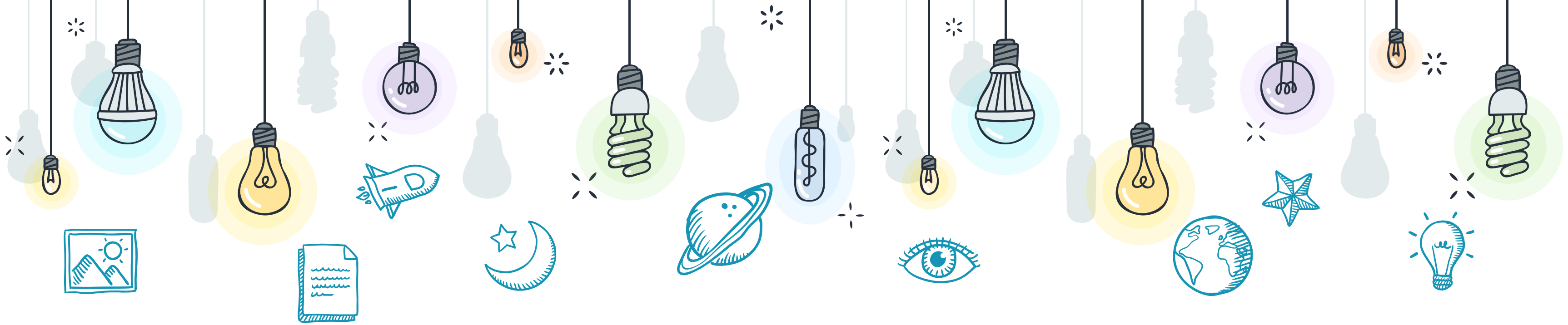


Evaluations and Benchmarks

in Context of Multimodal LLM

https://mllm2024.github.io/CVPR2025/

CVPR Nashville JUNE 11-15, 2025

CC BY 4.0 DEED
Attribution 4.0 International

This Keynote slide is licensed under a CC BY 4.0 International License.

Hao Fei

National University of Singapore

Xiang Yue

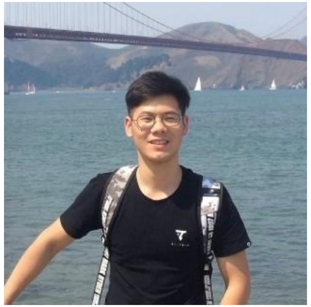
Carnegie Mellon University

Kaipeng Zhang

Shanghai AI Lab

Long Chen

HKUST



Jian Li

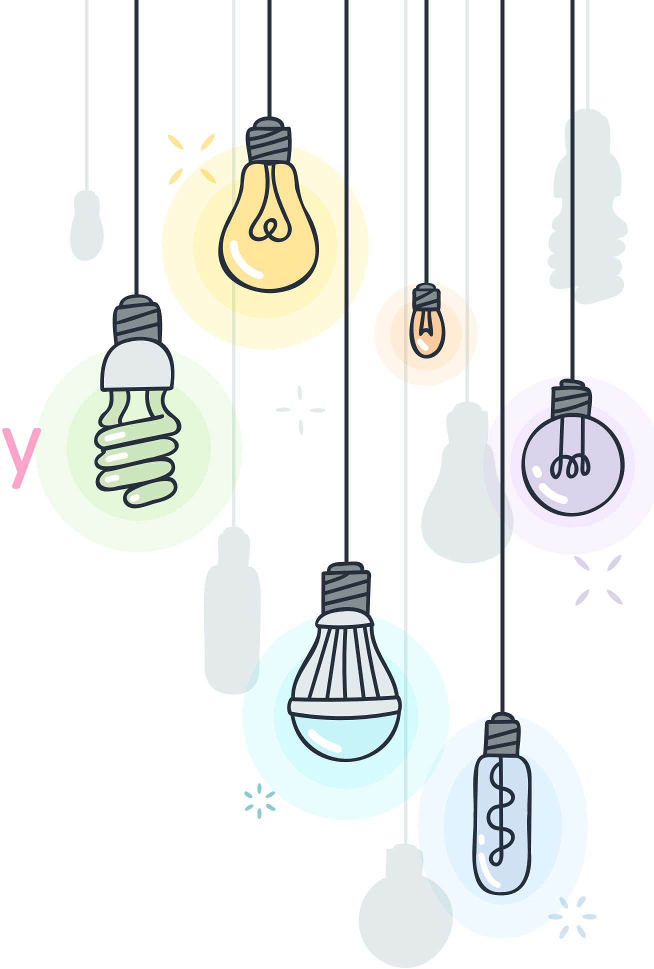
Tencent YoutuLab

Xinya Du

University of Texas at Dallas

Part-V

Expert-level Discipline Capability



Xiang Yue

Postdoc Researcher

Carnegie Mellon University

https://xiangyue9607.github.io/

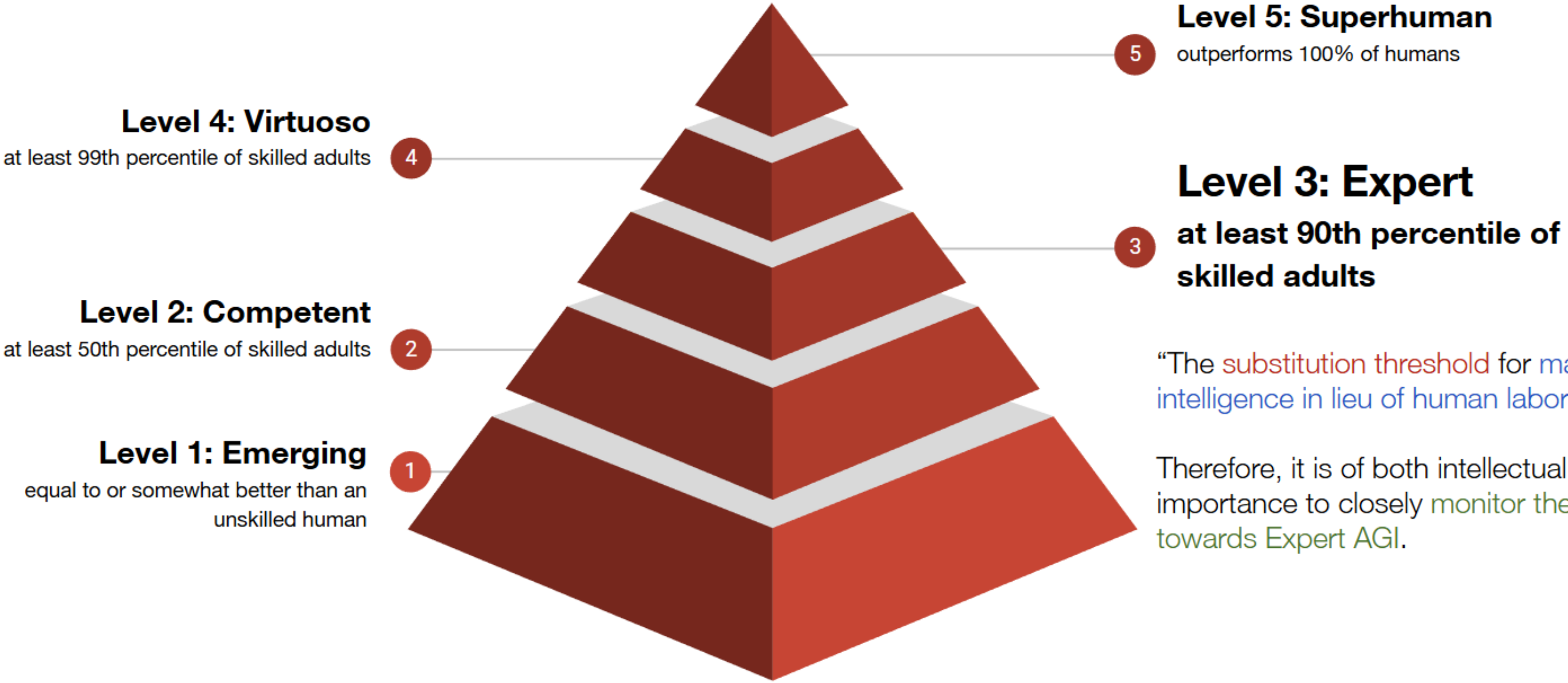
Table of Content

Expert-level Discipline Capability

- Motivation
- Key Benchmarks
  - General: MMMU, MMMU-Pro, Video-MMMU
  - Math: Mathvista, MathVerse, MATH-V
  - Science: ScienceQA, OlympiadBench
  - Medical: GMAI-MMBench, MedXpertQA
- Future Directions
  - Complex Agentic Tasks

Why Expert-Level Tasks?

- From a Artificial General Intelligence (AGI) perspective



Level 5: Superhuman
outperforms 100% of humans

Level 4: Virtuoso
at least 99th percentile of skilled adults

Level 3: Expert
at least 90th percentile of skilled adults

Level 2: Competent
at least 50th percentile of skilled adults

Level 1: Emerging
equal to or somewhat better than an unskilled human

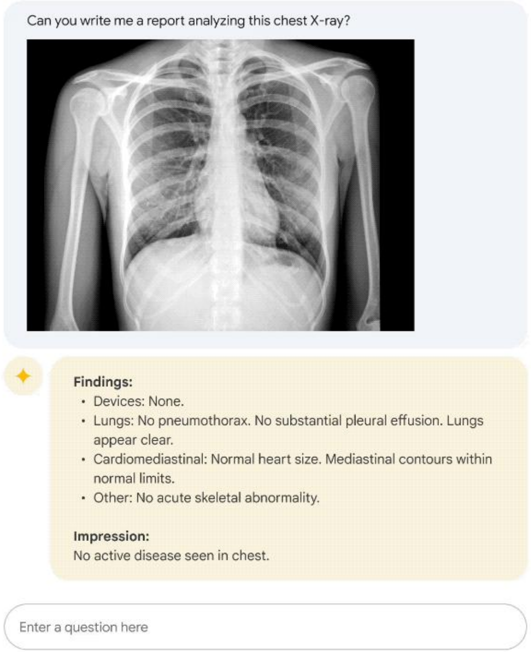
"The substitution threshold for machine intelligence in lieu of human labor"

Therefore, it is of both intellectual and societal importance to closely monitor the progress towards Expert AGI.

Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." *ICML 2024*

Why Expert-Level Tasks?

Unlocking Real-World Utility



Can you write me a report analyzing this chest X-ray?

Findings:
- Devices: None.
- Lungs: No pneumothorax. No substantial pleural effusion. Lungs appear clear.
- Cardiomediastinal: Normal heart size. Mediastinal contours within normal limits.
- Other: No acute skeletal abnormality.

Impression:
No active disease seen in chest.

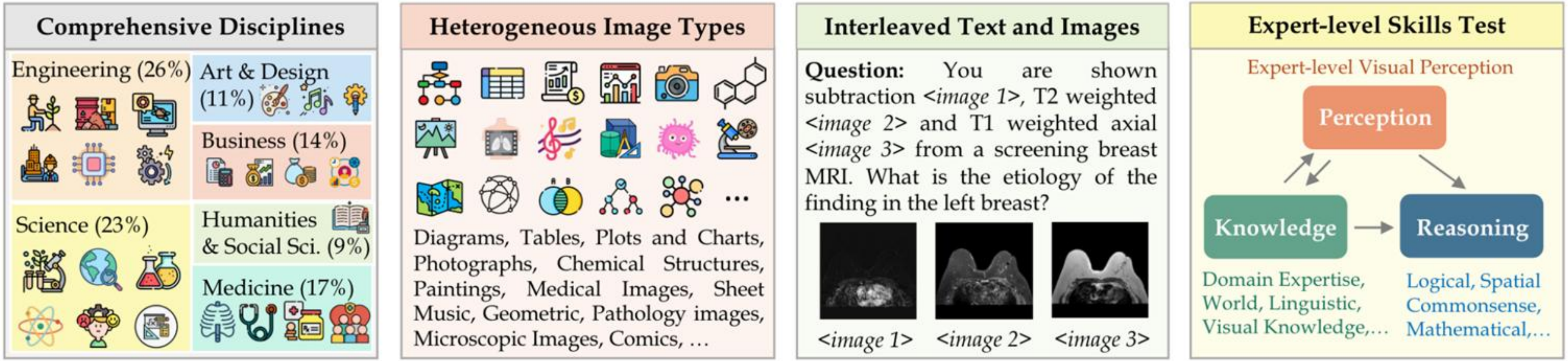
https://sites.research.google/med-palm/

Key Benchmarks

- General: MMMU, MMMU-Pro, Video-MMMU
- Math: Mathvista, MathVerse, MATH-V
- Science / STEM: ScienceQA, OlympiadBench
- Medical: GMAI-MMBench, MedXpertQA

MMMU

Massive Multi-discipline Multimodal Understanding and Reasoning



Comprehensive Disciplines

Engineering (26%)

Art & Design (11%)

Business (14%)

Science (23%)

Humanities & Social Sci. (9%)

Medicine (17%)

Heterogeneous Image Types

Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, …

Interleaved Text and Images

Question: You are shown subtraction *<image 1>*, T2 weighted *<image 2>* and T1 weighted axial *<image 3>* from a screening breast MRI. What is the etiology of the finding in the left breast?

<image 1> *<image 2>* *<image 3>*

Expert-level Skills Test

Expert-level Visual Perception

Perception

Knowledge → Reasoning

Domain Expertise, World, Linguistic, Visual Knowledge,…

Logical, Spatial Commonsense, Mathematical,…

(Breadth)

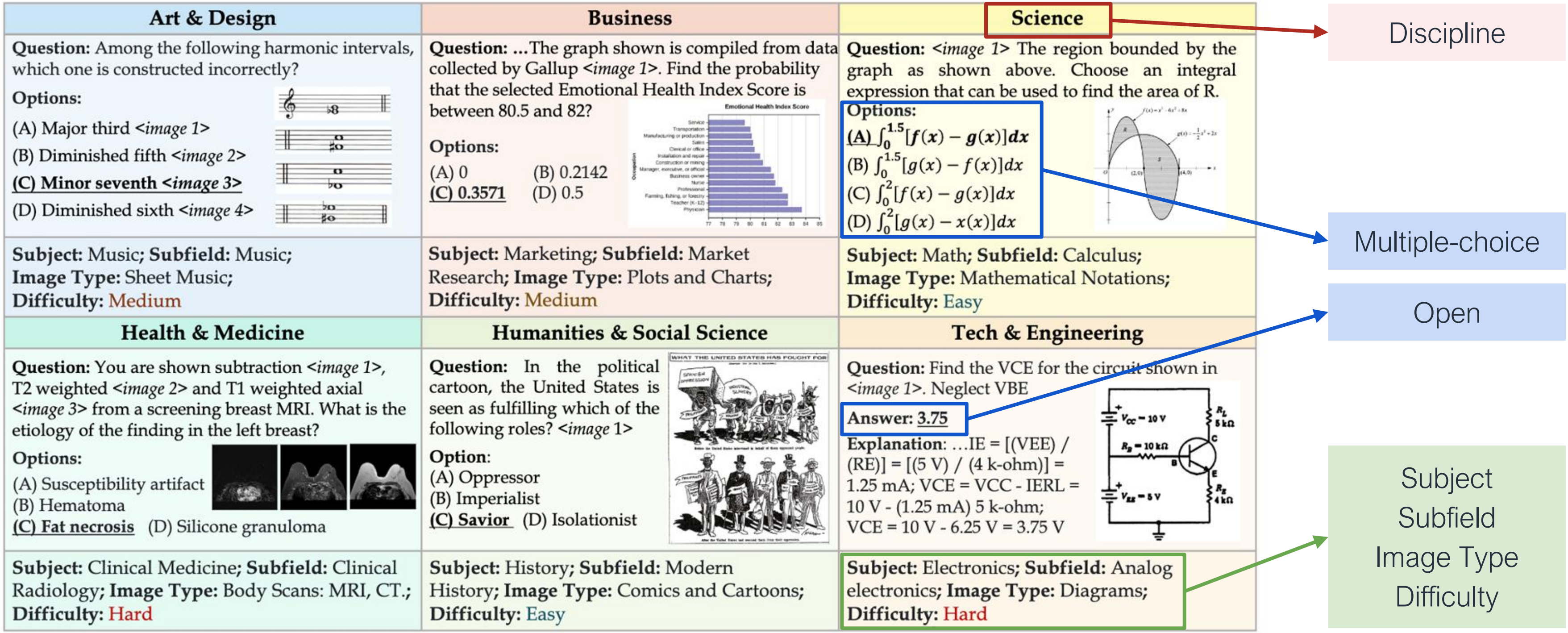
- 11.5K college-level problems across six broad disciplines and 30 college subjects
- 30 heterogeneous image types

(Depth)

- Interleaved text and (multiple) images
- Expert-level perception and reasoning rooted in deep subject knowledge

Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

Sampled MMMU examples from each discipline



Art & Design

Question: Among the following harmonic intervals, which one is constructed incorrectly?

Options:

(A) Major third <*image 1*>

(B) Diminished fifth <*image 2*>

(C) Minor seventh <*image 3*>

(D) Diminished sixth <*image 4*>

Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium

Business

Question: ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?

Options:

(A) 0 (B) 0.2142

(C) 0.3571 (D) 0.5

Subject: Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium

Science

Question: <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

Options:

(A) $\int_0^{1.5}[f(x)-g(x)]dx$

(B) $\int_0^{1.5}[g(x)-f(x)]dx$

(C) $\int_0^{2}[f(x)-g(x)]dx$

(D) $\int_0^{2}[g(x)-x(x)]dx$

Subject: Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy

Health & Medicine

Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast?

Options:

(A) Susceptibility artifact

(B) Hematoma

(C) Fat necrosis (D) Silicone granuloma

Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard

Humanities & Social Science

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*>

Option:

(A) Oppressor

(B) Imperialist

(C) Savior (D) Isolationist

Subject: History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy

Tech & Engineering

Question: Find the VCE for the circuit shown in <*image 1*>. Neglect VBE

Answer: 3.75

Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V

Subject: Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

Discipline

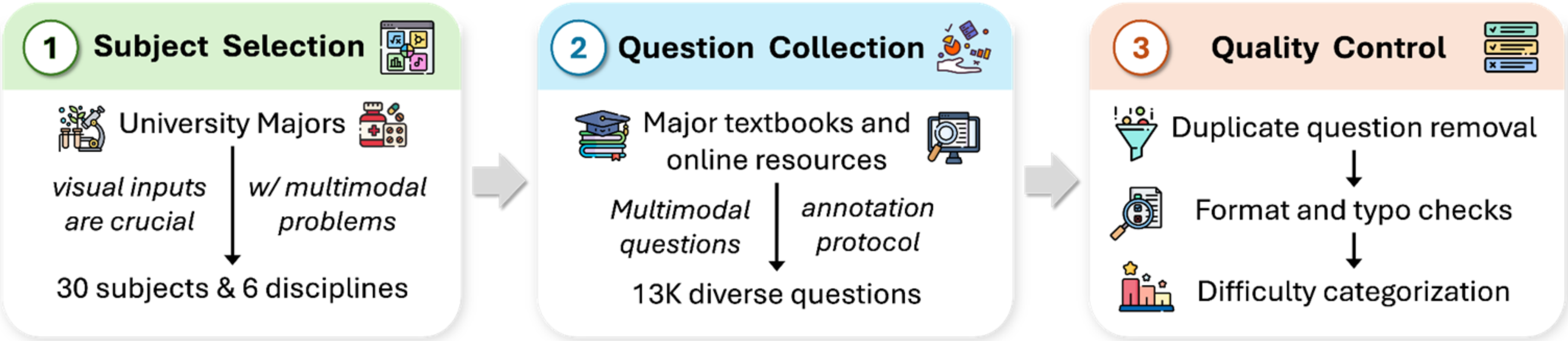
Multiple-choice

Open

Subject
Subfield
Image Type
Difficulty

Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

Data Curation Pipeline



① **Subject Selection**

University Majors

visual inputs are crucial ↓ *w/ multimodal problems*

30 subjects & 6 disciplines

② **Question Collection**

Major textbooks and online resources

Multimodal questions ↓ *annotation protocol*

13K diverse questions

③ **Quality Control**

Duplicate question removal

↓

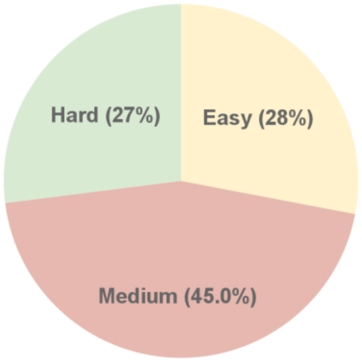
Format and typo checks

↓

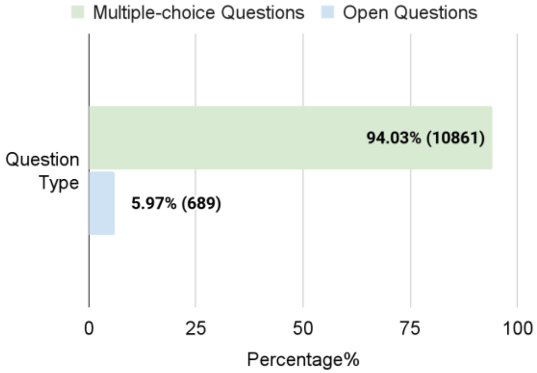
Difficulty categorization

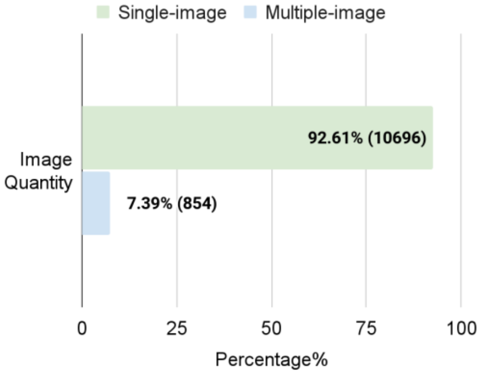
Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

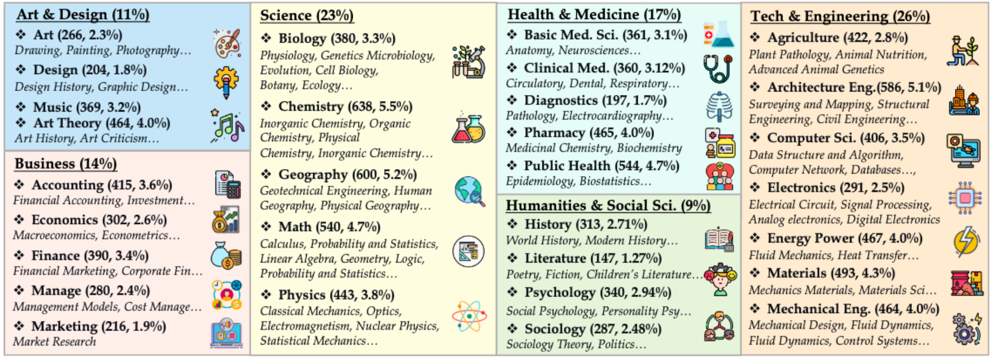
Statistics of MMMU



Question Difficulty







Art & Design (11%)

