# MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

**Xiang Yue**[*†,1], Yuansheng Ni[*2], Kai Zhang[*3], Tianyu Zheng[*4],
Ruoqi Liu[3], Ge Zhang[2], Samuel Stevens[3], Dongfu Jiang[2], Weiming Ren[2], Yuxuan Sun[4], Cong Wei[2], Botao Yu[3],
Ruibin Yuan[5], Renliang Sun[2], Ming Yin[7], Boyuan Zheng[3], Zhenzhu Yang[4], Yibo Liu[6], Wenhao Huang[4],
Huan Sun[*3], Yu Su[*†,3], Wenhu Chen[*†,2]

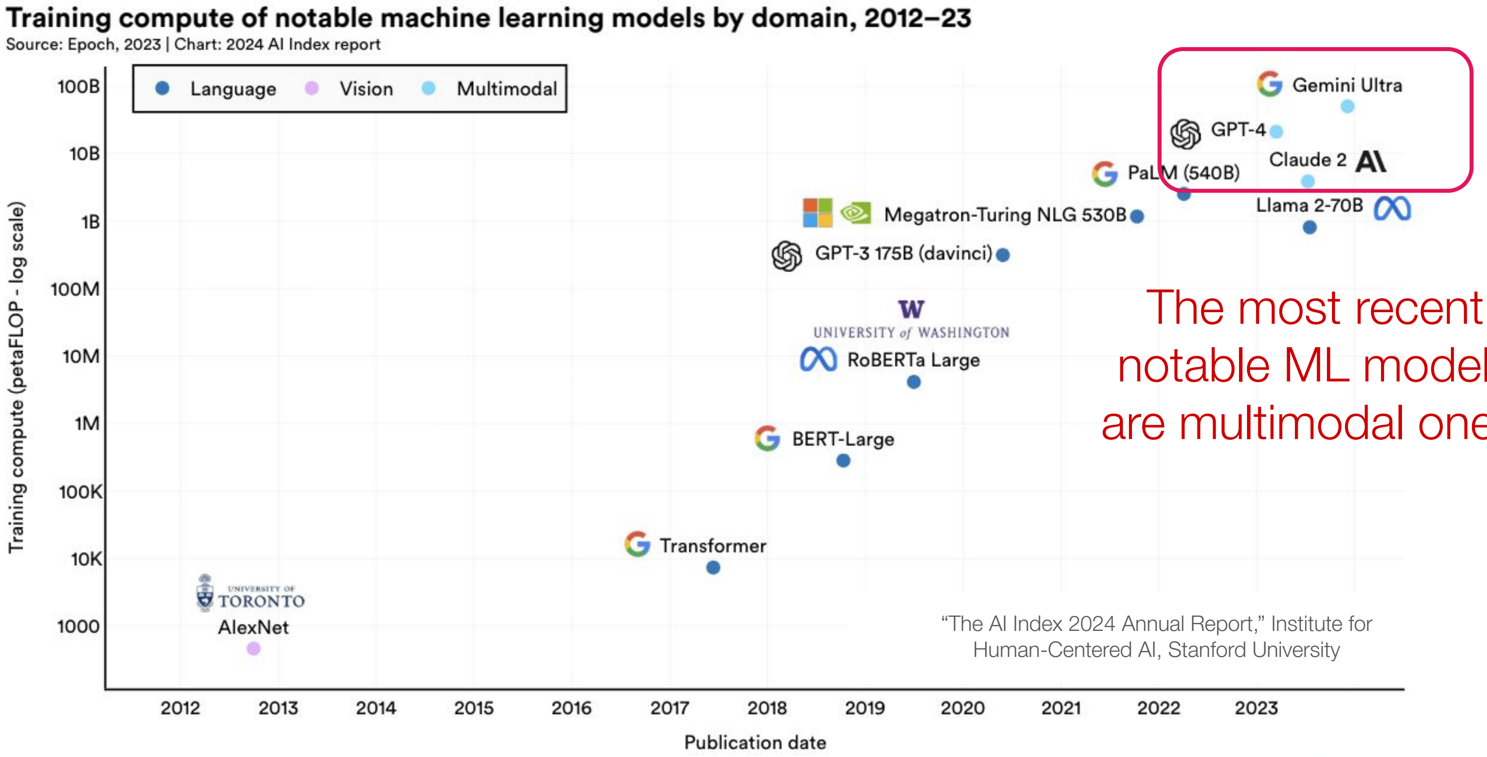
[1]IN.AI Research, [2]University of Waterloo, [3]The Ohio State University, [4]Independent,
[5]Carnegie Mellon University, [6]University of Victoria, [7]Princeton University

[*]Core Contributors
[†]Corresponding Authors

Slides of
this talk

CVPR 2024 Oral Presentation
🏆 - Award Candidate Paper

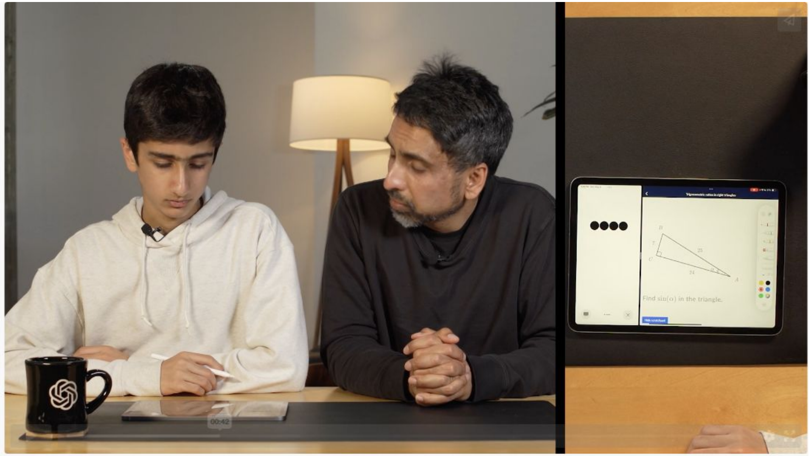
# Progress of Notable ML Models

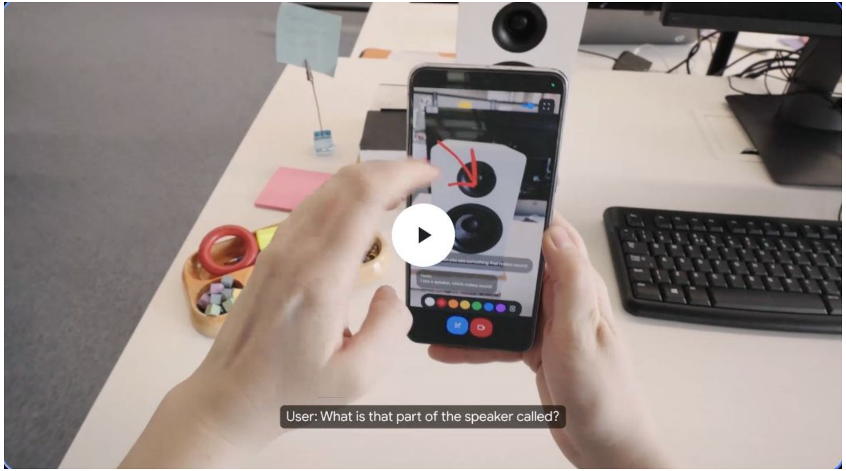


The most recent notable ML models are multimodal ones!

"The AI Index 2024 Annual Report," Institute for Human-Centered AI, Stanford University

# Multimodal Models Empower Real-world Apps

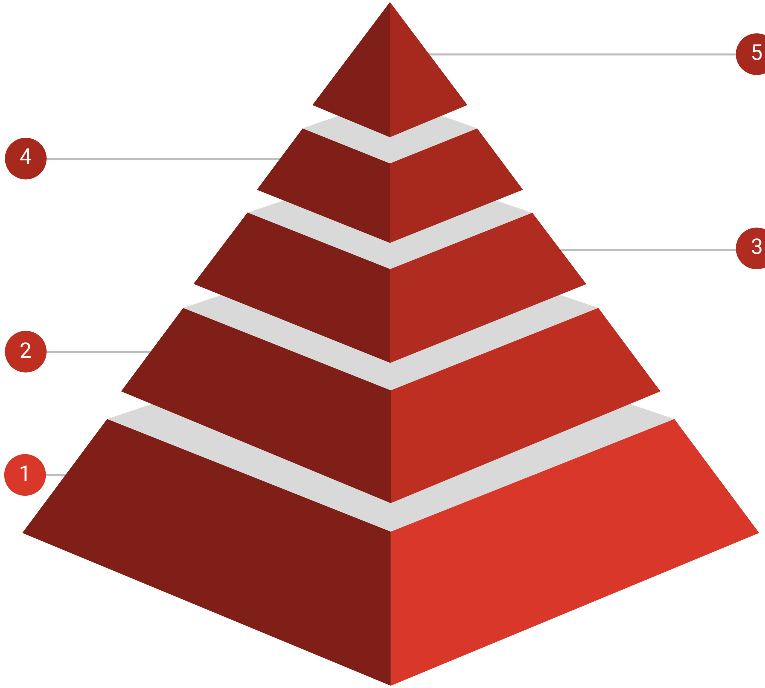

Math with Sal and Imran Khan.


User: What is that part of the speaker called?

https://openai.com/index/hello-gpt-4o/

https://deepmind.google/technologies/gemini/

# Levels of AGI



**Level 5: Superhuman**

⑤ outperforms 100% of humans

**Level 4: Virtuoso**

④ at least 99th percentile of skilled adults

**Level 3: Expert**

③ **at least 90th percentile of skilled adults**

**Level 2: Competent**

② at least 50th percentile of skilled adults

**Level 1: Emerging**

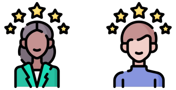
① equal to or somewhat better than an unskilled human

“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.

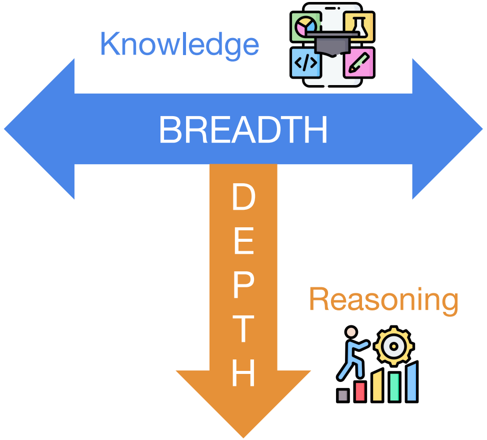
Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." *ICML 2024*

# Measuring Expert AGI?



Level 3: Expert AGI
(>=90% skilled adults)

?



Knowledge

BREADTH

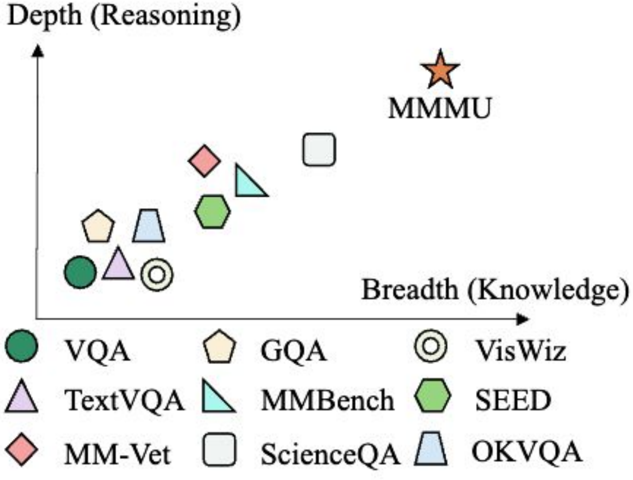
DEPTH

Reasoning

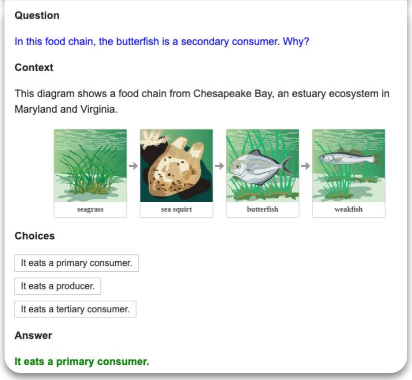
# Existing MM Benchmarks

## ● VQA
(Antol et al., 2015;
Goyal et al., 2017)

How many slices of pizza are there?
Is this a vegetarian pizza?



## □ ScienceQA
(Lu et al., 2022)



**Question**

In this food chain, the butterfish is a secondary consumer. Why?

**Context**

This diagram shows a food chain from Chesapeake Bay, an estuary ecosystem in Maryland and Virginia.

**Choices**

- It eats a primary consumer.
- It eats a producer.
- It eats a tertiary consumer.

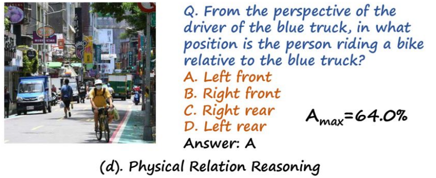
**Answer**

It eats a primary consumer.

## △ TextVQA
(Singh et al., 2019)

What is the top oz?

