

Learning to Reason with LLMs

Xiang Yue

Carnegie Mellon University



xyue2@andrew.cmu.edu



<https://xiangyue9607.github.io>



Solving Complex Reasoning Problems with LLMs

OPENAI / ARTIFICIAL INTELLIGENCE / TECH

OpenAI releases o1, its first
'reasoning' model

OpenAI o1 Model Sets New Math and

Understanding R1 and DeepSeek

R1 belongs to a new category of AI models known as "reasoning models," with OpenAI's o1 being the most well-known example. What makes reasoning models special is their approach to problem-solving. Rather than generating immediate responses, they employ an internal reasoning process that mirrors human trains of thought.

Image: The Verge

OpenAI is releasing a new model called o1, the first in a planned series of "reasoning" models that have been trained to answer more complex questions, faster than a human can. It's being released alongside o1-mini, a smaller, cheaper version. And yes, if you're steeped in AI rumors:



IMAGE CREDITS: PERESMEH / GETTY IMAGES

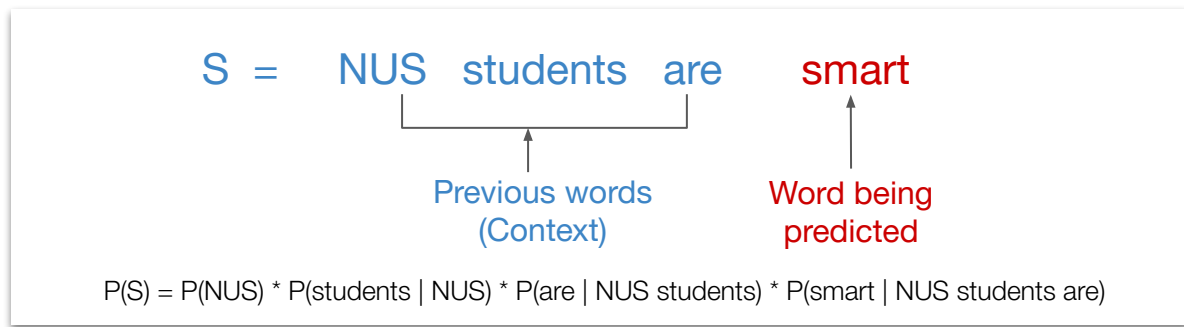
OpenAI's o1 on certain benchmarks

Kyle Wiggers — 2:27 PM PST · January 27, 2025

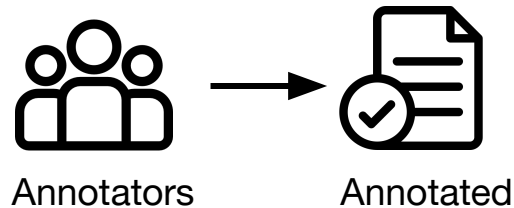


What are Large Language Models (LLMs)?

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^n p(w_i | w_1, \dots, w_{i-1})$$



Pre-training (~10 Trillion Tokens)



Post-training (~1M-10B tokens)



What is Reasoning?



“Reason is the capacity of **applying logic by drawing valid conclusions** from new or existing information”

- **Fact Retrieval**

Who was the 16th President of the United States?

- **Simple Paraphrasing**

*Rewrite this sentence in a formal tone:
"I gotta head out now."*

- **Grammar Corrections**

*"He go to store."
→ "He goes to the store."*



Not reasoning

- **Logical Deduction**

If Tom is taller than Jerry, and Jerry is taller than Ben, who is the shortest?

- **Mathematical Problem Solving**

The area of a circle is 100π square centimeters. Find the radius.

- **Multi-Step Planning**

A home assistant AI needs to prepare a house for an guest arriving in 2 hours.

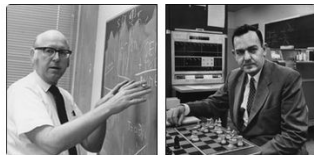


Reasoning



AI Reasoning Has a Long History

Allen Newell, Herbert Simon who created "Logic Theorist," 1st thinking machine in 1955

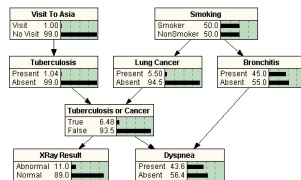


Prolog is a logic programming language that has automated theorem proving and computational linguistics

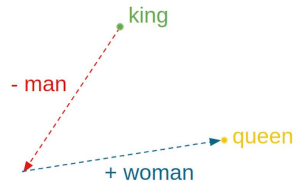
Sample Prolog Program:
grandmother.pl

```
grandmother(X, Y) :- mother(X, Z), parent(Z, Y).  
parent(X, Y) :- mother(X, Y).  
parent(X, Y) :- father(X, Y).  
  
mother(mary, stan).  
mother(gwen, alice).  
mother(valery, gwen).  
father(stan, alice).
```

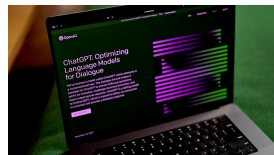
Bayesian networks and probabilistic models handle uncertainty and make informed decisions.



Neural networks advance semantic understanding and relational reasoning.



LLMs revolutionize natural language understanding and reasoning.



Foundational
Concepts and
Rule-based Systems
(1950s)

Knowledge
Representation &
Symbolic AI
(1970s-1990s)

Statistical and
Probabilistic
Methods
(2000s)

Neural
Reasoning
(2010s)

**LLM
Reasoning
(2020s)**

Why **LLM Reasoning**? What is Different Today?