- ❖ **Art (266, 2.3%)**
  Drawing, Painting, Photography…
- ❖ **Design (204, 1.8%)**
  Design History, Graphic Design…
- ❖ **Music (369, 3.2%)**
- ❖ **Art Theory (464, 4.0%)**
  Art History, Art Criticism…

Business (14%)

- ❖ **Accounting (415, 3.6%)**
  Financial Accounting, Investment…
- ❖ **Economics (302, 2.6%)**
  Macroeconomics, Econometrics…
- ❖ **Finance (390, 3.4%)**
  Financial Marketing, Corporate Fin…
- ❖ **Manage (280, 2.4%)**
  Management Models, Cost Manage…
- ❖ **Marketing (216, 1.9%)**
  Market Research

Science (23%)

- ❖ **Biology (380, 3.3%)**
  Physiology, Genetics Microbiology, Evolution, Cell Biology, Botany, Ecology…
- ❖ **Chemistry (638, 5.5%)**
  Inorganic Chemistry, Organic Chemistry, Physical Chemistry, Inorganic Chemistry…
- ❖ **Geography (600, 5.2%)**
  Geotechnical Engineering, Human Geography, Physical Geography…
- ❖ **Math (540, 4.7%)**
  Calculus, Probability and Statistics, Linear Algebra, Geometry, Logic, Probability and Statistics…
- ❖ **Physics (443, 3.8%)**
  Classical Mechanics, Optics, Electromagnetism, Nuclear Physics, Statistical Mechanics…

Health & Medicine (17%)

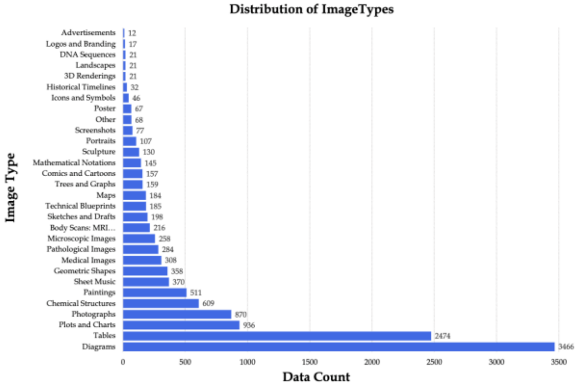
- ❖ **Basic Med. Sci. (361, 3.1%)**
  Anatomy, Neurosciences…
- ❖ **Clinical Med. (360, 3.12%)**
  Circulatory, Dental, Respiratory…
- ❖ **Diagnostics (197, 1.7%)**
  Pathology, Electrocardiography…
- ❖ **Pharmacy (465, 4.0%)**
  Medicinal Chemistry, Biochemistry
- ❖ **Public Health (544, 4.7%)**
  Epidemiology, Biostatistics…

Humanities & Social Sci. (9%)

- ❖ **History (313, 2.71%)**
  World History, Modern History…
- ❖ **Literature (147, 1.27%)**
  Poetry, Fiction, Children's Literature…
- ❖ **Psychology (340, 2.94%)**
  Social Psychology, Personality Psy…
- ❖ **Sociology (287, 2.48%)**
  Sociology Theory, Politics…

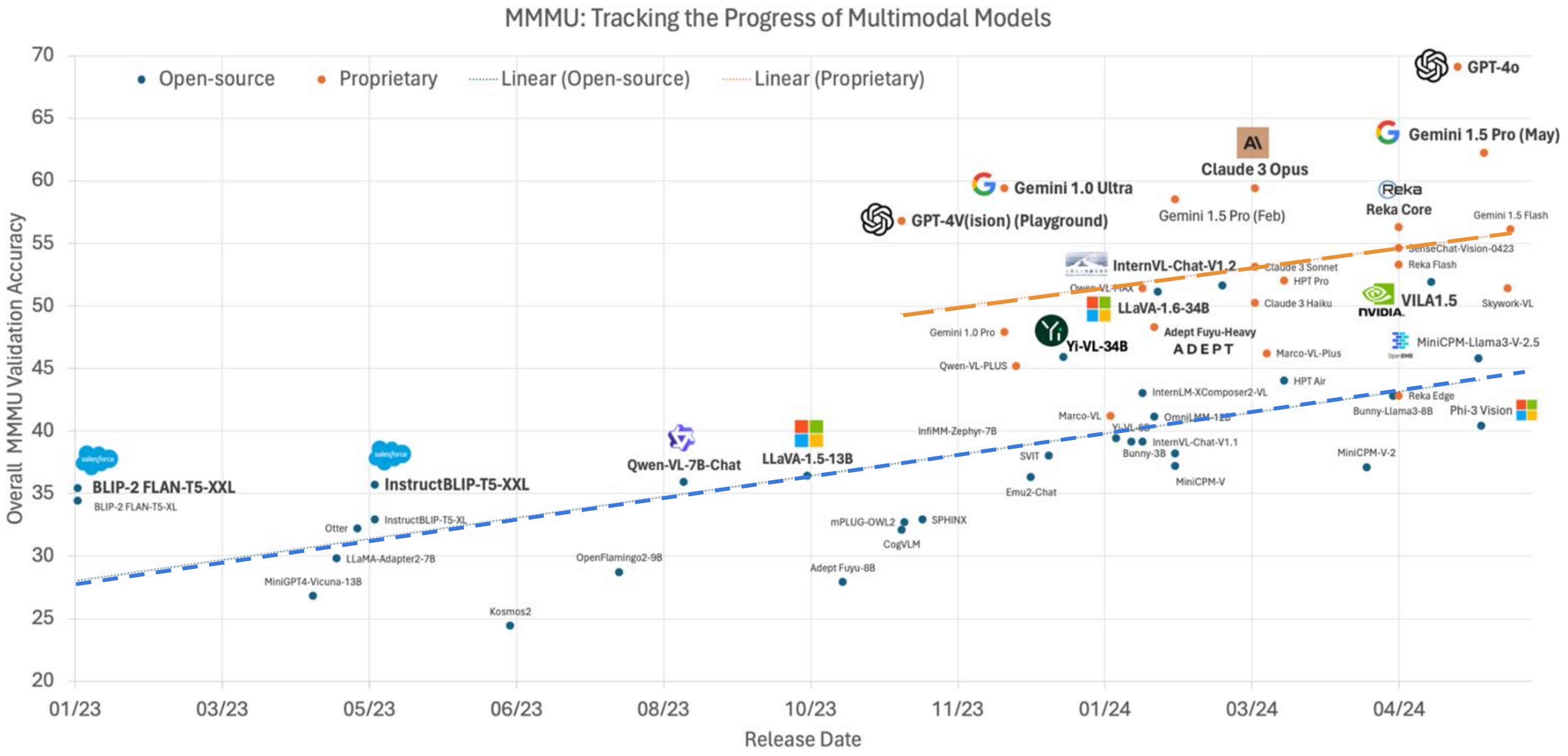
Tech & Engineering (26%)

- ❖ **Agriculture (422, 2.8%)**
  Plant Pathology, Animal Nutrition, Advanced Animal Genetics
- ❖ **Architecture Eng.(586, 5.1%)**
  Surveying and Mapping, Structural Engineering, Civil Engineering…
- ❖ **Computer Sci. (406, 3.5%)**
  Data Structure and Algorithm, Computer Network, Databases…,
- ❖ **Electronics (291, 2.5%)**
  Electrical Circuit, Signal Processing, Analog electronics, Digital Electronics
- ❖ **Energy Power (467, 4.0%)**
  Fluid Mechanics, Heat Transfer…
- ❖ **Materials (493, 4.3%)**
  Mechanics Materials, Materials Sci…
- ❖ **Mechanical Eng. (464, 4.0%)**
  Mechanical Design, Fluid Dynamics, Fluid Dynamics, Control Systems…



Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

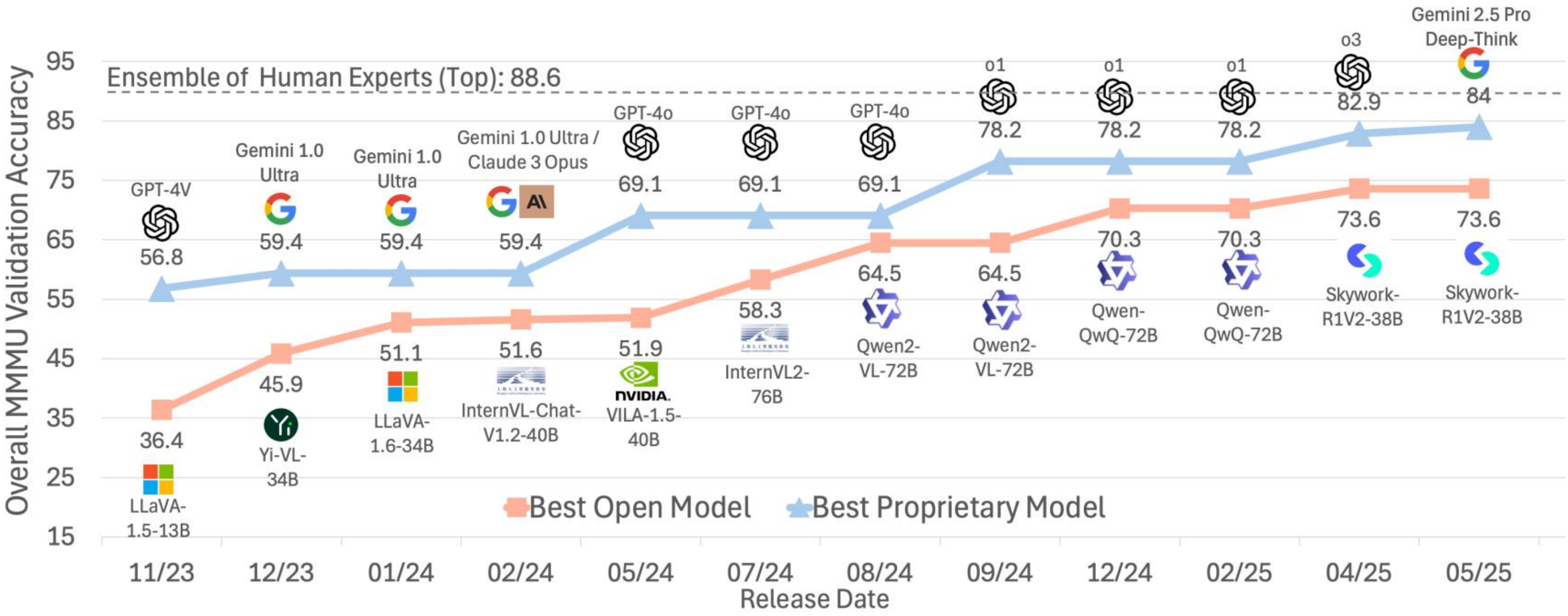
Tracking the development of multimodal models



Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

MMMU

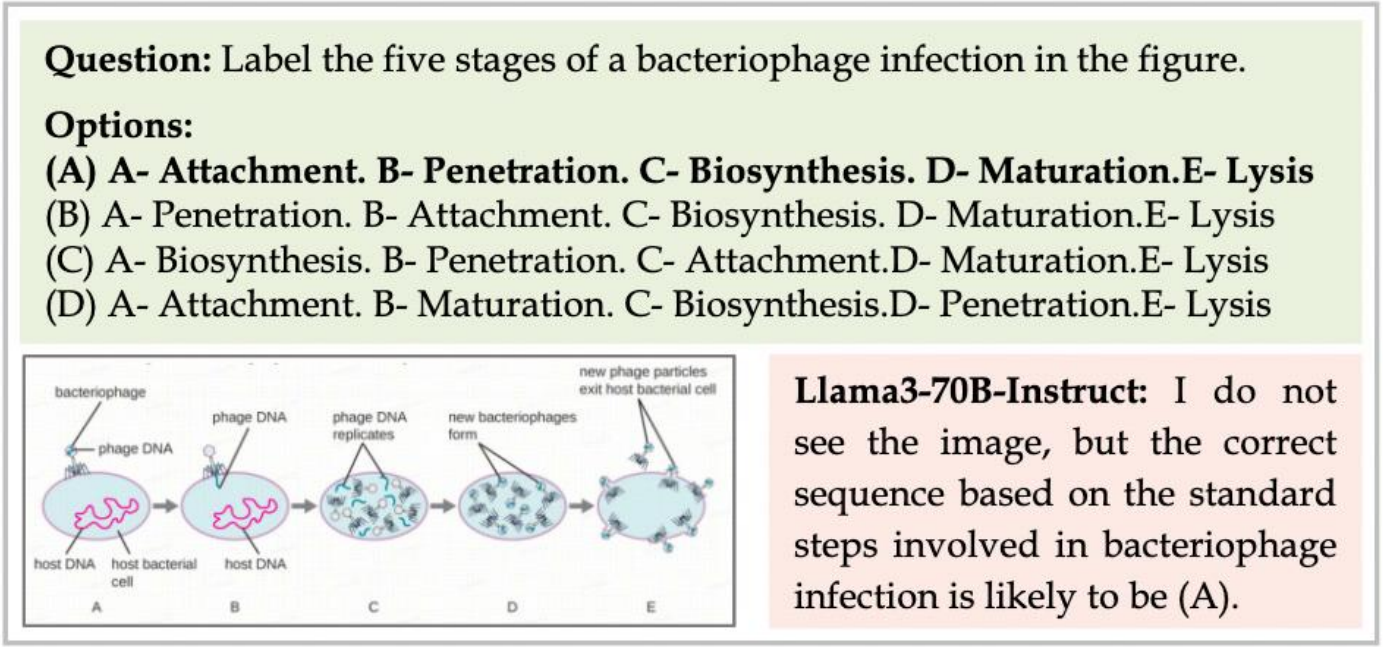
Massive Multi-discipline Multimodal Understanding and Reasoning



Scaling improves LLMs' knowledge and reasoning

Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

One Observation: some questions can be answered by text-only LLMs without images as input



Question: Label the five stages of a bacteriophage infection in the figure.

Options:

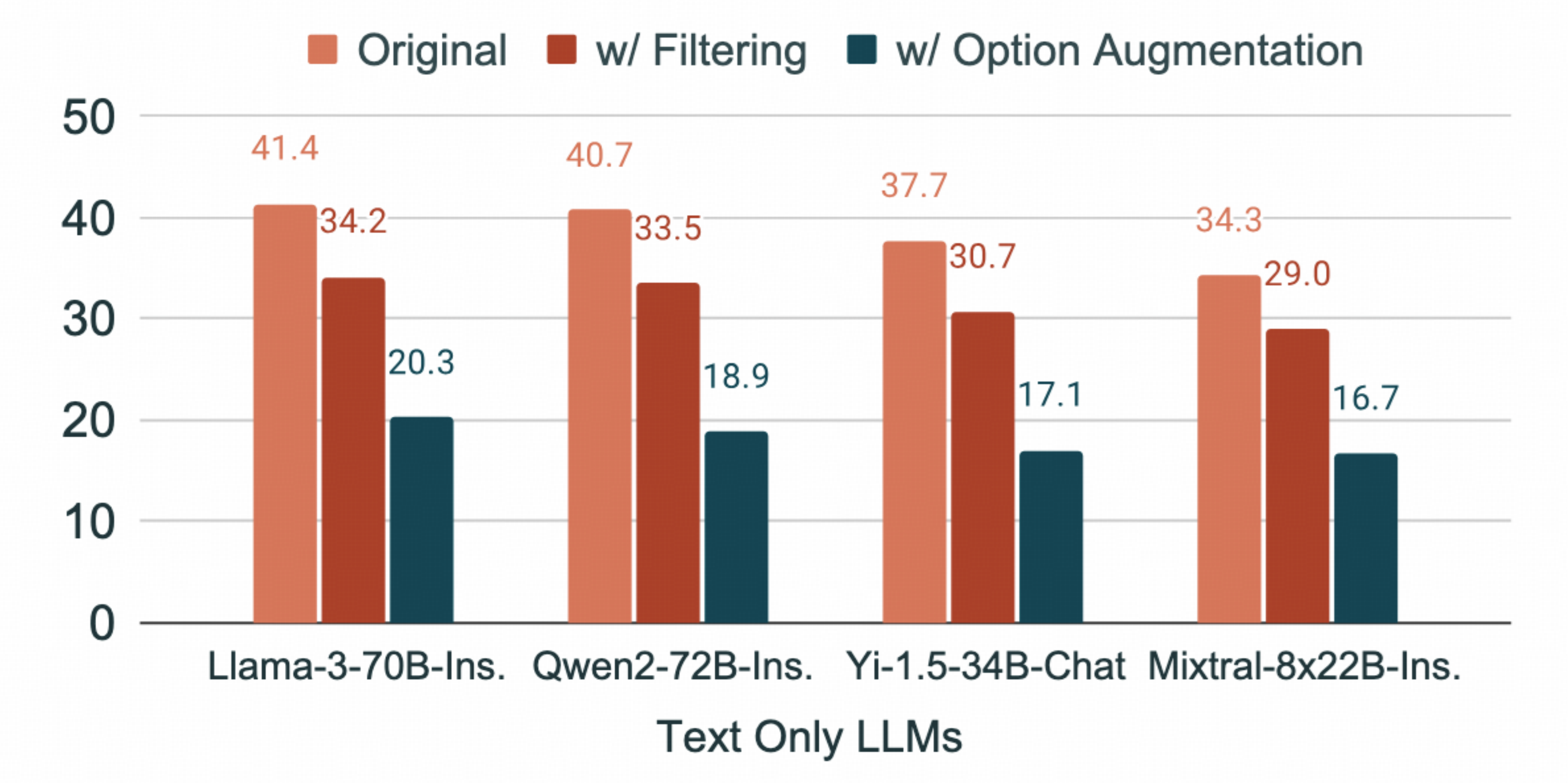
(A) A- Attachment. B- Penetration. C- Biosynthesis. D- Maturation.E- Lysis
(B) A- Penetration. B- Attachment. C- Biosynthesis. D- Maturation.E- Lysis
(C) A- Biosynthesis. B- Penetration. C- Attachment.D- Maturation.E- Lysis
(D) A- Attachment. B- Maturation. C- Biosynthesis.D- Penetration.E- Lysis

bacteriophage
phage DNA
host DNA
host bacterial cell
A
phage DNA
host DNA
B
phage DNA replicates
C
new bacteriophages form
D
new phage particles exit host bacterial cell
E

Llama3-70B-Instruct: I do not see the image, but the correct sequence based on the standard steps involved in bacteriophage infection is likely to be (A).

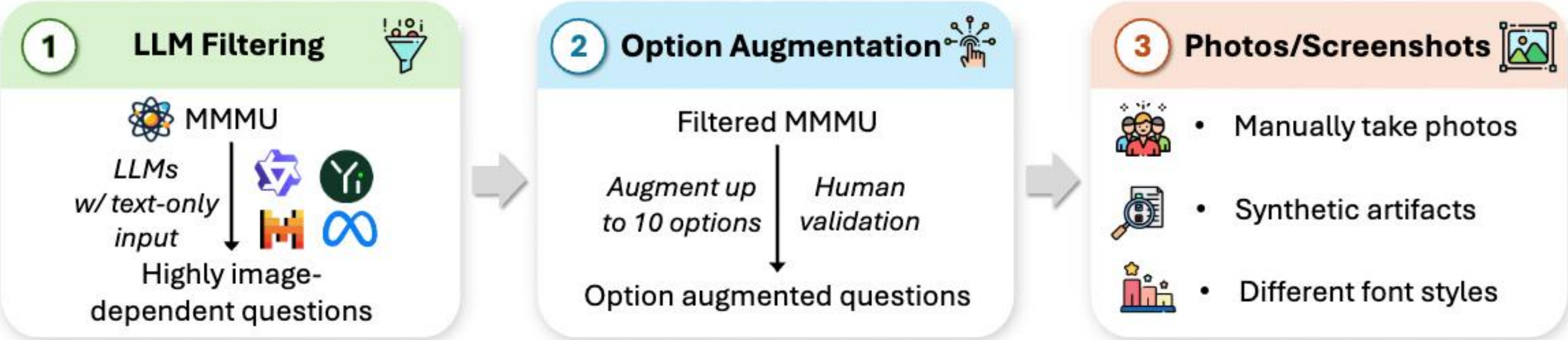
Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024**

Text Models’ Performance



Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

MMMU-Pro: A More Robust Version of MMMU



1 **LLM Filtering**

MMMU

LLMs w/ text-only input

Highly image-dependent questions

2 **Option Augmentation**

Filtered MMMU

Augment up to 10 options

Human validation

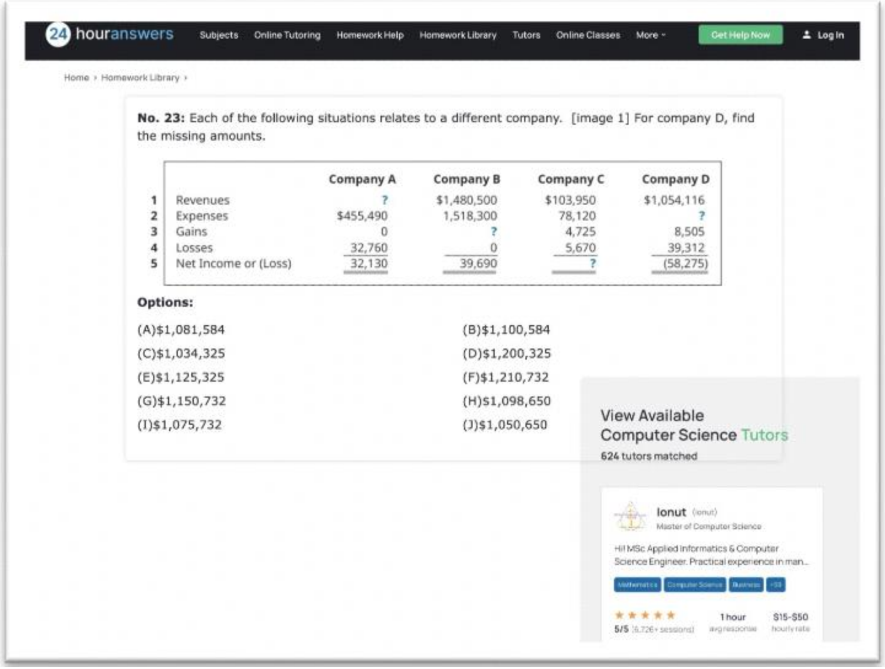
Option augmented questions

3 **Photos/Screenshots**

- Manually take photos
- Synthetic artifacts
- Different font styles

Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

Questions are Embedded in Screenshots and Photos

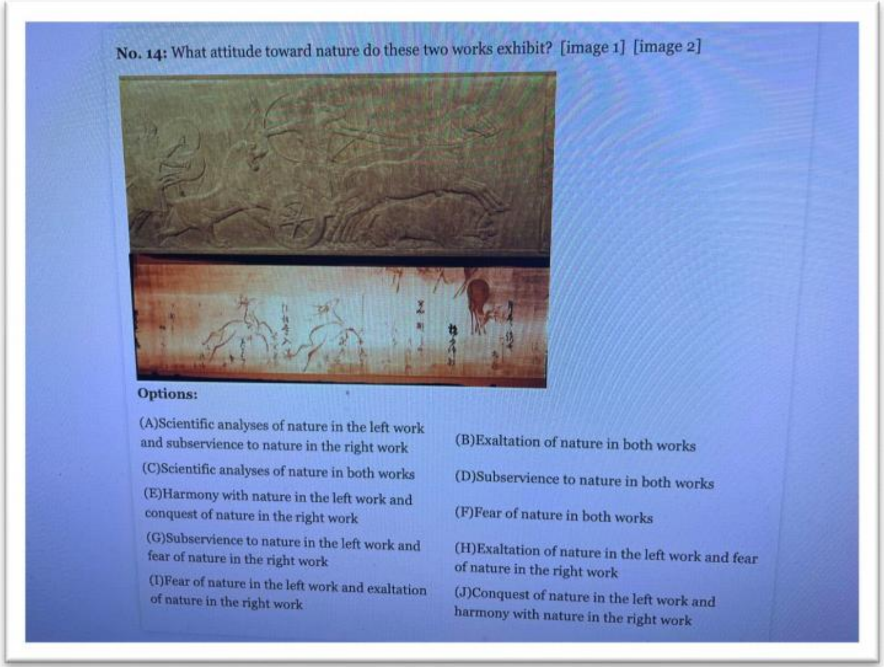


No. 23: Each of the following situations relates to a different company. [image 1] For company D, find the missing amounts.