## ◺ MMBench
(Liu et al., 2023)



Q. From the perspective of the driver of the blue truck, in what position is the person riding a bike relative to the blue truck?
A. Left front
B. Right front
C. Right rear
D. Left rear
Answer: A

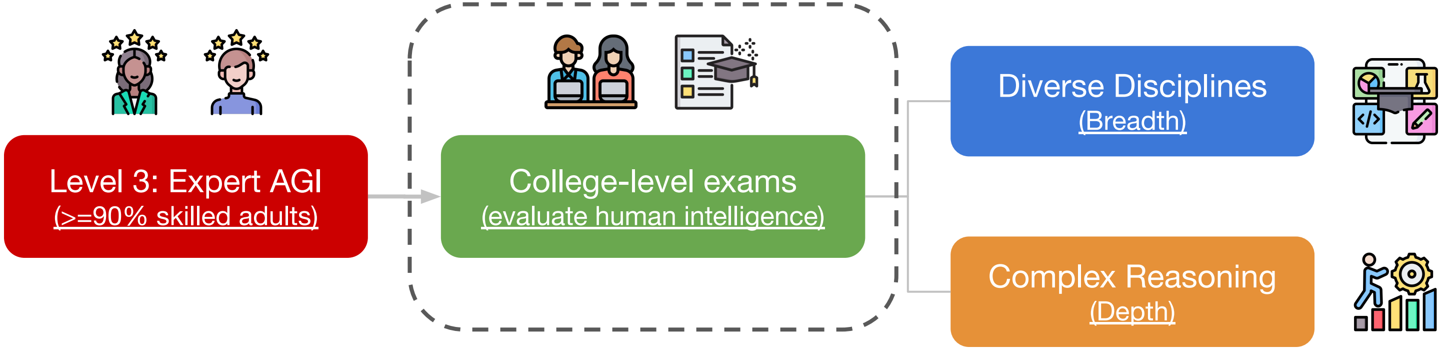
$A_{max}$=64.0%

(d). Physical Relation Reasoning

## ◇ MM-Vet
(Yu et al., 2023)

**Q**: What will the girl on the right write on the board?
**GT**: 14

# Measuring Expert AGI



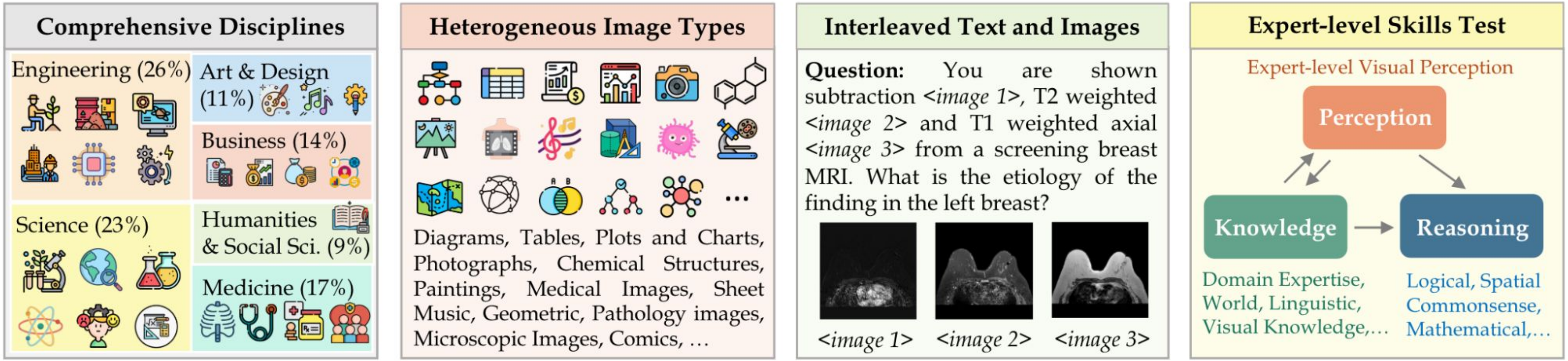
Level 3: Expert AGI
(>=90% skilled adults)

College-level exams
(evaluate human intelligence)

Diverse Disciplines
(Breadth)

Complex Reasoning
(Depth)

# MMMU: 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
<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,…

**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?

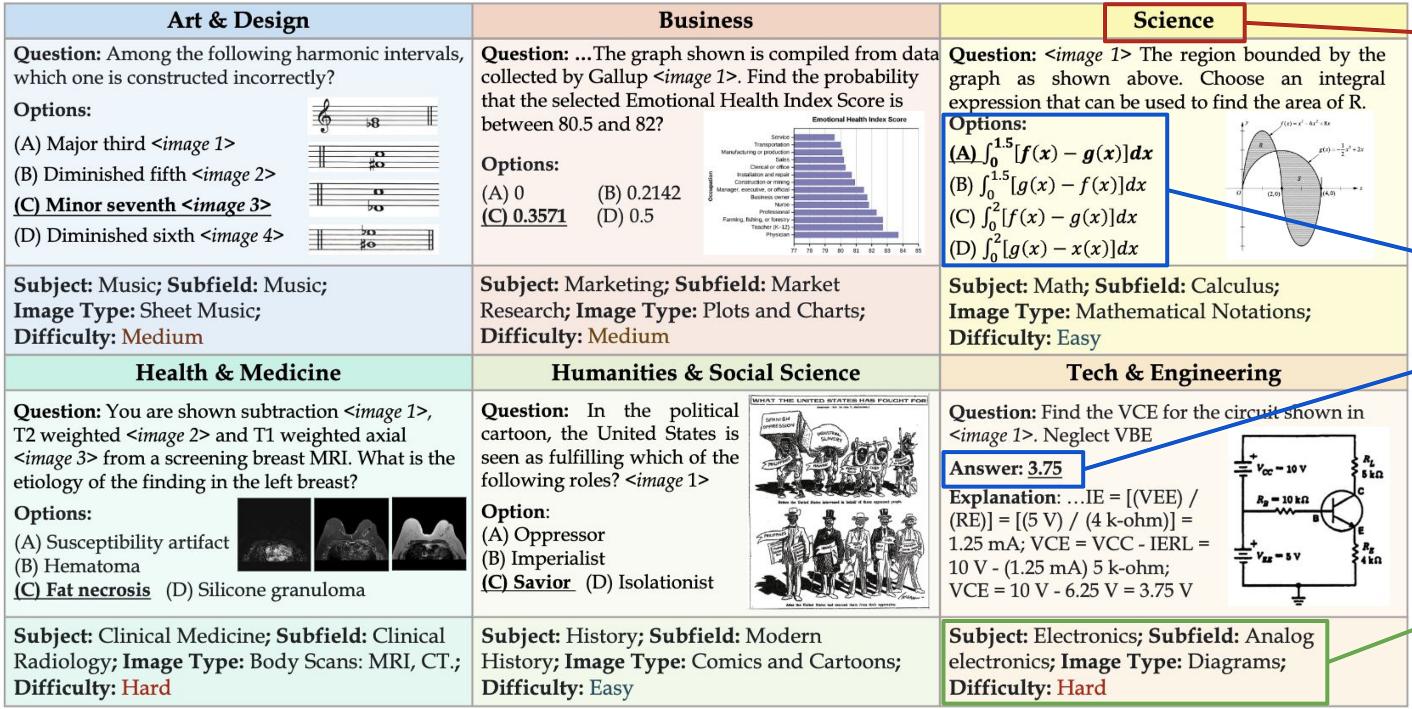
## (Breadth)

- **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**

# 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

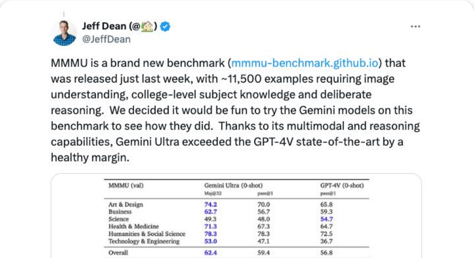
Discipline

Multiple-choice

Open

Subject
Subfield
Image Type
Difficulty

# Go-to-evaluation by Industrial Developers

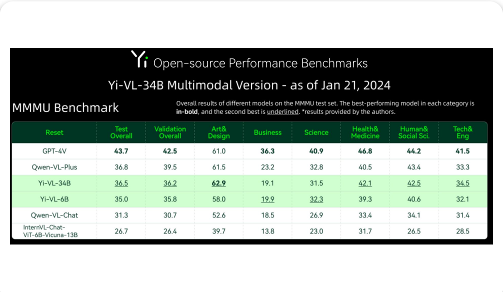


**Jeff Dean (@🏡)**
@JeffDean

MMMU is a brand new benchmark (mmmu-benchmark.github.io) that was released just last week, with ~11,500 examples requiring image understanding, college-level subject knowledge and deliberate reasoning. We decided it would be fun to try the Gemini models on this benchmark to see how they did. Thanks to its multimodal and reasoning capabilities, Gemini Ultra exceeded the GPT-4V state-of-the-art by a healthy margin.

| MMMU (val) | Gemini Ultra (0-shot) | | GPT-4V (0-shot) |
|---|---|---|---|
| | Maj@32 | pass@1 | pass@1 |
| Art & Design | **74.2** | 70.0 | 65.8 |
| Business | **62.7** | 56.7 | 59.3 |
| Science | 49.3 | 48.0 | **54.7** |
| Health & Medicine | **71.3** | 67.3 | 64.7 |
| Humanities & Social Science | **78.3** | 78.3 | 72.5 |
| Technology & Engineering | **53.0** | 47.1 | 36.7 |
| Overall | **62.4** | 59.4 | 56.8 |

Table 8 | **Gemini Ultra performance on the MMMU** benchmark (Yue et al., 2023) per discipline. Each discipline covers multiple subjects, requiring college-level knowledge and complex reasoning.



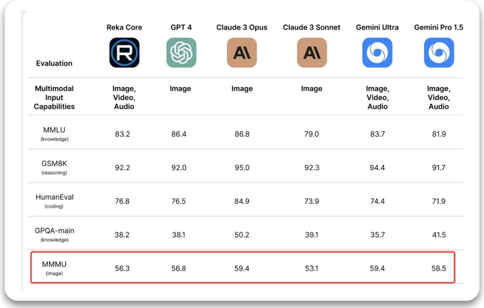
Yi Open-source Performance Benchmarks

Yi-VL-34B Multimodal Version - as of Jan 21, 2024

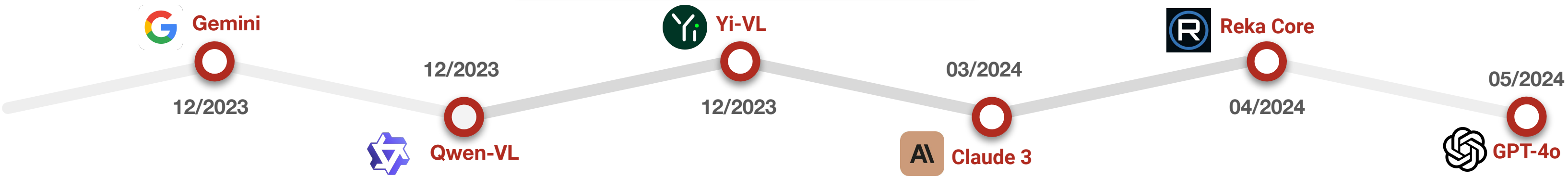
MMMU Benchmark

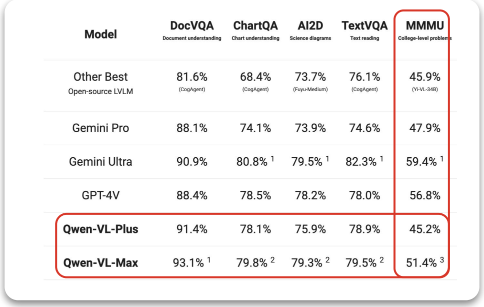
Overall results of different models on the MMMU test set. The best-performing model in each category is **in-bold**, and the second best is underlined. *results provided by the authors.

| Reset | Test Overall | Validation Overall | Art& Design | Business | Science | Health& Medicine | Human& Social Sci. | Tech& Eng |
|---|---|---|---|---|---|---|---|---|
| GPT-4V | **43.7** | **42.5** | 61.0 | **36.3** | **40.9** | **46.8** | **44.2** | **41.5** |
| Qwen-VL-Plus | 36.8 | 39.5 | 61.5 | 23.2 | 32.8 | 40.5 | 43.4 | 33.3 |
| Yi-VL-34B | 36.5 | 36.2 | **62.9** | 19.1 | 31.5 | 42.1 | 42.5 | 34.5 |
| Yi-VL-6B | 35.0 | 35.8 | 58.0 | 19.9 | 32.3 | 39.3 | 40.6 | 32.1 |
| Qwen-VL-Chat | 31.3 | 30.7 | 52.6 | 18.5 | 26.9 | 33.4 | 34.1 | 31.4 |
| InternVL-Chat-ViT-6B-Vicuna-13B | 26.7 | 26.4 | 39.7 | 13.8 | 23.0 | 31.7 | 26.5 | 28.5 |

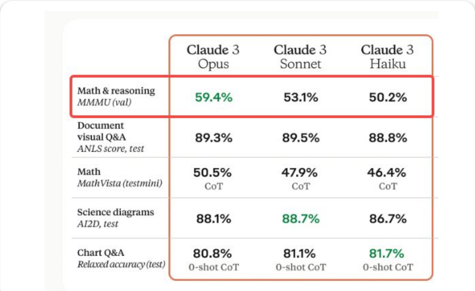


| Evaluation | Reka Core | GPT 4 | Claude 3 Opus | Claude 3 Sonnet | Gemini Ultra | Gemini Pro 1.5 |
|---|---|---|---|---|---|---|
| Multimodal Input Capabilities | Image, Video, Audio | Image | Image | Image | Image, Video, Audio | Image, Video, Audio |
| MMLU (knowledge) | 83.2 | 86.4 | 86.8 | 79.0 | 83.7 | 81.9 |
| GSM8K (reasoning) | 92.2 | 92.0 | 95.0 | 92.3 | 94.4 | 91.7 |
| HumanEval (coding) | 76.8 | 76.5 | 84.9 | 73.9 | 74.4 | 71.9 |
| GPQA-main (knowledge) | 38.2 | 38.1 | 50.2 | 39.1 | 35.7 | 41.5 |
| MMMU (image) | 56.3 | 56.8 | 59.4 | 53.1 | 59.4 | 58.5 |

