Benchmarking and Advancing Reasoning Capabilities in Foundation Models

Speaker: Wenhu Chen

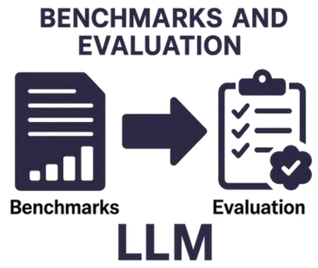
Brief Summary of Myself

- Graduated from PhD in 2021
- 2021 – 2022:
  - Building Multimodal RAG Models at Google Brain
- 2022 – early 2025:
  - 20% Part-time at Google Gemini for Image Generation and Evaluation.
- 2022 – Present:
  - Lead the TIGER-Lab at University of Waterloo

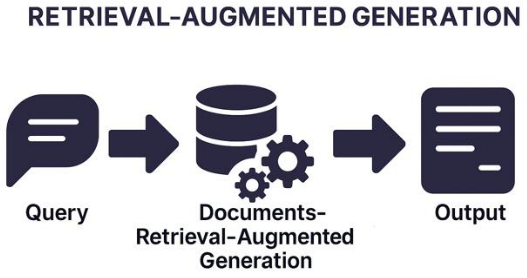
TIGER-Lab

- Text-and-Image GEneration Research

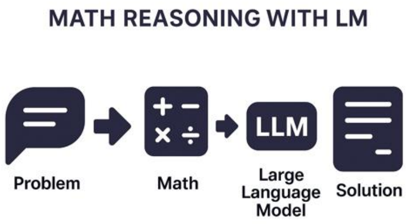
Evaluation:
MMMU, MMLU-Pro
MEGA-Bench



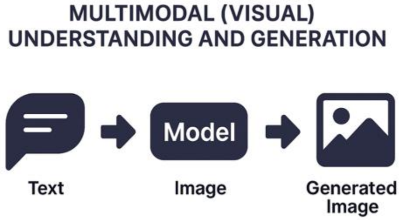
RAG:
UniIR, LM2Vec,
LongRAG



Reasoning:
MAmmoTH v1/v2,
General-Reasoner v1/v2



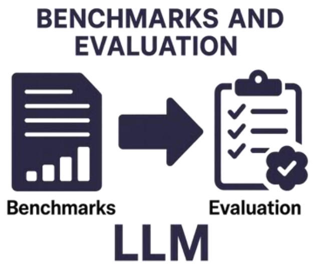
Multimodal:
SuTI, T2V-Turbo,
OmniEdit



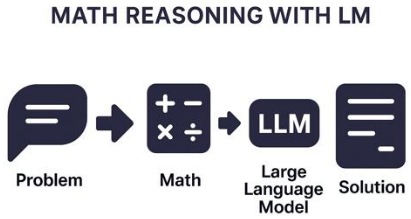
Talk Outline

- The talk outline for today:

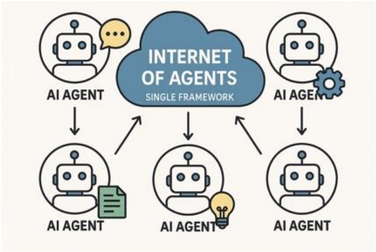
Evaluation:
MMMU, MMLU-Pro
MEGA-Bench



Reasoning:
MAmmoTH v1/v2,
General-Reasoner v1/v2

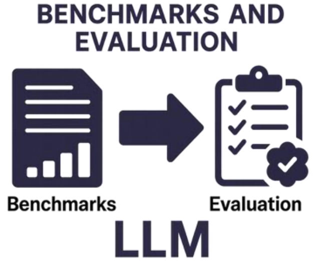


Vision:
Building Internet for AI



Section 1: Evaluation

Evaluation:
MMMU, MMLU-Pro
MEGA-Bench

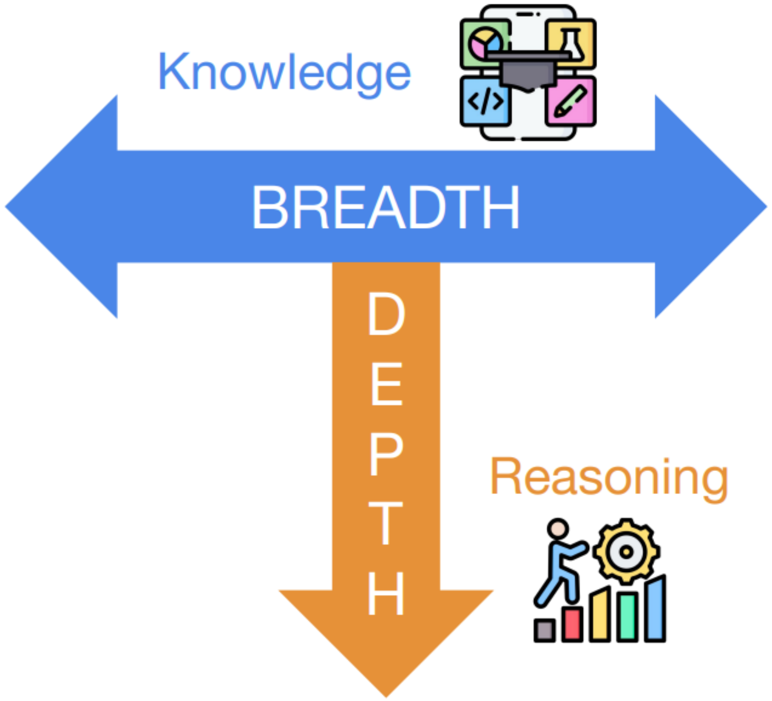


VL Benchmark: MMMU
LLM Benchmark: MMLU-Pro

Key Aspects in Expert-Level Benchmarks

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

?

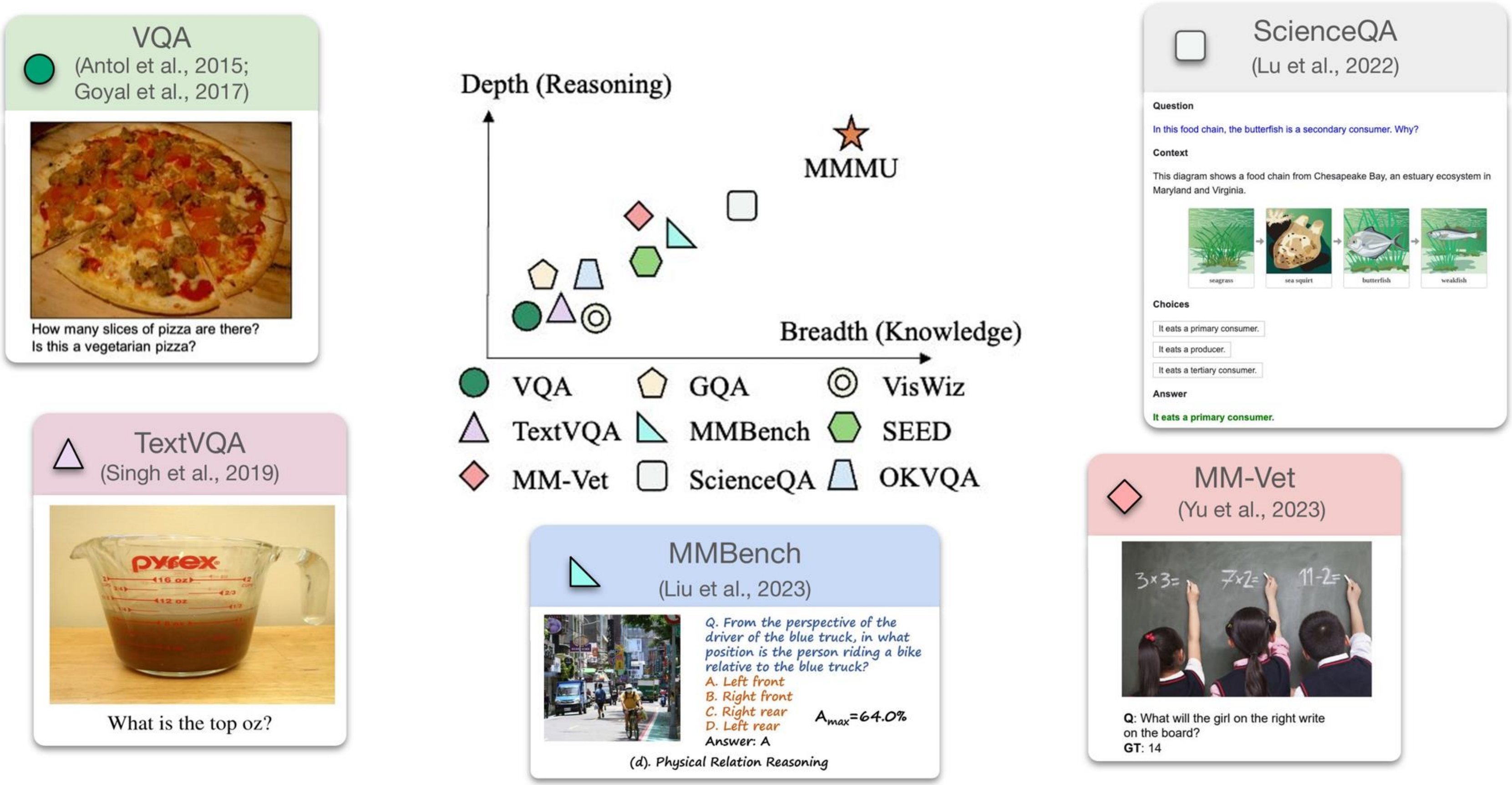


MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

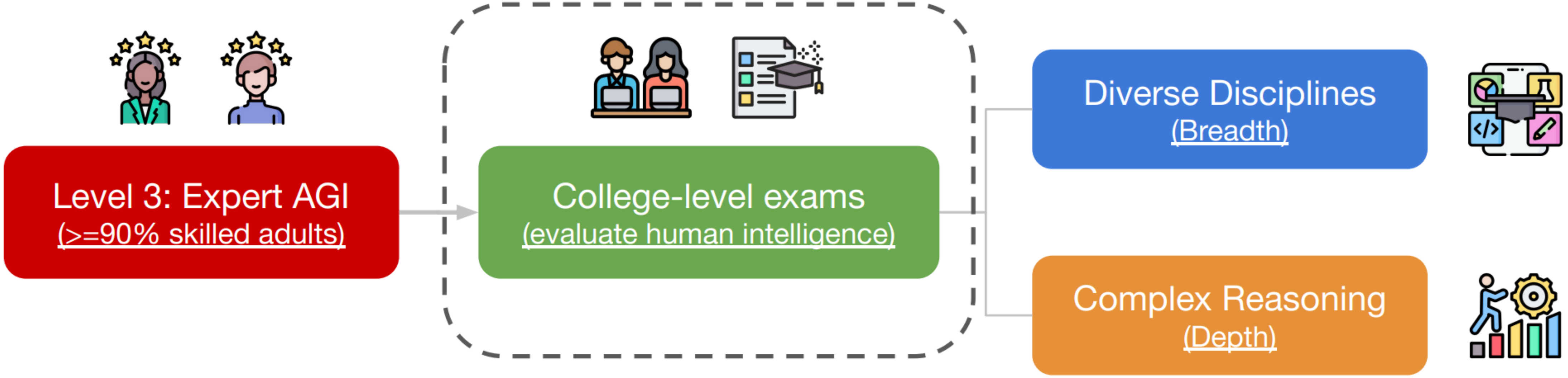
Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, Wenhu Chen

[CVPR 2024 Best Paper Finalist]

Existing VL Benchmarks (as of Oct 2023)