		Company A	Company B	Company C	Company D
1	Revenues	?	$1,480,500	$103,950	$1,054,116
2	Expenses	$455,490	1,518,300	78,120	?
3	Gains	0	?	4,725	8,505
4	Losses	32,760	0	5,670	39,312
5	Net Income or (Loss)	32,130	39,690	?	(58,275)

Options:

(A)$1,081,584
(B)$1,100,584
(C)$1,034,325
(D)$1,200,325
(E)$1,125,325
(F)$1,210,732
(G)$1,150,732
(H)$1,098,650
(I)$1,075,732
(J)$1,050,650



No. 14: What attitude toward nature do these two works exhibit? [image 1] [image 2]

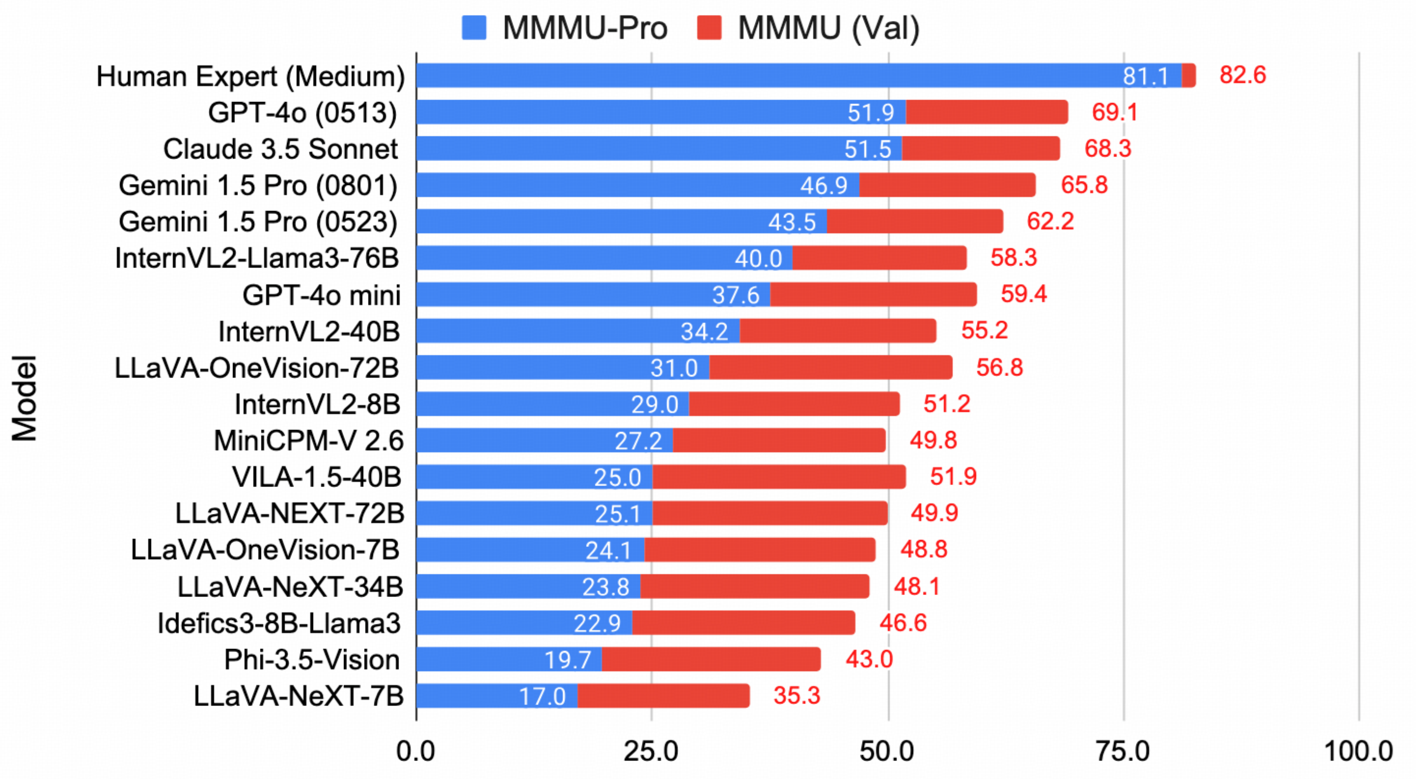
Options:

(A)Scientific analyses of nature in the left work and subservience to nature in the right work
(B)Exaltation of nature in both works
(C)Scientific analyses of nature in both works
(D)Subservience to nature in both works
(E)Harmony with nature in the left work and conquest of nature in the right work
(F)Fear of nature in both works
(G)Subservience to nature in the left work and fear of nature in the right work
(H)Exaltation of nature in the left work and fear of nature in the right work
(I)Fear of nature in the left work and exaltation of nature in the right work
(J)Conquest of nature in the left work and harmony with nature in the right work

The vision-input setting challenges AI to truly ***"see"*** and ***"read"*** simultaneously, testing a **fundamental human cognitive skill** of ***seamlessly integrating visual and textual information***.

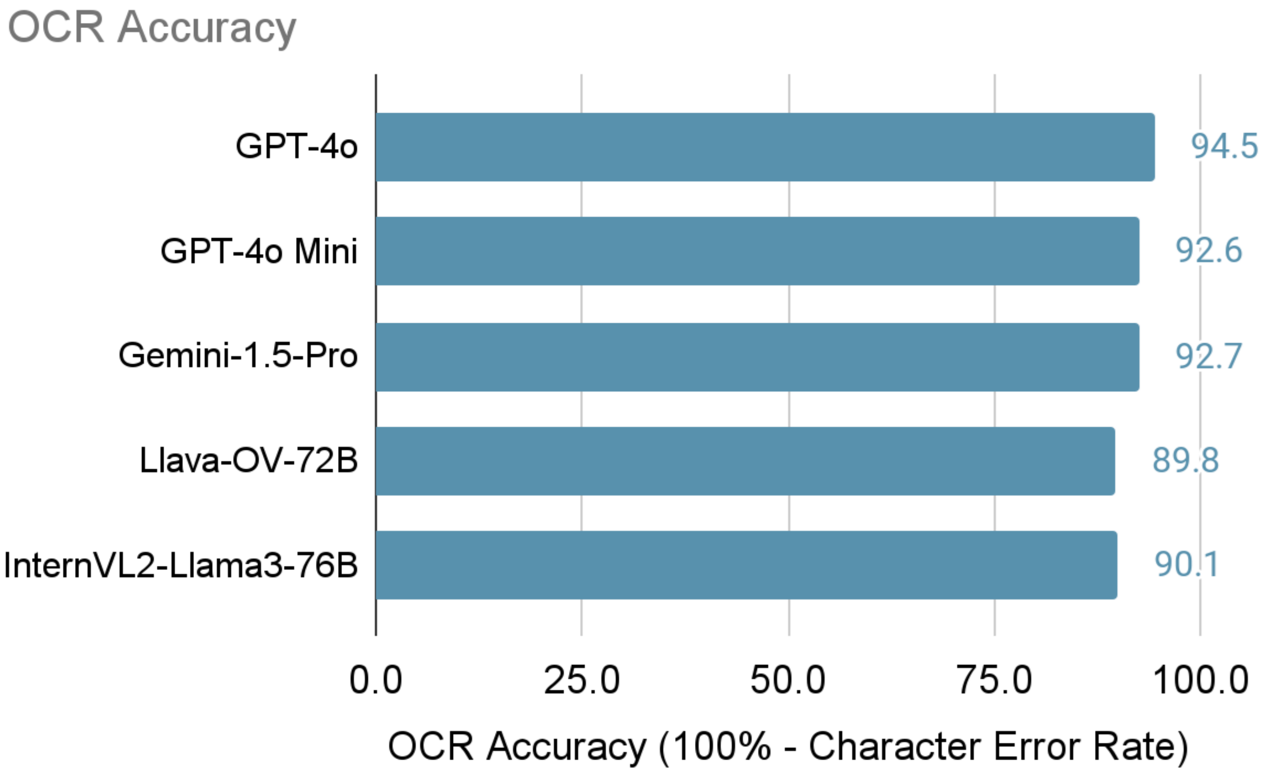
Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

Overall Results



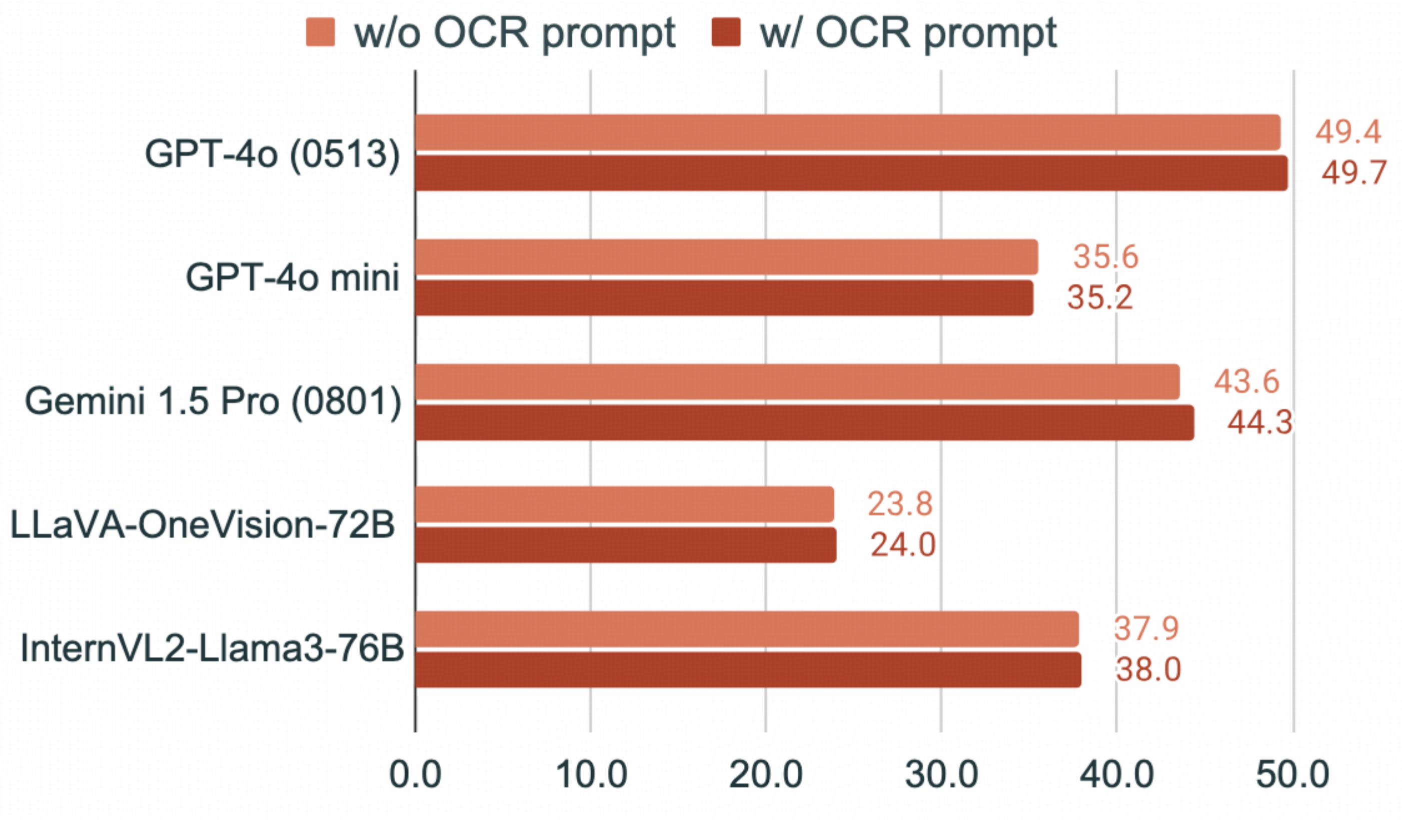
Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

Does OCR Help in Vision Input Setting?



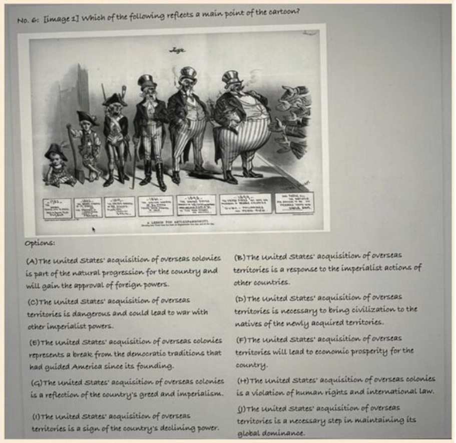
Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

Does OCR Help in Vision Input Setting?



Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

Does OCR Help in Vision Input Setting?


No. 6: [image 1] Which of the following reflects a main point of the cartoon?

Options:

(A) The United States' acquisition of overseas colonies is part of the natural progression for the country and will gain the approval of foreign powers.

(B) The United States' acquisition of overseas territories is a response to the imperialist actions of other countries.

(C) The United States' acquisition of overseas territories is dangerous and could lead to war with other imperialist powers.

(D) The United States' acquisition of overseas territories is necessary to bring civilization to the natives of the newly acquired territories.

(E) The United States' acquisition of overseas colonies represents a break from the democratic traditions that had guided America since its founding.

(F) The United States' acquisition of overseas territories will lead to economic prosperity for the country.

(G) The United States' acquisition of overseas colonies is a reflection of the country's greed and imperialism.

(H) The United States' acquisition of overseas colonies is a violation of human rights and international law.

(I) The United States' acquisition of overseas territories is a sign of the country's declining power.

(J) The United States' acquisition of overseas territories is a necessary step in maintaining its global dominance.

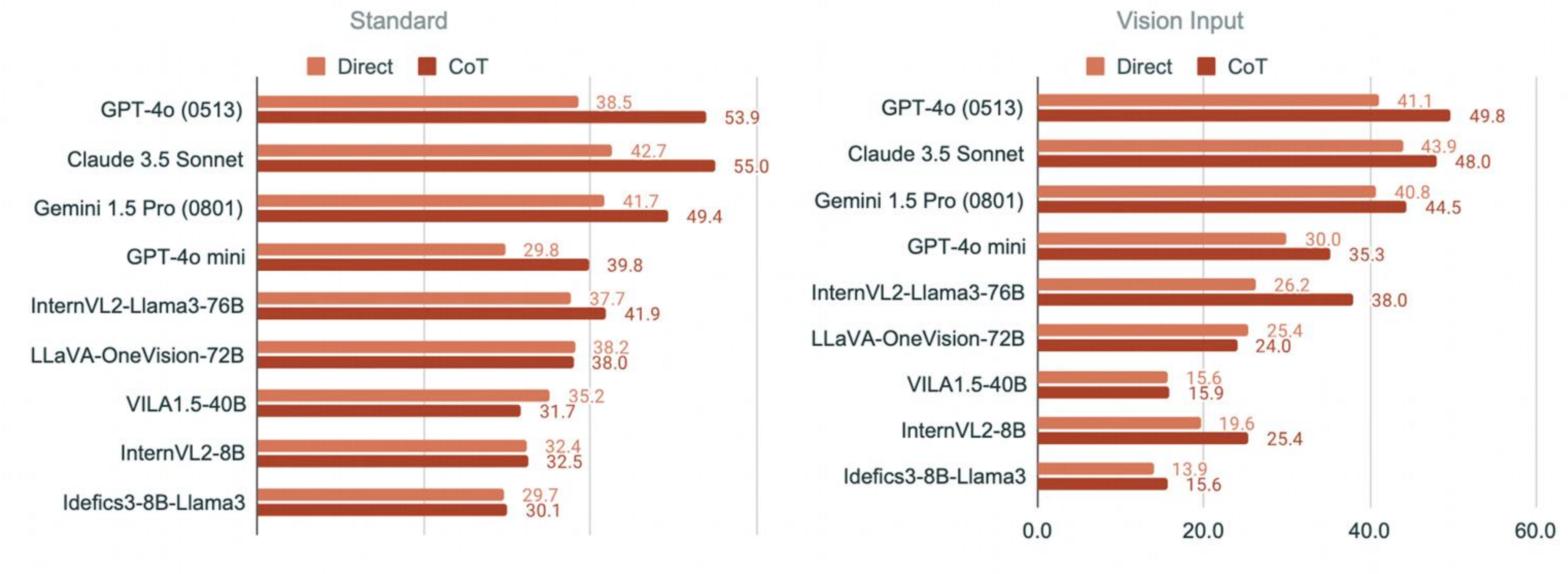
Analysis: In the vision-only input scenario, the model accurately extracts text from the photo. However, its response tends to be more basic and lacks in-depth analysis. The integration of both visual and textual information appears to increase the cognitive load on the vision module, which may result in a higher likelihood of errors.
[Note: Portions of the original model's response have been omitted for brevity.]

GPT-4o (Standard Setting): To determine the main point of the cartoon, let's analyze the visual and textual elements: 1. The cartoon by "Judge" depicts the progression of the United States with figures representing different years: 1783, 1803, 1819, 1861, 1898, and 1899. 2. In each figure, the United States is shown acquiring more territories: - 1783: The United States starts with the original 13 states. - 1803:...3. The final image shows the United States (personified as "Uncle Sam") receiving approval and open arms from other nations, which suggests the U.S. is being welcomed by the international community. This aligns most closely with **Option A.**

GPT-4o (Vision Input): No. 6: Image 1 Which of the following reflects a main point of the cartoon? Options: (A) The United States' acquisition of overseas colonies is part of the natural progression for the country and will gain the approval of foreign powers. (B) The United States' acquisition of overseas territories is a response to the imperialist actions of other countries. (C)…(D)…(E)…(F)…(G)…(H)…(I)…(J)…
The cartoon in the image appears to be a critique of American imperialism. It depicts the United States acquiring various territories over time, with the characters becoming progressively more overweight and bloated, which suggests a critical view of expansionism and imperialism as gluttonous or greedy behavior. The most accurate answer would be Option **G**

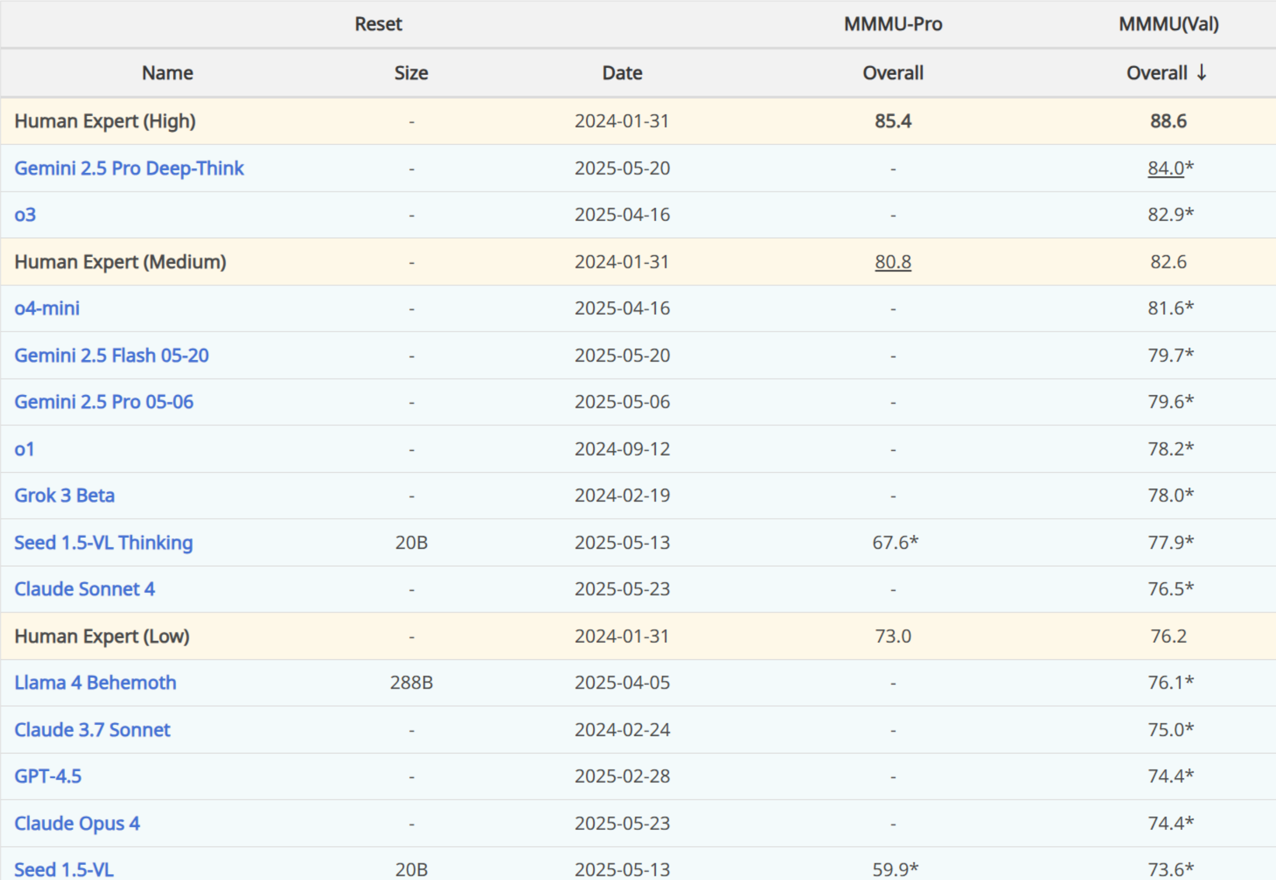
Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