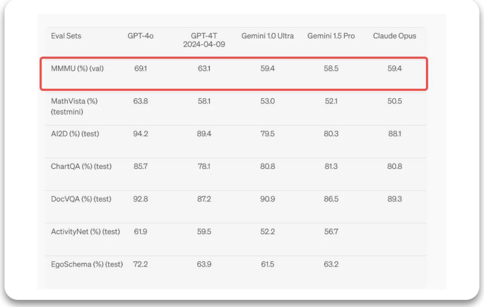




| Model | DocVQA<br>Document understanding | ChartQA<br>Chart understanding | AI2D<br>Science diagrams | TextVQA<br>Text reading | MMMU<br>College-level problems |
|---|---|---|---|---|---|
| Other Best<br>Open-source LVLM | 81.6%<br>(CogAgent) | 68.4%<br>(CogAgent) | 73.7%<br>(Fuyu-Medium) | 76.1%<br>(CogAgent) | 45.9%<br>(Yi-VL-34B) |
| Gemini Pro | 88.1% | 74.1% | 73.9% | 74.6% | 47.9% |
| Gemini Ultra | 90.9% | 80.8% [1] | 79.5% [1] | 82.3% [1] | 59.4% [1] |
| GPT-4V | 88.4% | 78.5% | 78.2% | 78.0% | 56.8% |
| **Qwen-VL-Plus** | 91.4% | 78.1% | 75.9% | 78.9% | 45.2% |
| **Qwen-VL-Max** | 93.1% [1] | 79.8% [2] | 79.3% [2] | 79.5% [2] | 51.4% [3] |



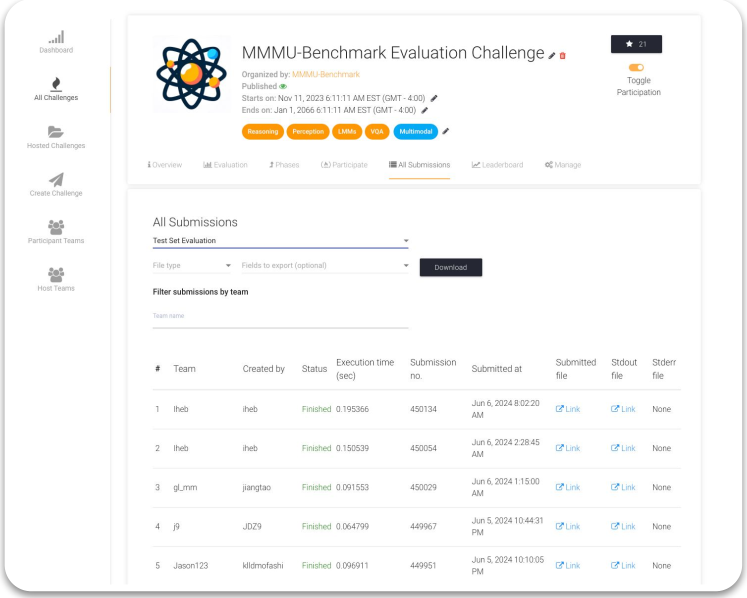
| | Claude 3<br>Opus | Claude 3<br>Sonnet | Claude 3<br>Haiku |
|---|---|---|---|
| Math & reasoning<br>*MMMU (val)* | **59.4%** | **53.1%** | **50.2%** |
| Document visual Q&A<br>*ANLS score, test* | **89.3%** | **89.5%** | **88.8%** |
| Math<br>*MathVista (testmini)* | **50.5%**<br>CoT | **47.9%**<br>CoT | **46.4%**<br>CoT |
| Science diagrams<br>*AI2D, test* | **88.1%** | **88.7%** | **86.7%** |
| Chart Q&A<br>*Relaxed accuracy (test)* | **80.8%**<br>0-shot CoT | **81.1%**<br>0-shot CoT | **81.7%**<br>0-shot CoT |



| Eval Sets | GPT-4o | GPT-4T<br>2024-04-09 | Gemini 1.0 Ultra | Gemini 1.5 Pro | Claude Opus |
|---|---|---|---|---|---|
| MMMU (%) (val) | 69.1 | 63.1 | 59.4 | 58.5 | 59.4 |
| MathVista (%) (testmini) | 63.8 | 58.1 | 53.0 | 52.1 | 50.5 |
| AI2D (%) (test) | 94.2 | 89.4 | 79.5 | 80.3 | 88.1 |
| ChartQA (%) (test) | 85.7 | 78.1 | 80.8 | 81.3 | 80.8 |
| DocVQA (%) (test) | 92.8 | 87.2 | 90.9 | 86.5 | 89.3 |
| ActivityNet (%) (test) | 61.9 | 59.5 | 52.2 | 56.7 | |
| EgoSchema (%) (test) | 72.2 | 63.9 | 61.5 | 63.2 | |

# MMMU in the Community

- EvalAI



MMMU-Benchmark Evaluation Challenge

Organized by: MMMU-Benchmark
Published
Starts on: Nov 11, 2023 6:11:11 AM EST (GMT - 4:00)
Ends on: Jan 1, 2066 6:11:11 AM EST (GMT - 4:00)

Reasoning · Perception · LMMs · VQA · Multimodal

All Submissions

Test Set Evaluation

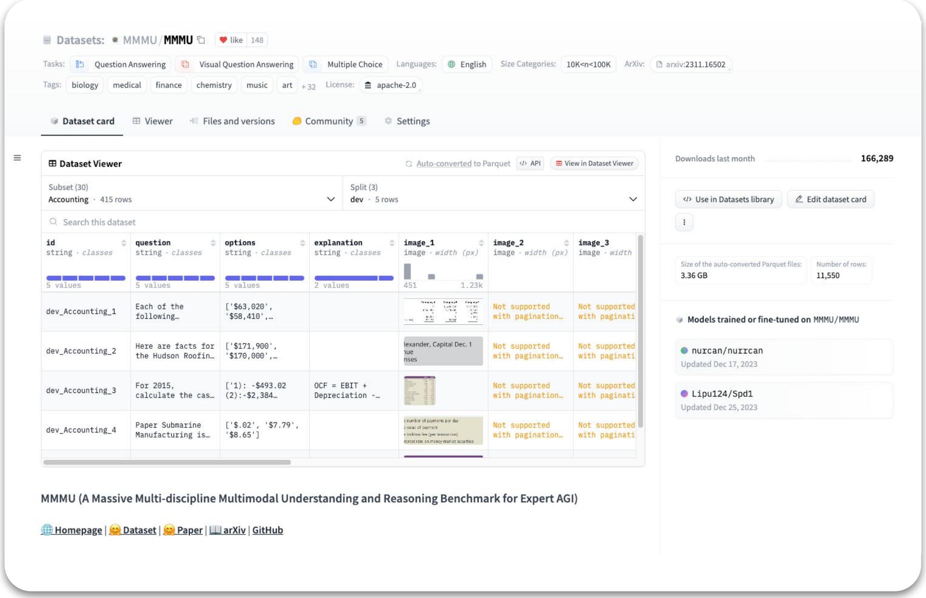
File type
Fields to export (optional)

Filter submissions by team

Team name

| # | Team | Created by | Status | Execution time (sec) | Submission no. | Submitted at | Submitted file | Stdout file | Stderr file |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Iheb | iheb | Finished | 0.195366 | 450134 | Jun 6, 2024 8:02:20 AM | Link | Link | None |
| 2 | Iheb | iheb | Finished | 0.150539 | 450054 | Jun 6, 2024 2:28:45 AM | Link | Link | None |
| 3 | gl_mm | jiangtao | Finished | 0.091553 | 450029 | Jun 6, 2024 1:15:00 AM | Link | Link | None |
| 4 | j9 | JDZ9 | Finished | 0.064799 | 449967 | Jun 5, 2024 10:44:31 PM | Link | Link | None |
| 5 | Jason123 | klldmofashi | Finished | 0.096911 | 449951 | Jun 5, 2024 10:10:05 PM | Link | Link | None |

- Hugging Face



Datasets: MMMU / MMMU  like 148

Tasks: Question Answering · Visual Question Answering · Multiple Choice  Languages: English  Size Categories: 10K<n<100K  ArXiv: arxiv:2311.16502

Tags: biology · medical · finance · chemistry · music · art · +32  License: apache-2.0

Dataset Viewer

Subset (30)
Accounting · 415 rows

Split (3)
dev · 5 rows

Search this dataset

| id<br>string · classes<br>5 values | question<br>string · classes<br>5 values | options<br>string · classes<br>5 values | explanation<br>string · classes<br>2 values | image_1<br>image · width (px)<br>451 1.23k | image_2<br>image · width (px) | image_3<br>image · width |
|---|---|---|---|---|---|---|
| dev_Accounting_1 | Each of the following… | ['$63,020', '$58,410',… | | | Not supported with pagination… | Not supported with paginati |
| dev_Accounting_2 | Here are facts for the Hudson Roofin… | ['$171,900', '$170,000',… | | | Not supported with pagination… | Not supported with paginati |
| dev_Accounting_3 | For 2015, calculate the cas… | ['1): -$493.02 (2):-$2,384… | OCF = EBIT + Depreciation -… | | Not supported with pagination… | Not supported with paginati |
| dev_Accounting_4 | Paper Submarine Manufacturing is… | ['$.02', '$7.79', '$8.65'] | | | Not supported with pagination… | Not supported with paginati |

Downloads last month 166,289

Size of the auto-converted Parquet files: 3.36 GB
Number of rows: 11,550

Models trained or fine-tuned on MMMU/MMMU

nurcan/nurrcan
Updated Dec 17, 2023

Lipu124/Spd1
Updated Dec 25, 2023

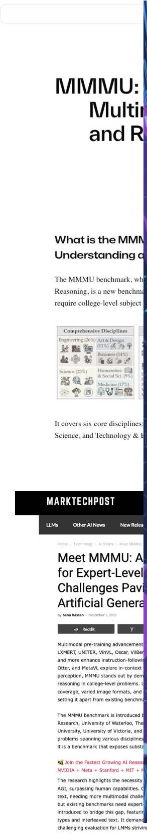
MMMU (A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI)

Homepage | Dataset | Paper | arXiv | GitHub

- **130+** Models
- **2500+** Submissions

- **160K+** Downloads Last Month
- **700K+** Downloads in Total

# Mentioned by AI blogs, posts, and reports



MMMU:
Multi
and R

What is the MMM
Understanding a

The MMMU benchmark, wh
Reasoning, is a new benchma
require college-level subject

It covers six core disciplines:
Science, and Technology & E

**MARKTECHPOST**

Meet MMMU: A
for Expert-Level
Challenges Pavi
Artificial Genera

By **Sana Hassan** - December 5, 2023

Multimodal pre-training advancement
LXMERT, UNITER, VinVL, Oscar, VilBer
and more enhance instruction-followi
Otter, and MetaVL explore in-context
perception, MMMU stands out by dem
reasoning in college-level problems. U
coverage, varied image formats, and
setting it apart from existing benchm

The MMMU benchmark is introduced b
Research, University of Waterloo, The
University, University of Victoria, and
problems spanning various disciplines
it is a benchmark that exposes substa

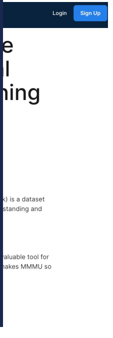
🐝 Join the Fastest Growing AI Resea
NVIDIA + Meta + Stanford + MIT + M

The research highlights the necessity
AGI, surpassing human capabilities. C
text, needing more multimodal challe
but existing benchmarks need expert-
introduced to bridge this gap, featuri
types and interleaved text. It demand
challenging evaluation for LMMs striving for advanced AI capabilities.

Artificial Intelligence Index Report 2024

Vision v1.5: CLIP-like Vision
Models that Can...

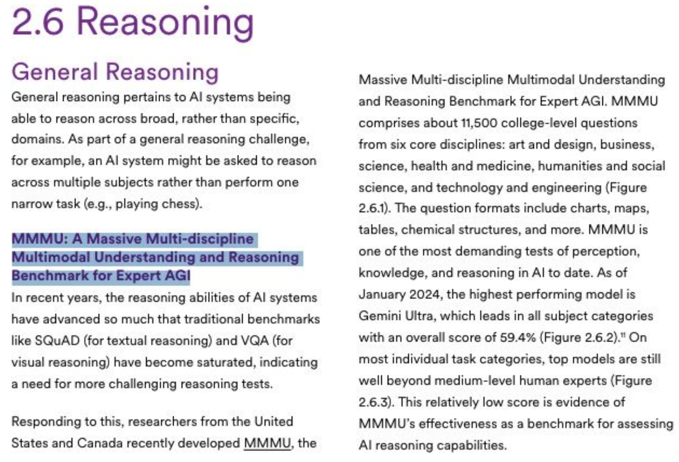
- **Art & Design:** Exploring visual arts, design principles, and aesthetic judgments.



e
l
ning

k) is a dataset
standing and

aluable tool for
nakes MMMU so



## 2.6 Reasoning

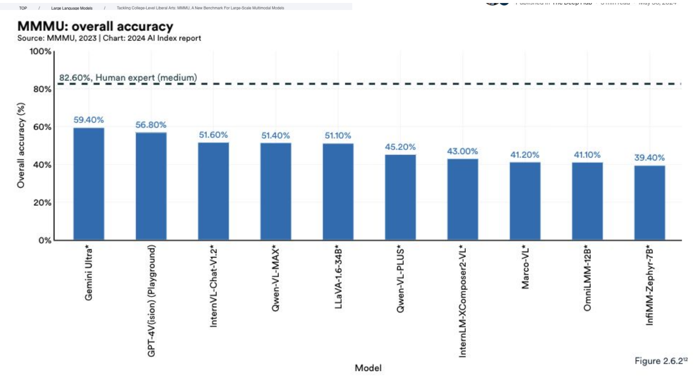
### General Reasoning

General reasoning pertains to AI systems being able to reason across broad, rather than specific, domains. As part of a general reasoning challenge, for example, an AI system might be asked to reason across multiple subjects rather than perform one narrow task (e.g., playing chess).

### MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

In recent years, the reasoning abilities of AI systems have advanced so much that traditional benchmarks like SQuAD (for textual reasoning) and VQA (for visual reasoning) have become saturated, indicating a need for more challenging reasoning tests.

