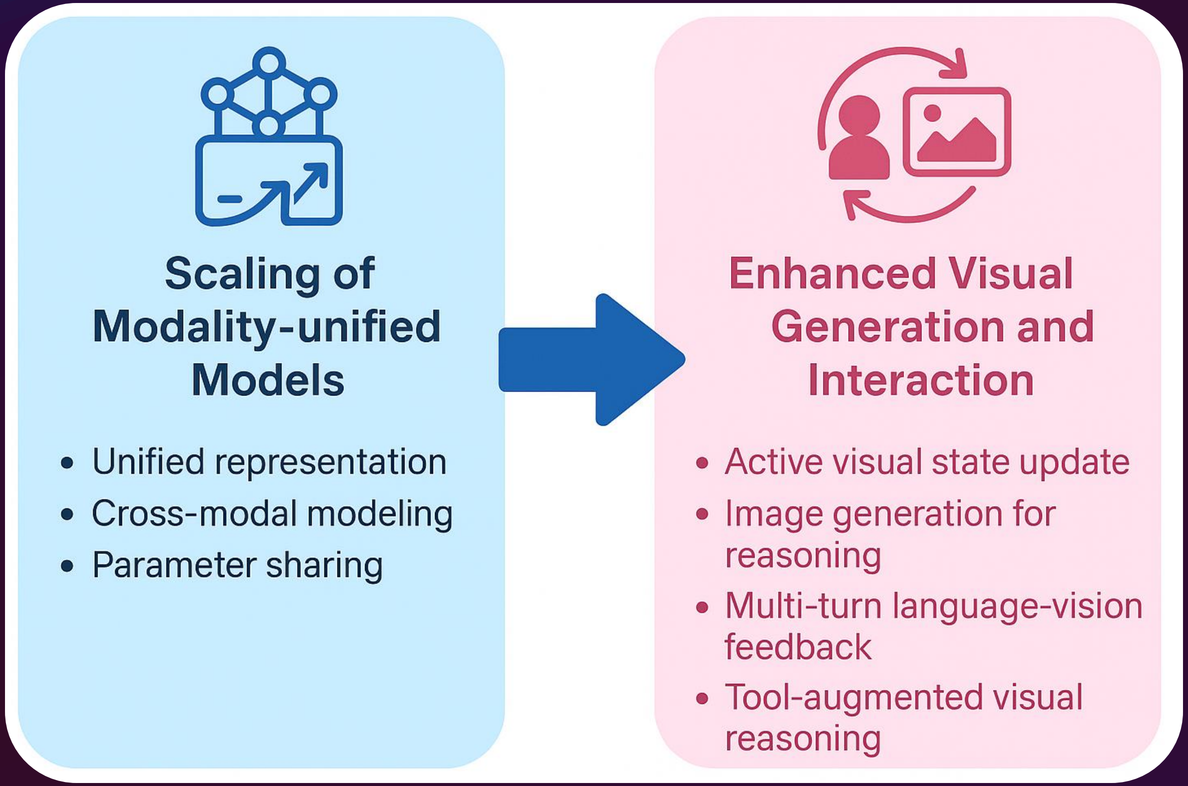
- **VL-Rethinker**: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning
- **SRPO**: Enhancing Multimodal LLM Reasoning via Reflection-Aware Reinforcement Learning
- ……

Future Work



- From language-driven to cross-modal collaboration for MM Reasoning
- Visual reasoning is a key bottleneck
- New paradigms needed!

Future Work



Scaling of Modality-unified Models

- Unified representation
- Cross-modal modeling
- Parameter sharing

Enhanced Visual Generation and Interaction

- Active visual state update
- Image generation for reasoning
- Multi-turn language-vision feedback
- Tool-augmented visual reasoning

- Scaling of Modality-unified Models
- Enhanced Visual Generation and Interaction

ICML
International Conference
On Machine Learning

Thanks!

Can MLLMs Reason/Think in Multimodality?

Twitter

EMMA

Homepage

Yunzhuo Hao

Jiawei Gu

Huichen Will Wang

Linjie Li

Zhengyuan Yang

Lijuan Wang

Yu Cheng