Why Reasoning with LLMs?

“Language” as a Universal Interface

Understand and
Interpret User Intents

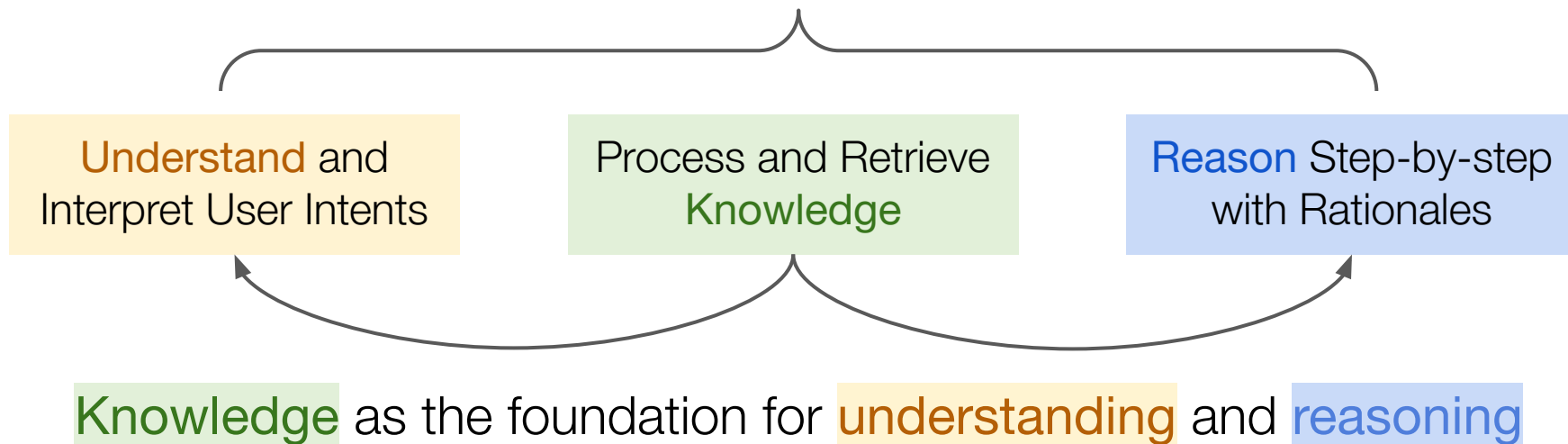
Process and Retrieve
Knowledge

Reason Step-by-step
with Rationales



Why Reasoning with LLMs?

“Language” as a Universal Interface



This enables models to generalize and reason over *unseen* questions



Challenges in LLM Reasoning



Understanding Reasoning

Lack *systematic* methods to **measure reasoning** and investigate the **factors that influence LLM reasoning**



Improving Reasoning

Lack **efficient** and **scalable post-training methods** to enhance LLM reasoning beyond pre-training



Responsible LMs

Enhanced reasoning capabilities may come at the cost of **reduced safety and security** in LLMs



My Research



Understanding Reasoning

Benchmarks:

- MMMU [🏆 CVPR'24 Best Paper Finalist]
- MMLU-Pro [NeurIPS'24a]
- MixEval(-X) [ICLR'25 Spotlight]
- MegaBench [ICLR'25b]

Probing & Ablation Study:

- Grokked Transformers [NeurIPS'24c]
- AI Debate [EMNLP'23 Findings]
- Demystify Long CoT [arXiv'25]



Improving Reasoning

Math and STEM Reasoning:

- MAMmoTH [ICLR'24 Spotlight]
- MAMmoTH-2 [NeurIPS'24d]
- Demystify Long CoT [arXiv'25]

Multimodal Reasoning:

- MAMmoTH-VL [arXiv'24]
- MultiUI [ICLR'25c]

Code Reasoning:

- OpenCodeInterpreter [ACL'24 Findings]



Responsible LMs

Multilinguality:

- Pangea [ICLR'25d]
- JMMMMU [NAACL'25]

Privacy and Security:

- Machine Unlearning [ACL'24]
- Differential Privacy [ACL'21; CCS'23; 🏆 ACL'23 Best Paper Honorable Mention]

AI for Healthcare:

[ACL'20; 🏆 BIBM'21 Best Paper; arXiv'24]



This Talk: Learning to Reason with LLMs



Understanding Reasoning

(~20min)



Improving Reasoning

(~20min)



Future Work

(~5min)



This Talk: Learning to Reason with LLMs



Understanding Reasoning

Reasoning Benchmarks:

- **MMMU** [🏆 CVPR'24 Best Paper Finalist]
- MMLU-Pro [NeurIPS'24a]
- MixEval(-X) [ICLR'25 Spotlight]

Constructing benchmarks that are **real**,
challenging, **robust**, and **dynamic**

Impact: Our benchmarks are heavily used by:



This Talk: Learning to Reason with LLMs



Understanding Reasoning

Reasoning Benchmarks:

- **MMMU** [🏆 CVPR'24 Best Paper Finalist]
- MMLU-Pro [NeurIPS'24a]
- MixEval(-X) [ICLR'25 Spotlight]

Factors Influencing Reasoning:

- **Demystify Long Chain-of-thought**
[ICML'25 Submission]
- Reasoning Learning Ability Gap [ACL'25 Submission]

Constructing benchmarks that are **real**,
challenging, **robust**, and **dynamic**

Impact: Our benchmarks are heavily used by:



How do **rationale patterns**, **model size**,
and **learning mechanisms** influence LLM
reasoning abilities?