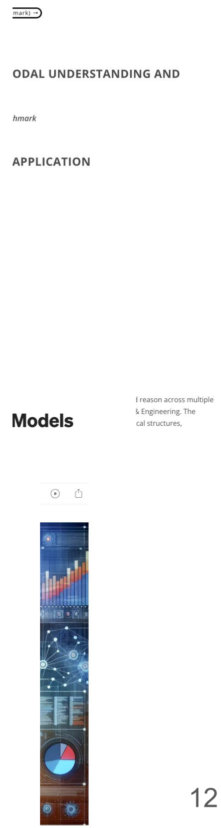
Responding to this, researchers from the United States and Canada recently developed MMMU, the Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. MMMU comprises about 11,500 college-level questions from six core disciplines: art and design, business, science, health and medicine, humanities and social science, and technology and engineering (Figure 2.6.1). The question formats include charts, maps, tables, chemical structures, and more. MMMU is one of the most demanding tests of perception, knowledge, and reasoning in AI to date. As of January 2024, the highest performing model is Gemini Ultra, which leads in all subject categories with an overall score of 59.4% (Figure 2.6.2).[11] On most individual task categories, top models are still well beyond medium-level human experts (Figure 2.6.3). This relatively low score is evidence of MMMU's effectiveness as a benchmark for assessing AI reasoning capabilities.



MMMU: overall accuracy

Source: MMMU, 2023 | Chart: 2024 AI Index report

Figure 2.6.2[12]



ODAL UNDERSTANDING AND

hmark

APPLICATION

**Models**

l reason across multiple
& Engineering. The
cal structures,

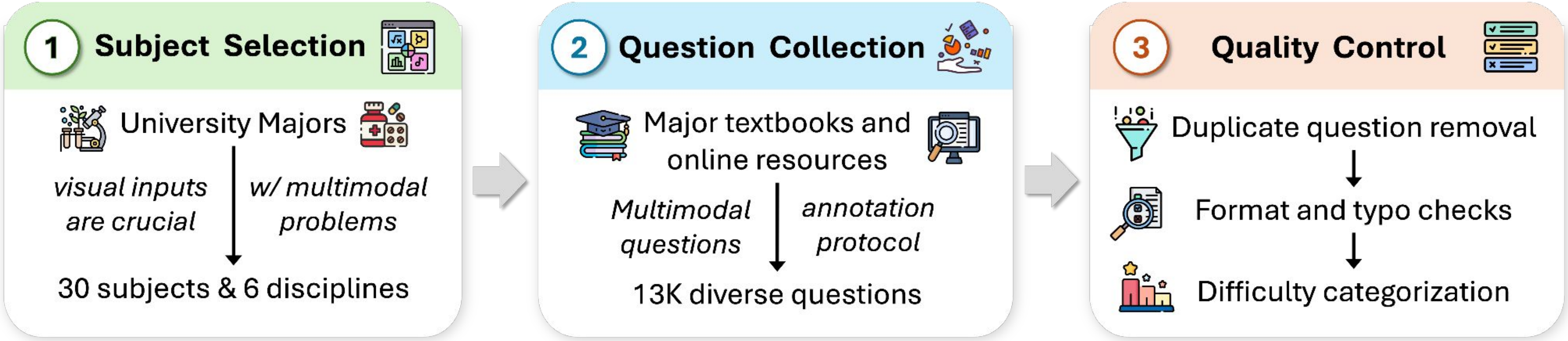
# Why MMMU?

- Rigorous Data Curation Process and High-Quality Data
- Effective Suite for Tracking Multimodal Model Development
- Excellent Model Diagnosis Tool

# Why MMMU?

- Rigorous Data Curation Process and High-Quality Data
- Effective Suite for Tracking Multimodal Model Development
- Excellent Model Diagnosis Tool

# 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

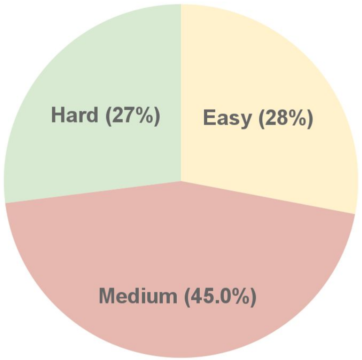
**3 Quality Control**

Duplicate question removal

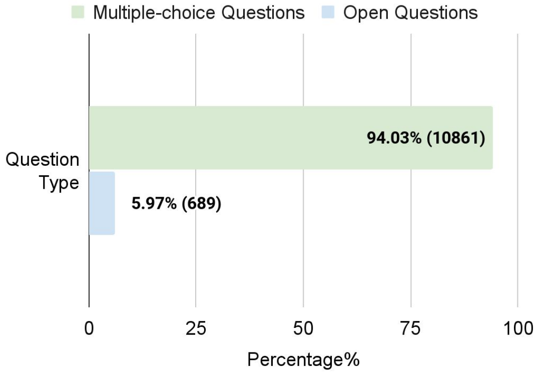
Format and typo checks

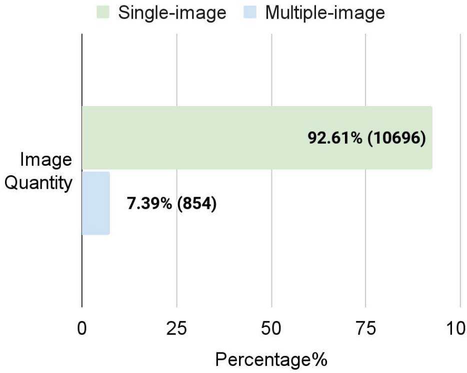
Difficulty categorization

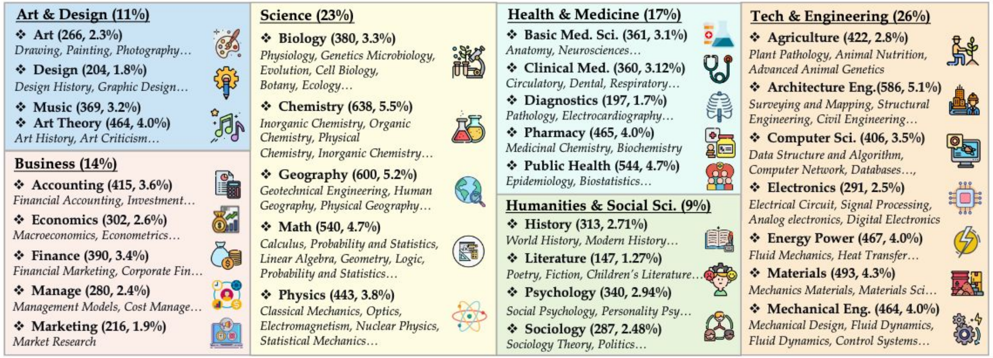
# Statistics of MMMU



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…*

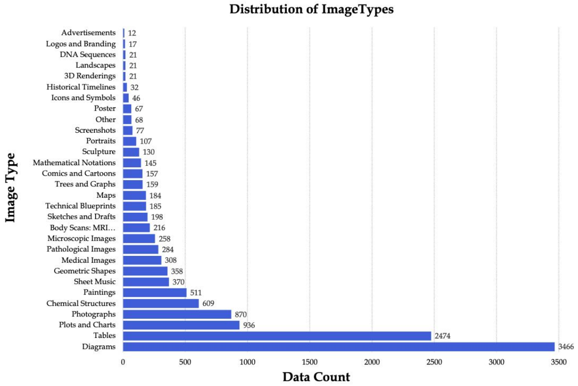
**Business (14%)**
- **Accounting (415, 3.6%)** *Financial Accounting, Investment…*
- **Economics (302, 2.6%)** *Macroeconomics, Econometrics…*
- **Finance (390, 3.4%)** *Financial Marketing, Corporate Fin…*
- **Manage (280, 2.4%)** *Management Models, Cost Manage…*
- **Marketing (216, 1.9%)** *Market 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…*

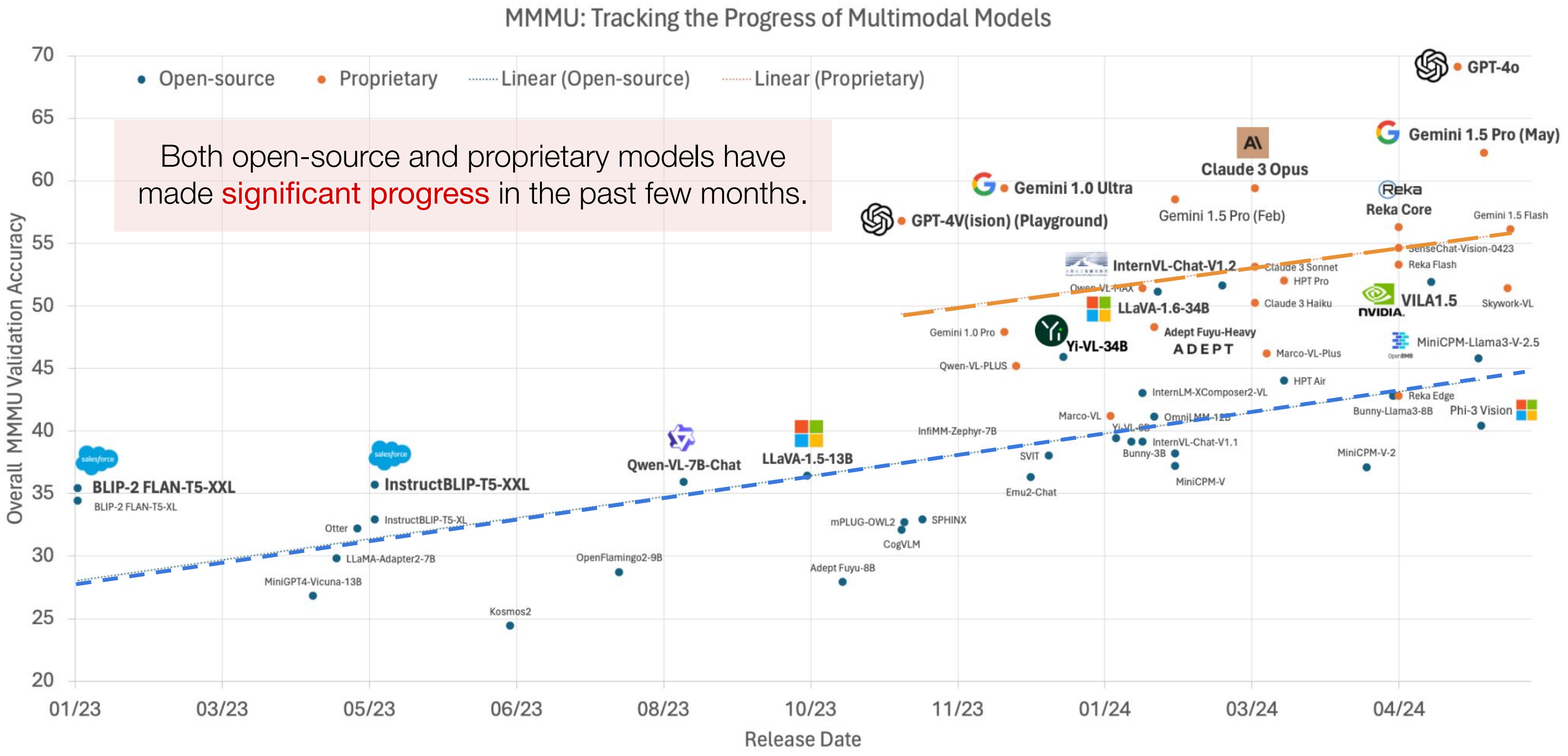
**Tech & Engineering (26%)**
- **Agriculture (422, 2.8%)** *Plant Pathology, Animal Nutrition, Advanced Animal Genetics*
- **Architecture Eng.(586, 5.1%)** *Surveying and Mapping, Structural Engineering, Civil Engineering…*
- **Computer Sci. (406, 3.5%)** *Data Structure and Algorithm, Computer Network, 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…*



# Why MMMU?

- Rigorous Data Curation Process and High-Quality Data
- Effective Suite for Tracking Multimodal Model Development
- Excellent Model Diagnosis Tool

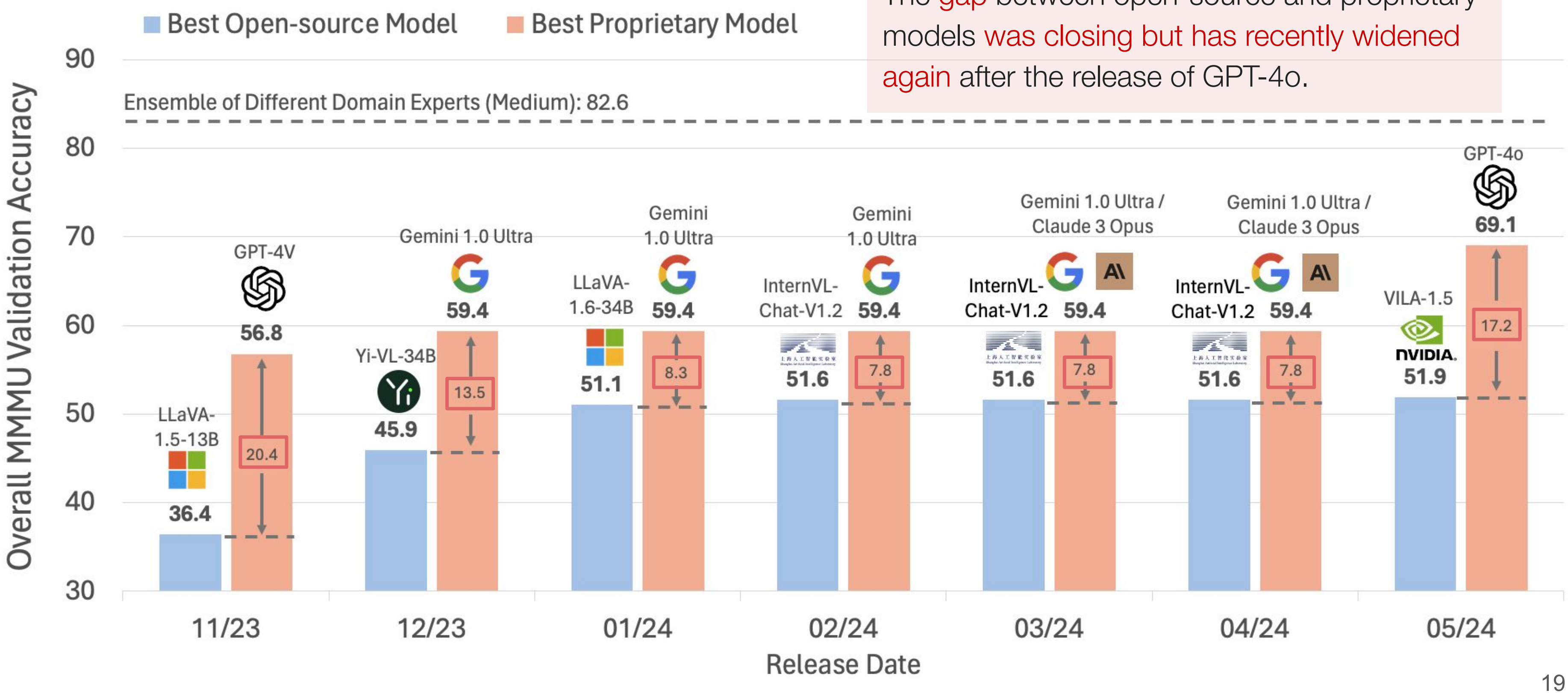
# Tracking the development of multimodal models



Both open-source and proprietary models have made significant progress in the past few months.