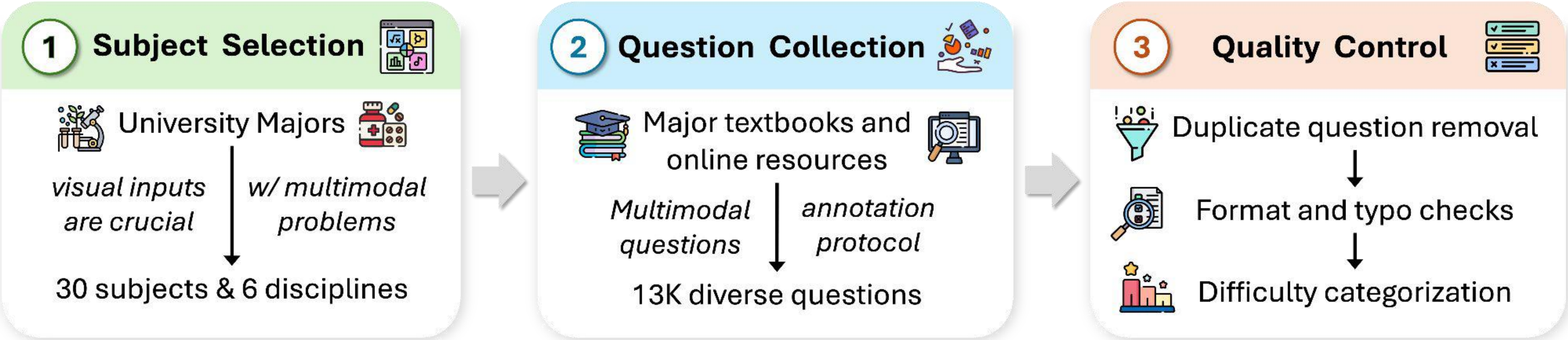


Measuring Expert AGI



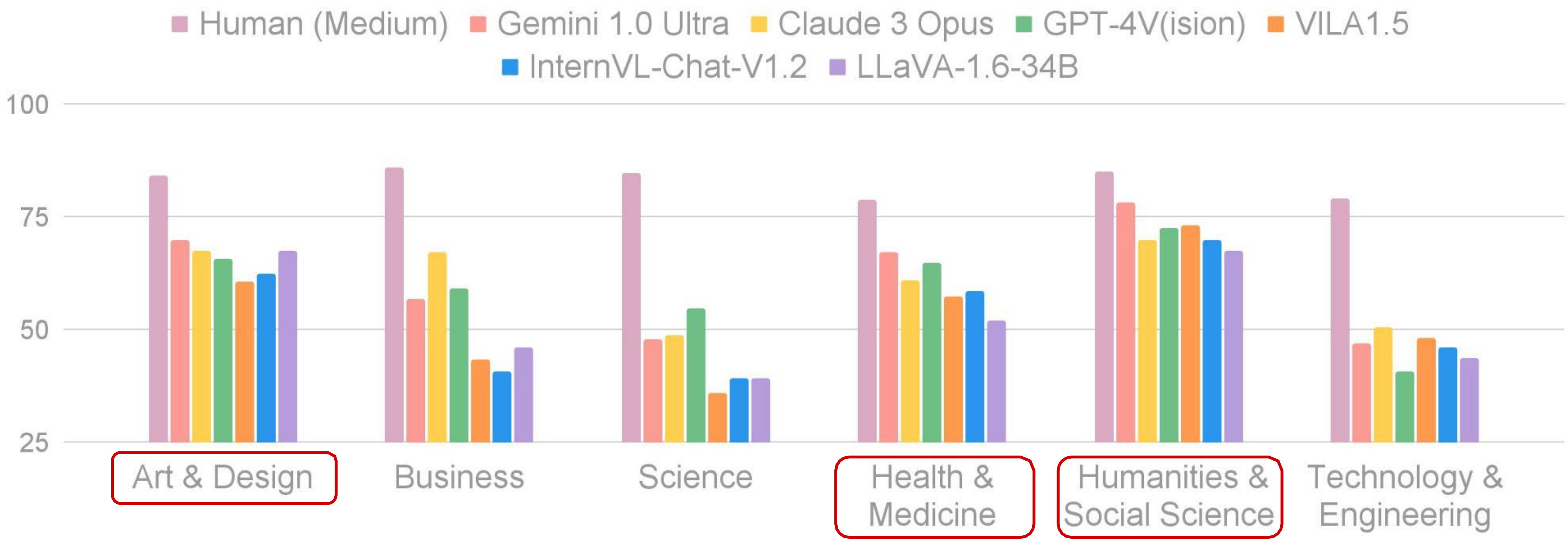
1) Rigorous Data Curation Process

Curation Pipeline



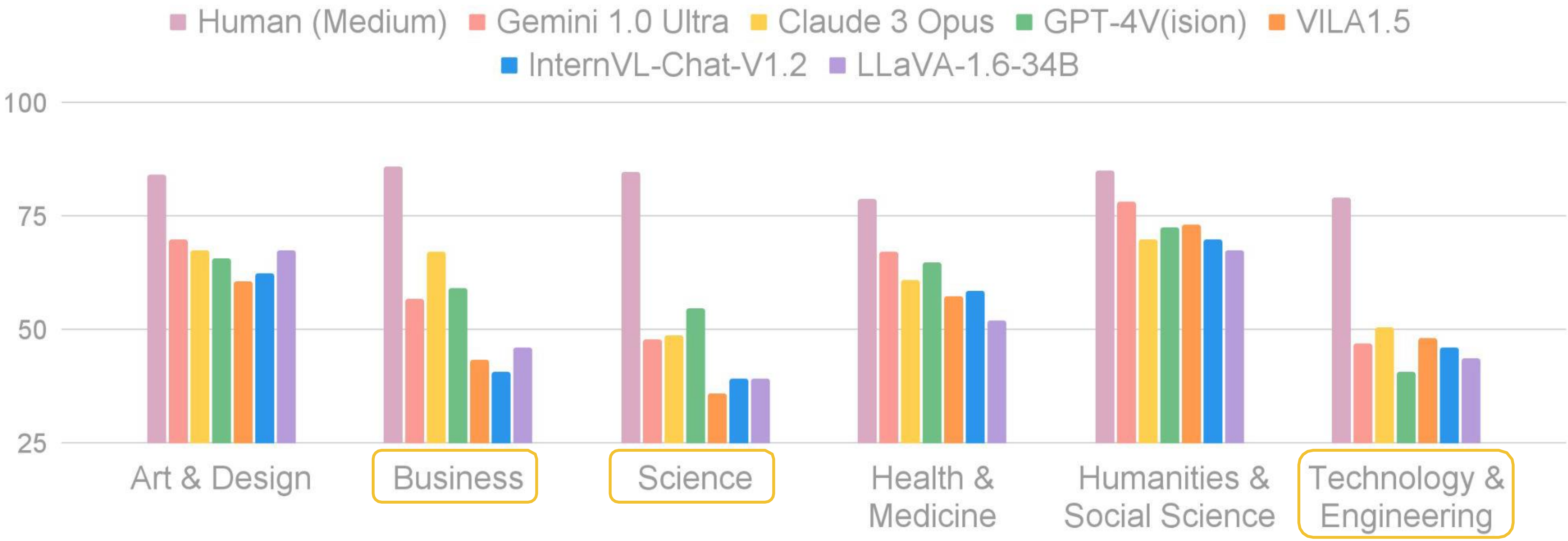
2) Model Diagnosis Tool

Subject-Specific Accuracy



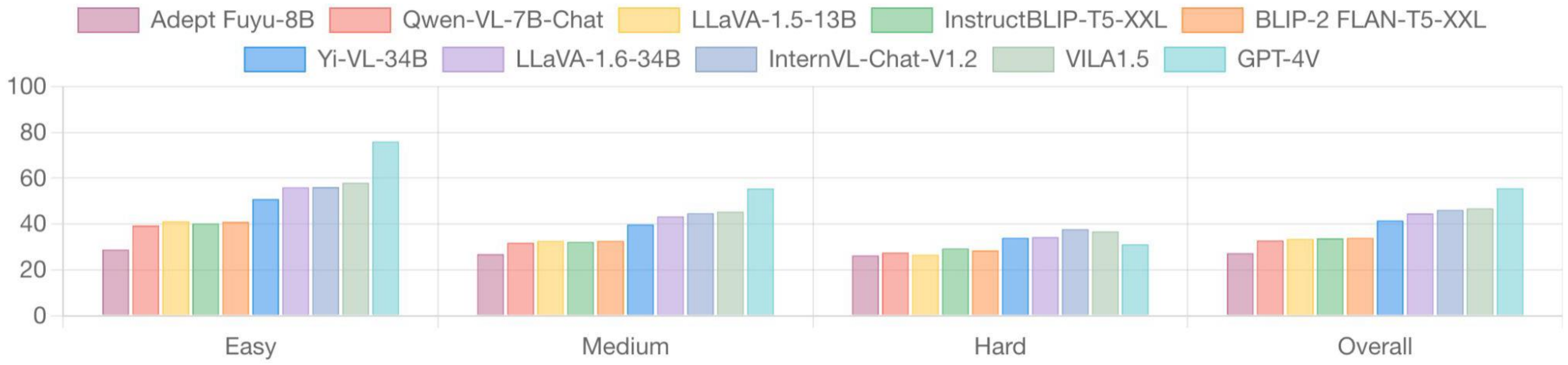
- The gap between the best models and human experts is not large.
- The difference between open-source and proprietary models is not significant.

Subject-Specific Accuracy



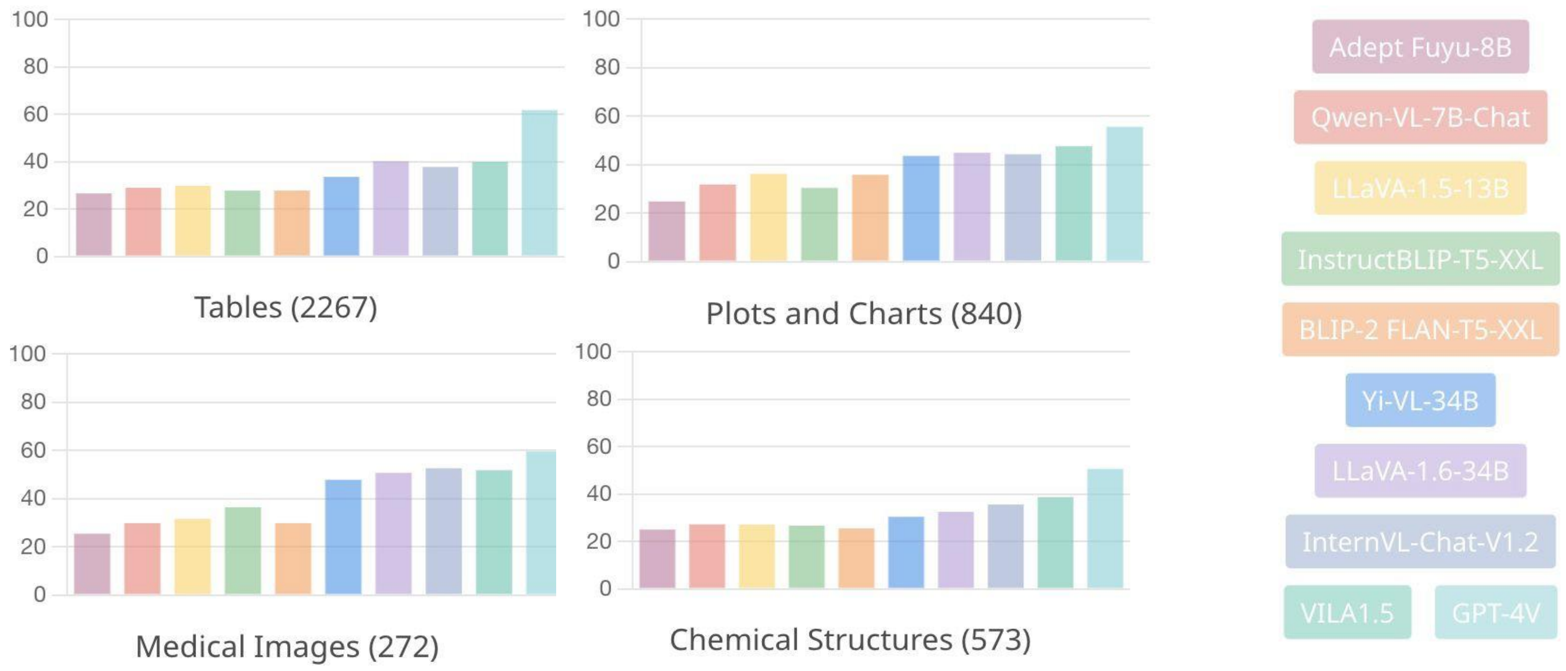
- The gap between the best models and human experts is significantly large.
- Models struggle with these subjects, which involve more complex reasoning questions