This Talk: Learning to Reason with LLMs



Improving Reasoning

Synthetic Reasoning Data:

- **MAmmoTH-2** [NeurIPS'24d]
- MAmmoTH [ICLR'24 Spotlight]
- MAmmoTH-VL [arXiv'24]
- Pangea [ICLR'25]

Efficient and scalable post-training methods

Fine-tuning with synthetic reasoning data

Impact: Our approaches are discussed and used in various leading foundation models:



This Talk: Learning to Reason with LLMs



Improving Reasoning

Synthetic Reasoning Data:

- **MAmmoTH-2** [NeurIPS'24d]
- MAmmoTH [ICLR'24 Spotlight]
- MAmmoTH-VL [arXiv'24]
- Pangea [ICLR'25]

Reward Shaping in RL:

- **Demystify Long CoT**
[ICML'25 Submission]

Efficient and scalable post-training methods

Fine-tuning with synthetic reasoning data

Impact: Our approaches are discussed and used in various leading foundation models:



How to **shape rewards** in RL training of reasoning LLMs to **control output** behaviors (e.g., length) for higher accuracy?



This Talk: Learning to Reason with LLMs



Understanding Reasoning

(~20min)



Improving Reasoning

(~20min)



Future Work

(~5min)

Understanding Reasoning

- **RQ1:** How to measure LLM reasoning ability in real-world scenarios?
- RQ2: What training factors influence LLM reasoning abilities?

A Multimodal Reasoning Example



Imagine a scenario where a student finishes a **complex accounting homework question**, and wants to verify their approach...

Instead of spending hours second-guessing, they simply **take a screenshot of the problem**, **upload it to a multimodal LLM**...

24 houranswers Subjects Online Tutoring Homework Help Homework Library Tutors Online Classes More ~ Get Help Now Log In

Home > Homework Library >

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) | <u>32,130</u> | <u>39,690</u> | <u>?</u> | <u>(58,275)</u> |

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 |

View Available
Computer Science Tutors
624 tutors matched



and instantly receive step-by-step guidance...

The formula to calculate Net Income (Loss):

$$\text{Net Income (Loss)} = \text{Revenues} - \text{Expenses} + \text{Gains} - \text{Losses}$$

Plugging in the values we have:

$$-58,275 = 1,054,116 - \text{Expenses} + 8,505 - 39,312$$

Simplify the equation:

$$-58,275 = (1,054,116 + 8,505 - 39,312) - \text{Expenses}$$

Calculate the values inside parentheses:

$$-58,275 = 1,023,309 - \text{Expenses}$$

Now, solve for Expenses:

$$\text{Expenses} = 1,023,309 + 58,275$$

$$\text{Expenses} = 1,081,584$$

Final Answer:

The missing amount (Expenses for Company D) is **\$1,081,584**. This corresponds to option (A).

That's what we expect AI can do:
reasoning over real-world complex queries

However, current benchmarks fail to capture the
complexity of real-world reasoning,
either in **breadth** or **depth**

Existing Multimodal Reasoning Benchmarks

VQA

(Antol et al., 2015;
Goyal et al., 2017)



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

TextVQA

(Singh et al., 2019)



What is the top oz?

MM-Vet

(Yu et al., 2023)

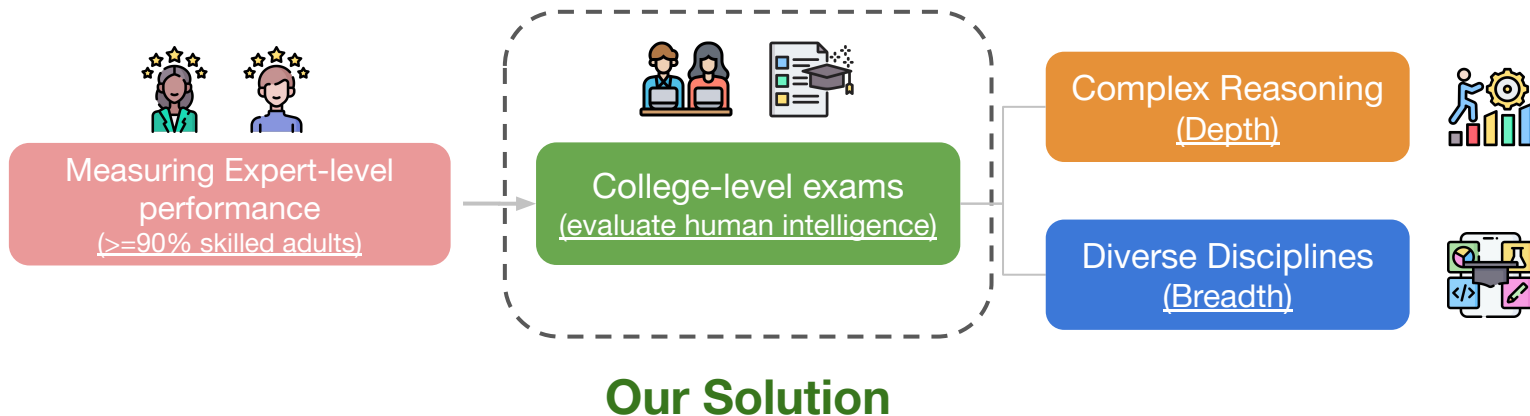


Q: What will the girl on the right write
on the board?