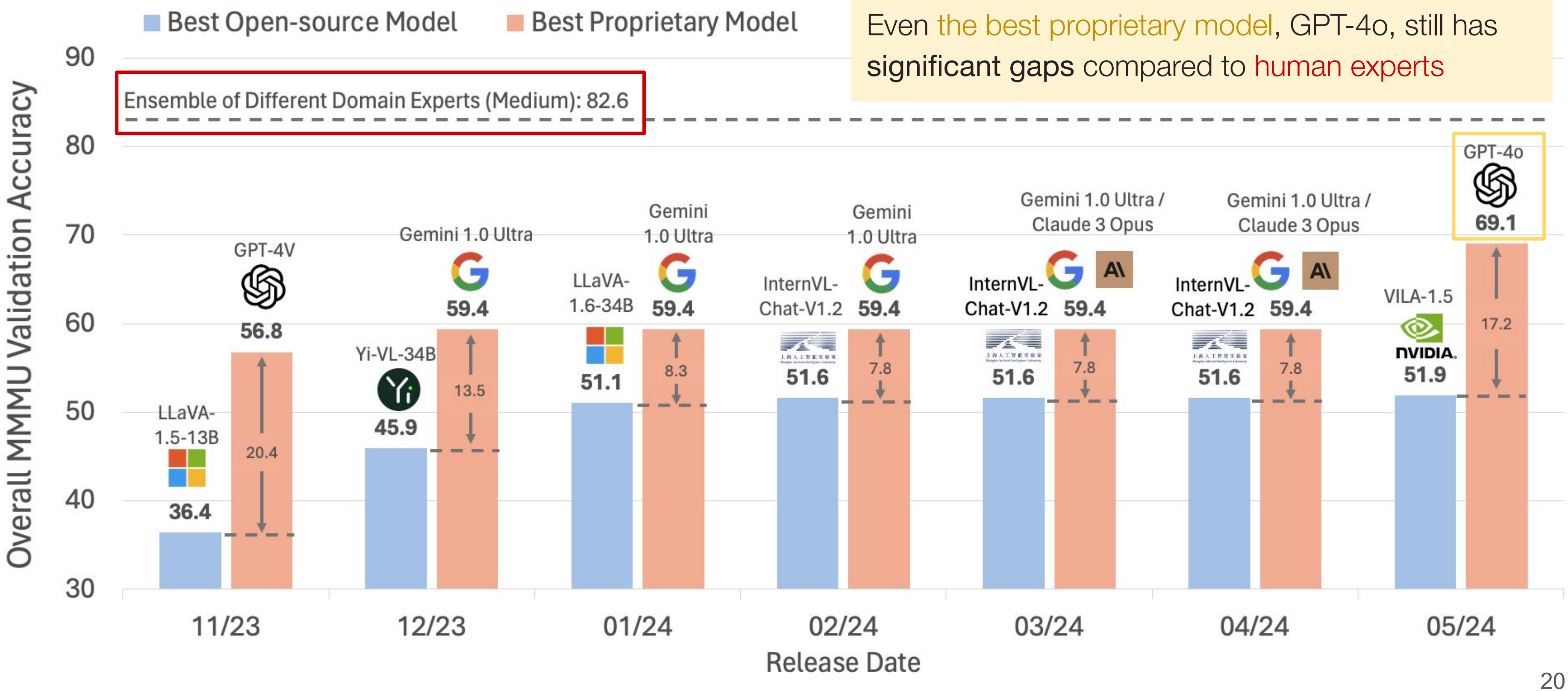
# Open-source VS. Proprietary

The gap between open-source and proprietary models was closing but has recently widened again after the release of GPT-4o.



# Open-source VS. Proprietary

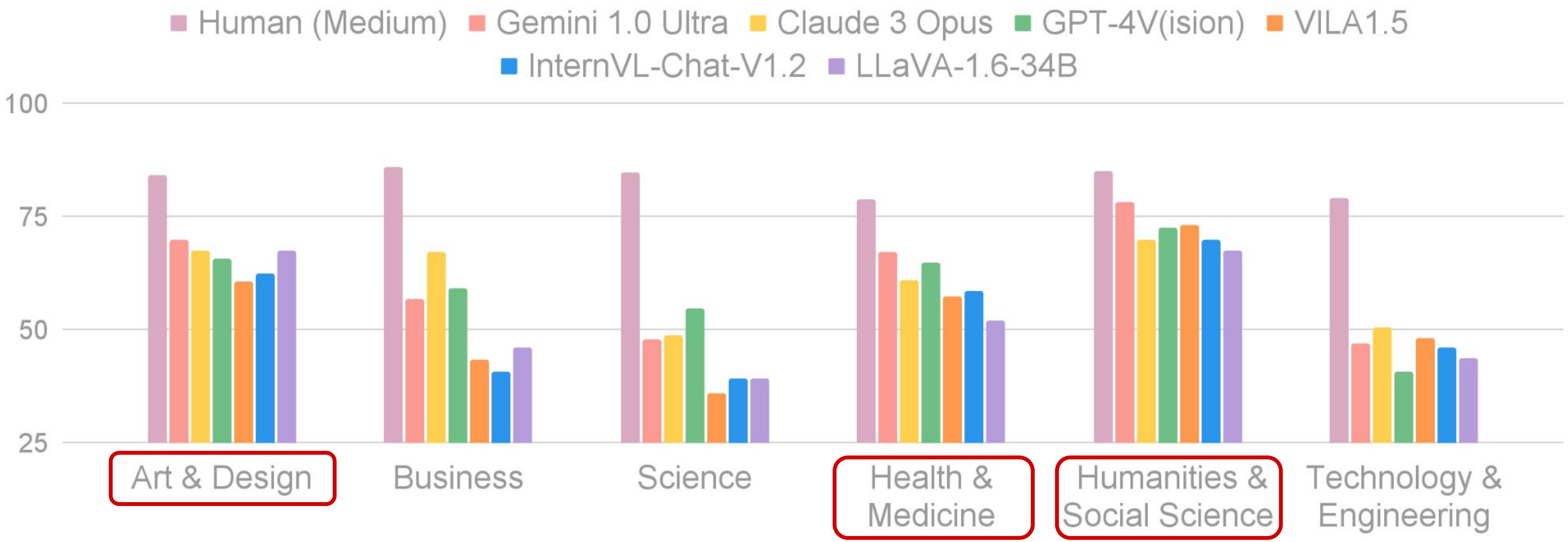
Even the best proprietary model, GPT-4o, still has **significant gaps** compared to human experts



# Why MMMU?

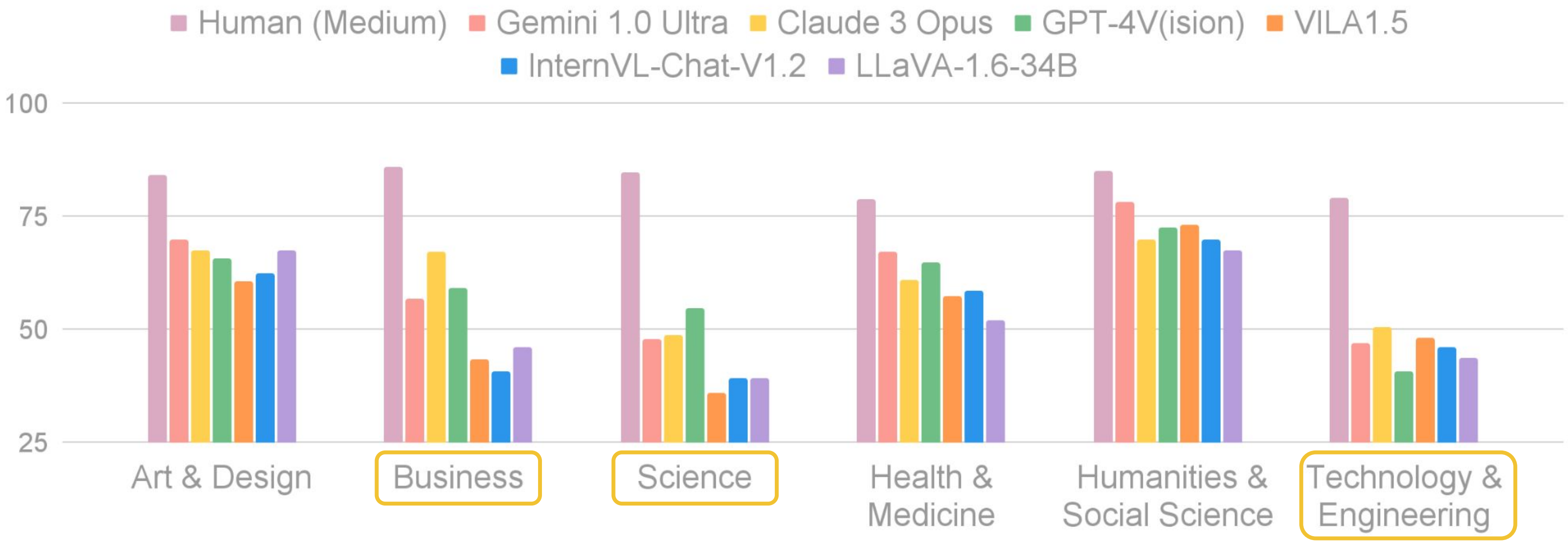
- Rigorous Data Curation Process and High-Quality Data
- Effective Suite for Tracking Multimodal Model Development
- Excellent Model Diagnosis Tool

# Subject-specific Accuracy



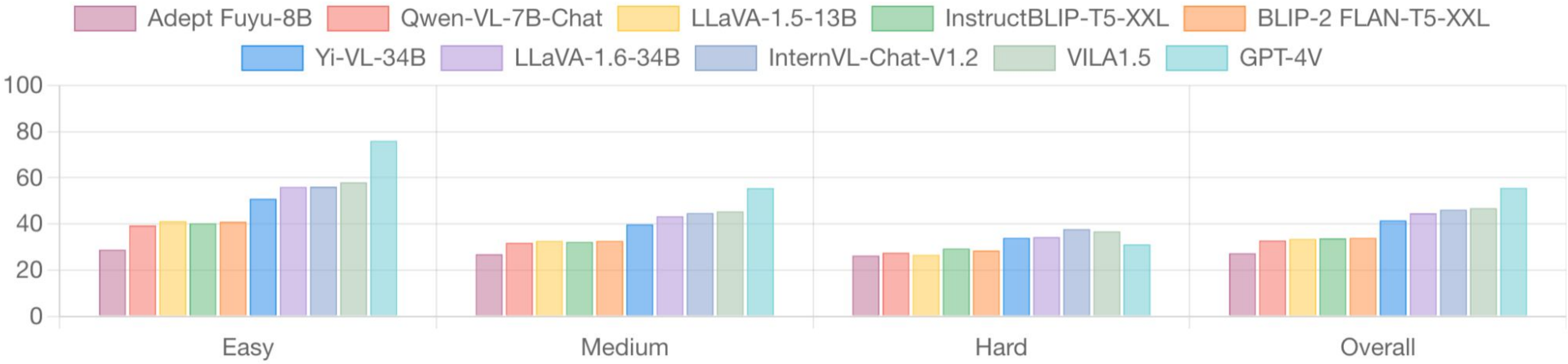
- ○ The gap between the best models and human experts is not large.
- ○ The difference between open-source and proprietary models is not significant.