Difficulty-Specific Accuracy



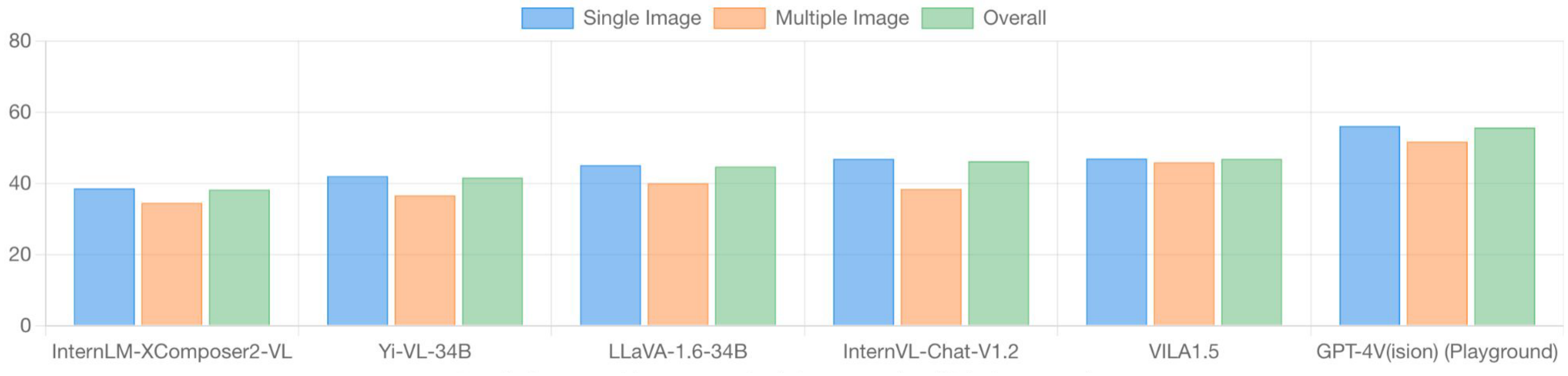
GPT-4V outperforms open-source models on easy and medium-level tasks, while all models struggle with hard examples.

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

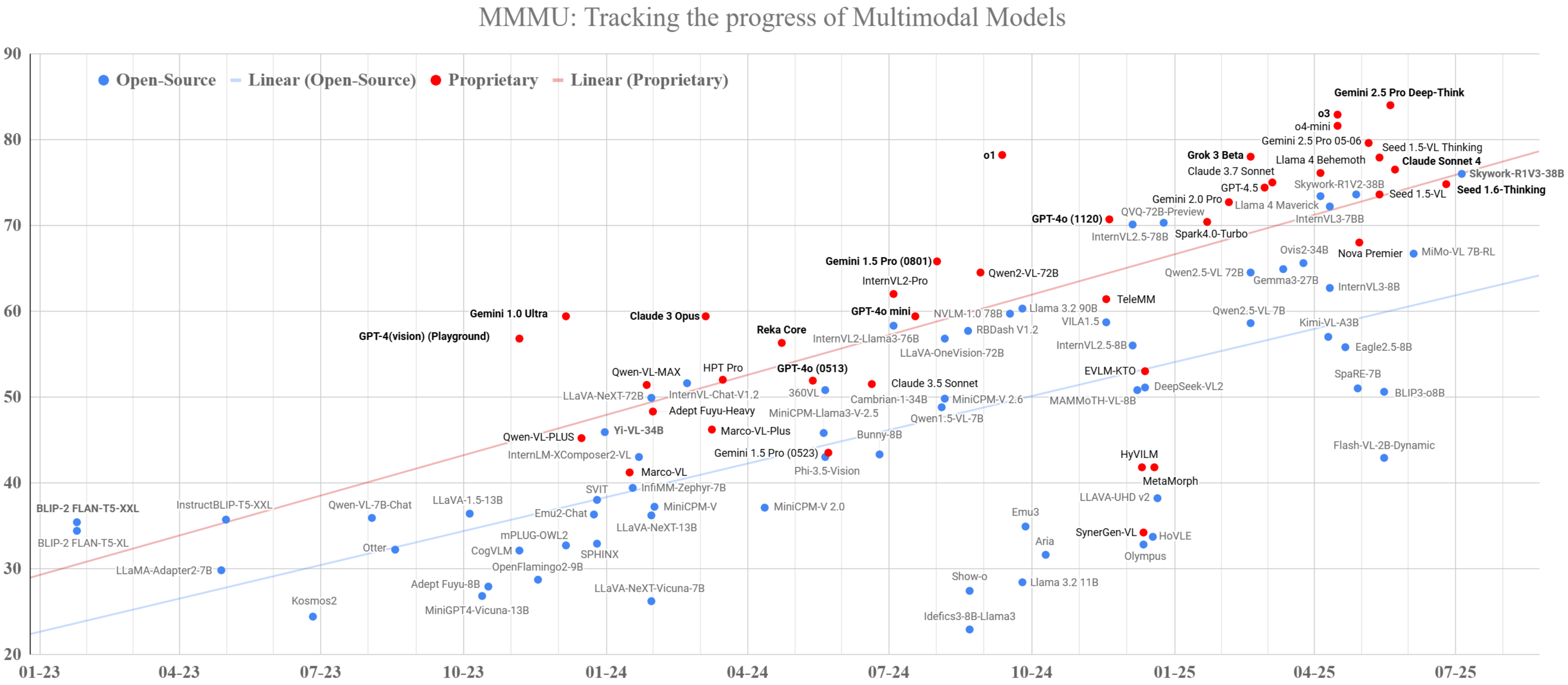


○ Models generally struggle with reasoning over multiple images

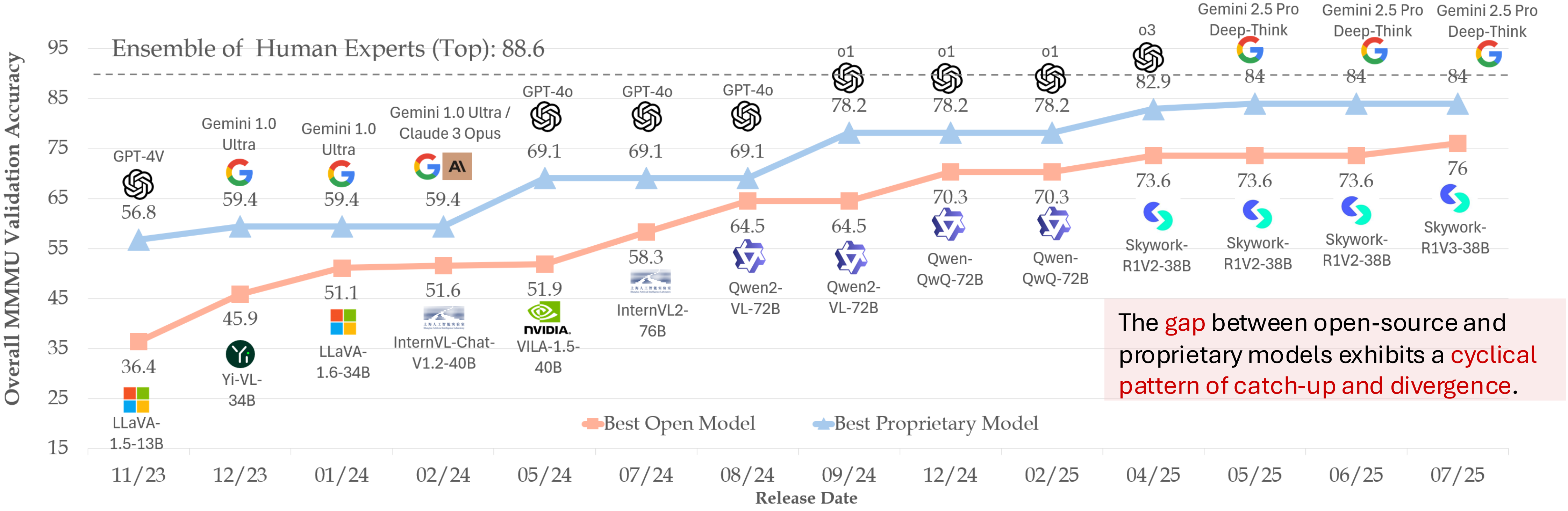
○ VILA performs notably better in this area

3) Comprehensive Evaluation

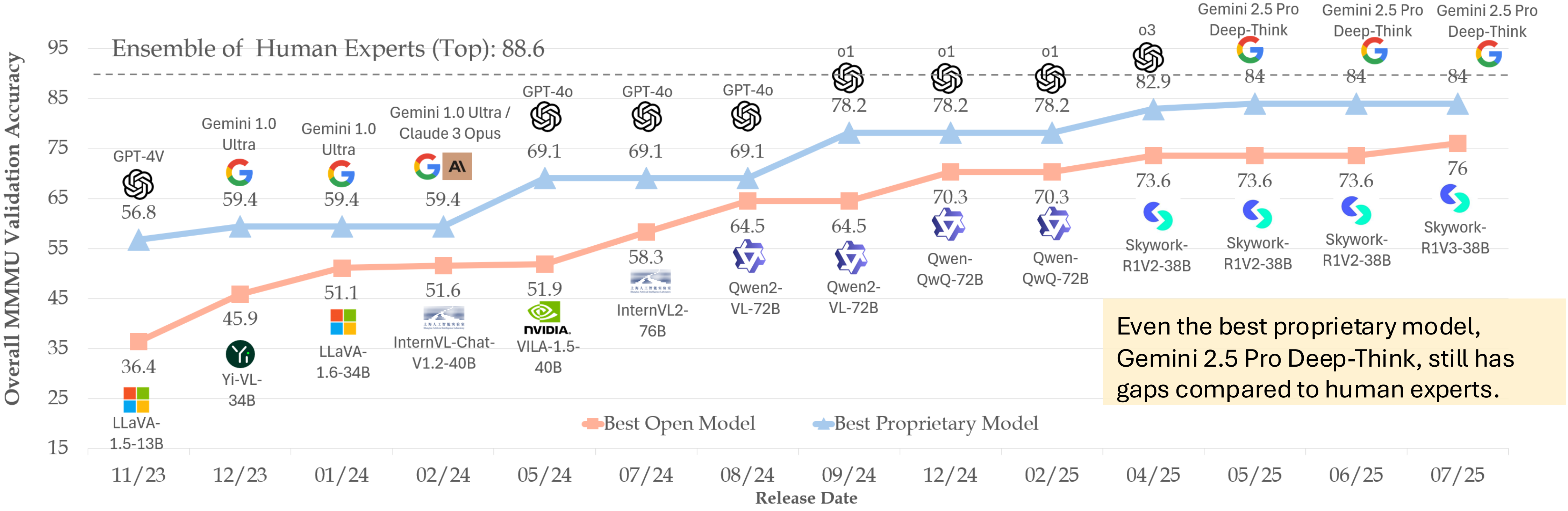
The Progress on MMMU



Open-Source VS. Proprietary



Open-Source VS. Proprietary



Even the best proprietary model, Gemini 2.5 Pro Deep-Think, still has gaps compared to human experts.

MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, Wenhu Chen

[NeurIPS 2024 Spotlight]

Existing LLM Benchmarks (as of March 2024)

MMLU

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23
Answer: B

Compute $i + i^2 + i^3 + \cdots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) i (D) $-i$
Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
(A) 28 (B) 21 (C) 40 (D) 30
Answer: C

Knowledge Intensive Benchmark

MATH

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?
Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: If $\sum_{n=0}^{\infty} \cos^{2n} \theta = 5$, what is $\cos 2\theta$?
Solution: This geometric series is $1 + \cos^2 \theta + \cos^4 \theta + \cdots = \frac{1}{1 - \cos^2 \theta} = 5$. Hence, $\cos^2 \theta = \frac{4}{5}$. Then $\cos 2\theta = 2\cos^2 \theta - 1 = \boxed{\frac{3}{5}}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.
Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}} \sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}} \sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right) \sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right) \sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1 + \cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1 - \sqrt{2}}{2}}$.

Math Reasoning

DROP

Reasoning	Passage (some parts shortened)	Question	Answer	BiDAF
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992

Common Sense Reasoning

Limitations of MMLU

1. **Performance saturation** (90%+) on MMLU limits differentiation between advanced models

2. **Knowledge-focused questions** with 4 options enable shortcut exploitation rather than understanding

3. **Dataset noise** creates artificial performance ceiling, reducing benchmark effectiveness

Dataset Construction

The dataset consolidates questions from several sources:

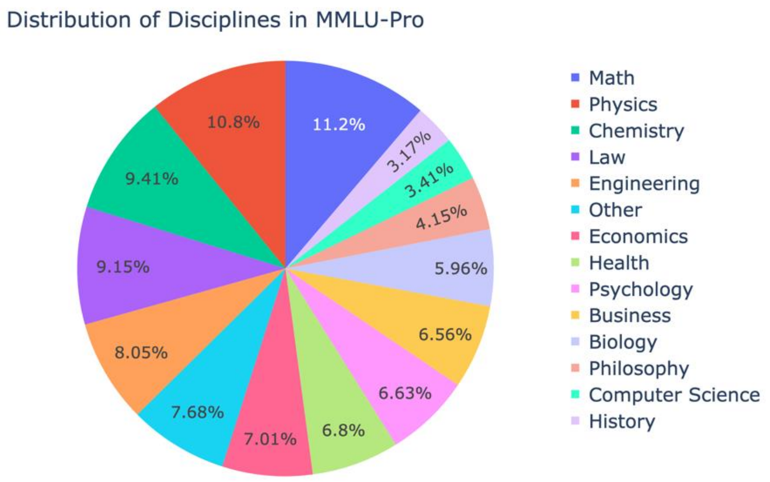
- **Original MMLU Questions:** Part of the dataset comes from the original MMLU dataset. We remove the trivial/ambiguous queries.

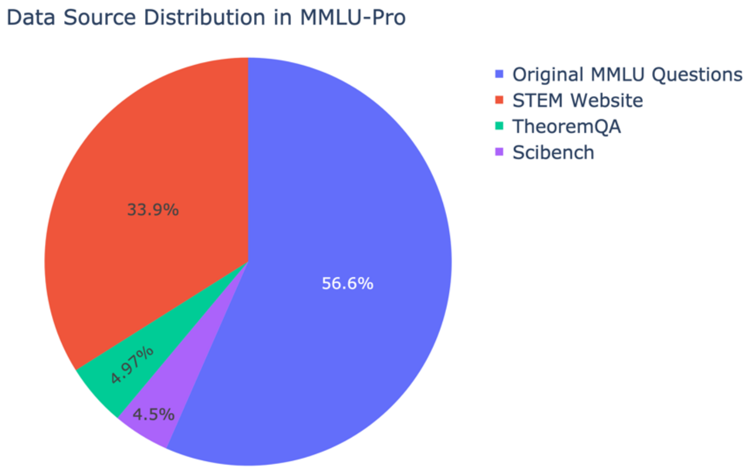
- **STEM & Non-STEM Website:** Hand-picking high-quality STEM problems from the Internet to augment the evaluation set.

- **Expanded answer choices from 4 to 10 options**, reducing random guess probability from 25% to 10%

Data Distribution

Discipline	Number of Questions	From Original MMLU	Newly Added
Math	1351	846	505
Physics	1299	411	888
Chemistry	1132	178	954
Law	1101	1101	0
Engineering	969	67	902
Other	924	924	0
Economics	844	444	400
Health	818	818	0
Psychology	798	493	305
Business	789	155	634
Biology	717	219	498
Philosophy	499	499	0
Computer Science	410	274	136
History	381	381	0
Total	**12032**	6810	5222

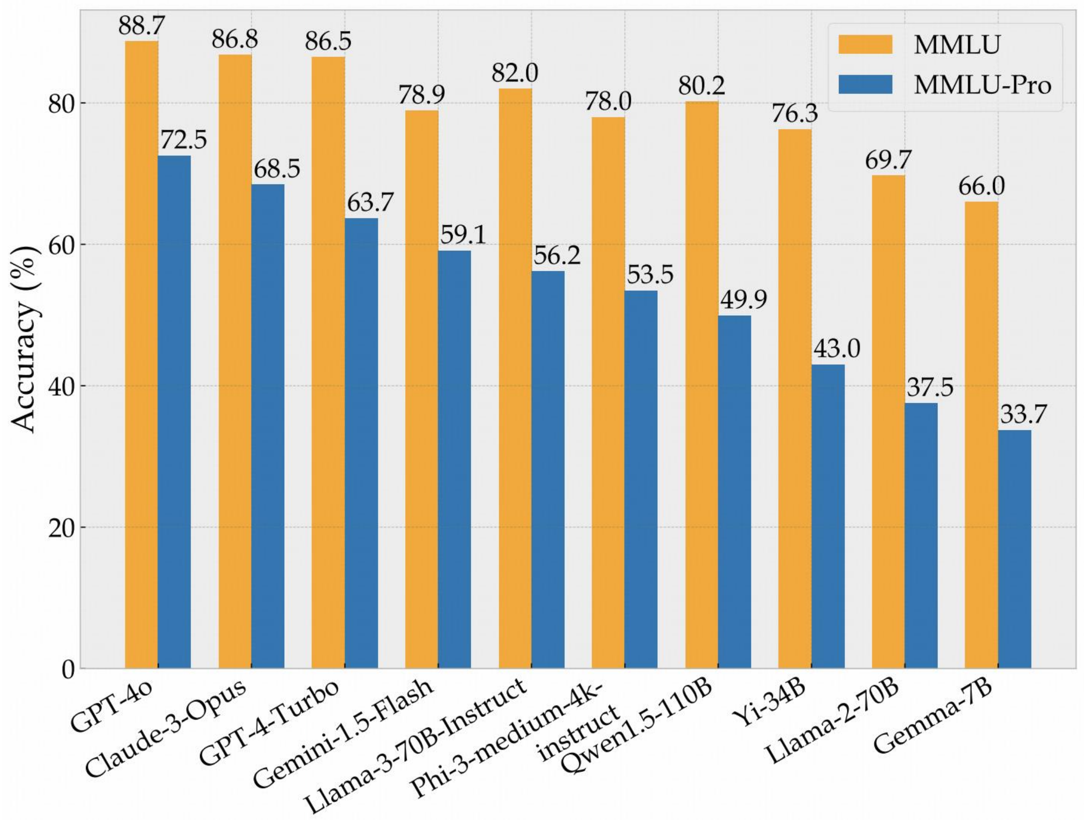




Analysis 1: Difficulty Level

MMLU vs MMLU-Pro Model Performance Analysis

- MMLU is Saturated
- Better Differentiation
- Room for Improvement



Analysis 2: Reasoning Level

Model Name	MMLU			MMLU-Pro		
	CoT	Direct Answer	CoT - DA	CoT	Direct Answer	CoT - DA
GPT-4o	88.7	87.2	**1.5**	72.6	53.5	**19.1**
GPT-4-Turbo	86.5	86.7	**-0.2**	63.7	48.4	**15.3**
Phi3-medium-4k-instruct	79.4	78.0	**1.4**	55.7	47.5	**8.2**
Llama-3-8B	62.7	66.6	**-3.9**	35.4	31.5	**3.9**
Gemma-7B	62.4	66.0	**-3.6**	33.7	27.0	**6.7**

CoT vs Direct Answering: Performance Analysis

- Overall Performance Trend
- Model-Specific Improvements

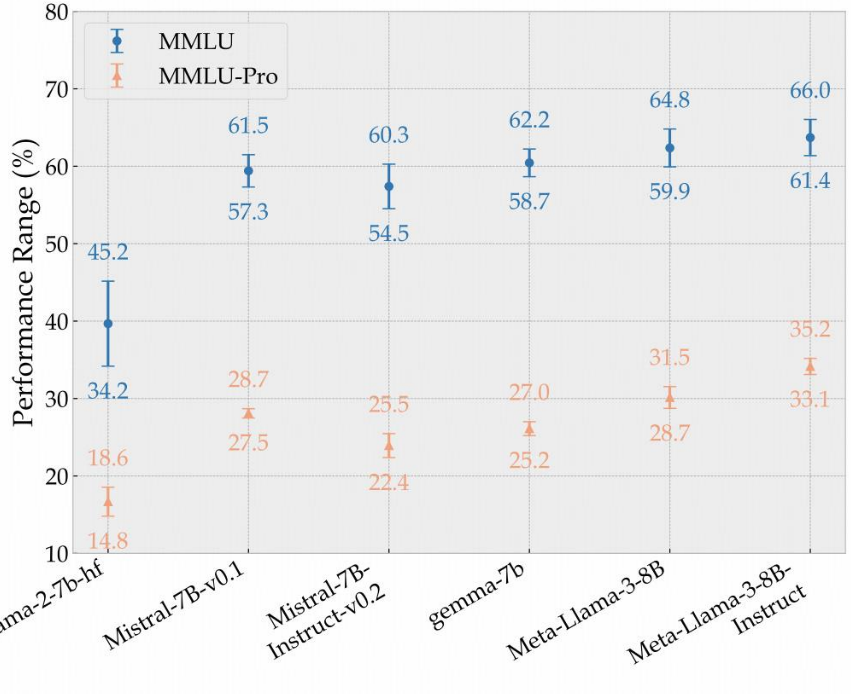
Analysis 3: Robustness Degree

- Tested using 24 different reasonable prompts

- Benchmark Comparison

  MMLU:
  - General variation: 4-5%
  - Maximum variation: 10.98%

  MMLU-Pro:
  - General variation: ~2%
  - Maximum variation: 3.74%



Performance Variability under Different Prompts on MMLU and MMLU-Pro

Error Analysis: GPT-4o

- Methodology
  - Analysis of 120 randomly selected errors
  - Evaluated by expert annotators
- Reasoning Errors: 39%
  - Logical inconsistencies
  - Pattern recognition vs true understanding
- Knowledge Gaps: 35%
  - Lack of specialized domain knowledge
  - Issues with technical applications
- Calculation Errors: 12%
  - Correct formulas but wrong computations

Impact of MMMU and MMLU-Pro

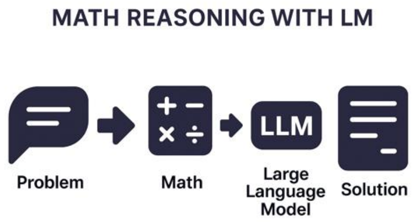
TITLE	CITED BY	YEAR
MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI X Yue, Y Ni, K Zhang, T Zheng, R Liu, G Zhang, S Stevens, D Jiang, ... CVPR 2024	1167	2023
MMLU-Pro: A more robust and challenging multi-task language understanding benchmark Y Wang, X Ma, G Zhang, Y Ni, A Chandra, S Guo, W Ren, A Arulraj, X He, ... NeurIPS 2024 (Spotlight)	642	2024

Adoption

Citations

Section 2: Reasoning

Reasoning:
MAmmoTH v1/v2,
General-Reasoner v1/v2



SFT Reasoning: MAmmoTH2
RL Reasoning: General-Reasoner

MAmmoTH2: Scaling Instructions from the Web

Xiang Yue, Tuney Zheng, Ge Zhang, Wenhu Chen

[NeurIPS 2024]

Instruction Tuning as Alignment

- A popular view claims that the instruction tuning is only for aligning the model.