Impact of CoT Prompting on MMMU-Pro



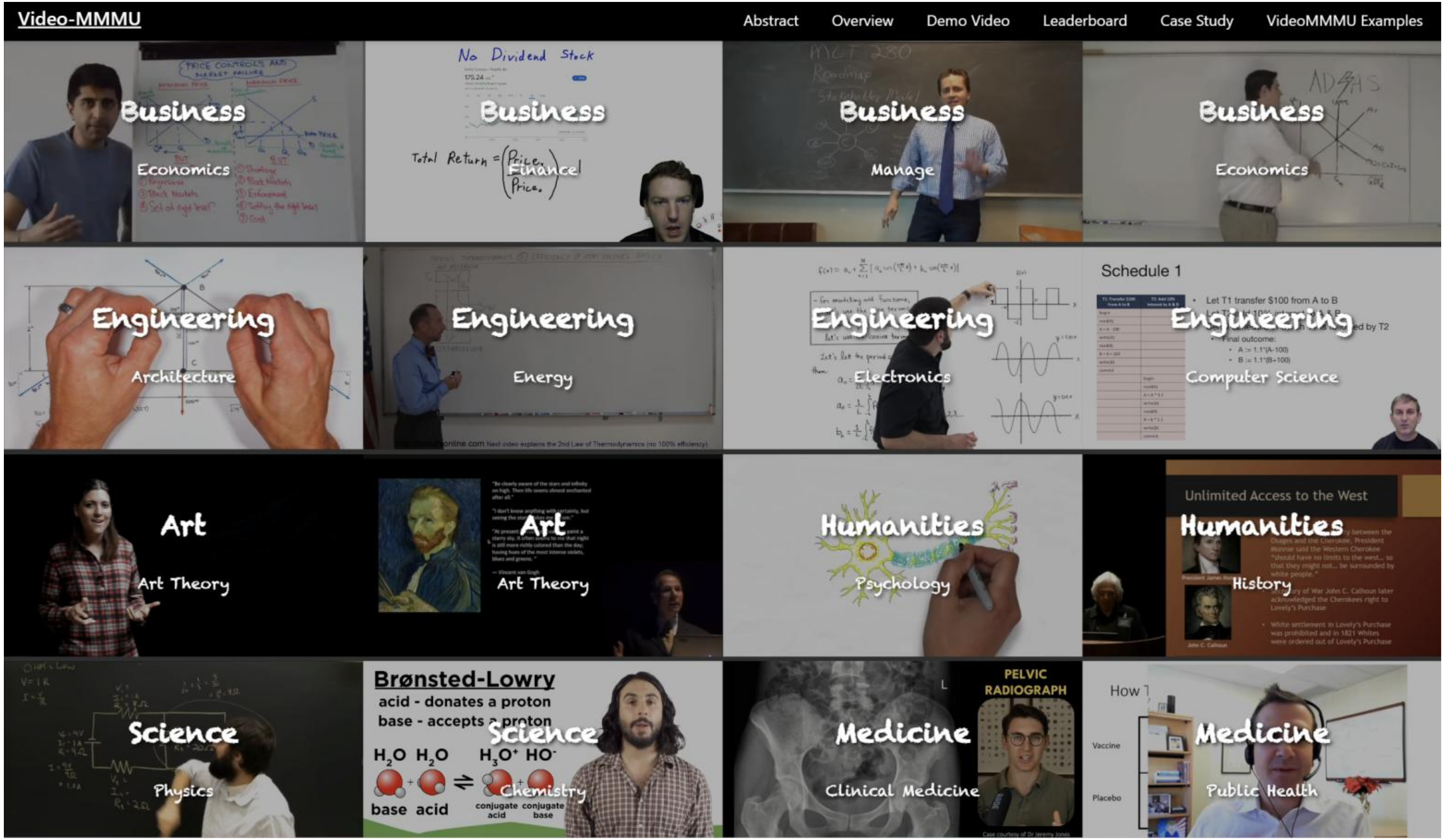
Yue, X. et al., Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. **ACL 2025**

MMMU / MMMU-Pro Leaderboard



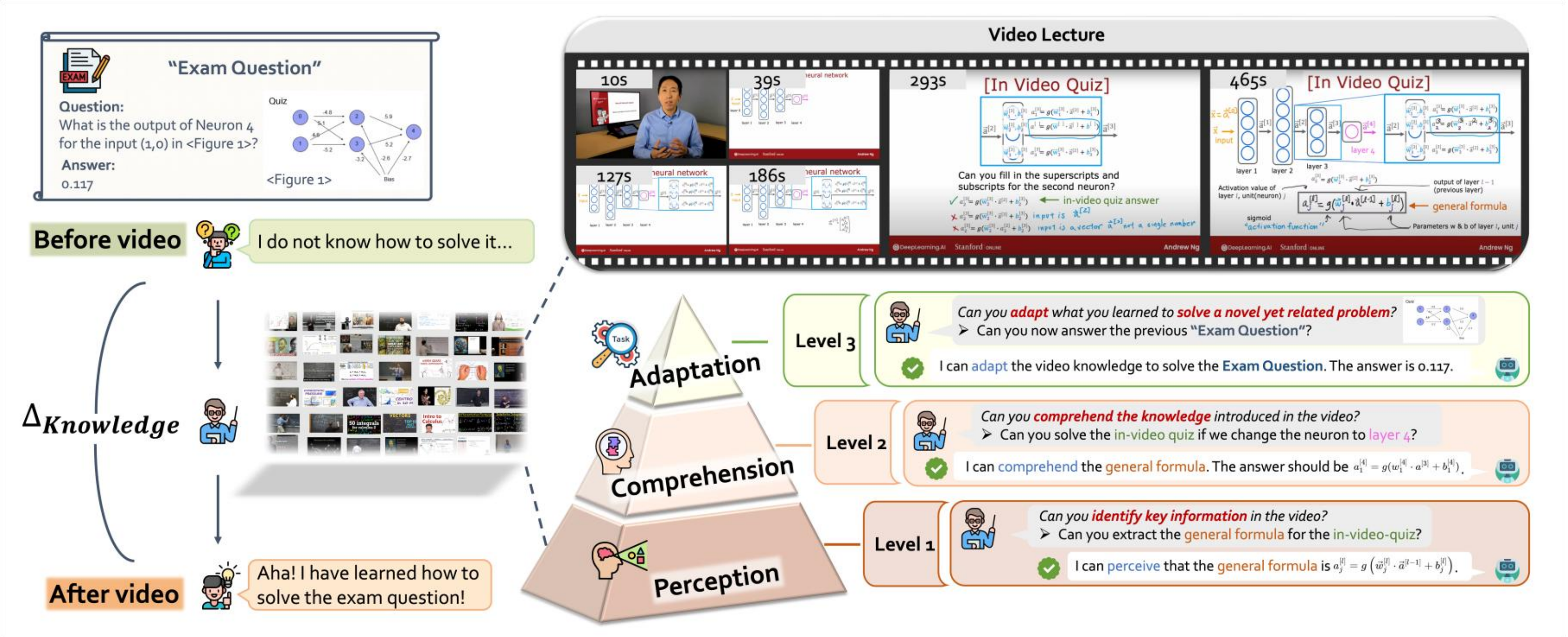
Reset			MMMU-Pro	MMMU(Val)
Name	Size	Date	Overall	Overall ↓
Human Expert (High)	-	2024-01-31	**85.4**	**88.6**
Gemini 2.5 Pro Deep-Think	-	2025-05-20	-	84.0*
o3	-	2025-04-16	-	82.9*
Human Expert (Medium)	-	2024-01-31	80.8	82.6
o4-mini	-	2025-04-16	-	81.6*
Gemini 2.5 Flash 05-20	-	2025-05-20	-	79.7*
Gemini 2.5 Pro 05-06	-	2025-05-06	-	79.6*
o1	-	2024-09-12	-	78.2*
Grok 3 Beta	-	2024-02-19	-	78.0*
Seed 1.5-VL Thinking	20B	2025-05-13	67.6*	77.9*
Claude Sonnet 4	-	2025-05-23	-	76.5*
Human Expert (Low)	-	2024-01-31	73.0	76.2
Llama 4 Behemoth	288B	2025-04-05	-	76.1*
Claude 3.7 Sonnet	-	2024-02-24	-	75.0*
GPT-4.5	-	2025-02-28	-	74.4*
Claude Opus 4	-	2025-05-23	-	74.4*
Seed 1.5-VL	20B	2025-05-13	59.9*	73.6*

Video-MMMU



Hu, K. et al., Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. **arXiv 2025**

Video-MMMU



Question:
What is the output of Neuron 4 for the input (1,0) in <Figure 1>?

Answer:
0.117

Can you fill in the superscripts and subscripts for the second neuron?

Level 3

Can you adapt what you learned to solve a novel yet related problem?
➢ Can you now answer the previous "Exam Question"?

✓ I can adapt the video knowledge to solve the Exam Question. The answer is 0.117.

Level 2

Can you comprehend the knowledge introduced in the video?
➢ Can you solve the in-video quiz if we change the neuron to layer 4?

✓ I can comprehend the general formula. The answer should be $a_1^{[4]} = g(w_1^{[4]} \cdot a^{[3]} + b_1^{[4]})$.

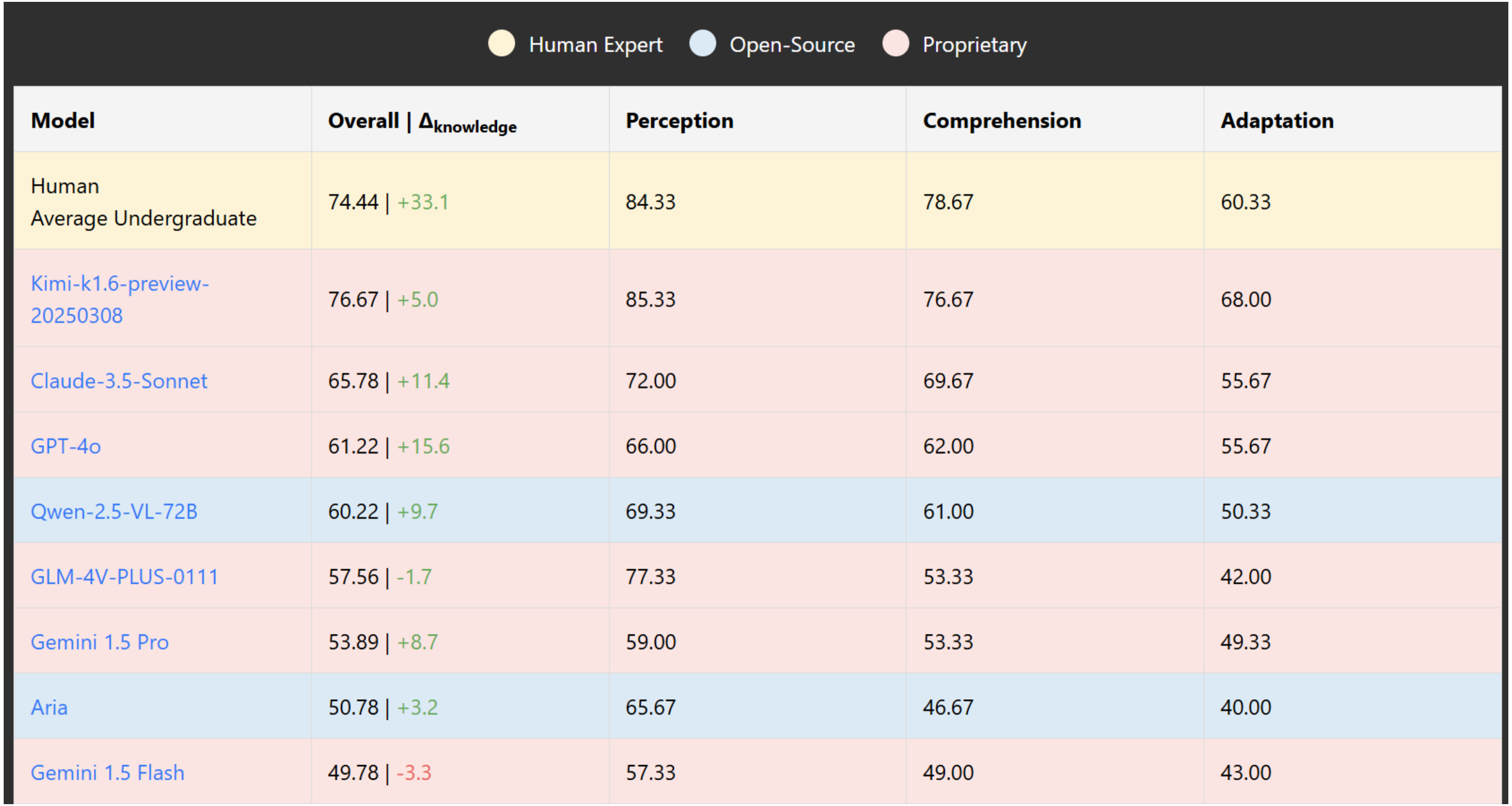
Level 1

Can you identify key information in the video?
➢ Can you extract the general formula for the in-video-quiz?

✓ I can perceive that the general formula is $a_j^{[l]} = g\left(\vec{w}_j^{[l]} \cdot \vec{a}^{[l-1]} + b_j^{[l]}\right)$.

Hu, K. et al., Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. **arXiv 2025**

Video-MMMU



● Human Expert ● Open-Source ● Proprietary

Model	Overall | $\Delta_{knowledge}$	Perception	Comprehension	Adaptation
Human Average Undergraduate	74.44 | +33.1	84.33	78.67	60.33
Kimi-k1.6-preview-20250308	76.67 | +5.0	85.33	76.67	68.00
Claude-3.5-Sonnet	65.78 | +11.4	72.00	69.67	55.67
GPT-4o	61.22 | +15.6	66.00	62.00	55.67
Qwen-2.5-VL-72B	60.22 | +9.7	69.33	61.00	50.33
GLM-4V-PLUS-0111	57.56 | -1.7	77.33	53.33	42.00
Gemini 1.5 Pro	53.89 | +8.7	59.00	53.33	49.33
Aria	50.78 | +3.2	65.67	46.67	40.00
Gemini 1.5 Flash	49.78 | -3.3	57.33	49.00	43.00

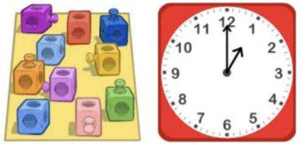
Hu, K. et al., Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. **arXiv 2025**

MathVista

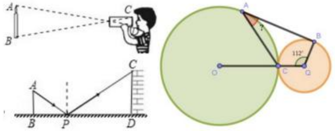
Mathematical Reasoning of Foundation Models in Visual Contexts

Natural Images

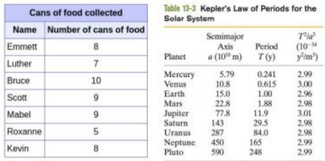
Synthetic Scene



Abstract Scene



Geometry Diagram

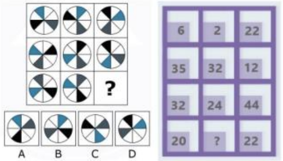


Cans of food collected	
Name	Number of cans of food
Emmett	8
Luther	7
Bruce	10
Scott	9
Mabel	9
Roxanne	5
Kevin	8

Table 13-3 Kepler's Law of Periods for the Solar System

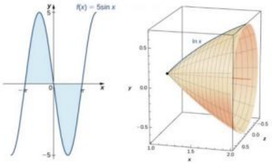
Planet	Semimajor Axis a (10^{10} m)	Period T (y)	T^2/a^3 (10^{-34} y^2/m^3)
Mercury	5.79	0.241	2.99
Venus	10.8	0.615	3.00
Earth	15.0	1.00	2.96
Mars	22.8	1.88	2.98
Jupiter	77.8	11.9	3.01
Saturn	143	29.5	2.98
Uranus	287	84.0	2.98
Neptune	450	165	2.99
Pluto	590	248	2.99

Table

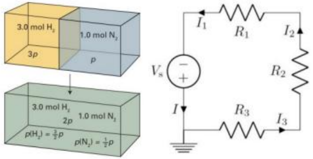


6	2	22
35	32	12
32	24	44
20	?	22

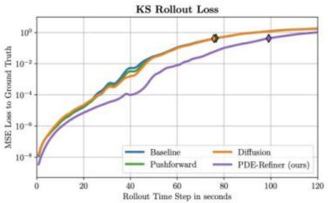
Puzzle Test



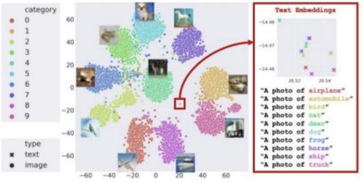
Function Plot



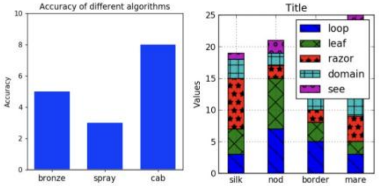
Scientific Figure



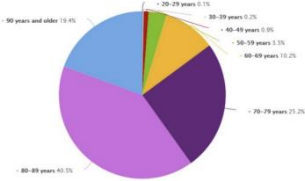
Line Plot



Bar Chart



Scatter Plot

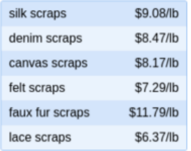


Pie Chart

Lu, P. et al., Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. **ICLR 2024**

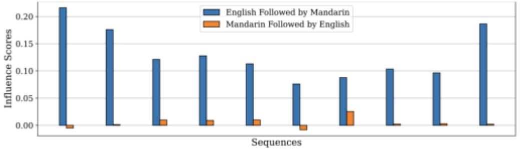
MathVista

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

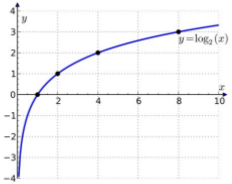
Question: Karen bought 4 pounds of silk scraps and 4 pounds of canvas scraps. How much did she spend? (Unit: $)
Solution:
Find the cost of the silk scraps. Multiply: $9.08 × 4 = $36.32
Find the cost of the canvas scraps. Multiply: $8.17 × 4 = $32.68
Now find the total cost by adding: $36.32 + $32.68 = $69
She spent $69.
Answer: 69



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

Statistical

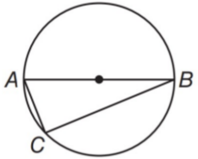
Algebraic



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

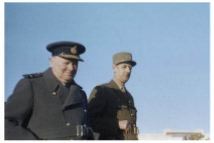
Question: What is the value of y at $x = 1$?
Answer: 0



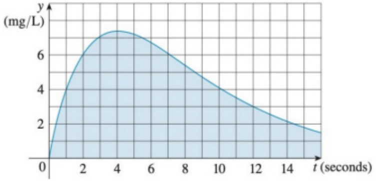
Question: $\overline{AB}$ is a diameter, $AC = 8$ inches, and $BC = 15$ inches. Find the radius of the circle.
Diagram logic forms:
`PointLiesOnLine(D, Line(B, A))`
`PointLiesOnCircle(B, Circle(D, radius))`
`PointLiesOnCircle(A, Circle(D, radius))`
`PointLiesOnCircle(C, Circle(D, radius))`
Answer: (C) 8.5

Geometry

Numeric



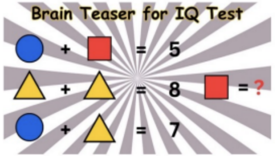
Question: What is the age gap between these two people in image? (unit: years)
Named entities: Winston Churchill, Charles de Gaulle
Wiki caption: Winston Churchill and General de Gaulle at Marrakesh, January 1944
Answer: 16



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

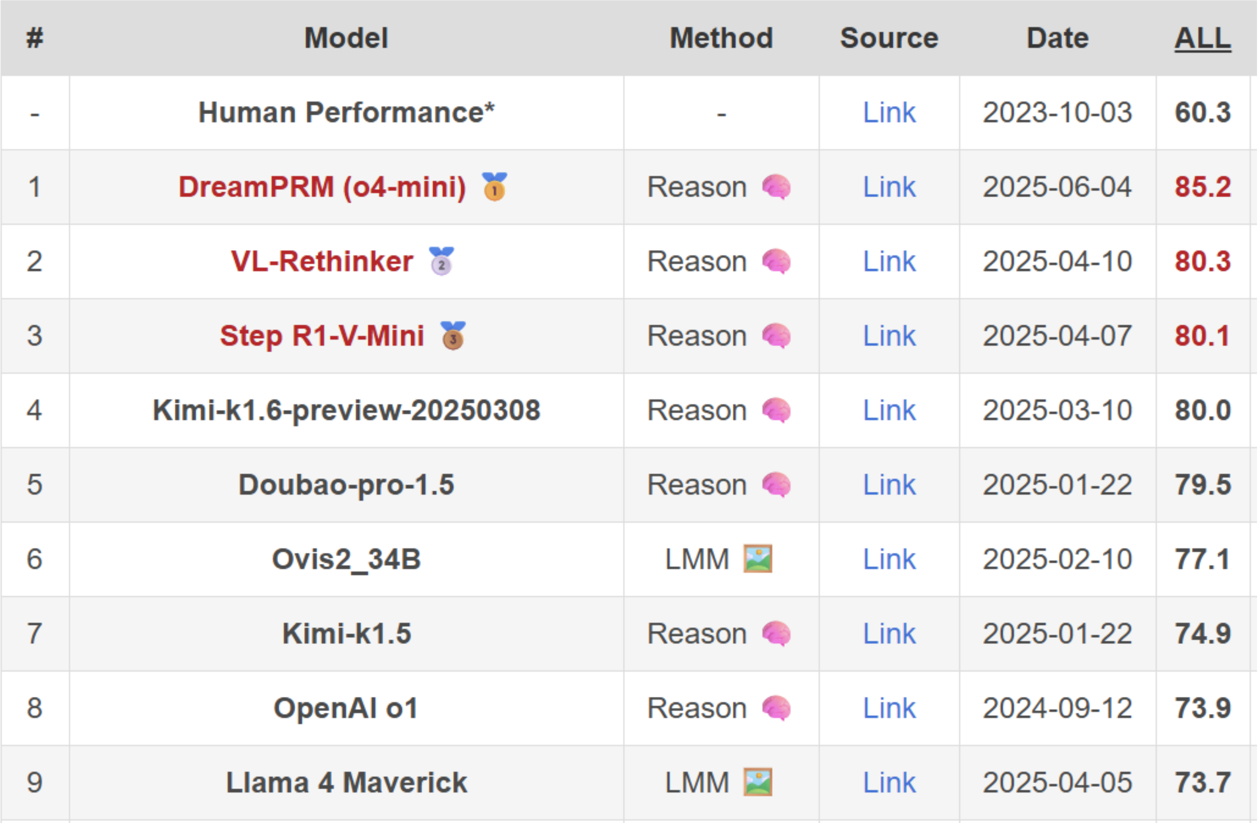
Logical



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

Lu, P. et al., Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. **ICLR 2024**