# Subject-specific Accuracy



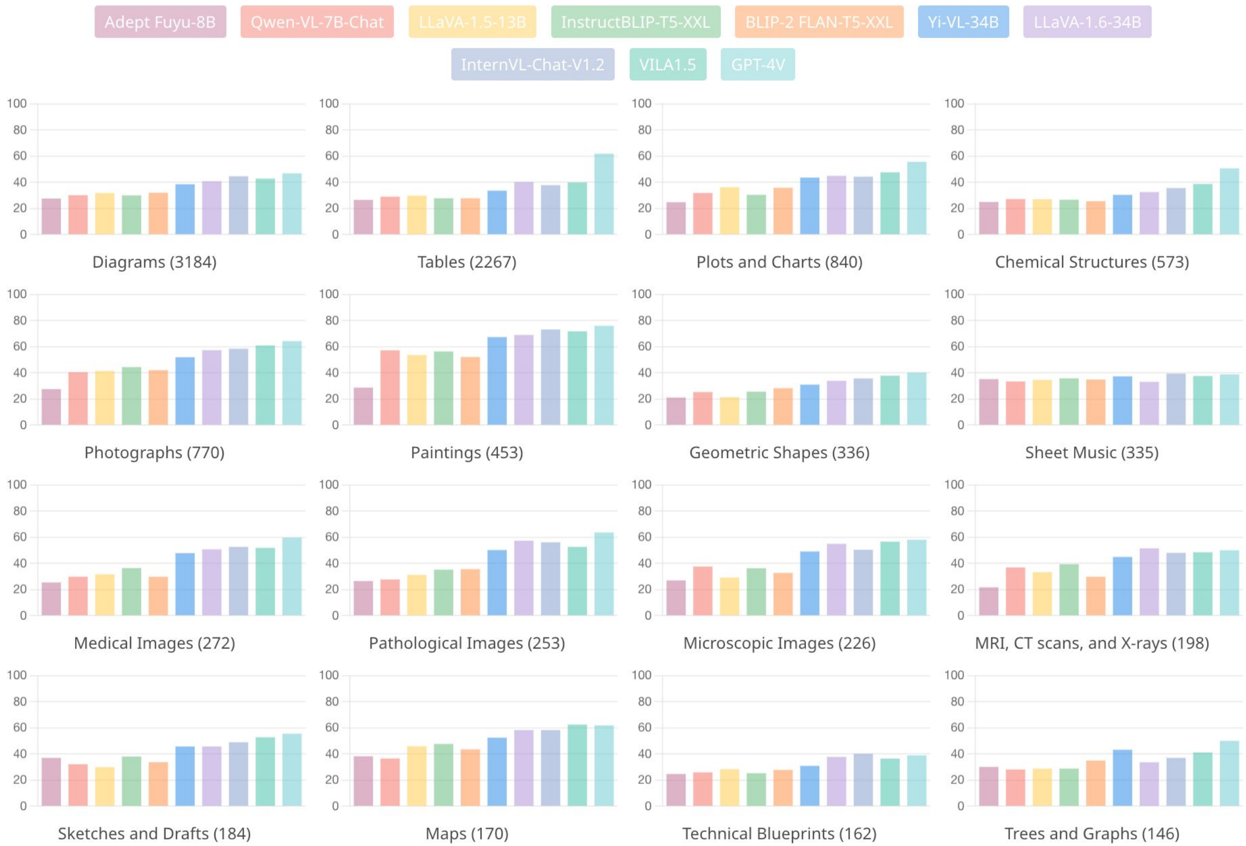
- The gap between the best models and human experts is significantly large.
- Models struggle with these subjects, which involve more complex reasoning questions

# Difficulty-specific Accuracy

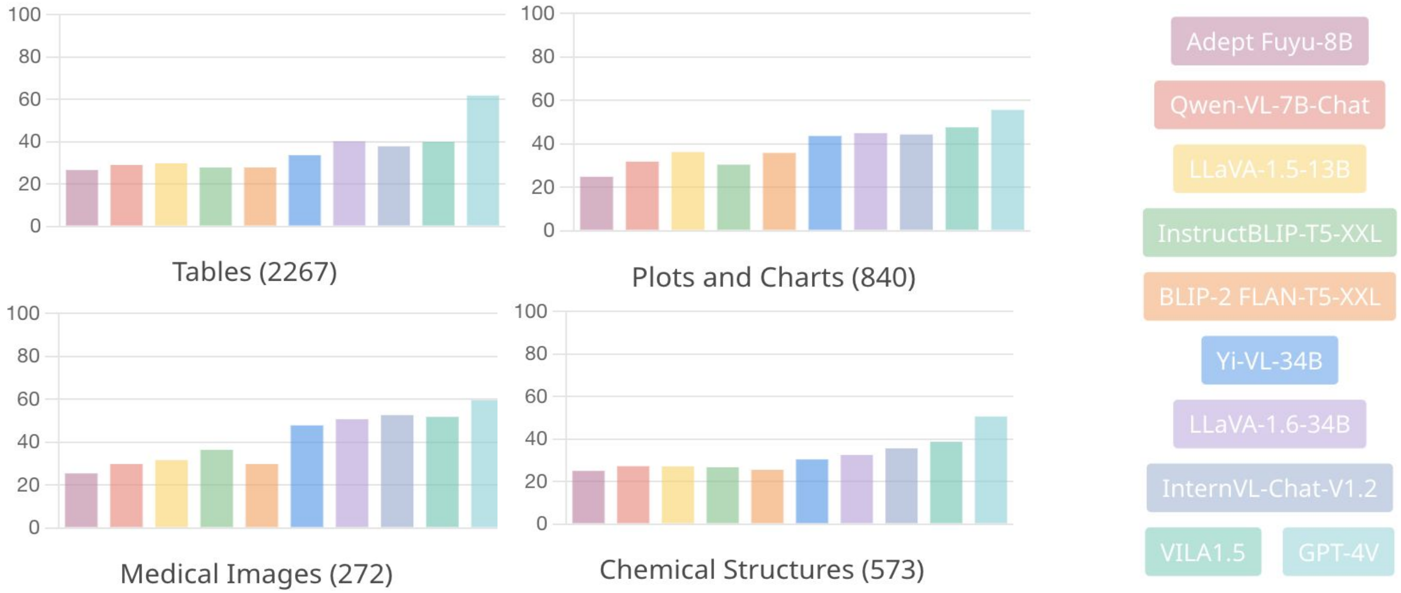


GPT-4V outperforms open-source models on easy and medium-level tasks, while all models struggle with hard examples.

# Image Type-specific Accuracy

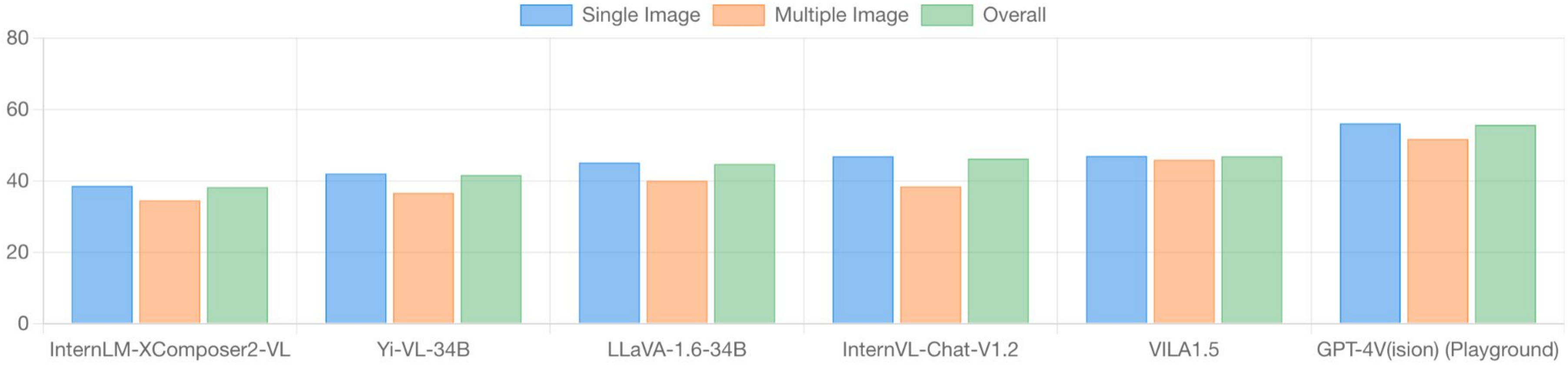


# Tables, plots, and domain-specific images



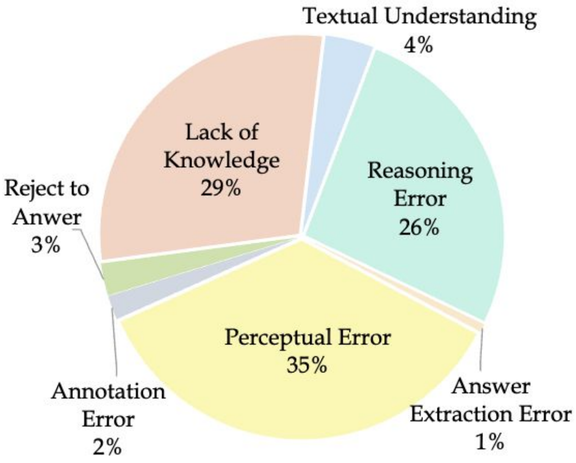
GPT-4V is better at comprehending tables, plots and domain-specific images compared with open-source models.

# Single-image V.S. Multiple-image



- Models generally struggle with reasoning over multiple images
- VILA performs notably better in this area

# Error Analysis



Error distribution over 150 annotated GPT-4V errors (root cause)

## Perceptual Errors (35%):

- Basic perceptual errors: fails in elementary visual interpretation
- Prioritizing textual over visual inputs
- Grounding and referring issues

## Lack of Knowledge (29%):

- Knowledge is the foundation of domain-specific perception and reasoning

## Reasoning Error (26%):

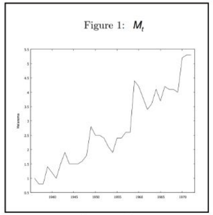
- Models struggle with complex reasoning and long calculation chains

# Language as Vehicle: Model can explain what it sees and the rationale for its prediction

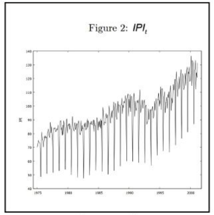
## ✓ Economics : Econometrics

**Question:** Figure 1 displays the number of cases of Melanoma (Mt) in the male population of a US State, from 1936 to 1972. Figure 2 shows the Spanish Industrial Production Index (IPIt), from January 1975 to March 2001. <image 1> <image 2> According to figures 1 and 2, which of the following statements is FALSE:

**Option:**
(A) IPIt displays a strong seasonality
(B) Both time series are mean stationary
(C) Both time series display a trending behavior



<image 1>



<image 2>

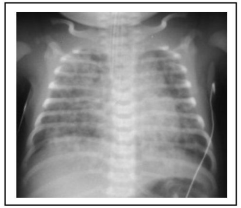
**GPT-4V:** Based on the provided figures:(A) IPIt displays a strong seasonality. - This statement appears to be TRUE. Figure 2 shows consistent and repeating peaks and valleys for the Spanish Industrial Production Index (IPIt), indicative of seasonality.(B) Both time series are mean stationary. - This statement appears to be FALSE. A mean stationary time series would have a constant mean (average) over time. Both figures show a changing mean over the given time frames, particularly Figure 1 where the mean value for Melanoma cases appears to be increasing over time.(C) Both time series display a trending behavior. - This statement appears to be TRUE. Figure 1 shows an upward trend in the number of Melanoma cases over time, and Figure 2 displays fluctuations but also an overall upward trend in the Industrial Production Index. Therefore, the FALSE statement is:
(B) Both time series are mean stationary.

**Ground Truth: (B) Both time series are mean stationary.**

## ✓ Clinical Medicine: Pathology

**Question:** The patient is a post-mature newborn. What is the most likely diagnosis? <image 1>

**Option:**
(A) Transient tachypnea of the newborn
(B) Congenital lobar emphysema
(C) Hyaline membrane disease
(D) Meconium aspiration syndrome

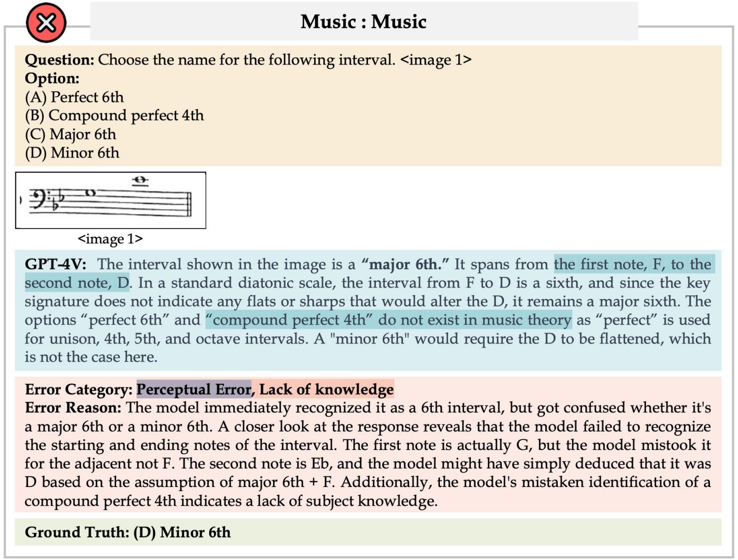


<image 1>

**GPT-4V:** Based on the image provided, which appears to be a chest X-ray, and the information that the patient is a post-mature newborn, the most suggestive diagnosis is: (D) Meconium aspiration syndrome. Post-mature newborns are at an increased risk of passing meconium (the baby's first stool) into the amniotic fluid before birth or during labor. If the baby breathes the meconium-stained amniotic fluid into their lungs, it can cause a blockage of the airways, leading to meconium aspiration syndrome. The chest X-ray of a newborn with meconium aspiration syndrome might show patchy or streaky areas on the lungs. However, it's important to consult with a pediatric radiologist or neonatologist for a definitive diagnosis based on clinical findings and the entire clinical picture.

**Ground Truth: (D) Meconium aspiration syndrome**

# Knowledge is the foundation for perception and reasoning



**Music : Music**

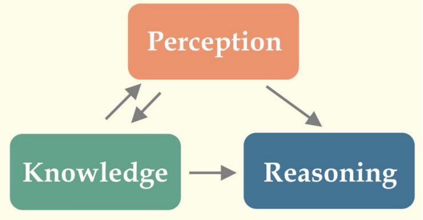
**Question:** Choose the name for the following interval. <image 1>
**Option:**
(A) Perfect 6th
(B) Compound perfect 4th
(C) Major 6th
(D) Minor 6th

<image 1>

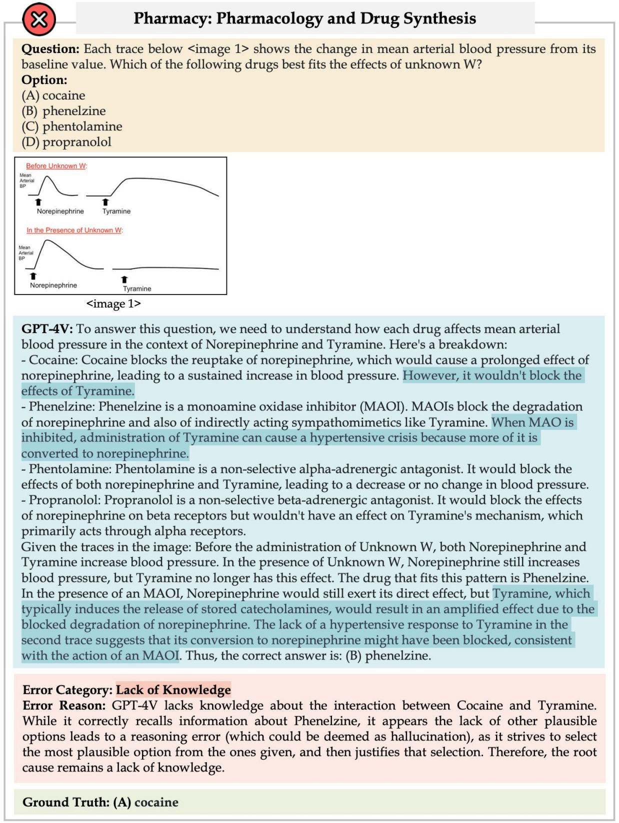
**GPT-4V:** The interval shown in the image is a **"major 6th."** It spans from the first note, F, to the second note, D. In a standard diatonic scale, the interval from F to D is a sixth, and since the key signature does not indicate any flats or sharps that would alter the D, it remains a major sixth. The options "perfect 6th" and "compound perfect 4th" do not exist in music theory as "perfect" is used for unison, 4th, 5th, and octave intervals. A "minor 6th" would require the D to be flattened, which is not the case here.