- Less is More: we can simply adopt a small dataset as few as 3K examples to align LLMs to downstream tasks.

- Common Beliefs: Instruction Tuning cannot improve models' general capabilities.

Exiting Datasets (as of Feb 2024)

Dataset	#Pairs	Domain	Format	Dataset Source
FLAN V2 (Chung et al., 2024)	100K	General	SFT	NLP data + Human CoT
Self-Instruct (Wang et al., 2023b)	82K	General	SFT	Generated by GPT3
GPT4-Alpaca (Taori et al., 2023)	52K	General	SFT	Generated by GPT4
SuperNI (Wang et al., 2022)	96K	General	SFT	NLP Datasets
Tora (Gou et al., 2023)	16K	Math	SFT	GSM+MATH Synthesis by GPT4
WizardMath (Luo et al., 2023)	96K	Math	SFT	GSM+MATH Synthesis by GPT4
MathInstruct (Yue et al., 2023b)	262K	Math	SFT	Math datasets Synthesis by GPT4
MetaMathQA (Yu et al., 2023)	395K	Math	SFT	GSM+MATH Synthesis by GPT3.5
XwinMath (Li et al., 2024a)	1.4M	Math	SFT	GSM+MATH Synthesis by GPT4
OpenMathInstruct (Toshniwal et al., 2024)	1.8M	Math	SFT	GSM+MATH Synthesis by Mixtral

- Diversity is low: it's mostly math only or compiled by several human-annotated ones.
- Scale is also low: the largest ones are around 1M, which are totally synthesized.

Can We Scale Up Instruction Tuning?

- We emphasize both quality and quantity.

- Previous work adopts:
  - Human labels
  - LLM Synthesis

- How to ensure quality and quantity?
  - Mine existing instruction pairs from the web.

Natural Instruction on the Web

- Available Resources: Forums, Educational Website, Quiz

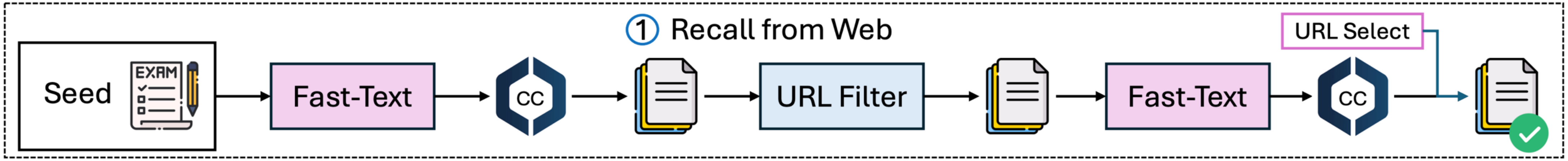
How to mine them?

- Highly dispersed across the web.
- Containing lots of unrelated information.
- Missing lots of useful information, with incomplete answers.

Pipeline

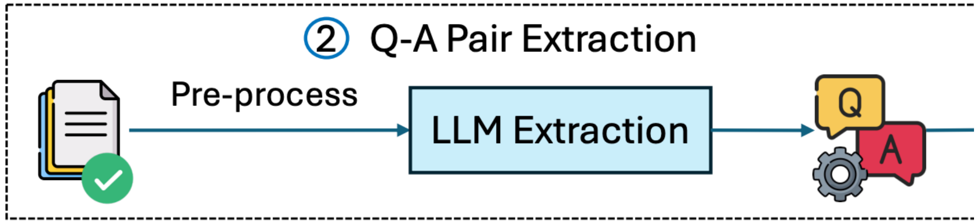
- Efficient classifier-based Recall:
  - Mine the useful URLs from Common Crawl.
  - Group and identify most useful domains.
- Web information Extraction:
  - Customized content extraction from raw web page.
- Information Completion:
  - Refine the extracted information from the web with LLMs.

1. Recall Step



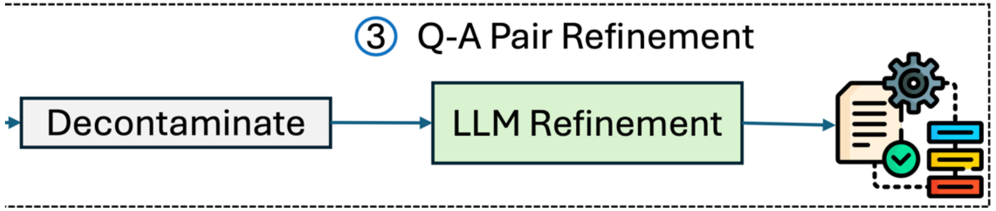
1. Identifying the seed data we desire.
2. Curating enough seed data for classifier training.
3. Train an initial version of classifier.
4. Recall from the web.
5. Group URLs based on domain and then use GPT-3.5 to select the most useful domains.
6. Gather the web page from the useful domains and re-train classifier with larger data.
7. Recall again from the web.
8. Group URLs based on domain and then use GPT-3.5 to select the final URL domains.

2. Extraction Step



1. Customized rule-based web parsing for the top domains.
2. Further utilize LLMs to select the span for instruction and response.

3. Refine Step



1. Utilizing LLMs like Mixtral-22B and Qwen-72B for refining
2. Refine the content format and remove noise.
3. Complete the response if it’s missing, especially in educational websites.

Pipeline Example

Raw Docs *Unformatted Text, Site Information, Ads*

Topics Science\nAnatomy&Physiology\nAstronomy\nAstrophysics \nBiology\nChemistry \n...Socratic Meta...Featured Answers How do you simplify #((u^4v^3)/(u^2v^-1)^4)^0# and write it using only positive exponents? Answer by NickTheTurtle (Apr 1, 2017) Explanation:\nAnything raised to the \(0^{th}\) power is simply 1. \n\nRelated Questions\n What is the quotient of powers property? \n\n\n\n How do you simplify expressions using the quotient rule?...\nImpact of this question\n1274 views around the world #Apps\niOS\nAndroid\nLinks\n[Privacy](#)\n[Terms](#)\n[Help](#)

Extracted QA *Formatted QA but lacking detailed solutions*

Question: How do you simplify $(u^4\, v^3/(u^2\, v^{-1}\,)^4)^0$ and write it using only positive exponents?
Answer: Explanation: Anything to the $0th$ power is just simply 1.

Rewritten QA *Formatted QA augmented with detailed solutions*

Question:
How do you simplify $(u^4\, v^3/(u^2\, v^{-1}\,)^4)^0$and write it using only positive exponents?

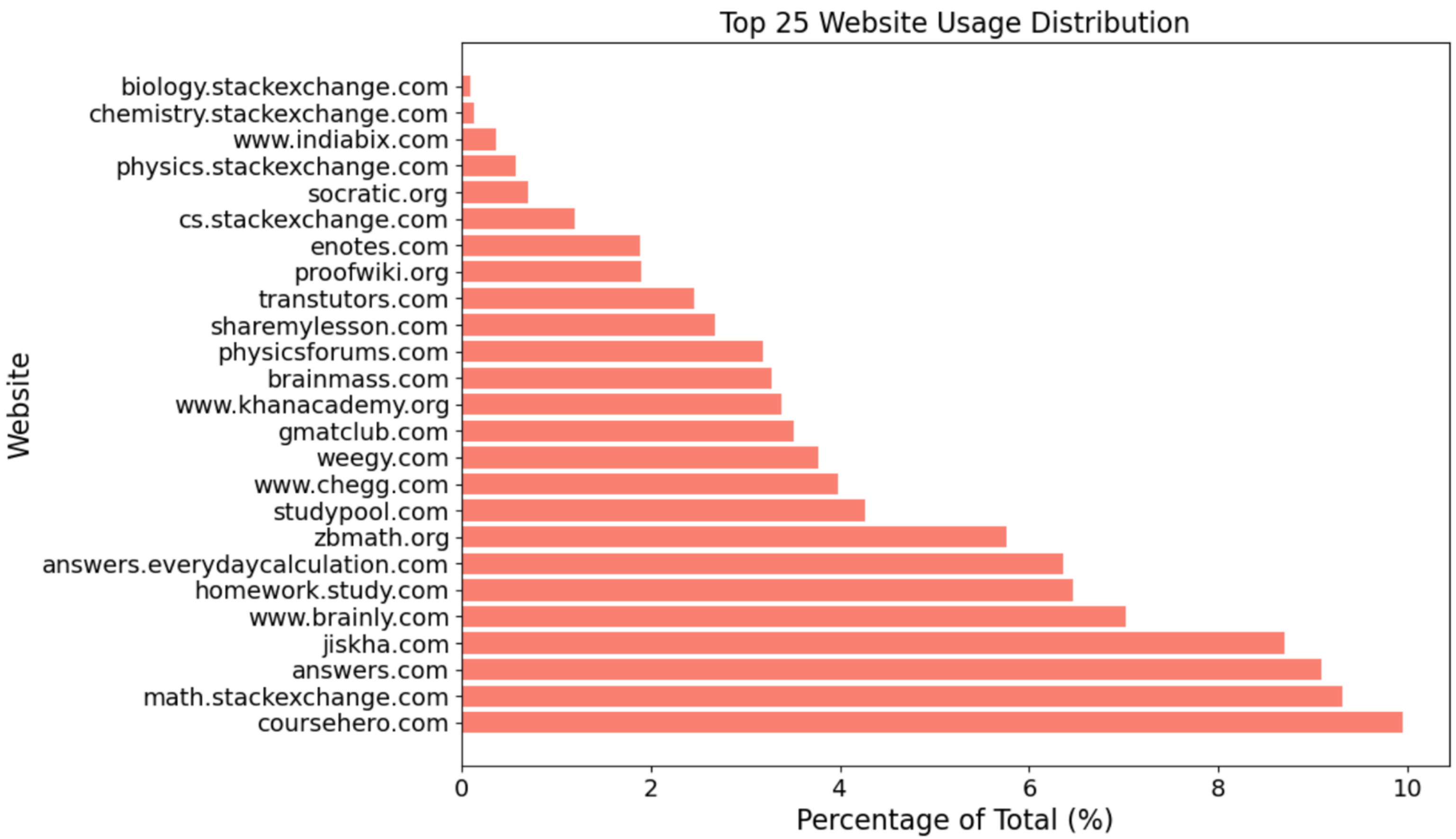
Answer:
To simplify the expression $(u^4\, v^3/(u^2\, v^{-1}\,)^4)^0$ and rewrite it using only positive exponents, we start by evaluating the expression from the innermost operation outward.