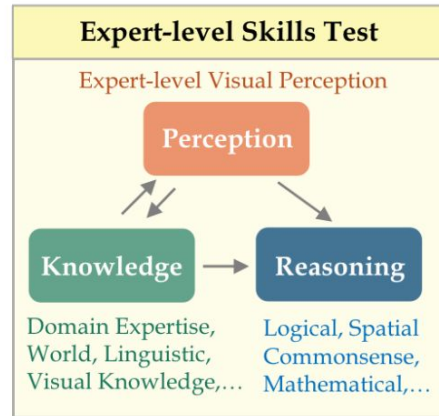
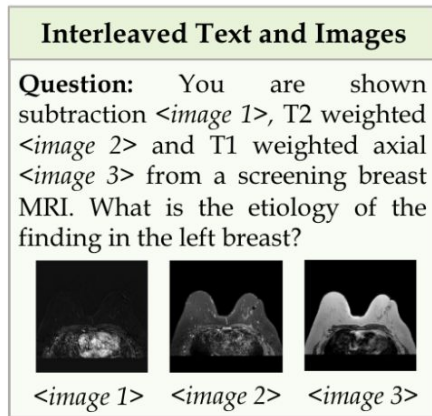
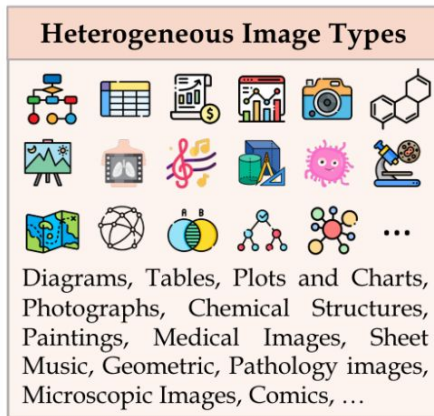
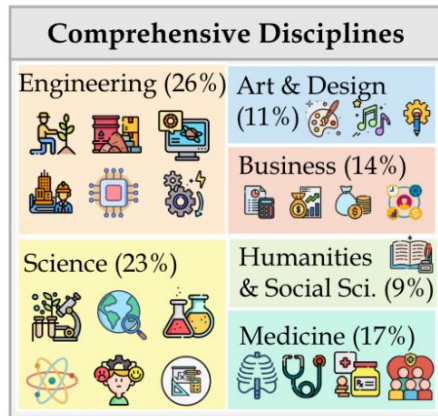
GT: 14

How to Measure Complex Multimodal Reasoning?

Technical Challenge: Effectively integrates diverse multimodal inputs with **expert-level** **complex reasoning** across **a broad range of disciplines**



MMMU: multi-discipline multimodal understanding and reasoning



(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



Impact: Go-to-evaluation

Jeff Dean (@jeffdean)

MMM is a brand new benchmark ([mmm-benchmark.github.io](https://github.com/google/mmmu)) 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.

| MMM (val) | Gemini Ultra (0-shot) | GPT-4V (0-shot) |
|-----------------------------|-----------------------|-----------------|
| Art & Design | 74.2 | 70.0 |
| Business | 68.2 | 58.9 |
| Science | 49.3 | 48.0 |
| Health & Medicine | 71.9 | 62.9 |
| Humanities & Social Science | 78.9 | 70.3 |
| Technology & Engineering | 59.0 | 47.1 |
| Overall | 62.4 | 59.4 |

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.

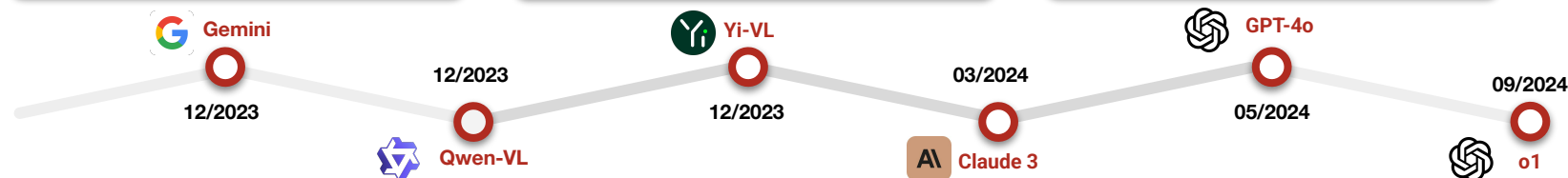
Yi Open-source Performance Benchmarks

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

MMM Benchmark

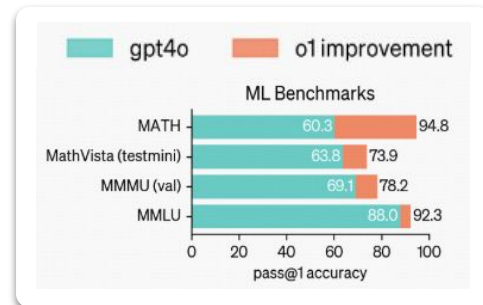
| Model | Text | Image | Table | Diagram | Science | Health | Human | Code |
|------------------|------|-------|-------|---------|---------|--------|-------|------|
| GPT-4V | 43.7 | 42.5 | 61.0 | 36.3 | 40.9 | 46.8 | 44.2 | 41.5 |
| Qwen-VL-Plus | 36.8 | 39.5 | 61.5 | 23.2 | 32.8 | 40.5 | 43.4 | 33.3 |
| Yi-VL-34B | 36.5 | 26.2 | 62.9 | 19.1 | 31.5 | 42.1 | 42.5 | 26.5 |
| Yi-VL-6B | 35.0 | 35.8 | 58.0 | 15.9 | 32.3 | 39.3 | 40.4 | 32.1 |
| Qwen-VL-Chat | 31.3 | 30.7 | 52.6 | 18.5 | 26.9 | 33.4 | 34.1 | 31.4 |
| Yi-6B-Vision-13B | 26.7 | 26.4 | 39.7 | 13.8 | 23.0 | 31.7 | 26.5 | 28.5 |