**Error Category: Perceptual Error, Lack of knowledge**
**Error Reason:** The model immediately recognized it as a 6th interval, but got confused whether it's a major 6th or a minor 6th. A closer look at the response reveals that the model failed to recognize the starting and ending notes of the interval. The first note is actually G, but the model mistook it for the adjacent not F. The second note is Eb, and the model might have simply deduced that it was D based on the assumption of major 6th + F. Additionally, the model's mistaken identification of a compound perfect 4th indicates a lack of subject knowledge.

**Ground Truth: (D) Minor 6th**



Lack of deep subject knowledge could lead to perception and reasoning errors



**Pharmacy: Pharmacology and Drug Synthesis**

**Question:** Each trace below <image 1> shows the change in mean arterial blood pressure from its baseline value. Which of the following drugs best fits the effects of unknown W?
**Option:**
(A) cocaine
(B) phenelzine
(C) phentolamine
(D) propranolol

Before Unknown W:
Mean Arterial BP
Norepinephrine
Tyramine
In the Presence of Unknown W:
Mean Arterial BP
Norepinephrine
Tyramine

<image 1>

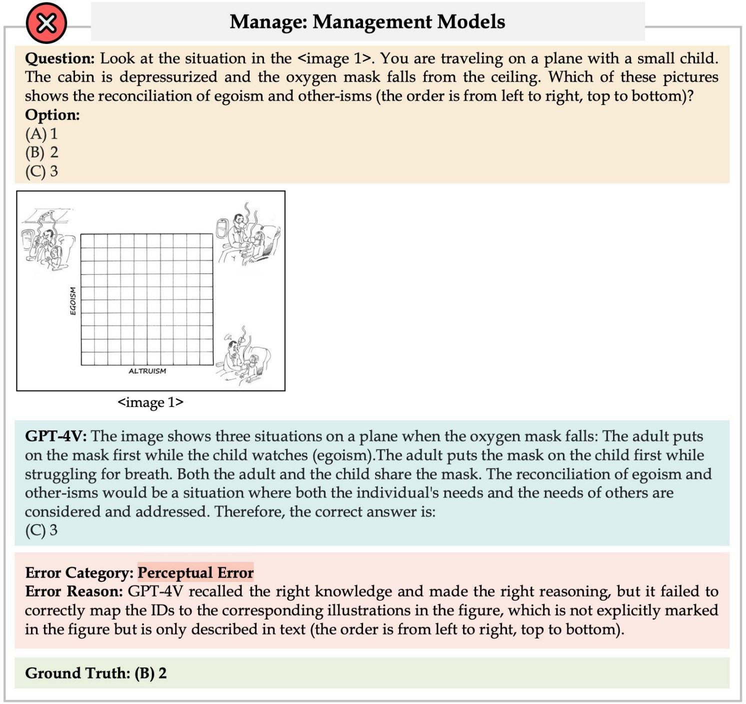
**GPT-4V:** To answer this question, we need to understand how each drug affects mean arterial blood pressure in the context of Norepinephrine and Tyramine. Here's a breakdown:
- Cocaine: Cocaine blocks the reuptake of norepinephrine, which would cause a prolonged effect of norepinephrine, leading to a sustained increase in blood pressure. However, it wouldn't block the effects of Tyramine.
- Phenelzine: Phenelzine is a monoamine oxidase inhibitor (MAOI). MAOIs block the degradation of norepinephrine and also of indirectly acting sympathomimetics like Tyramine. When MAO is inhibited, administration of Tyramine can cause a hypertensive crisis because more of it is converted to norepinephrine.
- Phentolamine: Phentolamine is a non-selective alpha-adrenergic antagonist. It would block the effects of both norepinephrine and Tyramine, leading to a decrease or no change in blood pressure.
- Propranolol: Propranolol is a non-selective beta-adrenergic antagonist. It would block the effects of norepinephrine on beta receptors but wouldn't have an effect on Tyramine's mechanism, which primarily acts through alpha receptors.

Given the traces in the image: Before the administration of Unknown W, both Norepinephrine and Tyramine increase blood pressure. In the presence of Unknown W, Norepinephrine still increases blood pressure, but Tyramine no longer has this effect. The drug that fits this pattern is Phenelzine. In the presence of an MAOI, Norepinephrine would still exert its direct effect, but Tyramine, which typically induces the release of stored catecholamines, would result in an amplified effect due to the blocked degradation of norepinephrine. The lack of a hypertensive response to Tyramine in the second trace suggests that its conversion to norepinephrine might have been blocked, consistent with the action of an MAOI. Thus, the correct answer is: (B) phenelzine.

**Error Category: Lack of Knowledge**
**Error Reason:** GPT-4V lacks knowledge about the interaction between Cocaine and Tyramine. While it correctly recalls information about Phenelzine, it appears the lack of other plausible options leads to a reasoning error (which could be deemed as hallucination), as it strives to select the most plausible option from the ones given, and then justifies that selection. Therefore, the root cause remains a lack of knowledge.

**Ground Truth: (A) cocaine**

# Grounding and referring is challenging



**Manage: Management Models**

**Question:** Look at the situation in the <image 1>. You are traveling on a plane with a small child. The cabin is depressurized and the oxygen mask falls from the ceiling. Which of these pictures shows the reconciliation of egoism and other-isms (the order is from left to right, top to bottom)?

**Option:**
(A) 1
(B) 2
(C) 3

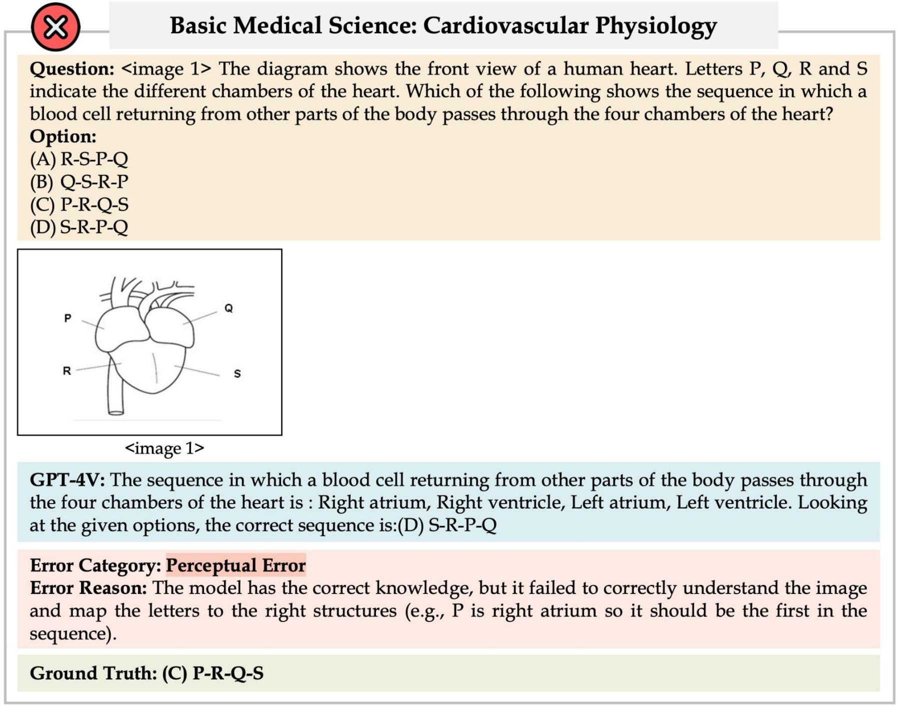
EGOISM
ALTRUISM

<image 1>

**GPT-4V:** The image shows three situations on a plane when the oxygen mask falls: The adult puts on the mask first while the child watches (egoism).The adult puts the mask on the child first while struggling for breath. Both the adult and the child share the mask. The reconciliation of egoism and other-isms would be a situation where both the individual's needs and the needs of others are considered and addressed. Therefore, the correct answer is:
(C) 3

**Error Category:** **Perceptual Error**
**Error Reason:** GPT-4V recalled the right knowledge and made the right reasoning, but it failed to correctly map the IDs to the corresponding illustrations in the figure, which is not explicitly marked in the figure but is only described in text (the order is from left to right, top to bottom).

**Ground Truth: (B) 2**



**Basic Medical Science: Cardiovascular Physiology**

**Question:** <image 1> The diagram shows the front view of a human heart. Letters P, Q, R and S indicate the different chambers of the heart. Which of the following shows the sequence in which a blood cell returning from other parts of the body passes through the four chambers of the heart?

**Option:**
(A) R-S-P-Q
(B) Q-S-R-P
(C) P-R-Q-S
(D) S-R-P-Q

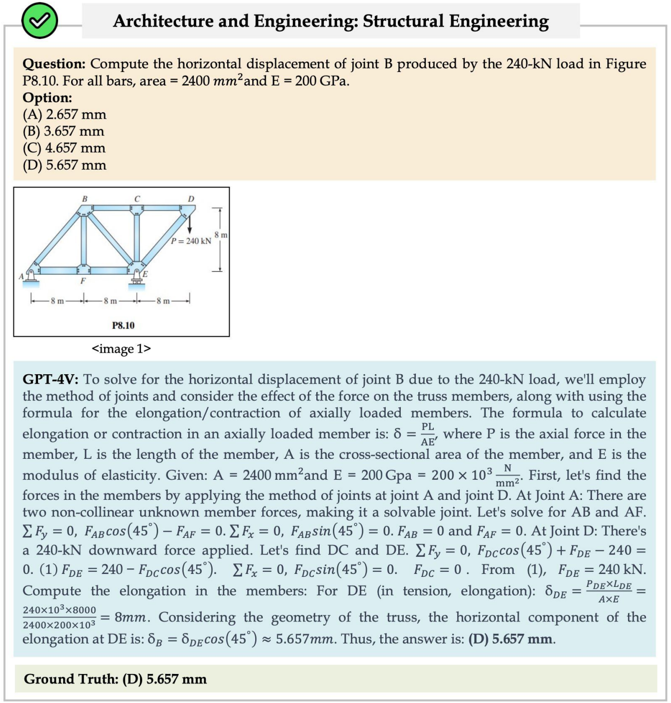
<image 1>

**GPT-4V:** The sequence in which a blood cell returning from other parts of the body passes through the four chambers of the heart is : Right atrium, Right ventricle, Left atrium, Left ventricle. Looking at the given options, the correct sequence is:(D) S-R-P-Q

**Error Category:** **Perceptual Error**
**Error Reason:** The model has the correct knowledge, but it failed to correctly understand the image and map the letters to the right structures (e.g., P is right atrium so it should be the first in the sequence).

**Ground Truth: (C) P-R-Q-S**

# Complex Reasoning



### Architecture and Engineering: Structural Engineering

**Question:** Compute the horizontal displacement of joint B produced by the 240-kN load in Figure P8.10. For all bars, area = 2400 $mm^2$ and E = 200 GPa.

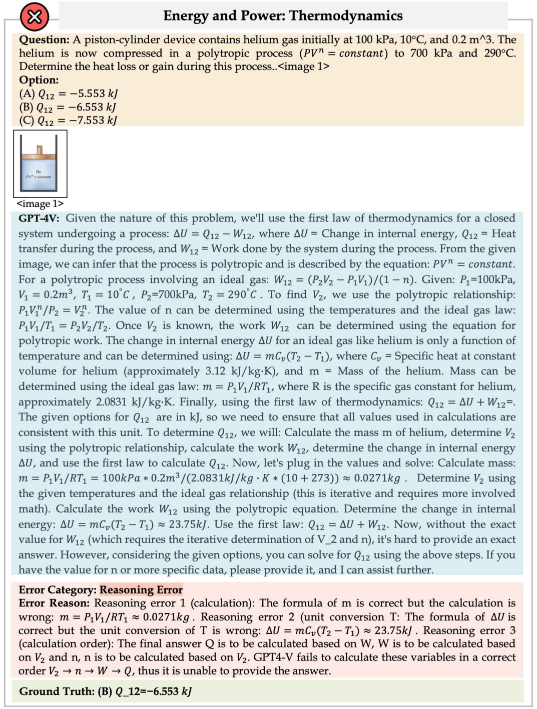
**Option:**
(A) 2.657 mm
(B) 3.657 mm
(C) 4.657 mm
(D) 5.657 mm