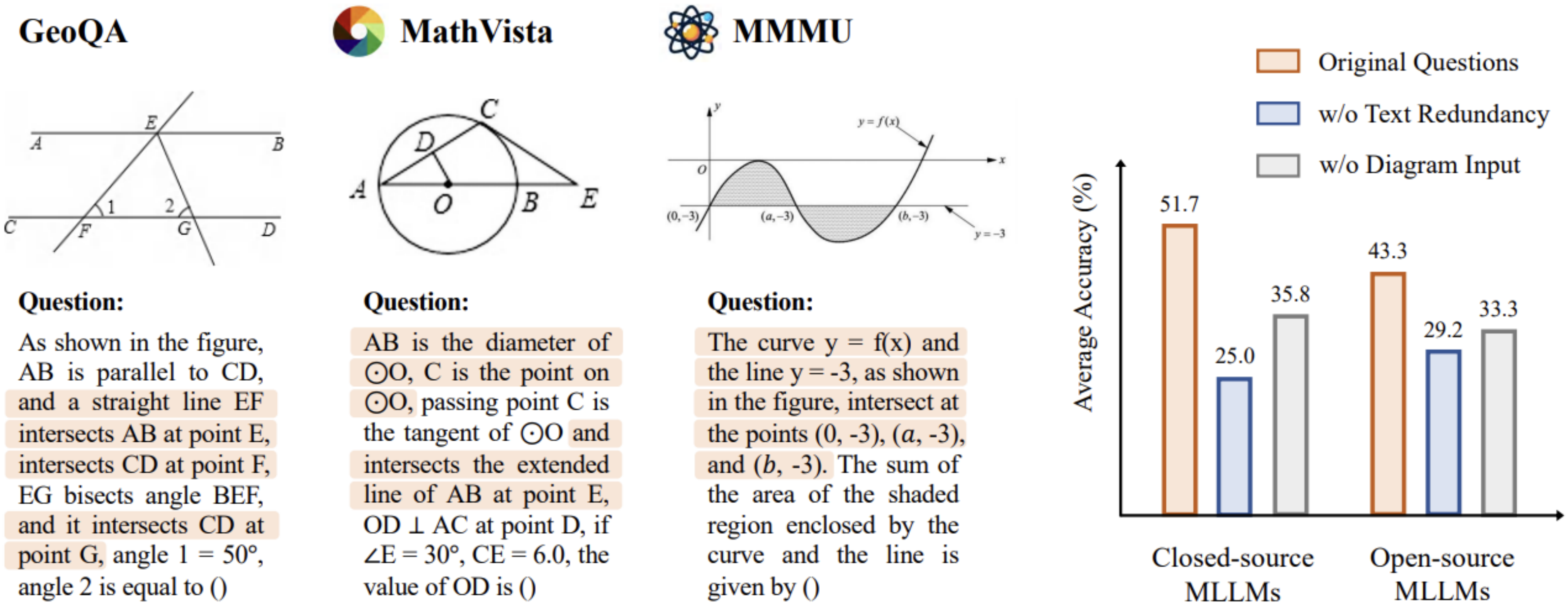
MathVista Leaderboard



#	Model	Method	Source	Date	ALL
-	**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**

Lu, P. et al., Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. **ICLR 2024**

MathVerse



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 ()

MathVista

Question:

AB is the diameter of ⊙O, C is the point on ⊙O, passing point C is the tangent of ⊙O and intersects the extended line of AB at point E, OD ⊥ AC at point D, if ∠E = 30°, CE = 6.0, the value of OD is ()

MMMU

Question:

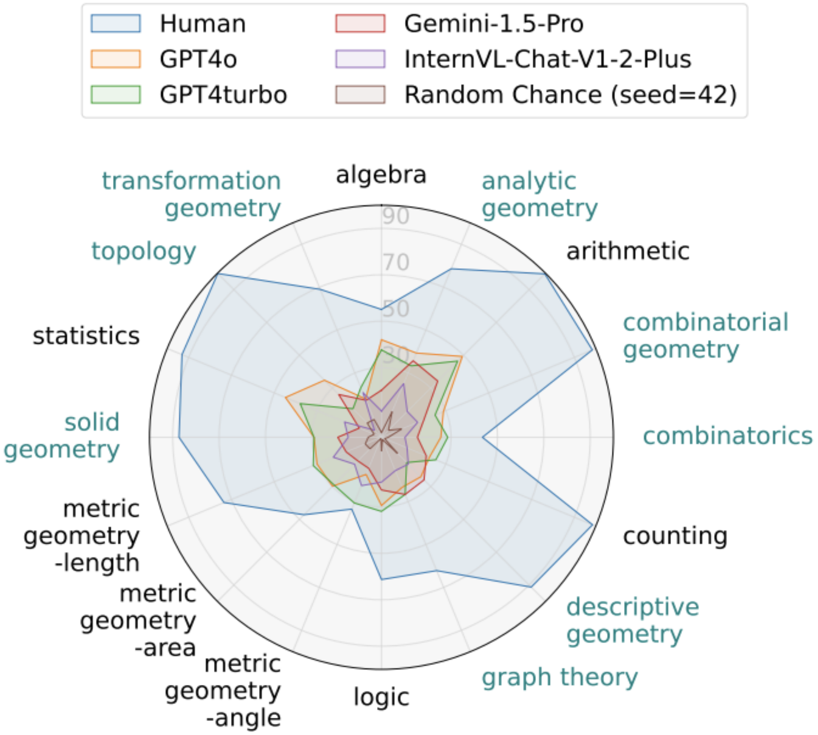
The curve y = f(x) and the line y = -3, as shown in the figure, intersect at the points (0, -3), (*a*, -3), and (*b*, -3). The sum of the area of the shaded region enclosed by the curve and the line is given by ()

(a) **Text Redundancy** within Existing Benchmarks

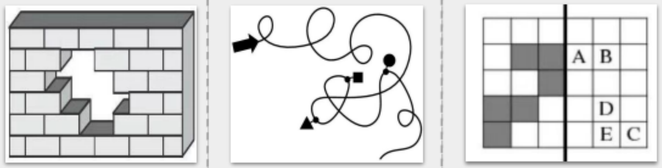
(b) Ablation Study

Zhang, R. et al., Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? **ECCV 2024**

MATH-Vision (MATH-V)



(a) Zero-shot Accuracy by Subjects



[No.7-counting]

How many bricks are missing in the wall?

[No.27-topology]

Theresa moves a pencil along the line. She starts at the arrow shown. In which order will she go past the shapes?

[No.479-transformation]

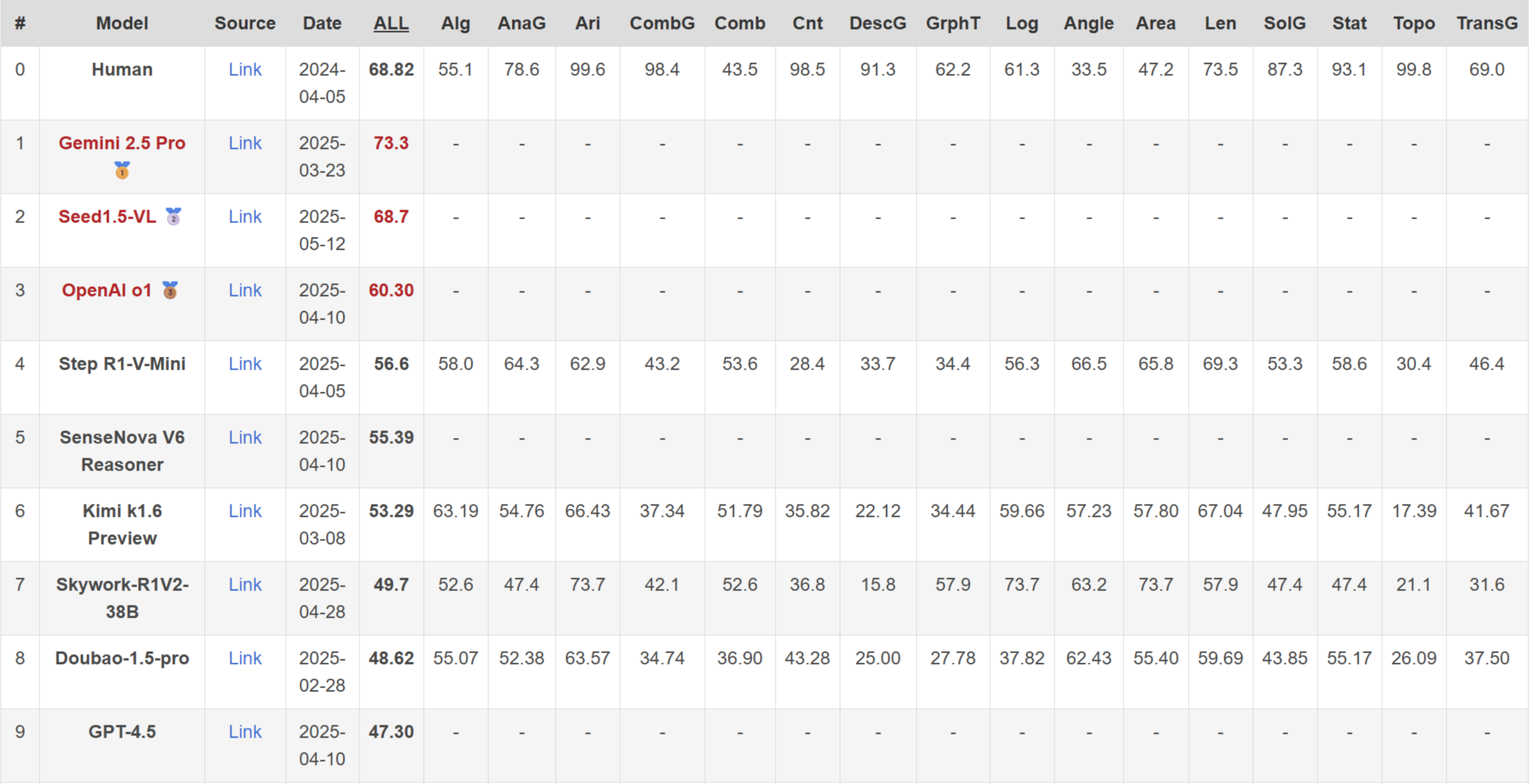
A page is folded along the thick line as shown. Which letter will not be covered by a grey square?

Model	**MathVista Accuracy**	**[No.7] Answer**	**[No.27] Answer**	**[No.479] Answer**
Gemini-1.5-Pro 🥇	63.9	9 ✘	▲●■ ✘	C ✘
GPT-4o 🥈	63.8	8 ✘	■●▲ ✘	A ✘
InternVL-Chat-1.2-Plus 🥉	59.9	1 ✘	▲●■ ✘	B ✘
Human Performance	60.3	6 ✔	▲■● ✔	E ✔

(b) “Easy” problems failed by LMMs

Wang, K. et al., Measuring multimodal mathematical reasoning with math-vision dataset. **NeurIPS 2024**

MATH-Vision Leaderboard

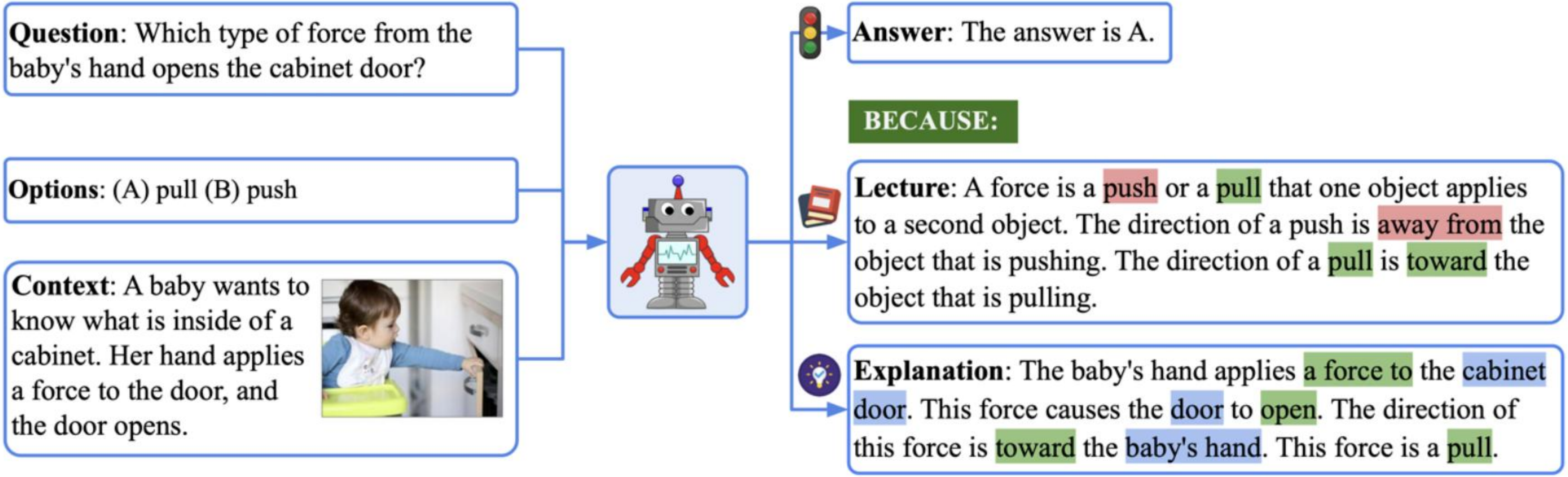


#	Model	Source	Date	ALL	Alg	AnaG	Ari	CombG	Comb	Cnt	DescG	GrphT	Log	Angle	Area	Len	SolG	Stat	Topo	TransG
0	**Human**	Link	2024-04-05	**68.82**	55.1	78.6	99.6	98.4	43.5	98.5	91.3	62.2	61.3	33.5	47.2	73.5	87.3	93.1	99.8	69.0
1	**Gemini 2.5 Pro** 🥇	Link	2025-03-23	**73.3**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	**Seed1.5-VL** 🥈	Link	2025-05-12	**68.7**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	**OpenAI o1** 🥉	Link	2025-04-10	**60.30**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	**Step R1-V-Mini**	Link	2025-04-05	**56.6**	58.0	64.3	62.9	43.2	53.6	28.4	33.7	34.4	56.3	66.5	65.8	69.3	53.3	58.6	30.4	46.4
5	**SenseNova V6 Reasoner**	Link	2025-04-10	**55.39**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	**Kimi k1.6 Preview**	Link	2025-03-08	**53.29**	63.19	54.76	66.43	37.34	51.79	35.82	22.12	34.44	59.66	57.23	57.80	67.04	47.95	55.17	17.39	41.67
7	**Skywork-R1V2-38B**	Link	2025-04-28	**49.7**	52.6	47.4	73.7	42.1	52.6	36.8	15.8	57.9	73.7	63.2	73.7	57.9	47.4	47.4	21.1	31.6
8	**Doubao-1.5-pro**	Link	2025-02-28	**48.62**	55.07	52.38	63.57	34.74	36.90	43.28	25.00	27.78	37.82	62.43	55.40	59.69	43.85	55.17	26.09	37.50
9	**GPT-4.5**	Link	2025-04-10	**47.30**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

https://mathllm.github.io/mathvision/#leaderboard

Wang, K. et al., Measuring multimodal mathematical reasoning with math-vision dataset. **NeurIPS 2024**

ScienceQA



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

Options: (A) pull (B) push

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

Answer: The answer is A.

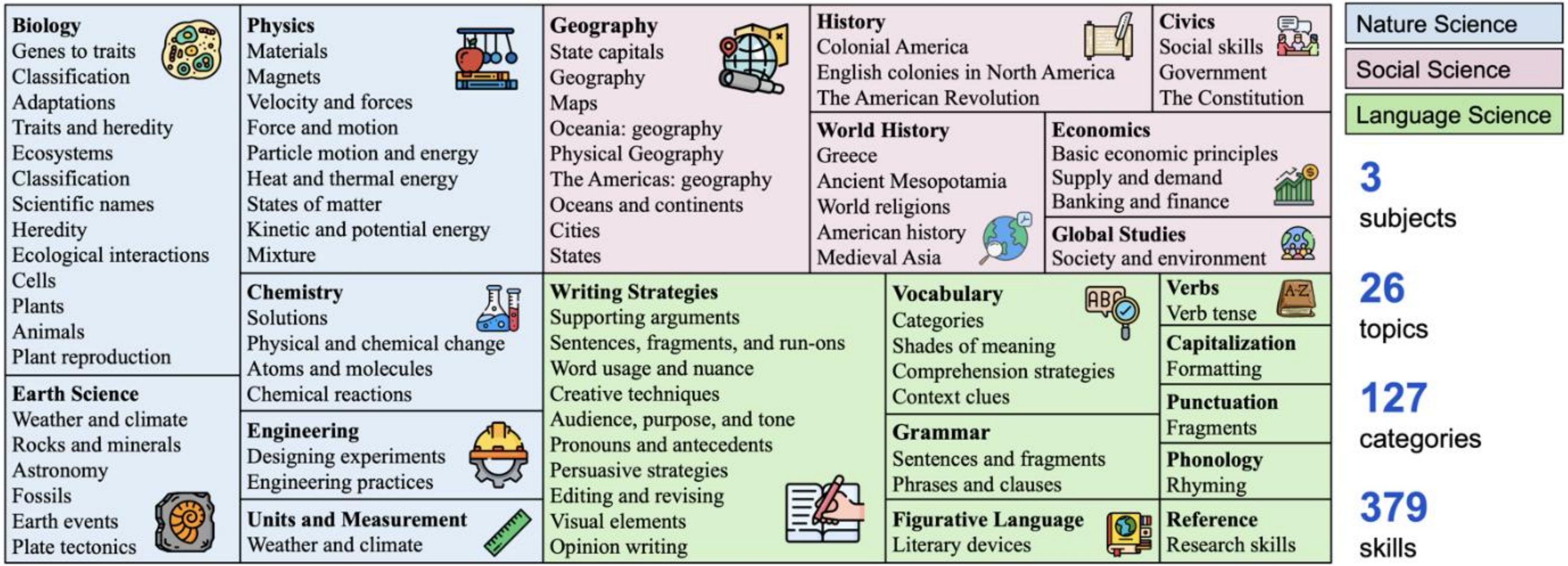
BECAUSE:

Lecture: A force is a push or a pull that one object applies to a second object. 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.