| Eval Sets | GPT-4o | GPT-4T 2024-04-09 | Gemini 1.0 Ultra | Gemini 1.5 Pro | Claude Opus |
|----------------------------|--------|----------------------|------------------|----------------|-------------|
| MMM (val) | 69.1 | 63.1 | 59.4 | 58.5 | 59.4 |
| MathVista (val) (testmini) | 63.8 | 58.1 | 53.0 | 52.1 | 50.5 |
| AIGD (val) (test) | 94.2 | 89.4 | 79.5 | 80.3 | 88.1 |
| ChartQA (val) (test) | 85.7 | 78.1 | 80.8 | 81.3 | 80.8 |
| DocVQA (val) (test) | 92.8 | 87.2 | 90.9 | 86.5 | 85.3 |
| ActivityNet (val) (test) | 61.9 | 59.5 | 52.2 | 56.7 | |
| EgoSchema (val) (test) | 72.2 | 63.9 | 61.5 | 63.2 | |

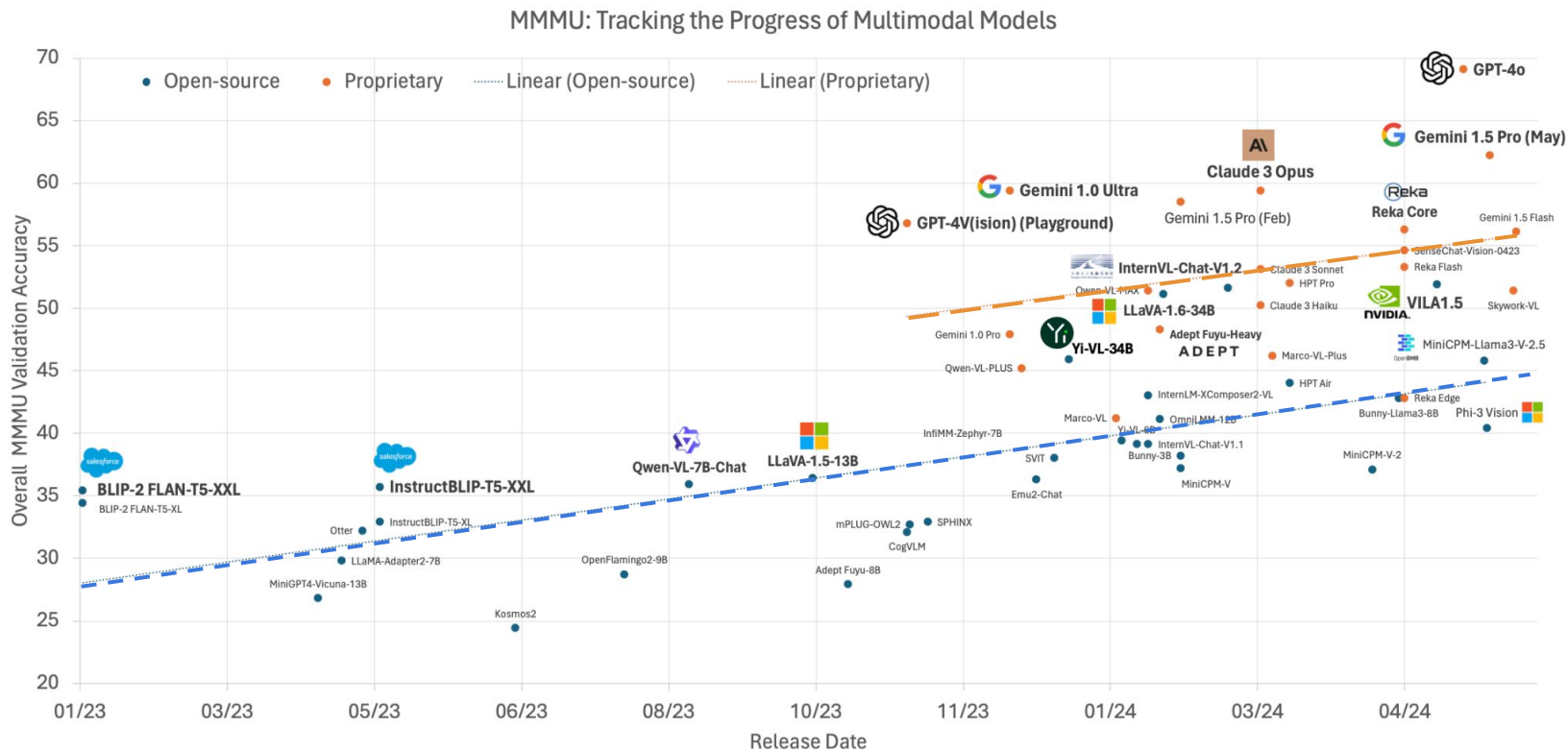


| Model | DocVQA Document understanding | ChartQA Chart understanding | AIGD Science diagrams | TextVQA Text reading | MMM College-level problem |
|-------------------------------|----------------------------------|--------------------------------|--------------------------|-------------------------|------------------------------|
| Other Best Open-source LLM | 81.6% (CogAgent) | 68.4% (CogAgent) | 73.7% (EvoMedius) | 76.1% (CogAgent) | 45.9% (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 Opus | Claude 3 Sonnet | Claude 3 Haiku |