B
C
D
A
F
E
P = 240 kN
8 m
8 m
8 m
8 m
P8.10

<image 1>

**GPT-4V:** To solve for the horizontal displacement of joint B due to the 240-kN load, we'll employ the method of joints and consider the effect of the force on the truss members, along with using the formula for the elongation/contraction of axially loaded members. The formula to calculate elongation or contraction in an axially loaded member is: $\delta = \frac{PL}{AE}$, where P is the axial force in the member, L is the length of the member, A is the cross-sectional area of the member, and E is the modulus of elasticity. Given: $A = 2400\,mm^2$ and $E = 200\,Gpa = 200 \times 10^3 \frac{N}{mm^2}$. First, let's find the forces in the members by applying the method of joints at joint A and joint D. At Joint A: There are two non-collinear unknown member forces, making it a solvable joint. Let's solve for AB and AF. $\sum F_y = 0,\ F_{AB}cos(45^{\circ}) - F_{AF} = 0$. $\sum F_x = 0,\ F_{AB}sin(45^{\circ}) = 0$. $F_{AB} = 0$ and $F_{AF} = 0$. At Joint D: There's a 240-kN downward force applied. Let's find DC and DE. $\sum F_y = 0,\ F_{DC}cos(45^{\circ}) + F_{DE} - 240 = 0$. (1) $F_{DE} = 240 - F_{DC}cos(45^{\circ})$. $\sum F_x = 0,\ F_{DC}sin(45^{\circ}) = 0$. $F_{DC} = 0$. From (1), $F_{DE} = 240$ kN. Compute the elongation in the members: For DE (in tension, elongation): $\delta_{DE} = \frac{P_{DE} \times L_{DE}}{A \times E} = \frac{240 \times 10^3 \times 8000}{2400 \times 200 \times 10^3} = 8mm$. Considering the geometry of the truss, the horizontal component of the elongation at DE is: $\delta_B = \delta_{DE}cos(45^{\circ}) \approx 5.657mm$. Thus, the answer is: **(D) 5.657 mm.**

**Ground Truth: (D) 5.657 mm**



### Energy and Power: Thermodynamics

**Question:** A piston-cylinder device contains helium gas initially at 100 kPa, 10°C, and 0.2 m^3. The helium is now compressed in a polytropic process ($PV^n = constant$) to 700 kPa and 290°C. Determine the heat loss or gain during this process..<image 1>

**Option:**
(A) $Q_{12} = -5.553\ kJ$
(B) $Q_{12} = -6.553\ kJ$
(C) $Q_{12} = -7.553\ kJ$

He
$PV^n$ = constant

<image 1>

**GPT-4V:** Given the nature of this problem, we'll use the first law of thermodynamics for a closed system undergoing a process: $\Delta U = Q_{12} - W_{12}$, where $\Delta U$ = Change in internal energy, $Q_{12}$ = Heat transfer during the process, and $W_{12}$ = Work done by the system during the process. From the given image, we can infer that the process is polytropic and is described by the equation: $PV^n = constant$. For a polytropic process involving an ideal gas: $W_{12} = (P_2V_2 - P_1V_1)/(1-n)$. Given: $P_1$=100kPa, $V_1 = 0.2m^3$, $T_1 = 10^{\circ}C$, $P_2$=700kPa, $T_2 = 290^{\circ}C$. To find $V_2$, we use the polytropic relationship: $P_1V_1^n/P_2 = V_2^n$. The value of n can be determined using the temperatures and the ideal gas law: $P_1V_1/T_1 = P_2V_2/T_2$. Once $V_2$ is known, the work $W_{12}$ can be determined using the equation for polytropic work. The change in internal energy $\Delta U$ for an ideal gas like helium is only a function of temperature and can be determined using: $\Delta U = mC_v(T_2 - T_1)$, where $C_v$ = Specific heat at constant volume for helium (approximately 3.12 kJ/kg·K), and m = Mass of the helium. Mass can be determined using the ideal gas law: $m = P_1V_1/RT_1$, where R is the specific gas constant for helium, approximately 2.0831 kJ/kg·K. Finally, using the first law of thermodynamics: $Q_{12} = \Delta U + W_{12}$=. The given options for $Q_{12}$ are in kJ, so we need to ensure that all values used in calculations are consistent with this unit. To determine $Q_{12}$, we will: Calculate the mass m of helium, determine $V_2$ using the polytropic relationship, calculate the work $W_{12}$, determine the change in internal energy $\Delta U$, and use the first law to calculate $Q_{12}$. Now, let's plug in the values and solve: Calculate mass: $m = P_1V_1/RT_1 = 100kPa * 0.2m^3/(2.0831kJ/kg \cdot K * (10+273)) \approx 0.0271kg$. Determine $V_2$ using the given temperatures and the ideal gas relationship (this is iterative and requires more involved math). Calculate the work $W_{12}$ using the polytropic equation. Determine the change in internal energy: $\Delta U = mC_v(T_2 - T_1) \approx 23.75kJ$. Use the first law: $Q_{12} = \Delta U + W_{12}$. Now, without the exact value for $W_{12}$ (which requires the iterative determination of V_2 and n), it's hard to provide an exact answer. However, considering the given options, you can solve for $Q_{12}$ using the above steps. If you have the value for n or more specific data, please provide it, and I can assist further.

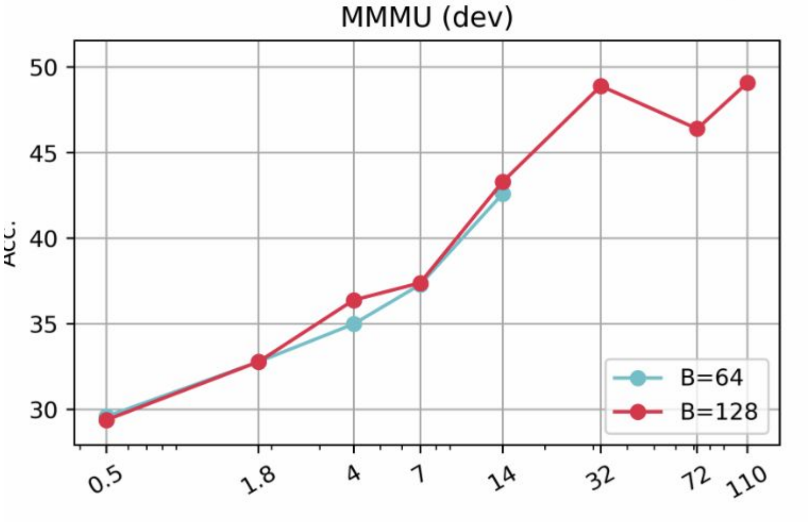
**Error Category: Reasoning Error**
**Error Reason:** Reasoning error 1 (calculation): The formula of m is correct but the calculation is wrong: $m = P_1V_1/RT_1 \approx 0.0271kg$. Reasoning error 2 (unit conversion T: The formula of $\Delta U$ is correct but the unit conversion of T is wrong: $\Delta U = mC_v(T_2 - T_1) \approx 23.75kJ$. Reasoning error 3 (calculation order): The final answer Q is to be calculated based on W, W is to be calculated based on $V_2$ and n, n is to be calculated based on $V_2$. GPT4-V fails to calculate these variables in a correct order $V_2 \rightarrow n \rightarrow W \rightarrow Q$, thus it is unable to provide the answer.

**Ground Truth: (B) $Q\_12 = -6.553\ kJ$**

GPT-4V demonstrates moments of excellence in complex reasoning but tends to falter with extended reasoning chains or excessive calculations.

# How to improve the model's reasoning capability?

- Larger and more capable LLM



LLaVA NEXT with Different Sizes of LLMs

https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/

| Models | Image Encoders | LLMs | MMMU (test) |
|---|---|---|---|
| VILA1.5-3B | SigLIP-so400 | Sheared-LLaMA-2.7B | 30.8 |
| VILA1.5-8B | SigLIP-so400 | Llama3-8B | 36.0 |
| VILA1.5-13B | SigLIP-so400 | Vicuna-13B | 33.6 |
| VILA1.5-40B | InternViT | Yi-34B | 46.9 |

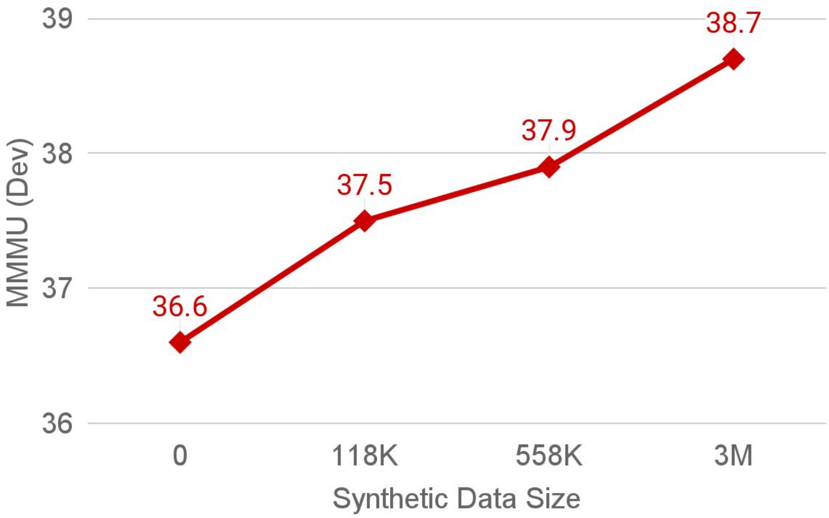
VILA 1.5 with Different Image Encoders and LLMs

https://github.com/NVlabs/VILA

The MMMU performance shows a clear **scaling effect** with the **LLM size**

# How to improve the model's reasoning capability?

- High-quality (synthetic) data



(Figure is re-plotted based on a LLaVA 1.6 ablation result)

- Synthetic captions by LLaVA-NEXT-34B
- Images are from COCO118K, BLIP558K, and CC3M
- Fine-tune a 7B model on the synthetic data

The MMMU performance shows a clear **scaling effect** with an increased number of **high-quality (synthetic) data**.

# How to improve the model’s reasoning capability?

- Image Encoders, Resolution, Visual Tokens

| Vision Encoder | Encoder size | Visual Tokens | Res. (px) | MMMU (dev) |
|---|---|---|---|---|
| CLIP-L | 0.3B | 256 | 448 | **38.2** |
| EVA-02-E | 4.7B | 256 | 448 | 33.6 |
| EVA-8B | 8B | 256 | 448 | 35.0 |

(Numbers are copied from LLaVA 1.6 ablation result)

| Vision Encoder | Encoder size | Resolution (px) | MMMU (dev) |
|---|---|---|---|
| CLIP-L | 0.3B | 336 | 38.1 |
| SigLIP | 0.4B | 384 | 40.1 |
| Idefics 2 | 0.4B | Up to 980 | **41.1** |

(Numbers are copied from the MANTIS paper)

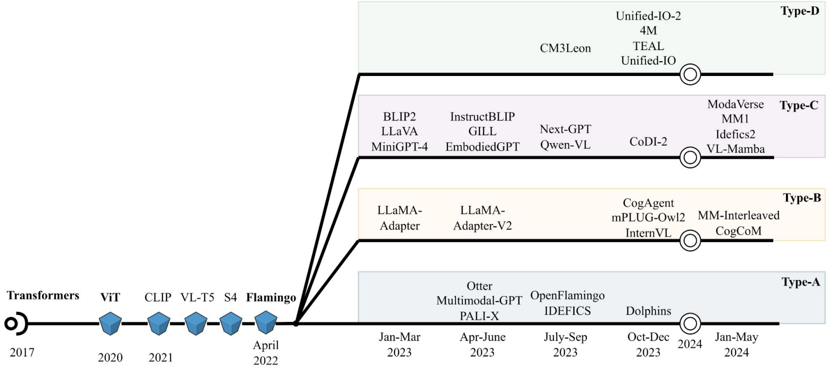
- Good vision encoders, utilizing better pretrained data and higher resolution, improve performance
- Scaling vision encoders has a minimal impact
- Increasing visual tokens offers little benefit

| Vision Encoder | Model size | Visual Tokens | MMMU (dev/test) |
|---|---|---|---|
| Idefics 2 | 0.4B | 64 | **43.5/37.9** |
| Idefics 2 | 0.4B | 320 | 43.0/37.7 |

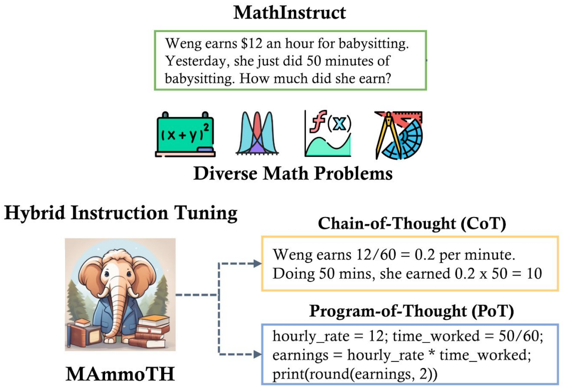
(Numbers are copied from the IDEFICS2 paper)

# How to improve the model’s reasoning capability?

- Different VLM architectures
  - Better vision and language alignment
- Augment multimodal models with tools
  - Combine natural language with code for reasoning



Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." arXiv preprint arXiv:2405.17927 (2024).



Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Weng earns 12/60 = 0.2 per minute. Doing 50 mins, she earned 0.2 x 50 = 10

```
hourly_rate = 12; time_worked = 50/60;
earnings = hourly_rate * time_worked;
print(round(earnings, 2))
```

Yue, Xiang, et al. "MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning." ICLR 2024.

# Conclusion

- The MMMU benchmark has proven to be an essential tool for evaluating and guiding the development of multimodal models, providing critical insights into model strengths and weaknesses.
- Despite progress, models still face significant challenges, particularly in domain-specific visual perception, grounding, and handling complex reasoning tasks.
- We shall caution that MMMU is ***not a sufficient test*** for Expert AGI. However, we believe it should be ***necessary*** for an Expert AGI to achieve strong performance on MMMU to demonstrate their broad and deep subject knowledge as well as expert-level understanding and reasoning capabilities.
- Future efforts should aim at closing the gap between multimodal models and human experts, working towards expert AGI.

# Acknowledgement

- We would like to thank all the coauthors from six institutes and over 50 volunteer annotators who collectively spent more than 1,000 hours to make this project possible!
- We are grateful to the model developers and the community, especially the authors of LLaVA, VILA, InternVL, and Yi-VL, for sharing their model diagnostic experience on MMMU.
- We appreciate all the developers who used MMMU and reported issues, helping us fix small problems and improve the benchmark.

IN.AI Research

UNIVERSITY OF WATERLOO

THE OHIO STATE UNIVERSITY

Carnegie Mellon University

University of Victoria

PRINCETON UNIVERSITY

# Thank you!

Questions?

Homepage &
Leaderboard

Paper

Dataset

Contact: xyue2@andrew.cmu.edu