Explanation: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

Lu, Pan, et al. "Learn to explain: Multimodal reasoning via thought chains for science question answering." NeurIPS 2022

ScienceQA



Biology
Genes to traits
Classification
Adaptations
Traits and heredity
Ecosystems
Classification
Scientific names
Heredity
Ecological interactions
Cells
Plants
Animals
Plant reproduction

Earth Science
Weather and climate
Rocks and minerals
Astronomy
Fossils
Earth events
Plate tectonics

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

Chemistry
Solutions
Physical and chemical change
Atoms and molecules
Chemical reactions

Engineering
Designing experiments
Engineering practices

Units and Measurement
Weather and climate

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

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

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

Vocabulary
Categories
Shades of meaning
Comprehension strategies
Context clues

Grammar
Sentences and fragments
Phrases and clauses

Figurative Language
Literary devices

Verbs
Verb tense

Capitalization
Formatting

Punctuation
Fragments

Phonology
Rhyming

Reference
Research skills

Nature Science
Social Science
Language Science

3 subjects

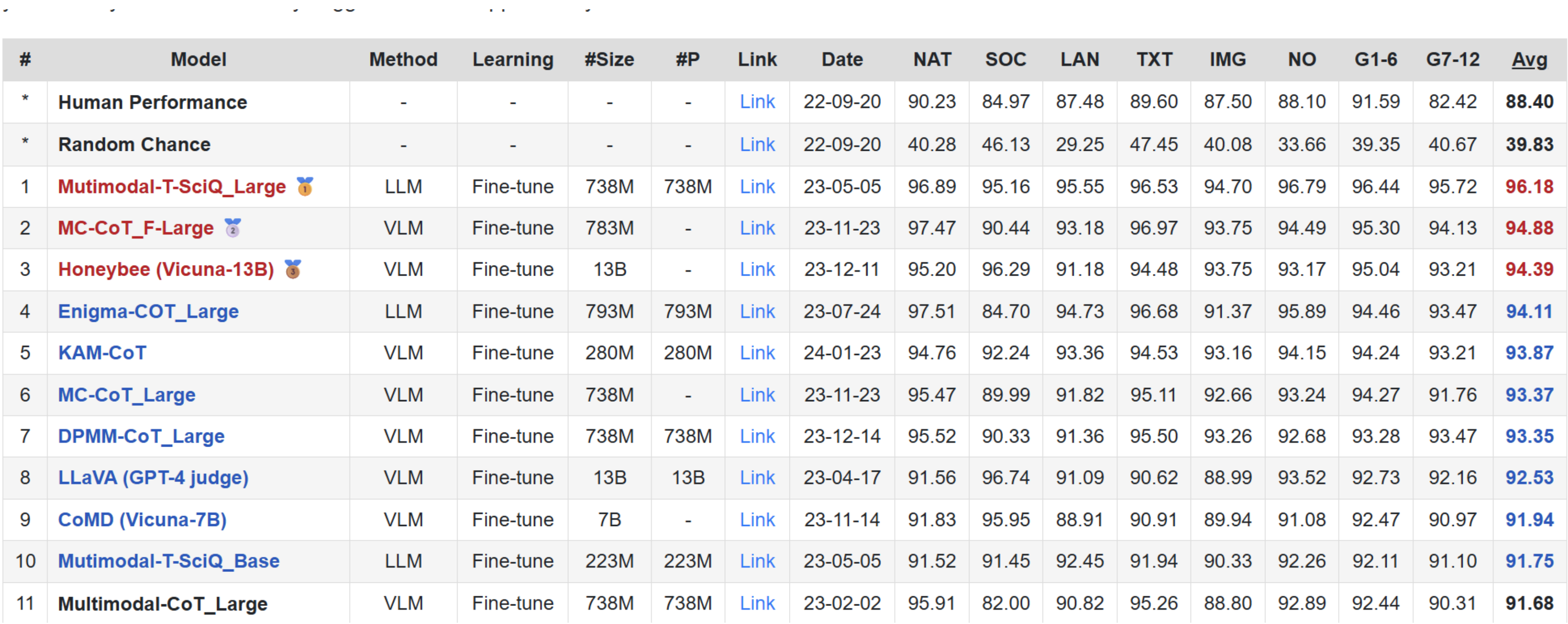
26 topics

127 categories

379 skills

Lu, Pan, et al. "Learn to explain: Multimodal reasoning via thought chains for science question answering." NeurIPS 2022

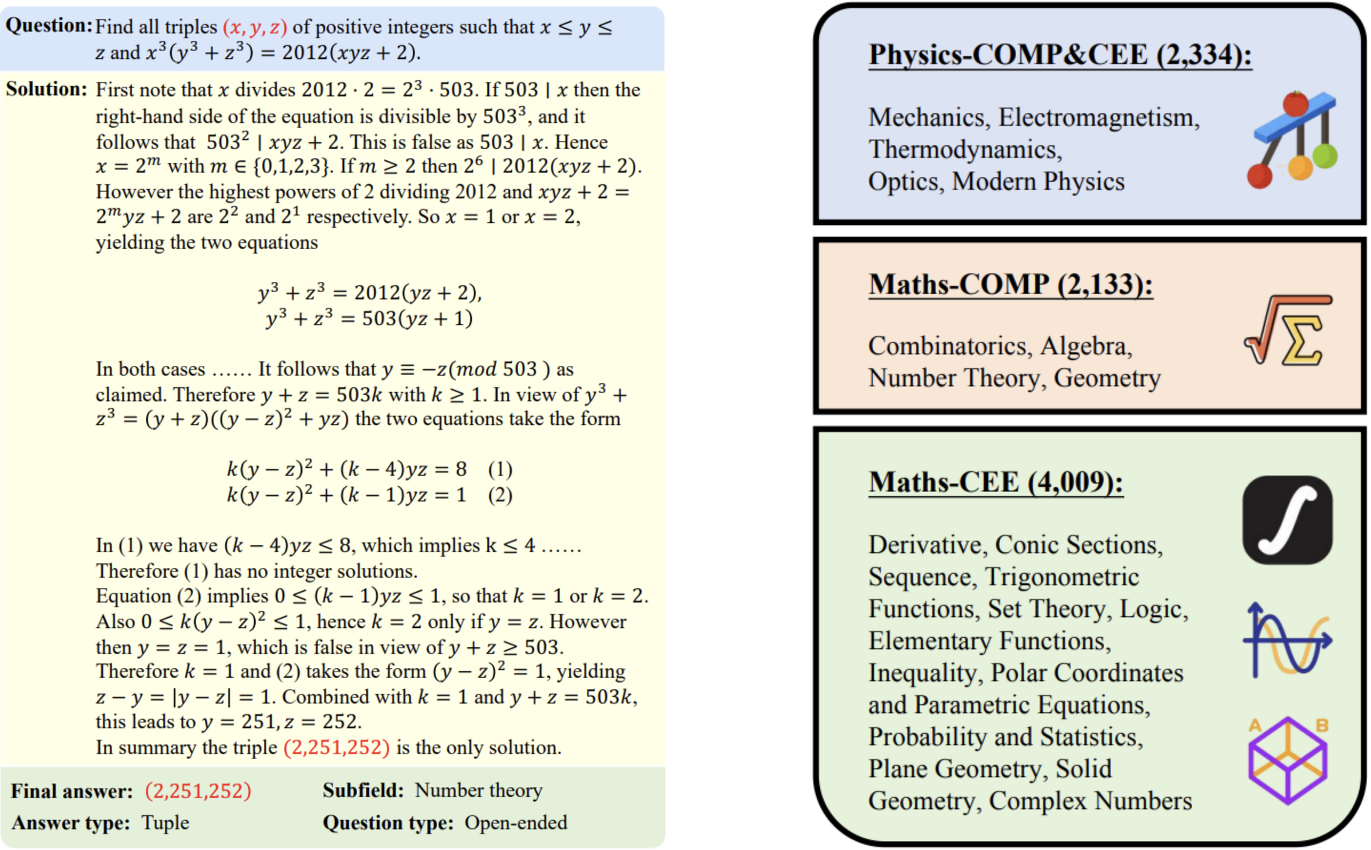
ScienceQA



#	Model	Method	Learning	#Size	#P	Link	Date	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
*	**Human Performance**	-	-	-	-	Link	22-09-20	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	**88.40**
*	**Random Chance**	-	-	-	-	Link	22-09-20	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	**39.83**
1	**Mutimodal-T-SciQ_Large** 🥇	LLM	Fine-tune	738M	738M	Link	23-05-05	96.89	95.16	95.55	96.53	94.70	96.79	96.44	95.72	**96.18**
2	**MC-CoT_F-Large** 🥈	VLM	Fine-tune	783M	-	Link	23-11-23	97.47	90.44	93.18	96.97	93.75	94.49	95.30	94.13	**94.88**
3	**Honeybee (Vicuna-13B)** 🥉	VLM	Fine-tune	13B	-	Link	23-12-11	95.20	96.29	91.18	94.48	93.75	93.17	95.04	93.21	**94.39**
4	**Enigma-COT_Large**	LLM	Fine-tune	793M	793M	Link	23-07-24	97.51	84.70	94.73	96.68	91.37	95.89	94.46	93.47	**94.11**
5	**KAM-CoT**	VLM	Fine-tune	280M	280M	Link	24-01-23	94.76	92.24	93.36	94.53	93.16	94.15	94.24	93.21	**93.87**
6	**MC-CoT_Large**	VLM	Fine-tune	738M	-	Link	23-11-23	95.47	89.99	91.82	95.11	92.66	93.24	94.27	91.76	**93.37**
7	**DPMM-CoT_Large**	VLM	Fine-tune	738M	738M	Link	23-12-14	95.52	90.33	91.36	95.50	93.26	92.68	93.28	93.47	**93.35**
8	**LLaVA (GPT-4 judge)**	VLM	Fine-tune	13B	13B	Link	23-04-17	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**
9	**CoMD (Vicuna-7B)**	VLM	Fine-tune	7B	-	Link	23-11-14	91.83	95.95	88.91	90.91	89.94	91.08	92.47	90.97	**91.94**
10	**Mutimodal-T-SciQ_Base**	LLM	Fine-tune	223M	223M	Link	23-05-05	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	**91.75**
11	**Multimodal-CoT_Large**	VLM	Fine-tune	738M	738M	Link	23-02-02	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	**91.68**

Lu, Pan, et al. "Learn to explain: Multimodal reasoning via thought chains for science question answering." NeurIPS 2022

OlympiadBench



Question: Find all triples (x, y, z) of positive integers such that $x \le y \le z$ and $x^3(y^3 + z^3) = 2012(xyz + 2)$.

Solution: First note that x divides $2012 \cdot 2 = 2^3 \cdot 503$. If $503 \mid x$ then the right-hand side of the equation is divisible by 503^3, and it follows that $503^2 \mid xyz + 2$. This is false as $503 \mid x$. Hence $x = 2^m$ with $m \in \{0,1,2,3\}$. If $m \ge 2$ then $2^6 \mid 2012(xyz + 2)$. However the highest powers of 2 dividing 2012 and $xyz + 2 = 2^m yz + 2$ are 2^2 and 2^1 respectively. So $x = 1$ or $x = 2$, yielding the two equations

$$y^3 + z^3 = 2012(yz + 2),$$
$$y^3 + z^3 = 503(yz + 1)$$

In both cases …… It follows that $y \equiv -z (mod\ 503)$ as claimed. Therefore $y + z = 503k$ with $k \ge 1$. In view of $y^3 + z^3 = (y + z)((y - z)^2 + yz)$ the two equations take the form

$$k(y - z)^2 + (k - 4)yz = 8 \quad (1)$$
$$k(y - z)^2 + (k - 1)yz = 1 \quad (2)$$

In (1) we have $(k - 4)yz \le 8$, which implies $k \le 4$ ……
Therefore (1) has no integer solutions.
Equation (2) implies $0 \le (k - 1)yz \le 1$, so that $k = 1$ or $k = 2$.
Also $0 \le k(y - z)^2 \le 1$, hence $k = 2$ only if $y = z$. However then $y = z = 1$, which is false in view of $y + z \ge 503$.
Therefore $k = 1$ and (2) takes the form $(y - z)^2 = 1$, yielding $z - y = |y - z| = 1$. Combined with $k = 1$ and $y + z = 503k$, this leads to $y = 251, z = 252$.
In summary the triple $(2,251,252)$ is the only solution.

Final answer: (2,251,252) **Subfield:** Number theory
Answer type: Tuple **Question type:** Open-ended

Physics-COMP&CEE (2,334):

Mechanics, Electromagnetism,
Thermodynamics,
Optics, Modern Physics

Maths-COMP (2,133):

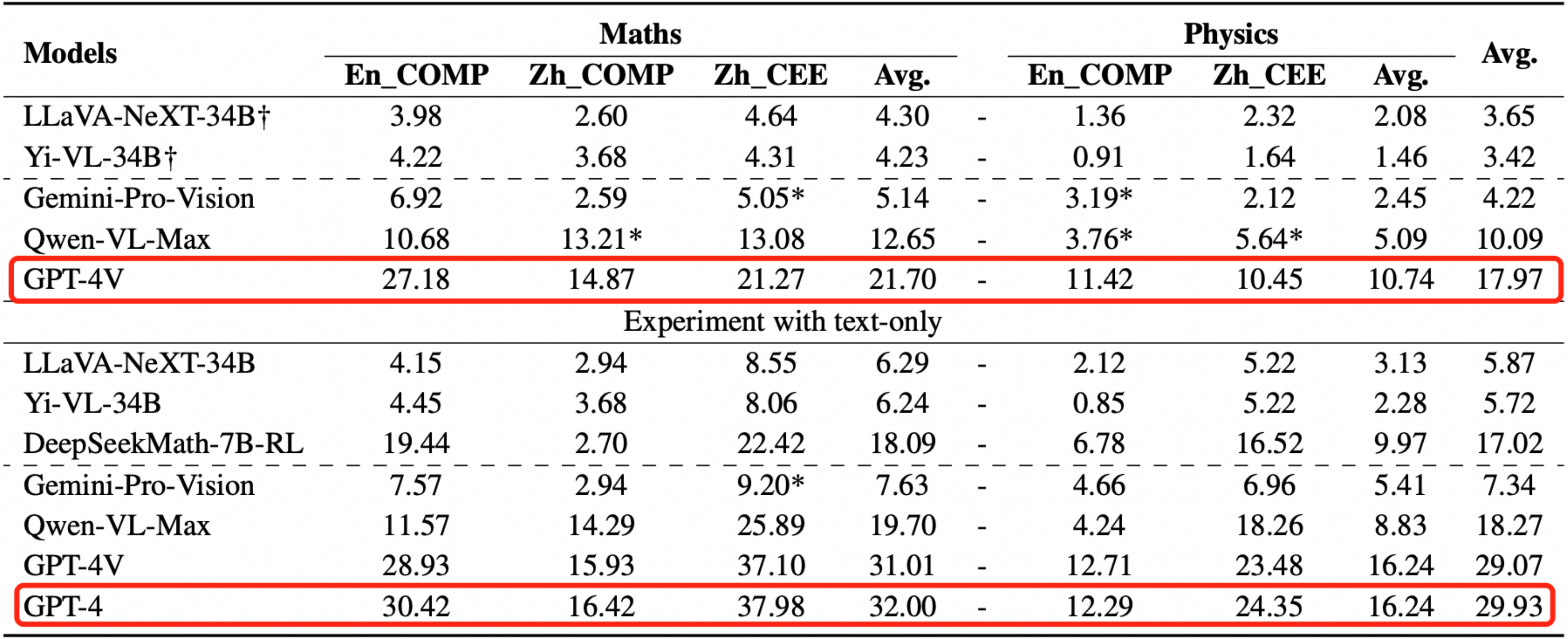
Combinatorics, Algebra,
Number Theory, Geometry

Maths-CEE (4,009):

Derivative, Conic Sections, Sequence, Trigonometric Functions, Set Theory, Logic, Elementary Functions, Inequality, Polar Coordinates and Parametric Equations, Probability and Statistics, Plane Geometry, Solid Geometry, Complex Numbers

He, Chaoqun, et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. ACL 2024.

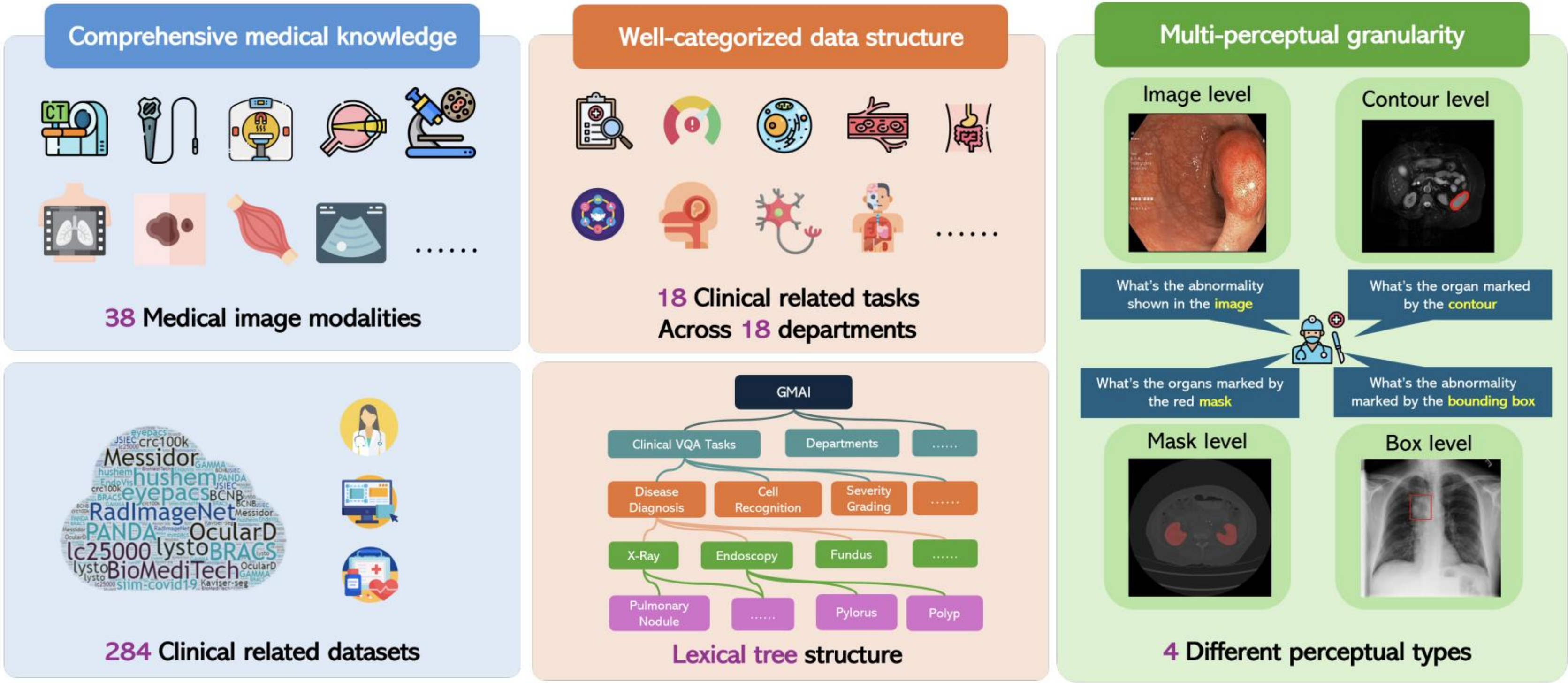
OlympiadBench



Models	Maths					Physics			Avg.
	En_COMP	Zh_COMP	Zh_CEE	Avg.		En_COMP	Zh_CEE	Avg.	
LLaVA-NeXT-34B†	3.98	2.60	4.64	4.30	-	1.36	2.32	2.08	3.65
Yi-VL-34B†	4.22	3.68	4.31	4.23	-	0.91	1.64	1.46	3.42
Gemini-Pro-Vision	6.92	2.59	5.05*	5.14	-	3.19*	2.12	2.45	4.22
Qwen-VL-Max	10.68	13.21*	13.08	12.65	-	3.76*	5.64*	5.09	10.09
GPT-4V	27.18	14.87	21.27	21.70	-	11.42	10.45	10.74	17.97
Experiment with text-only									
LLaVA-NeXT-34B	4.15	2.94	8.55	6.29	-	2.12	5.22	3.13	5.87
Yi-VL-34B	4.45	3.68	8.06	6.24	-	0.85	5.22	2.28	5.72
DeepSeekMath-7B-RL	19.44	2.70	22.42	18.09	-	6.78	16.52	9.97	17.02
Gemini-Pro-Vision	7.57	2.94	9.20*	7.63	-	4.66	6.96	5.41	7.34
Qwen-VL-Max	11.57	14.29	25.89	19.70	-	4.24	18.26	8.83	18.27
GPT-4V	28.93	15.93	37.10	31.01	-	12.71	23.48	16.24	29.07
GPT-4	30.42	16.42	37.98	32.00	-	12.29	24.35	16.24	29.93

He, Chaoqun, et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. ACL 2024.

GMAI-MMBench


Comprehensive medical knowledge
CT
38 Medical image modalities
crc100k
Messidor
hushem
eyepacs
RadImageNet
PANDA
OcularD
lc25000
lysto
BRACS
BioMediTech
BCNB
JSIEC
GAMMA
siim-covid19
Kaviser-seg
284 Clinical related datasets
Well-categorized data structure
18 Clinical related tasks
Across 18 departments
GMAI
Clinical VQA Tasks
Departments
……
Disease Diagnosis
Cell Recognition
Severity Grading
……
X-Ray
Endoscopy
Fundus
……
Pulmonary Nodule
……
Pylorus
Polyp
Lexical tree structure
Multi-perceptual granularity
Image level
Contour level
What's the abnormality shown in the image
What's the organ marked by the contour
What's the organs marked by the red mask
What's the abnormality marked by the bounding box
Mask level
Box level
4 Different perceptual types

Ye, Jin, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. NeurIPS 2024

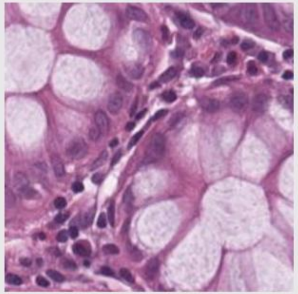
GMAI-MMBench

Image level

Question: Determine which option best matches the content displayed in the histology image.

Options:
A. debris
B. lymphocyte
C. normal colonic mucosa
D. smooth muscle

Please select the correct answer from the options above

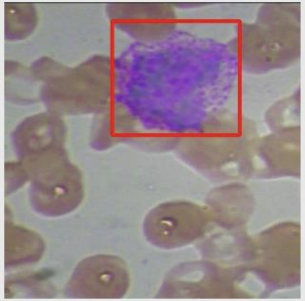


Box level

Question: Observe the microscopy image. Can you identify the target within the outlined box?

Options:
A. red blood cell
B. white blood cell
C. platelet
D. mycobacterium tuberculosis

Please select the correct answer from the options above

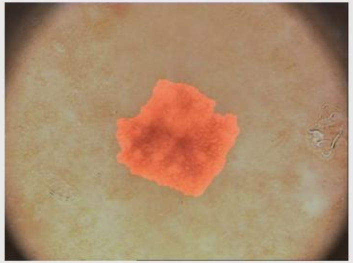


Mask level

Question: Observe the Dermoscopy image. What is the most likely abnormality shown in the highlight area?

Options:
A. pleural effusion
B. esophageal cancer
C. globules skin lesion
D. lung consolidation
E. melanocytic lesions

Please select the correct answer from the options above

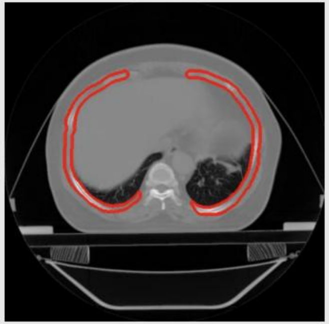


Contour level

Question: Observe the CT image. Can you identify the organ in the highlight area?