|--|---------------------|---------------------|---------------------|
| Math & reasoning MMM (val) | 59.4% | 53.1% | 50.2% |
| Document visual Q&A ANLS score, test | 89.3% | 89.5% | 88.8% |
| Math MathVista (testmini) CoT | 50.5% | 47.9% CoT | 46.4% CoT |
| Science diagrams AIGD, test | 88.1% | 88.7% | 86.7% |
| Chart Q&A Relaxed accuracy (test) | 80.8% 0-shot CoT | 81.1% 0-shot CoT | 81.7% 0-shot CoT |



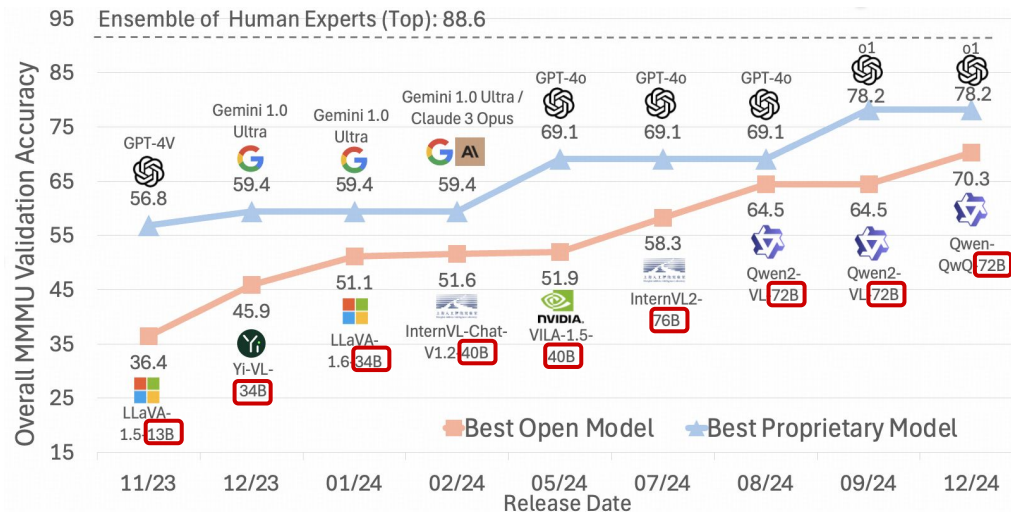
Track Progress of Multimodal LLMs with MMMU



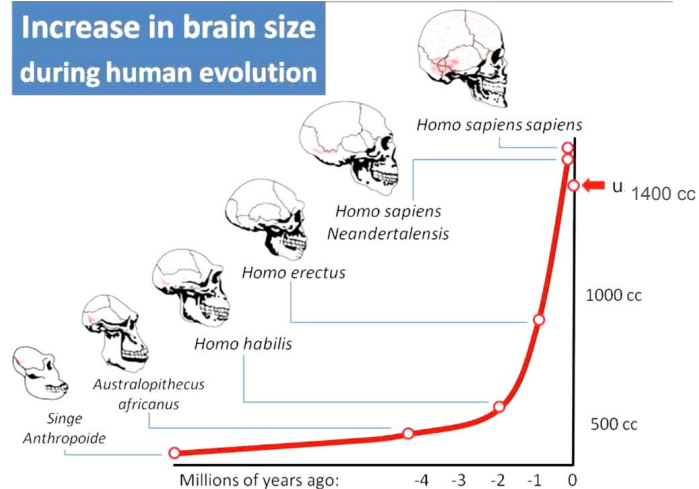
Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024 Oral (Best Paper Finalist)**



MMMU: Push the Scaling of Multimodal LLMs



Scaling improves LLMs' knowledge and reasoning



(image source: <https://aquatic-human-ancestor.org/anatomy/brain.html>)

“A larger brain results in larger capacity for adaptive knowledge”

Muthukrishna M et al. (2018) "The Cultural Brain Hypothesis: How culture drives brain expansion, sociality, and life history". *PLOS Computational Biology*.