First, consider the exponent of zero on the entire fraction:
Any expression raised to the power of zero is equal to 1 (provided the base is not zero). *Therefore*, $(u^4\, v^3/(u^2\, v^{-1}\,)^4)^0 = 1$

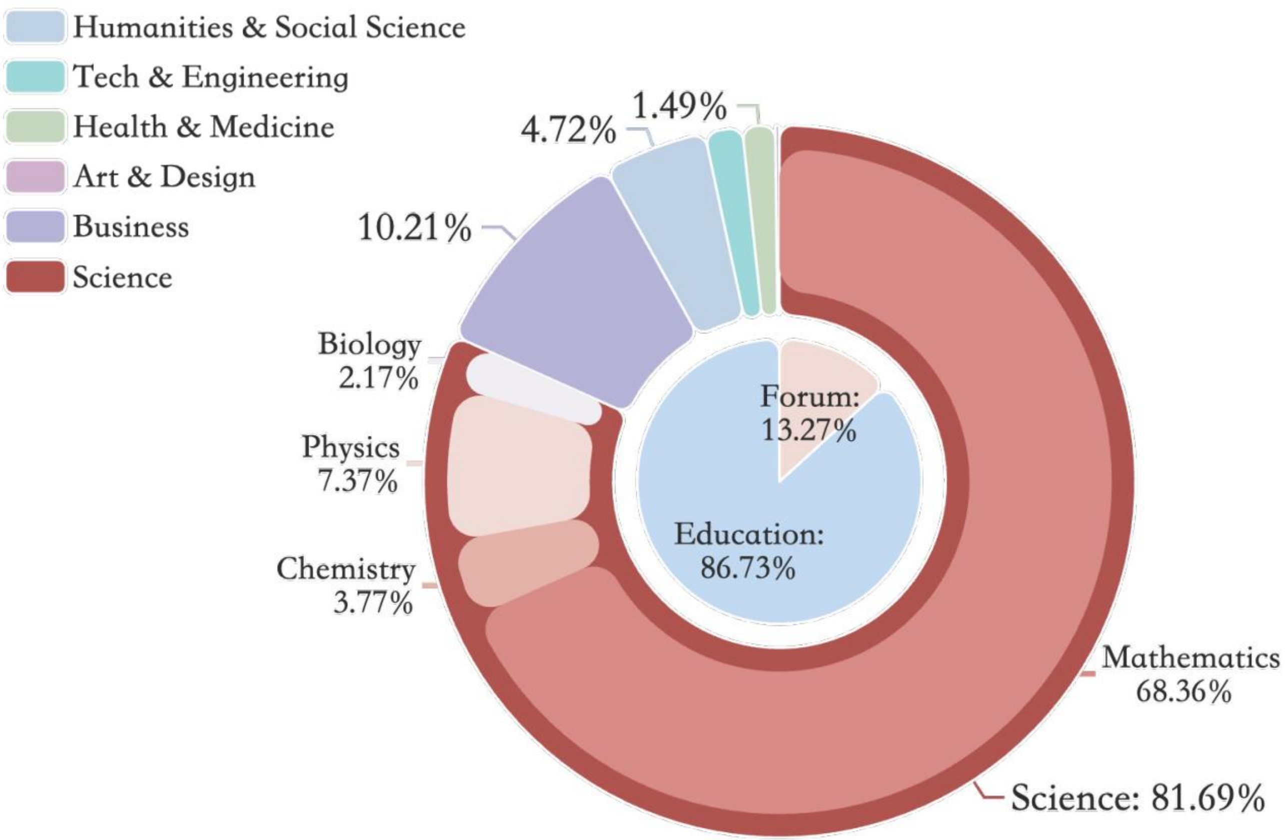
This simplification makes the other calculations unnecessary because raising any non-zero expression to the power of zero will always result in 1.

So, the simplified expression is simply: 1

Top Domains


Top 25 Website Usage Distribution
Website
biology.stackexchange.com
chemistry.stackexchange.com
www.indiabix.com
physics.stackexchange.com
socratic.org
cs.stackexchange.com
enotes.com
proofwiki.org
transtutors.com
sharemylesson.com
physicsforums.com
brainmass.com
www.khanacademy.org
gmatclub.com
weegy.com
www.chegg.com
studypool.com
zbmath.org
answers.everydaycalculation.com
homework.study.com
www.brainly.com
jiskha.com
answers.com
math.stackexchange.com
coursehero.com
0
2
4
6
8
10
Percentage of Total (%)

Domain Type Distribution



WebInstruct vs. Existing Datasets

Dataset	#Pairs	Domain	Format	Dataset Source
FLAN V2 (Chung et al., 2024)	100K	General	SFT	NLP data + Human CoT
Self-Instruct (Wang et al., 2023b)	82K	General	SFT	Generated by GPT3
GPT4-Alpaca (Taori et al., 2023)	52K	General	SFT	Generated by GPT4
SuperNI (Wang et al., 2022)	96K	General	SFT	NLP Datasets
Tora (Gou et al., 2023)	16K	Math	SFT	GSM+MATH Synthesis by GPT4
WizardMath (Luo et al., 2023)	96K	Math	SFT	GSM+MATH Synthesis by GPT4
MathInstruct (Yue et al., 2023b)	262K	Math	SFT	Math datasets Synthesis by GPT4
MetaMathQA (Yu et al., 2023)	395K	Math	SFT	GSM+MATH Synthesis by GPT3.5
XwinMath (Li et al., 2024a)	1.4M	Math	SFT	GSM+MATH Synthesis by GPT4
OpenMathInstruct (Toshniwal et al., 2024)	1.8M	Math	SFT	GSM+MATH Synthesis by Mixtral
WEBINSTRUCT	(10M) 5B	Math & Sci.	SFT	Recall and Extracted from Web

- Diversity is high: WebInstruct covers broader disciplines
- Scale is high: WebInstruct is at least 3x larger than the existing SFT datasets.

Experimental Results (Reasoning)

Model	TheoremQA	MATH	GSM8K	GPQA	MMLU-ST	BBH	ARC-C	AVG
GPT-4-Turbo-0409	48.4	69.2	94.5	46.2	76.5	86.7	93.6	73.6
Deepseek-7B	15.7	6.4	17.4	25.7	43.1	42.8	47.8	28.4
Qwen-1.5-7B	14.2	13.3	54.1	26.7	45.4	45.2	75.6	39.2
Mistral-7B	19.2	11.2	36.2	24.7	50.1	55.7	74.2	38.8
Gemma-7B	21.5	24.3	46.4	25.7	53.3	57.4	72.5	43.0
Llemma-7B	17.2	18.0	36.4	23.2	45.2	44.9	50.5	33.6
WizardMath-7B-1.1	11.7	33.0	83.2	28.7	52.7	56.7	76.9	49.0
Abel-7B-002	19.3	29.5	83.2	30.3	29.7	32.7	72.5	42.5
Intern-Math-7B	13.2	34.6	78.1	22.7	41.1	48.1	59.8	42.5
Rho-1-Math-7B	21.0	31.0	66.9	29.2	53.1	57.7	72.7	47.3
Deepseek-Math-7B	25.3	34.0	64.2	29.2	56.4	59.5	67.8	48.0
Deepseek-Math-Instruct	23.7	44.3	82.9	31.8	59.3	55.4	70.1	52.5
Llama-3-8B	20.1	21.3	54.8	27.2	55.6	61.1	78.6	45.5
Llama-3-8B-Instruct	22.8	30.0	79.5	34.5	60.2	66.0	80.8	53.4
Trained only with WEBINSTRUCT (All evaluations are held-out)								
MAmmoTH2-7B	29.0	36.7	68.4	32.4	62.4	58.6	81.7	52.8
Δ over Mistral	+9.8	+25.5	+32.2	+7.7	+12.3	+2.9	+7.5	+14.0
MAmmoTH2-8B	32.2	35.8	70.4	35.2	64.2	62.1	82.2	54.3
Δ over Llama3	+12.2	+14.5	+15.6	+8.0	+8.6	+1.0	+3.6	+8.8
Continue trained with additional instruction datasets (All held-out except MATH and GSM8K)								
MAmmoTH2-7B-Plus	29.2	**45.0**	**84.7**	36.8	64.5	63.1	83.0	58.0
MAmmoTH2-8B-Plus	**32.5**	42.8	84.1	**37.3**	**65.7**	**67.8**	**83.4**	**59.1**
Δ over best baseline	+7.2	+0.7	+1.5	+2.8	+5.5	+1.8	+2.6	+5.7

Experimental Results (Reasoning)

Model	TheoremQA	MATH	GSM8K	GPQA	MMLU-ST	BBH	ARC-C	AVG
GPT-4-Turbo-0409	48.4	69.2	94.5	46.2	76.5	86.7	93.6	73.6
Qwen-1.5-110B	34.9	49.6	85.4	35.9	73.4	74.8	91.6	63.6
Qwen-1.5-72B	29.3	46.8	77.6	36.3	68.5	68.0	92.2	59.8
Deepseek-LM-67B	25.3	15.9	66.5	31.8	57.4	71.7	86.8	50.7
Yi-34B	23.2	15.9	67.9	29.7	62.6	66.4	89.5	50.7
Llemma-34B	21.1	25.0	71.9	29.2	54.7	48.4	69.5	45.7
Mixtral-8×7B	23.2	28.4	74.4	29.7	59.7	66.8	84.7	52.4
Mixtral-8×7B-Instruct	25.3	22.1	71.7	32.4	61.4	57.3	84.7	50.7
Intern-Math-20B	17.1	37.7	82.9	28.9	50.1	39.3	68.6	46.4
Trained only with WEBINSTRUCT (All evaluations are held-out)								
MAmmoTH2-34B	30.4	35.0	75.6	31.8	64.5	68.0	90.0	56.4
Δ over Yi	+7.2	+19.1	+7.7	+2.1	+2.9	+1.2	+0.5	+5.8
MAmmoTH2-8x7B	32.2	39.0	75.4	36.8	67.4	71.1	87.5	58.9
Δ over Mixtral	+9.2	+10.6	+1.0	+7.1	+7.4	+3.3	+2.8	+6.5
Continue trained with additional instruction datasets (All held-out except MATH and GSM8K)								
MAmmoTH2-8x7B-Plus	**34.1**	**47.0**	**86.4**	**37.8**	**72.4**	**74.1**	**88.4**	**62.9**
Δ over Qwen-1.5-110B	-0.8	-2.6	+1.0	+1.5	-1.0	-0.7	-4.0	-0.7

Experimental Results (General)

	Code Generation	MT-Bench	Alpaca Eval 2.0	Arena Hard	MMLU	MMLU-Pro
GPT-4-1106-preview	85.6 (77.5)	9.32	50.0	-	-	-
GPT-3.5-Turbo-1106	79.7 (70.2)	8.32	19.3	18.9	-	-
GPT-3.5-Turbo-0301	-	7.94	18.1	18.1	70.0	-
Tulu-2-DPO-70B	51.2 (43.0)	7.89	21.2	15.0	67.8	40.5
Llama-2-70b-chat	31.4 (26.5)	6.86	14.7	11.6	63.0	33.6
Yi-34B-Chat	38.7 (32.6)	7.86	27.2	23.1	73.5	42.1
Mistral-7B-Instruct-v0.2	43.4 (36.5)	7.60	17.1	12.6	60.8	30.8
Llama-3-8B-Instruct	65.8 (58.0)	8.02	22.9	20.6	67.2	40.9
Mixtral-8×7B-Instruct-v0.1	52.3 (44.7)	**8.30**	23.7	23.4	**70.6**	41.0
MAmmoTH2-7B-Plus	**66.1** (**58.2**)	7.88	23.4	14.6	63.3	40.9
MAmmoTH2-8B-Plus	61.9 (53.3)	7.95	18.5	16.6	64.6	43.4
MAmmoTH2-8x7B-Plus	63.3 (55.3)	8.20	**33.8**	**32.6**	68.3	**50.4**