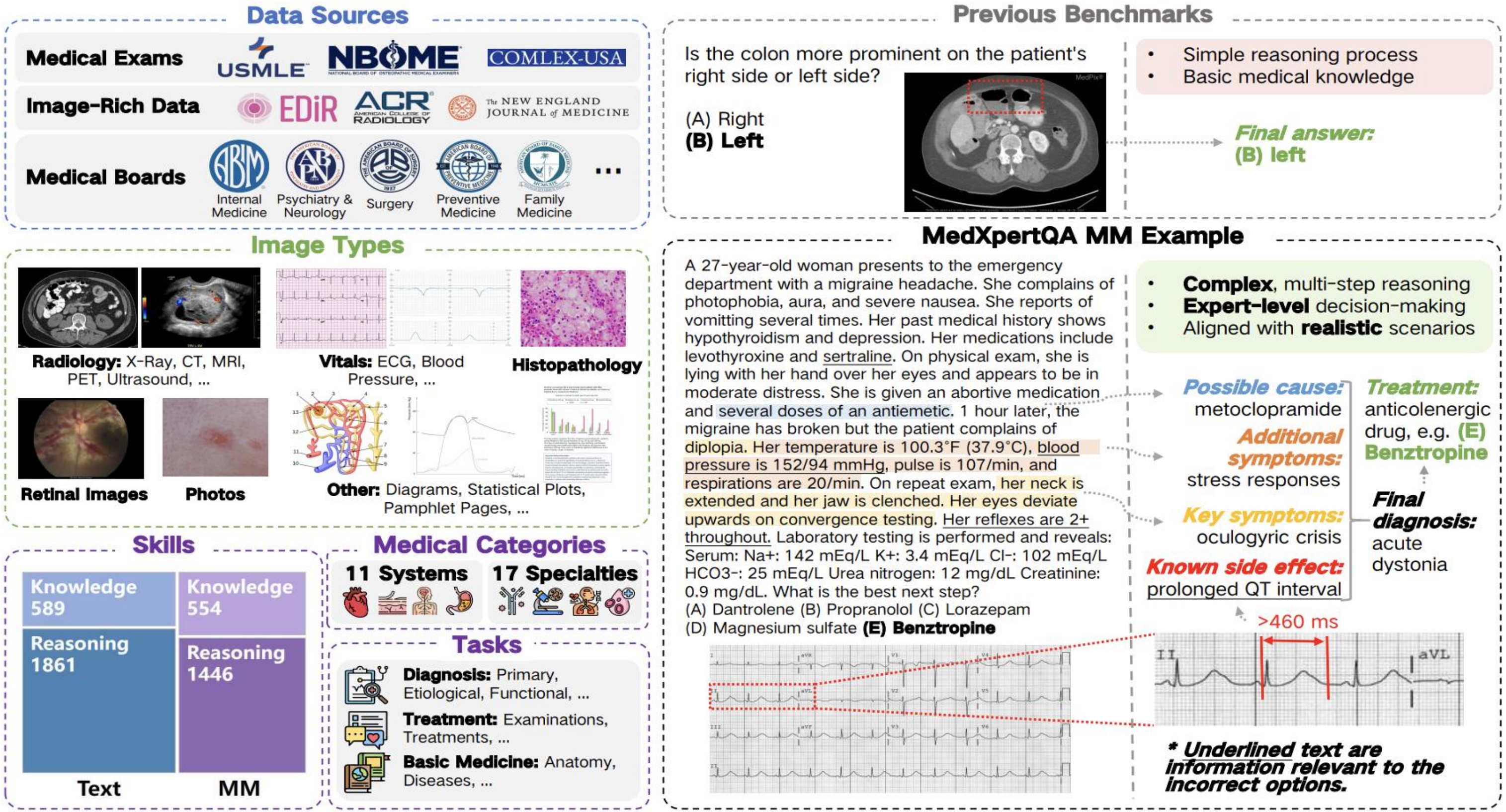
Options:
A. spinal cord
B. pulmonary artery
C. chest wall
D. Esophagus

Please select the correct answer from the options above



Ye, Jin, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. NeurIPS 2024

MedXpertQA



Data Sources

- **Medical Exams**: USMLE, NBOME, COMLEX-USA
- **Image-Rich Data**: EDiR, ACR American College of Radiology, The New England Journal of Medicine
- **Medical Boards**: Internal Medicine, Psychiatry & Neurology, Surgery, Preventive Medicine, Family Medicine, ...

Image Types

- **Radiology:** X-Ray, CT, MRI, PET, Ultrasound, ...
- **Vitals:** ECG, Blood Pressure, ...
- **Histopathology**
- **Retinal Images**
- **Photos**
- **Other:** Diagrams, Statistical Plots, Pamphlet Pages, ...

Skills

	Text	MM
Knowledge	589	554
Reasoning	1861	1446

Medical Categories

11 Systems **17 Specialties**

Tasks

- **Diagnosis:** Primary, Etiological, Functional, ...
- **Treatment:** Examinations, Treatments, ...
- **Basic Medicine:** Anatomy, Diseases, ...

Previous Benchmarks

Is the colon more prominent on the patient's right side or left side?

(A) Right
(B) Left

- Simple reasoning process
- Basic medical knowledge

Final answer: (B) left

MedXpertQA MM Example

A 27-year-old woman presents to the emergency department with a migraine headache. She complains of photophobia, aura, and severe nausea. She reports of vomitting several times. Her past medical history shows hypothyroidism and depression. Her medications include levothyroxine and sertraline. On physical exam, she is lying with her hand over her eyes and appears to be in moderate distress. She is given an abortive medication and several doses of an antiemetic. 1 hour later, the migraine has broken but the patient complains of diplopia. Her temperature is 100.3˚F (37.9˚C), blood pressure is 152/94 mmHg, pulse is 107/min, and respirations are 20/min. On repeat exam, her neck is extended and her jaw is clenched. Her eyes deviate upwards on convergence testing. Her reflexes are 2+ throughout. Laboratory testing is performed and reveals: Serum: Na+: 142 mEq/L K+: 3.4 mEq/L Cl-: 102 mEq/L HCO3-: 25 mEq/L Urea nitrogen: 12 mg/dL Creatinine: 0.9 mg/dL. What is the best next step?

(A) Dantrolene (B) Propranolol (C) Lorazepam
(D) Magnesium sulfate **(E) Benztropine**

- **Complex**, multi-step reasoning
- **Expert-level** decision-making
- Aligned with **realistic** scenarios

Possible cause: metoclopramide

Additional symptoms: stress responses

Key symptoms: oculogyric crisis

Known side effect: prolonged QT interval

Treatment: anticolenergic drug, e.g. (E) Benztropine

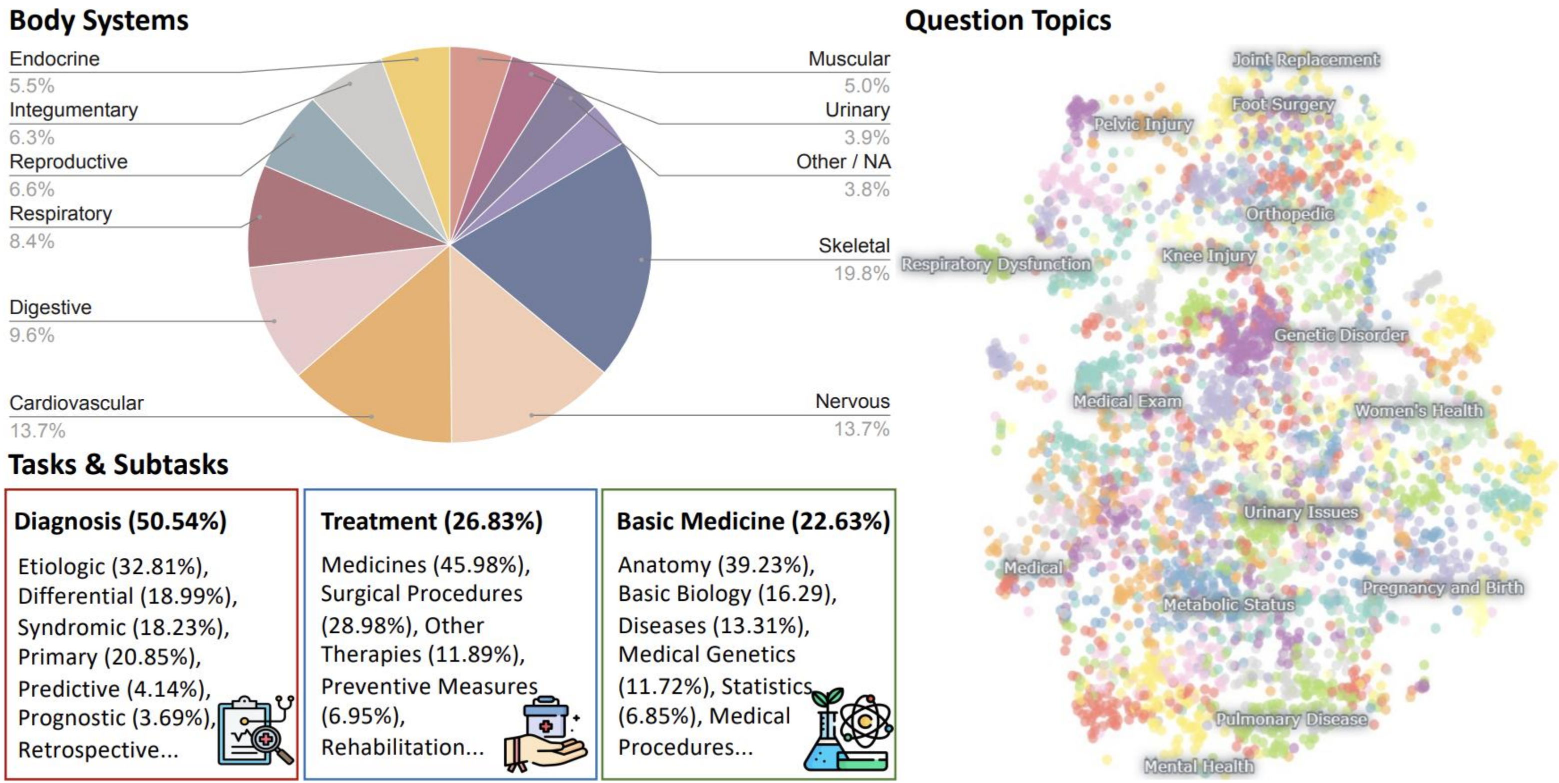
Final diagnosis: acute dystonia

>460 ms

** Underlined text are information relevant to the incorrect options.*

Zuo, Yuxin, et al. "MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding." *arXiv preprint arXiv:2501.18362* (2025).

MedXpertQA



Tasks & Subtasks

Diagnosis (50.54%)

Etiologic (32.81%), Differential (18.99%), Syndromic (18.23%), Primary (20.85%), Predictive (4.14%), Prognostic (3.69%), Retrospective...

Treatment (26.83%)

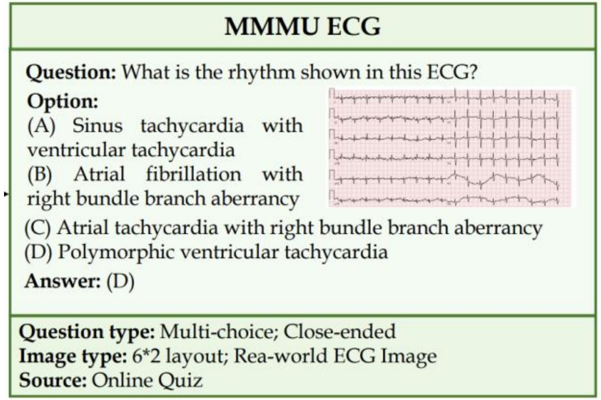
Medicines (45.98%), Surgical Procedures (28.98%), Other Therapies (11.89%), Preventive Measures (6.95%), Rehabilitation...

Basic Medicine (22.63%)

Anatomy (39.23%), Basic Biology (16.29), Diseases (13.31%), Medical Genetics (11.72%), Statistics (6.85%), Medical Procedures...

Zuo, Yuxin, et al. "MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding." *arXiv preprint arXiv:2501.18362* (2025).

Other specialized medical tasks



MMMU ECG

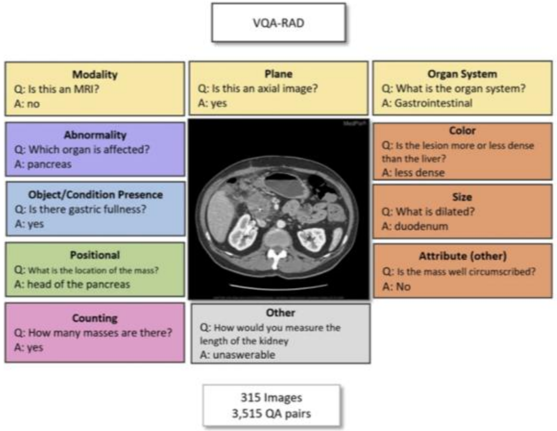
Question: What is the rhythm shown in this ECG?

Option:
(A) Sinus tachycardia with ventricular tachycardia
(B) Atrial fibrillation with right bundle branch aberrancy
(C) Atrial tachycardia with right bundle branch aberrancy
(D) Polymorphic ventricular tachycardia

Answer: (D)

Question type: Multi-choice; Close-ended
Image type: 6*2 layout; Rea-world ECG Image
Source: Online Quiz

Liu, Ruoqi, et al. Teach Multimodal LLMs to Comprehend Electrocardiographic Images. *arXiv* 2024



VQA-RAD

Modality
Q: Is this an MRI?
A: no

Plane
Q: Is this an axial image?
A: yes

Organ System
Q: What is the organ system?
A: Gastrointestinal

Abnormality
Q: Which organ is affected?
A: pancreas

Color
Q: Is the lesion more or less dense than the liver?
A: less dense

Object/Condition Presence
Q: Is there gastric fullness?
A: yes

Size
Q: What is dilated?
A: duodenum

Positional
Q: What is the location of the mass?
A: head of the pancreas

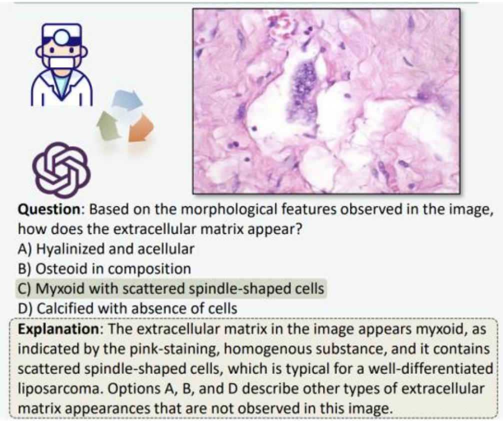
Attribute (other)
Q: Is the mass well circumscribed?
A: No

Counting
Q: How many masses are there?
A: yes

Other
Q: How would you measure the length of the kidney
A: unaswerable

315 Images
3,515 QA pairs

Lau, Jason J., et al. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 2018

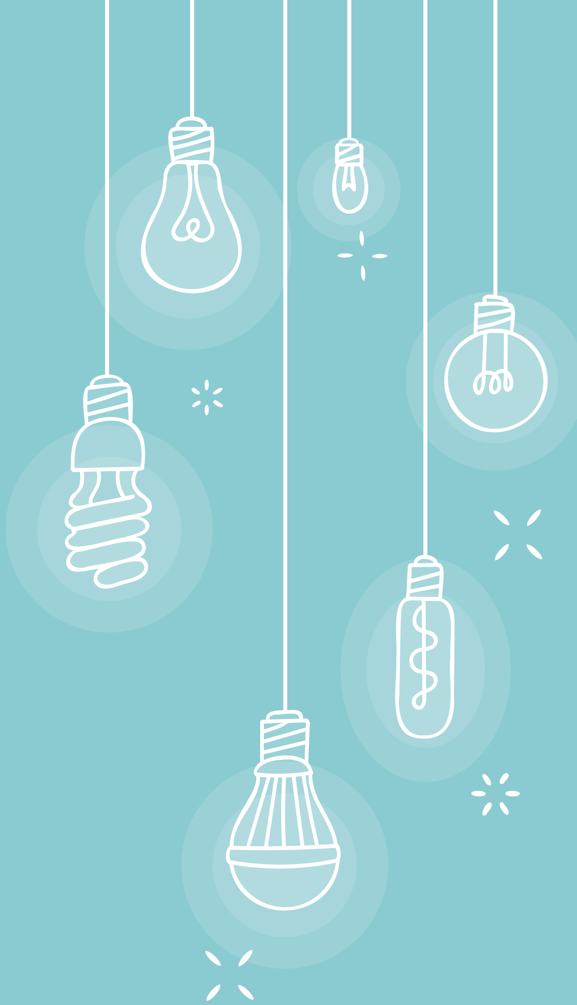


Question: Based on the morphological features observed in the image, how does the extracellular matrix appear?
A) Hyalinized and acellular
B) Osteoid in composition
C) Myxoid with scattered spindle-shaped cells
D) Calcified with absence of cells

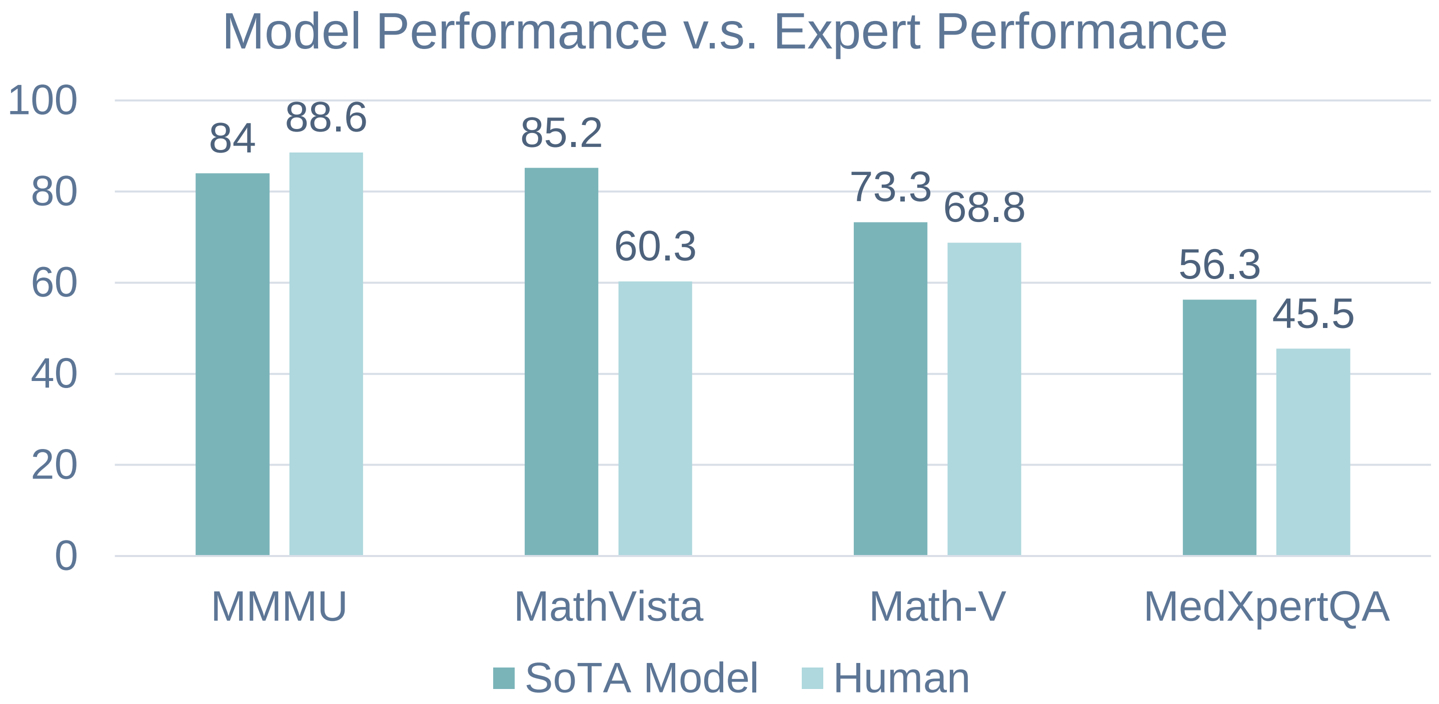
Explanation: The extracellular matrix in the image appears myxoid, as indicated by the pink-staining, homogenous substance, and it contains scattered spindle-shaped cells, which is typical for a well-differentiated liposarcoma. Options A, B, and D describe other types of extracellular matrix appearances that are not observed in this image.

Sun, Yuxuan, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. *ECCV*, 2024.

Discussions and Future Directions



Performance on Expert-level Tasks



What’s next?

All existing benchmarks, even expert-level tasks, are either saturated or approaching saturation

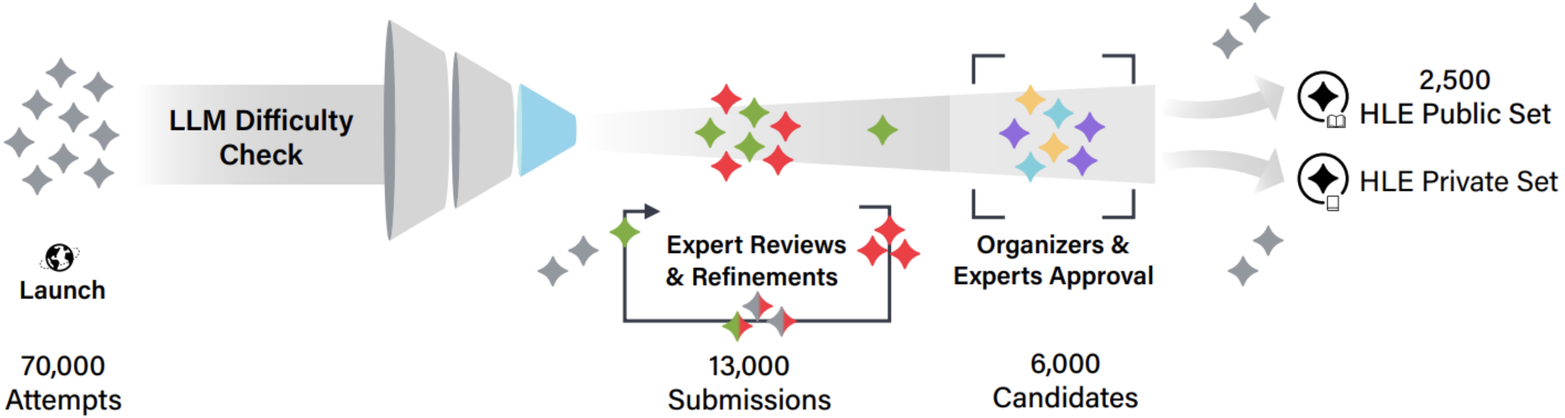
What’s next?