Impact: Use in Community

EvalAI



Hugging Face



MMMU-Benchmark Evaluation Challenge

Organized by: MMMU Benchmark
Published: Nov 11, 2023 6:11:11 AM EST (GMT - 4:00)
Starts on: Nov 11, 2023 6:11:11 AM EST (GMT - 4:00)
Ends on: Jan 1, 2026 6:11:11 AM EST (GMT - 4:00)

Overview | Evaluation | Phases | Participate | **All Submissions** | Leaderboard | Manage

All Submissions

Test Set Evaluation

File type: Fields to export (optional) Download

Filter submissions by team

| # | Team | Created by | Status | Execution time (sec) | Submission no. | Submitted at | Submitted file | Stdout file | Stderr file |
|---|----------|------------|----------|----------------------|----------------|-------------------------|----------------------|----------------------|-------------|
| 1 | ihb | ihb | Finished | 0.195366 | 450134 | Jun 6, 2024 8:02:20 AM | Link | Link | None |
| 2 | ihb | ihb | Finished | 0.150539 | 450254 | Jun 6, 2024 2:28:45 AM | Link | Link | None |
| 3 | gl_mn | jiangtao | Finished | 0.091553 | 450029 | Jun 6, 2024 1:15:00 AM | Link | Link | None |
| 4 | j9 | J029 | Finished | 0.064799 | 449967 | Jun 5, 2024 10:44:31 PM | Link | Link | None |
| 5 | Jason123 | kldmofashi | Finished | 0.096911 | 449951 | Jun 5, 2024 10:10:05 PM | Link | Link | None |

Datasets: MMMU Like 148

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

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

Dataset card | Viewer | Files and versions | Community | Settings

Dataset Viewer

Subset (30) Accounting - 415 rows

Search this dataset

| id | question | options | explanation | image_1 | image_2 | image_3 |
|------------------|---|--|-------------------------------|---|----------------------------------|----------------------------------|
| string - classes | string - classes | string - classes | string - classes | image - width (px) | image - width (px) | image - width |
| dev_Accounting_1 | Each of the following... | ['\$63,020', '\$68,410', ...] | | AS1 | 1,234 | |
| dev_Accounting_2 | Here are facts for the Hudson Roofin... | ['\$171,900', '\$170,000', ...] | | founder, Capital Dec. 1 | Not supported with pagination... | Not supported with pagination... |
| dev_Accounting_3 | For 2010, calculate the cas... | ['1', '\$493.82', '2', '\$2,384', ...] | DCF = EBIT + Depreciation ... | | Not supported with pagination... | Not supported with pagination... |
| dev_Accounting_4 | Paper Submarine Manufacturing is... | ['\$.02', '\$7.79', '\$8.65'] | | Chapter 10: Submarine and the Chapter 10: Submarine and the Chapter 10: Submarine and the | Not supported with pagination... | Not supported with pagination... |

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

[Homepage](#) | [Dataset](#) | [Paper](#) | [arXiv](#) | [GitHub](#)

Downloads last month: 166,289

Use in Datasets library | Edit dataset card

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

Models trained or fine-tuned on MMMU/MMMU

- nurcan/nurcan Updated Dec 17, 2023
- Lipu124/Spd1 Updated Dec 25, 2023

150+ Models, 3000+ Submissions



1.5M+ Downloads in Total

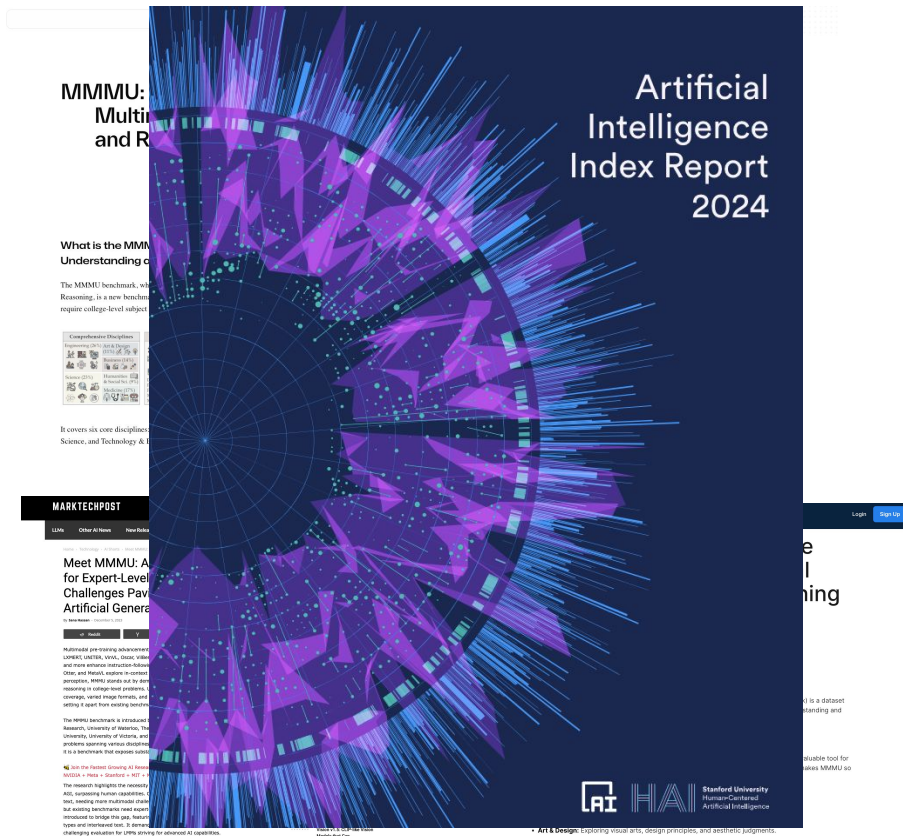


Yue, X. et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. **CVPR 2024 Oral (Best Paper Finalist)**

Understanding Reasoning



Impact: Media and Report Coverage



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).¹ 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.