Takeaways

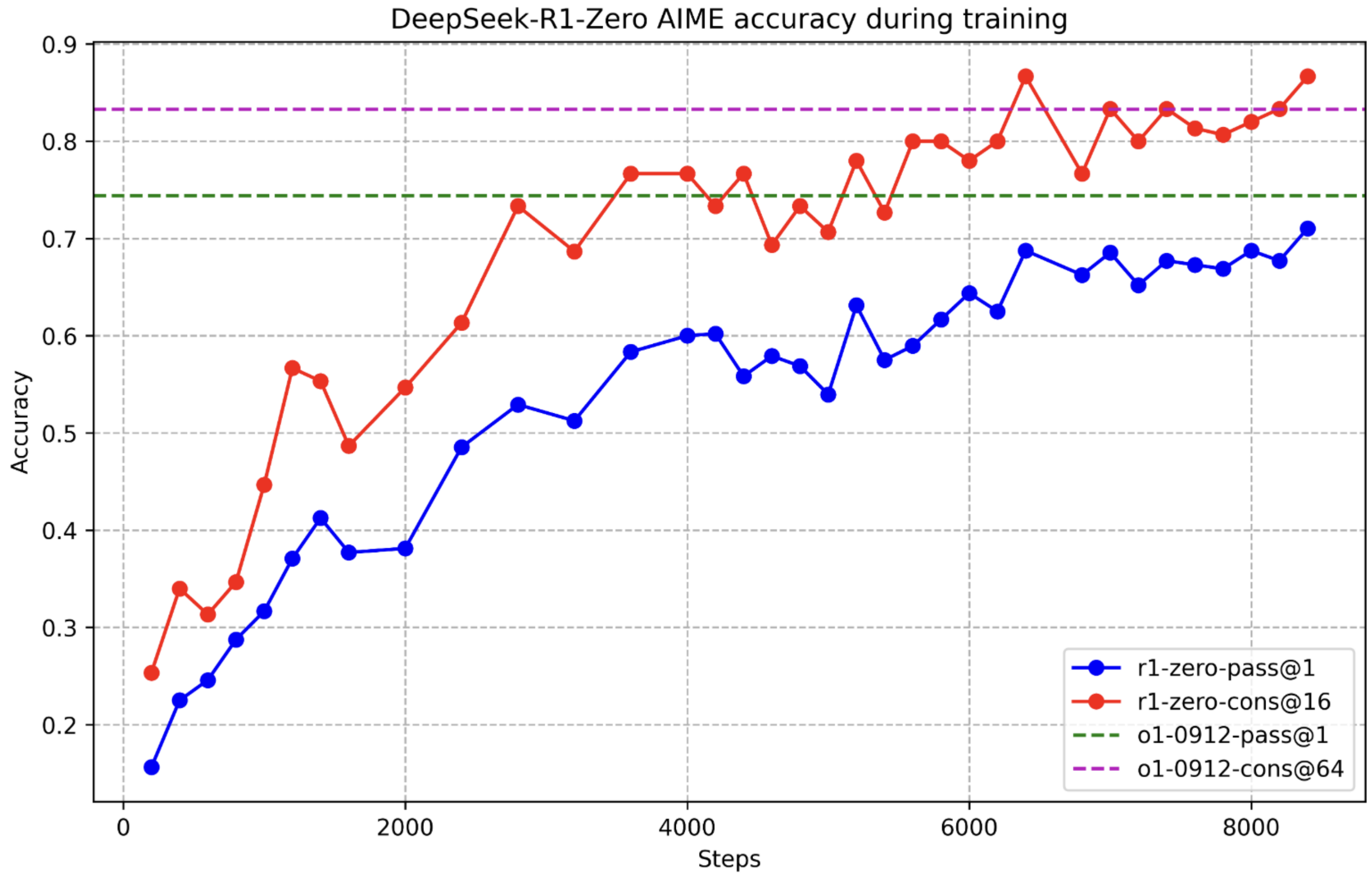
- Scaling up instruction tuning data is important.
- Extraction and Refining are necessary steps to improve perf.
- SFT loss is more effective than LM loss.
- Utilizing more capable models in the middle could lead to further improvement.

General-Reasoner: Advancing LLM Reasoning Across All Domains

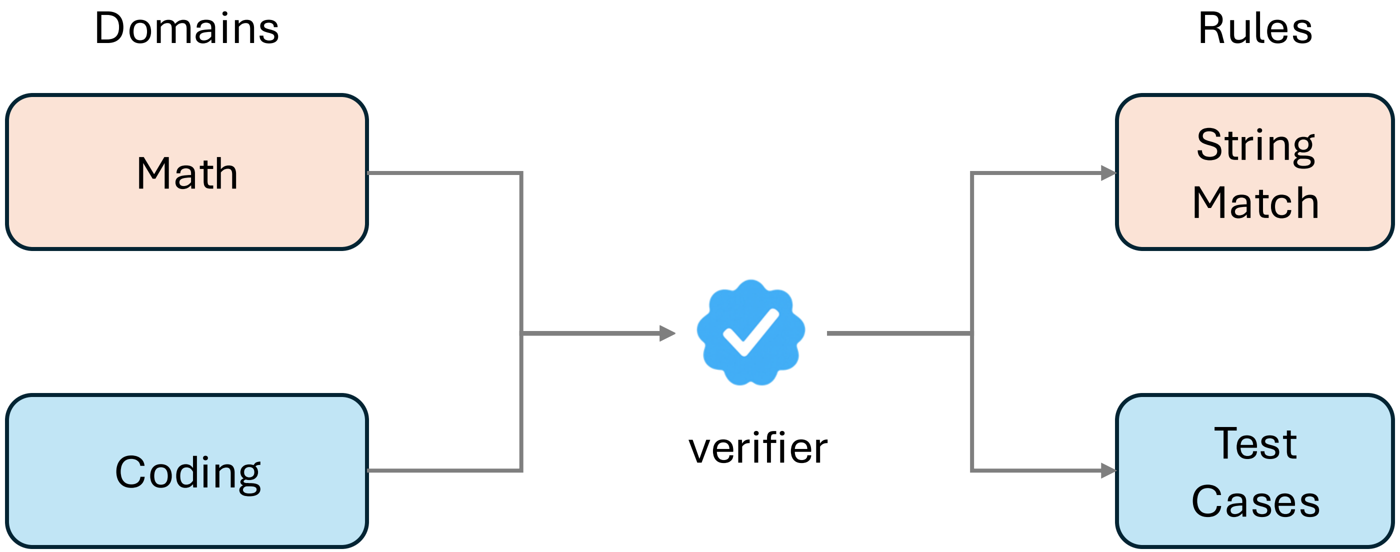
Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, Wenhu Chen

[Arxiv 2025]

R1-style Training



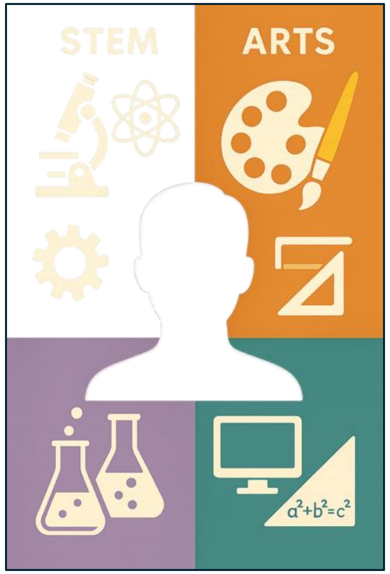
Limitations of Existing R1Trainig



- The output length becomes much longer and the model hallucinates more!
- The general capabilities are not improved, MMLU-Pro normally drops by 4%+.

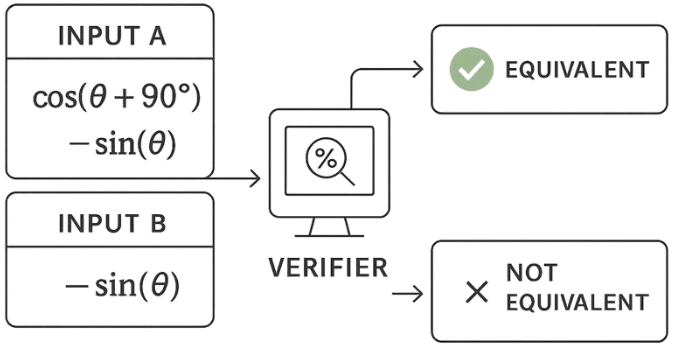
Towards General R1-Training

Math + STEM + Arts



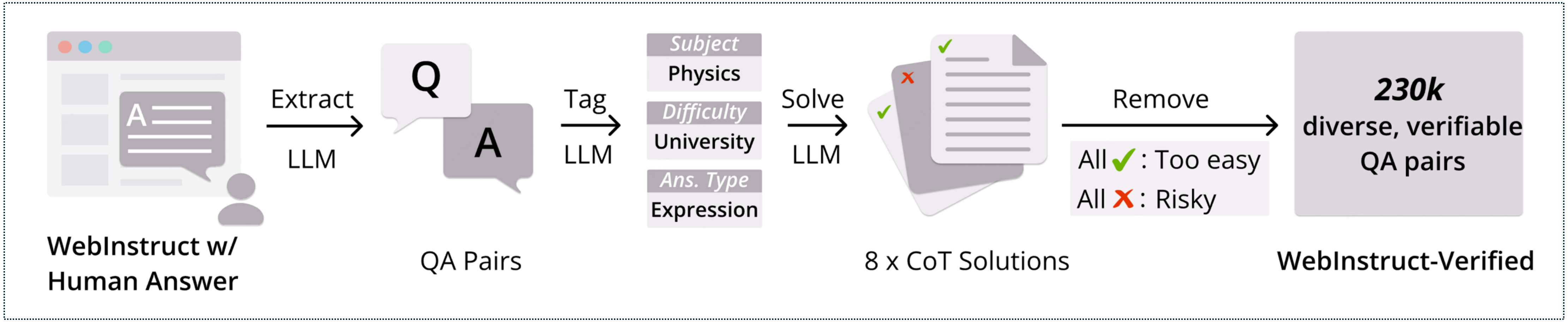
Diverse Data

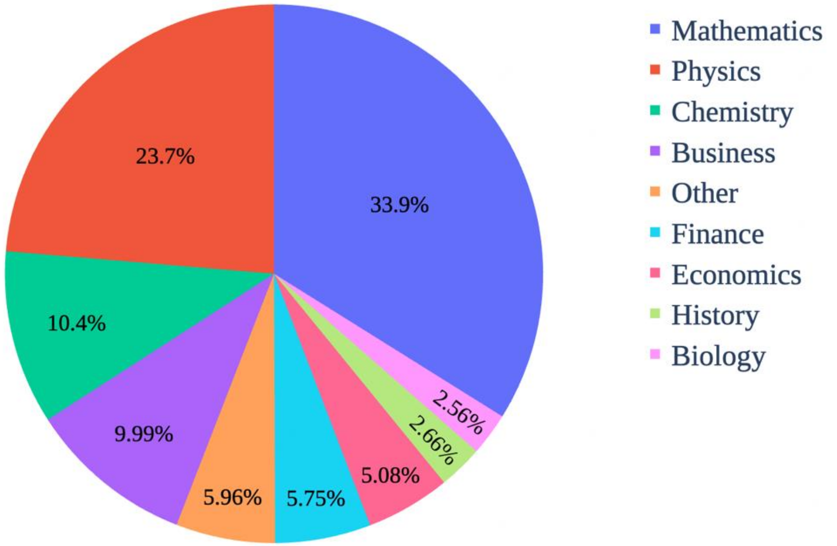
Beyond String Matching



General Verifier

Data: WebInstruct-verified





Verifier: General Verifier

- Given Q, prompt open models to generate $\hat{A}$.
  - Prompt Gemini-2.0-Pro to Generate CoT to compare A and $\hat{A}$
  - Synthesize large-scale inp-output: $(Q, A, \hat{A}) \Rightarrow (CoT, V)$
  - V is the verdict (equal or not equal)

- Distill the inp-output $(Q, A, \hat{A}) \Rightarrow (CoT, V)$
  - We adopt Qwen-2.5-3B to distill the judgement data.
  - It reaches 88% agreement rate with Gemini-2.0-Pro.
  - It’s can be served with minimum GPU for RL training.

General Verifier vs. Rule-based Verifier

Table 1: Examples of reasoning questions where the model provides correct answers, but the rule-based verifier fails to recognize their correctness, while the model-based verifier succeeds.