All existing benchmarks, even expert-level tasks, are either saturated or approaching saturation

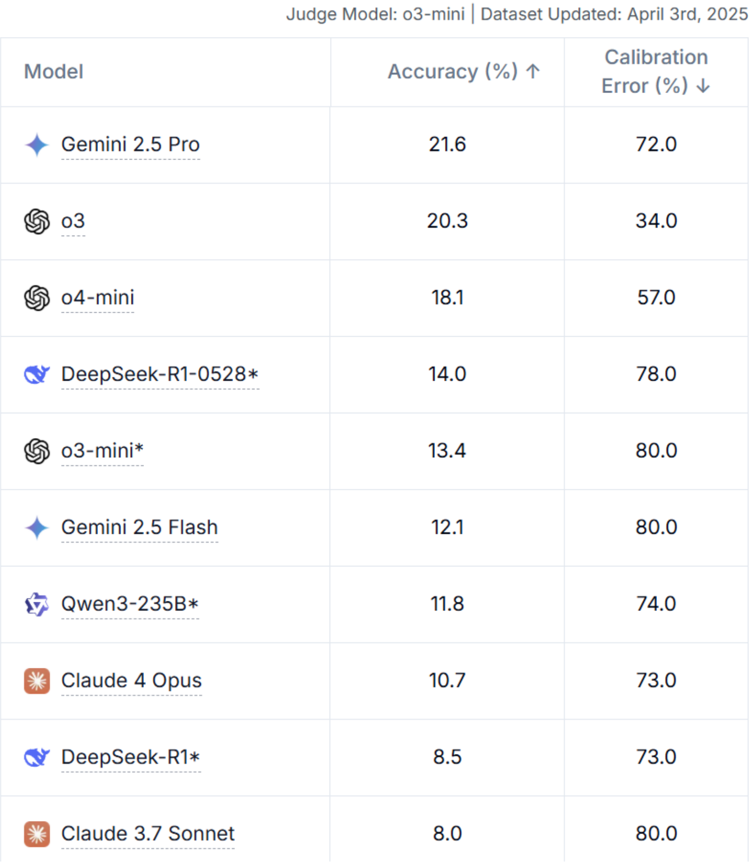
- Challenging
- Real-world
- Dynamic

More Challenging Benchmark

Humanity’s Last Exam (HLE)



More Challenging Benchmark

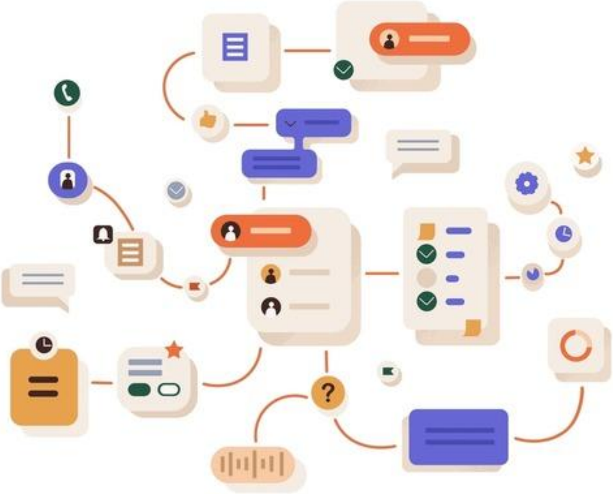


Judge Model: o3-mini | Dataset Updated: April 3rd, 2025

Model	Accuracy (%) ↑	Calibration Error (%) ↓
Gemini 2.5 Pro	21.6	72.0
o3	20.3	34.0
o4-mini	18.1	57.0
DeepSeek-R1-0528*	14.0	78.0
o3-mini*	13.4	80.0
Gemini 2.5 Flash	12.1	80.0
Qwen3-235B*	11.8	74.0
Claude 4 Opus	10.7	73.0
DeepSeek-R1*	8.5	73.0
Claude 3.7 Sonnet	8.0	80.0

Gaps between exam questions and real expert working scenarios

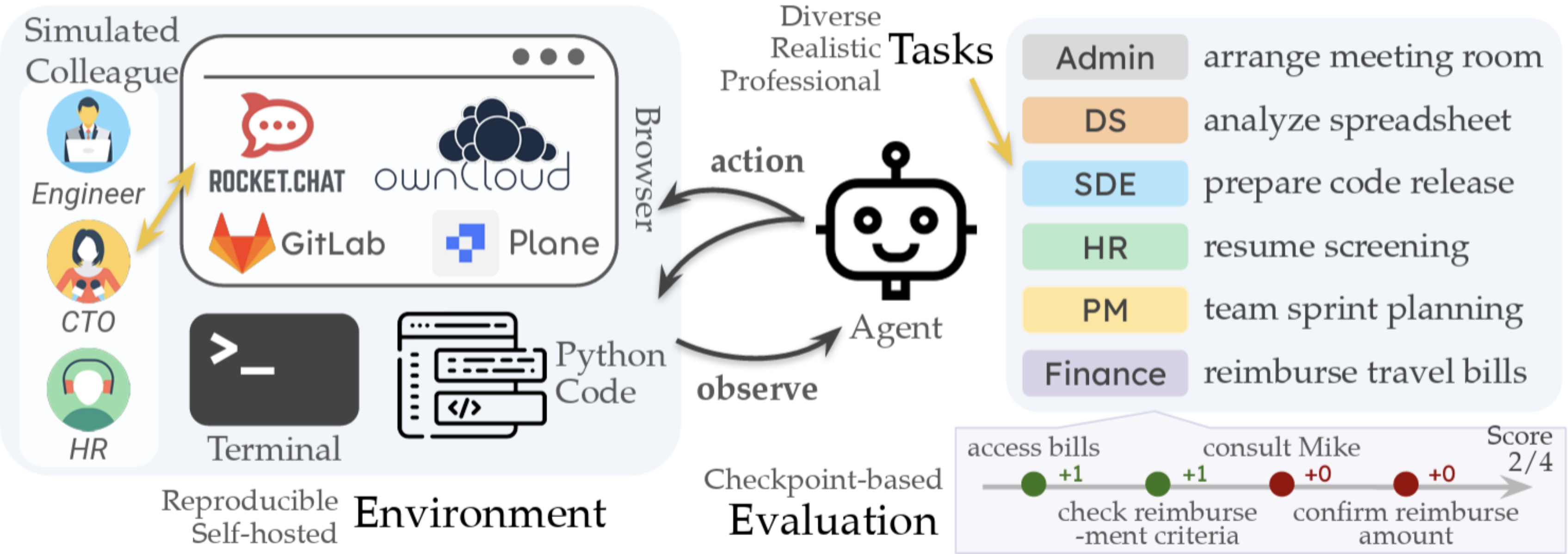
Exam Style Questions



Real Expert Workflows

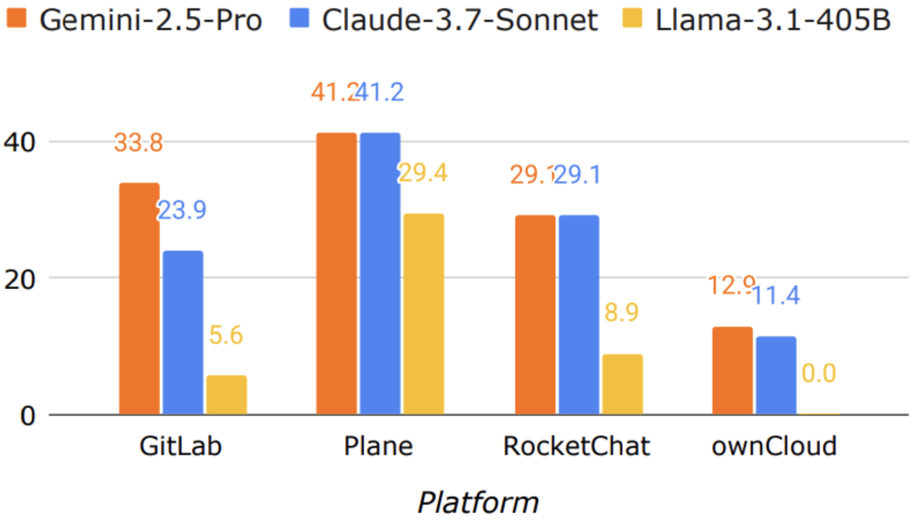
More Realistic Tasks

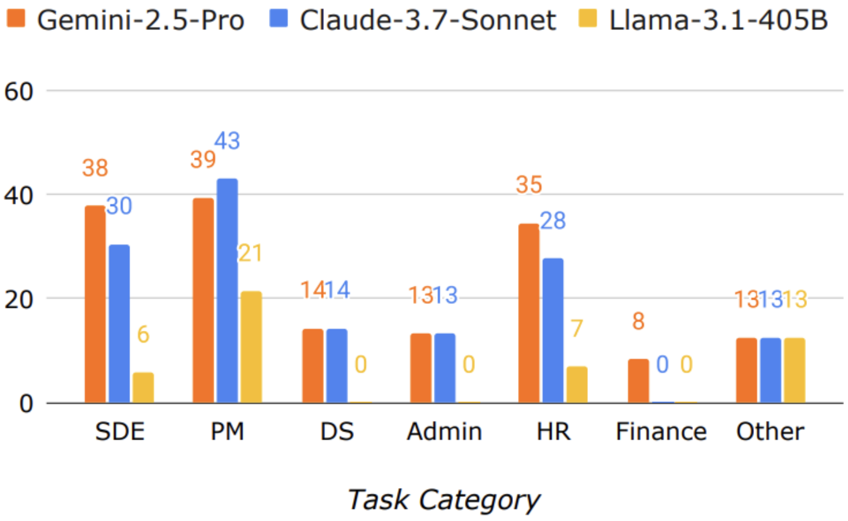
The Agent Company



Xu, Frank F., et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv* 2024

Performance on The Agent Company





The most competitive agent can complete 30% of tasks autonomously

Xu, Frank F., et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv* 2024

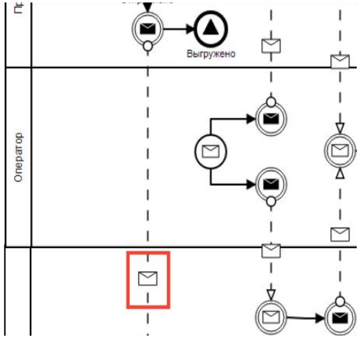
SWE-bench Multimodal

Diagramming

Show message element name

Currently, names of message elements on message flows are not rendered

Given this example diagram [Image] ...



bpmn-js

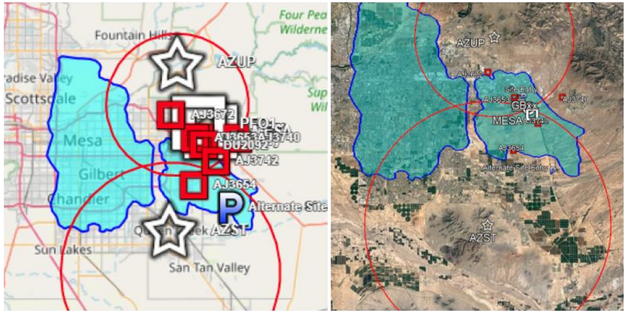
Interactive Mapping

KML Symbol Align/Placement/Size

There is a bug with the anchor point for some symbols

I've attached a screen clipping from Google Earth to show how it is supposed to look.

[Right Image] ...



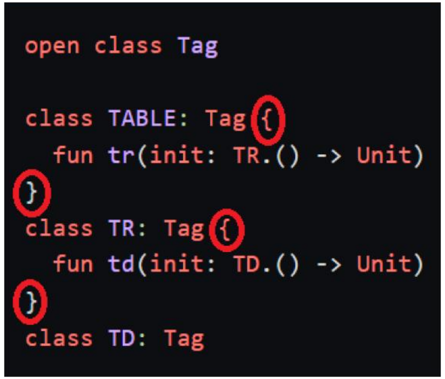
openlayers

Syntax Highlighting

Bracket highlighted with different color in class inheritance context.

– Reproduced in JSFiddle: https://jsfiddle.net/kkangmj/e7h48w36/7/

[Image] ...



```
open class Tag

class TABLE: Tag{
  fun tr(init: TR.() -> Unit)
}
class TR: Tag{
  fun td(init: TD.() -> Unit)
}
class TD: Tag
```

highlight.js

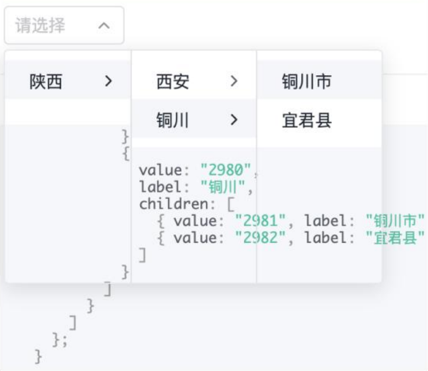
Web Frameworks

[CascaderSelect]使用虚拟滚动时背景色异常

Component: CascaderSelect

Steps to reproduce

[Image] ...



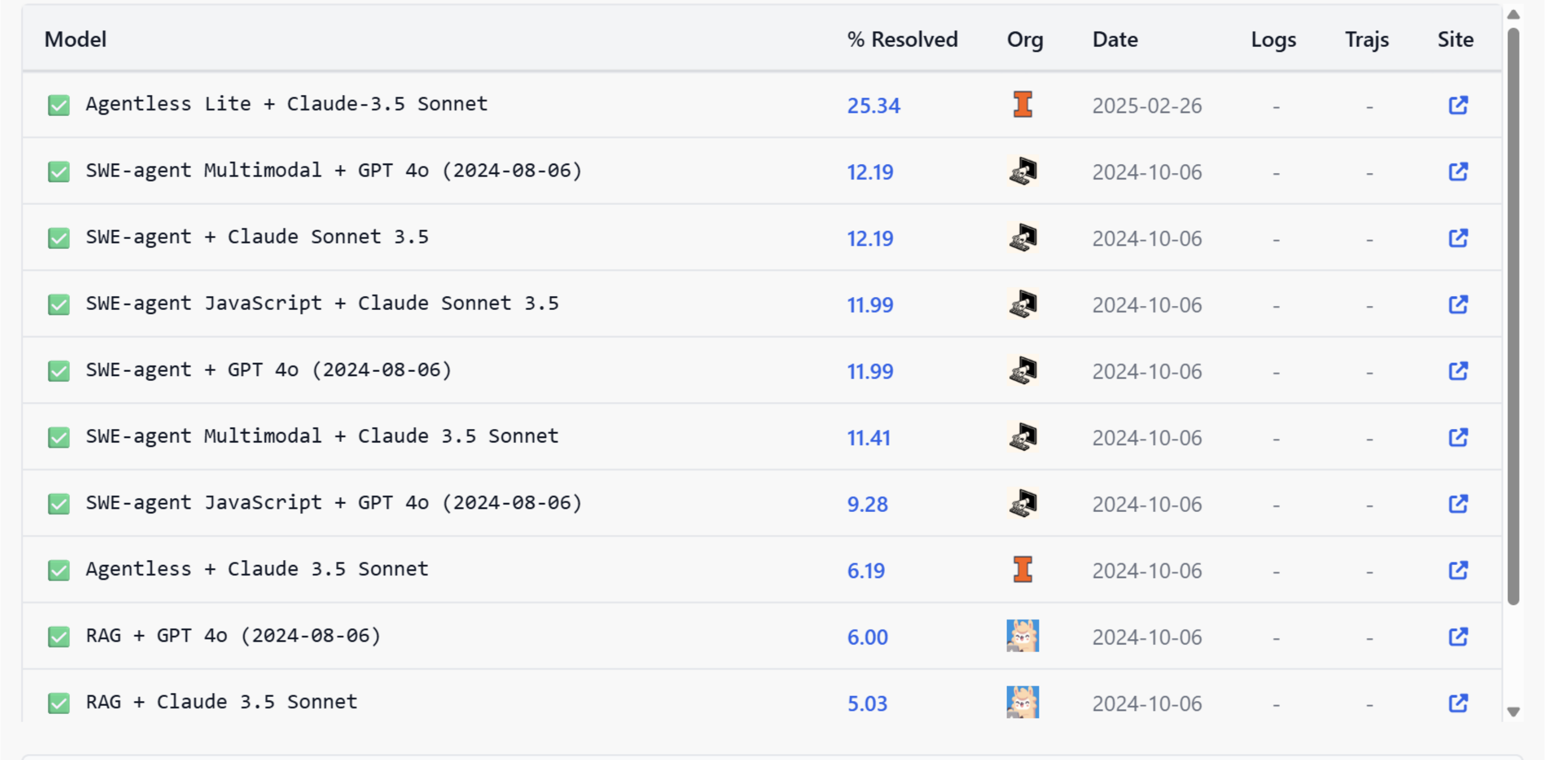
```
            }
            {
              value: "2980",
              label: "铜川",
              children: [
                { value: "2981", label: "铜川市"
                { value: "2982", label: "宜君县"
              ]
            }
          ]
        }
      ]
    };
}
```

next

Evaluate systems on their ability to fix bugs in visual, user-facing JavaScript software

Yang, John, et al. "SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains?." *ICLR 2025*

SWE-bench Multimodal Leaderboard



Model	% Resolved	Org	Date	Logs	Trajs	Site
☑ Agentless Lite + Claude-3.5 Sonnet	25.34		2025-02-26	-	-	↗
☑ SWE-agent Multimodal + GPT 4o (2024-08-06)	12.19		2024-10-06	-	-	↗
☑ SWE-agent + Claude Sonnet 3.5	12.19		2024-10-06	-	-	↗
☑ SWE-agent JavaScript + Claude Sonnet 3.5	11.99		2024-10-06	-	-	↗
☑ SWE-agent + GPT 4o (2024-08-06)	11.99		2024-10-06	-	-	↗
☑ SWE-agent Multimodal + Claude 3.5 Sonnet	11.41		2024-10-06	-	-	↗
☑ SWE-agent JavaScript + GPT 4o (2024-08-06)	9.28		2024-10-06	-	-	↗
☑ Agentless + Claude 3.5 Sonnet	6.19		2024-10-06	-	-	↗
☑ RAG + GPT 4o (2024-08-06)	6.00		2024-10-06	-	-	↗
☑ RAG + Claude 3.5 Sonnet	5.03		2024-10-06	-	-	↗

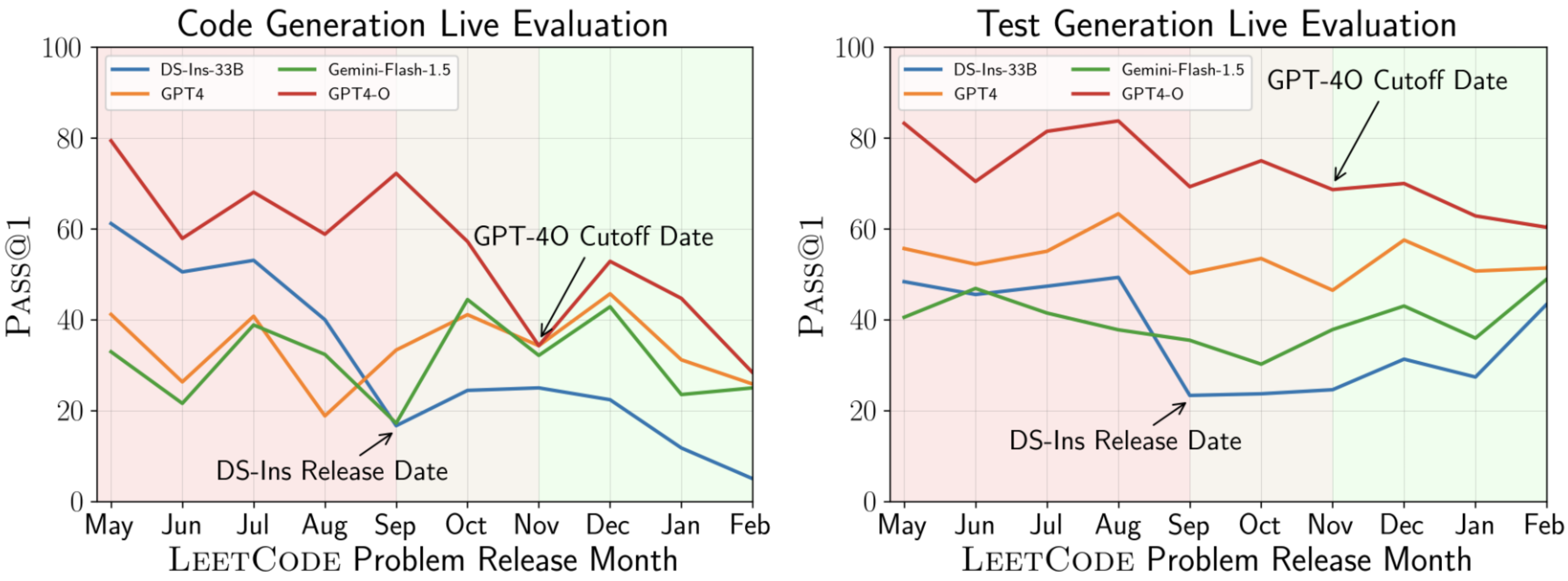
More Dynamic Evaluations

Increasing concerns on overfitting and contamination issues of benchmarks

↓

Can we construct dynamic evaluations?

LiveCodeBench

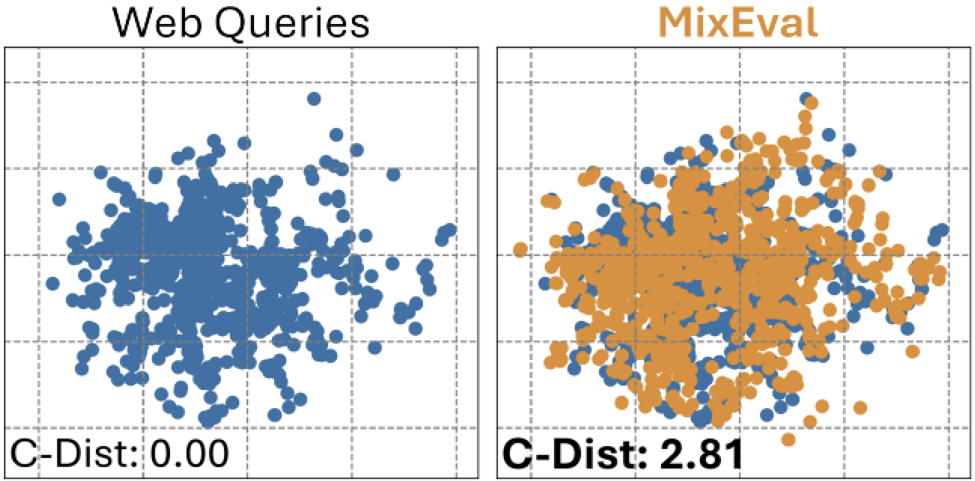


Jain, Naman, et al. "LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code." *ICLR 2025*

MixEval / MixEval-X

Benchmark 1
Benchmark 2
Benchmark 3
Benchmark 4
Benchmark 5
…

Sampling



Ni, Jinjie, et al. "MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures." NeurIPS *2024*
Ni, Jinjie, et al. "MixEval-X: Any-to-Any Evaluations from Real-World Data Mixtures." ICLR 2025

Conclusion

- Expert-level evaluations are essential for assessing AI capabilities in real-world applications.
- Current benchmarks provide valuable insights into AI performance across various expert domains.
- Future evaluations should aim to be: More challenging, realistic, and dynamic.

Thanks!

Any questions?

Xiang Yue
Postdoc Researcher
Carnegie Mellon University
https://xiangyue9607.github.io/
Twitter/X: @xiangyue96
Email: xyue2@andrew.cmu.edu