	Example 1	**Example 2**	**Example 3**
Question	Consider the line perpendicular to the surface $z = x^2 + y^2$ at the point where $x = 4$ and $y = 1$. Find a vector parametric equation for this line in terms of the parameter t.	Find the partial pressure in a solution containing ethanol and 1-propanol with a total vapor pressure of 56.3 torr. The pure vapor pressures are 100.0 torr and 37.6 torr, respectively, and the solution has a mole fraction of 0.300 of ethanol.	What is the work done to push a 1 kg box horizontally for 1 meter on a surface with a coefficient of friction of 0.5?
Ground Truth Answer	x = 4 + 8t, y = 1 + 2t, z = 17 - t	30.0 torr, 26.3 torr	4.9 J
Student Answer	4 + 8t, 1 + 2t, 17 - t	The partial pressure of ethanol is 30.0 torr and the partial pressure of 1-propanol is 26.32 torr.	4.9 N·m
Rule Based Verifier	False	False	False
Model Based Verifier	True	True	True

Our Training Framework

WebInstruct-verified

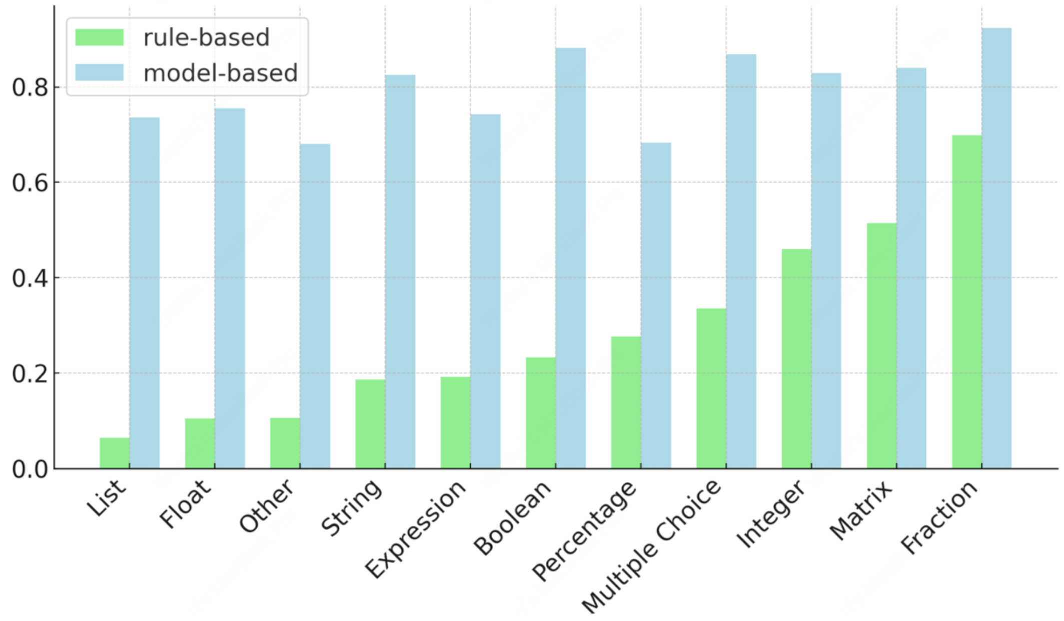
(q, a_{ref})

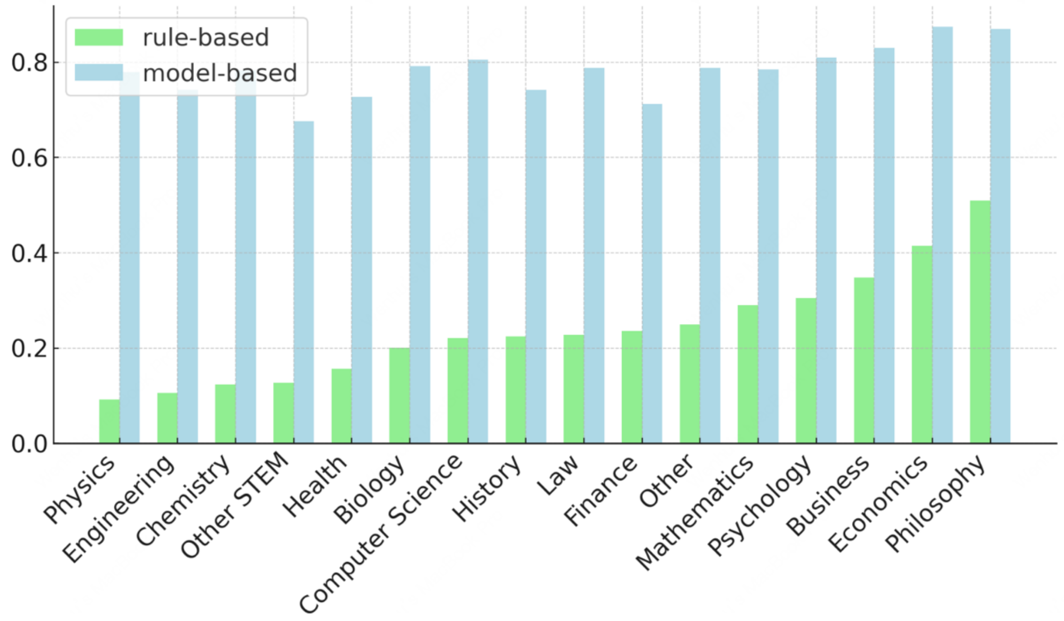
$$J = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}(\min(\frac{\pi_\theta(o_{i,t}|q,..)}{\pi_{old}(o_{i,t}|q,..)}A_{i,t}, clip(\frac{\pi_\theta(o_{i,t}|q,..)}{\pi_{old}(o_{i,t}|q,..)}, 1-\epsilon, 1+\epsilon)A_{i,t}))$$

normalize

$(q, a_{ref}, o_i) \Longrightarrow \quad \Longrightarrow R_i$

Impact of General Verifier





Impact of General Verifier

Table 5: Zero RL training using our model-based verifier versus the rule-based verifier on the Qwen3-4B-Base model for 120 step.

Dataset	Model-Based	Rule-Based
MMLU-Pro	60.1	58.1
GPQA	39.4	37.9
SuperGPQA	30.5	30.1
Math-Related	50.4	50.0

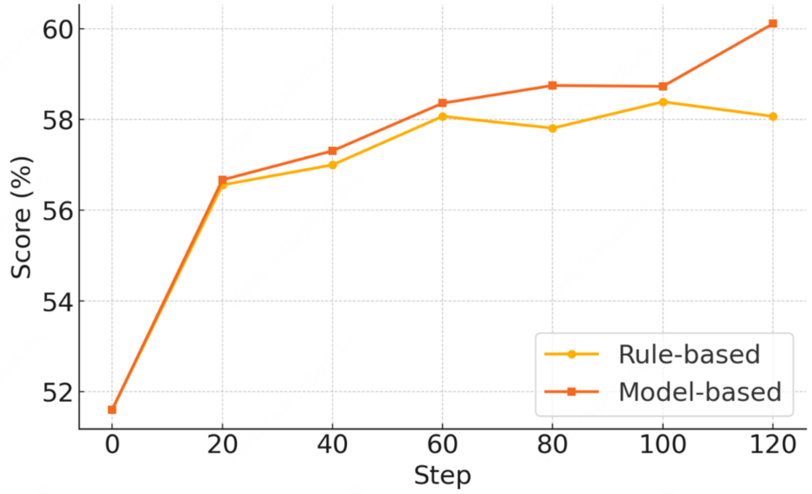


Figure 4: MMLU-Pro evaluation score at different training step using model-based verifier and rule-based verifier.

Experimental Results (General)

Model Name Metric	Backbone	MMLU-Pro Micro	GPQA-D Acc	SuperGPQA Macro (discipline)	TheoremQA Acc	BBEH Micro
MiMo-RL	MiMo-Base	58.6	54.4	40.5	38.8	11.4
QwQ-32B	Qwen2.5-32B-Inst	52.0	54.5	43.6	48.4	22.6
GPT-4o	-	74.6	50.0	46.3	43.6	22.3
o1-mini	-	80.3	60.0	45.2	53.1	-
DeepSeek-R1	DeepSeek-V3	84.0	71.5	59.9	59.1	34.9
		4B Models				
Qwen3-4B-Base	-	51.6	26.3	25.4	34.8	8.1
Qwen3-4B-Instruct (non-think)	Qwen3-4B-Base	61.8	41.7	32.1	42.0	**14.9**
GENERAL-REASONER-4B	Qwen3-4B-Base	**62.8**	**42.9**	**32.5**	**48.3**	12.2
		7B Models				
Qwen2.5-7B-Base	-	47.7	29.3	26.7	29.1	8.0
Qwen2.5-7B-Instruct	Qwen2.5-7B-Base	57.0	33.8	30.7	36.6	12.2
Open-Reasoner-Zero	Qwen2.5-7B-Base	**59.4**	36.6	32.8	37.4	12.2
Nemotron-CrossThink	Qwen2.5-7B-Base	57.8	38.5	29.1	-	-
SimpleRL-Qwen2.5-7B-Zoo	Qwen2.5-7B-Base	51.5	24.2	29.9	38.0	11.9
GENERAL-REASONER-7B	Qwen2.5-7B-Base	58.9	**38.8**	**34.2**	**45.3**	**12.5**
		14B Models				
Qwen2.5-14B-Base	-	53.3	32.8	30.7	33.0	10.8
Qwen2.5-14B-Instruct	Qwen2.5-14B-Base	62.7	41.4	35.8	41.9	15.2
SimpleRL-Qwen2.5-14B-Zoo	Qwen2.5-14B-Base	64.0	39.4	35.7	40.8	13.6
GENERAL-REASONER-Qw2.5-14B	Qwen2.5-14B-Base	66.6	43.4	39.5	44.3	15.2
Qwen3-14B-Base	-	64.2	45.9	36.5	44.0	13.0
Qwen3-14B-Instruct (non-think)	Qwen3-14B-Base	**70.9**	54.8	39.8	42.4	**19.2**
GENERAL-REASONER-Qw3-14B	Qwen3-14B-Base	70.3	**56.1**	**39.9**	**54.4**	17.3

Experimental Results (Math)

Model Name	AVG	MATH-500	Olympiad	Minerva	GSM8K	AMC	AIME24	AIME25
4B Models								
Qwen3-4B-Base	40.3	68.2	34.8	42.3	72.6	47.5	10.3	6.7
Qwen3-4B-Instruct (non-think)	**54.2**	80.4	**49.0**	57.0	92.0	**62.5**	**22.5**	**16.1**
General-Reasoner-4B	53.4	**80.6**	47.7	**57.7**	**92.2**	60.0	20.0	15.4
7B Models								
Qwen2.5-7B-Base	34.7	60.2	28.6	36.0	83.1	30.0	3.8	1.4
Qwen2.5-7B-Instruct	46.3	75.0	39.4	45.2	90.9	52.5	12.5	8.5
SimpleRL-Qwen2.5-7B-Zoo	48.4	74.0	**41.9**	49.6	90.7	**60.0**	**15.2**	7.5
General-Reasoner-7B	**48.5**	**76.0**	37.9	**54.0**	**92.7**	55.0	13.8	**10.4**
14B Models								
Qwen2.5-14B-Base	37.0	65.4	33.5	24.3	91.6	37.5	3.6	2.9
Qwen2.5-14B-Instruct	49.9	77.4	44.7	52.2	**94.5**	57.5	12.2	11.0
SimpleRL-Qwen2.5-14B-Zoo	50.7	77.2	44.6	54.0	94.2	60.0	12.9	11.8
General-Reasoner-Qw2.5-14B	53.9	78.6	42.1	58.1	94.2	70.0	17.5	16.9
Qwen3-14B-Base	49.9	74.6	44.3	55.9	93.2	55.0	14.7	11.4
Qwen3-14B-Instruct (non-think)	57.0	82.0	**52.4**	59.9	93.9	57.5	**28.5**	**25.1**
General-Reasoner-Qw3-14B	**58.8**	**83.8**	51.9	**68.0**	**94.4**	**70.0**	24.4	19.2

Impact of All-Domain Dataset

Table 4: Model performance trained with the diverse domain reasoning data vs. math-only data.

Backbone	Data	MMLU-Pro	GPQA	SuperGPQA	Math-Related
Qwen2.5-7B-Base	Full	58.9	34.3	34.2	48.5
Qwen2.5-7B-Base	Math Only	56.9	32.8	29.8	49.1
Qwen2.5-14B-Base	Full	66.6	43.4	39.5	53.9
Qwen2.5-14B-Base	Math Only	64.8	38.9	35.6	48.6

Diverse-domain Dataset can not only improves general reasoning but also math-only.

Takeaways

- Scaling up RL data is important
- Cross-domain generalization is essential for LLMs.
- Model-based Verifier can provide more dense rewards.