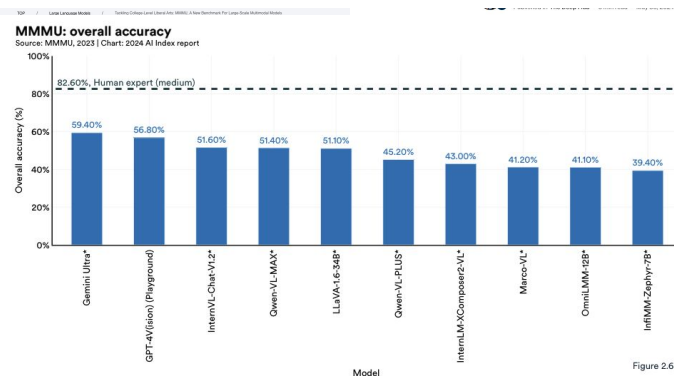


Figure 2.6.2¹

MARKTECHPOST

ODAL UNDERSTANDING AND

Mark

APPLICATION

Models

Reason across multiple
Engineering. The
cal structures.



Impact: A Most Influential CVPR 2024 Paper

Paper Digest Team
New York City, New York, 10017
team@paperdigest.org

TABLE 1: Most Influential CVPR Papers (2024-09)

| YEAR | RANK | PAPER | AUTHOR(S) |
|------|------|---|---|
| 2024 | 1 | Improved Baselines with Visual Instruction Tuning IF:8 Related Papers Related Patents Related Grants Related Venues Related Experts View <i>Highlight: In this paper we present the first systematic study to investigate the design choices of LMMs in a controlled setting under the LLaVA framework.</i> | Haotian Liu; Chunyuan Li; Yuheng Li; Yong Jae Lee; |
| 2024 | 2 | MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI IF:5 Related Papers Related Patents Related Grants Related Venues Related Experts View <i>Highlight: We introduce MMMU: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning.</i> | XIANG YUE et. al. |



CITED BY



Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi

650

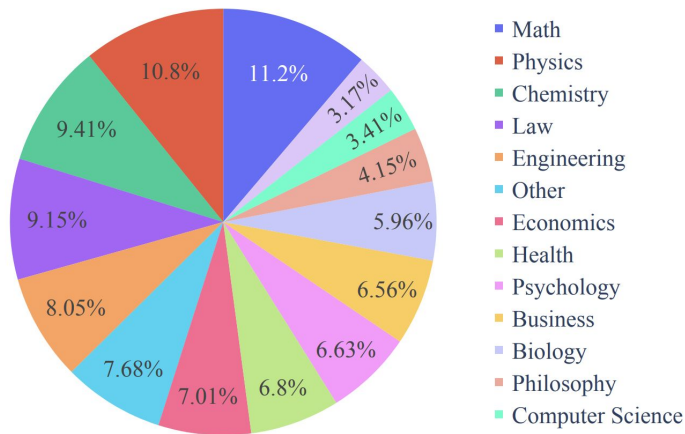
X Yue, Y Ni, K Zhang, T Zheng, R Liu, G Zhang, S Stevens, D Jiang, ...
CVPR 2024 (Oral); Best Paper Candidate, 2024



CVPR 2024 Best Paper Final List



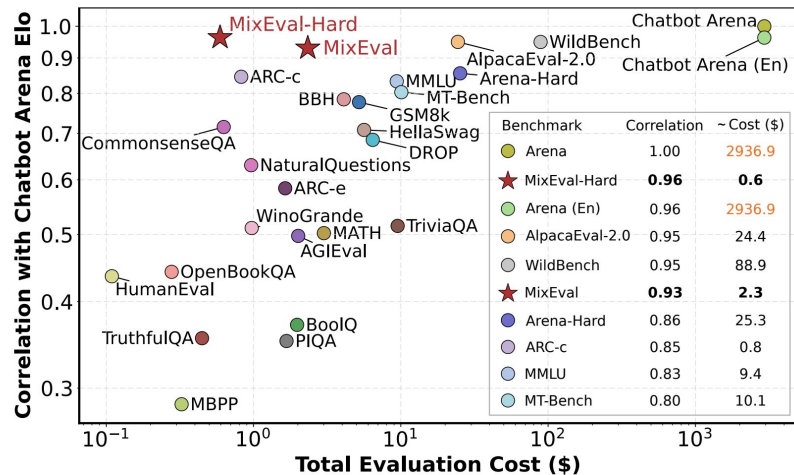
Other Widely-adopted LLM Benchmarks



MMLU-Pro [NeurIPS'24a Spotlight]

Robust multi-task reasoning,
featuring STEM subjects

Wang, Y., Ma, X., , ... & **Yue, X.**, Chen, W. (2024). MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *NeurIPS 2024 **Spotlight***



MixEval [NeurIPS'24b]

Efficient and dynamic evaluation by
mixing existing benchmarks

Ni, J.*, Xue, F.*, **Yue, X.***, Deng, Y., Shah, M., Jain, K., ... & You, Y. (2024). Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *NeurIPS 2024 (*: Core Contributors)*

Understanding Reasoning

- **RQ1:** How to measure LLM reasoning ability in real-world scenarios?
- **RQ2:** What training factors influence LLM reasoning ability?

Reason with Rationales: Chain-of-thought (CoT)

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Wei, J. (2022). Chain-of-thought prompting elicits reasoning in large language models. NeurIPS 2022




Scaling up *Inference* Compute with Long CoT

~200 Tokens

GPT-4o

such that $1 \leq |k| \leq n$.

Find all other real numbers x for which $p\left(\frac{1}{x}\right) = x^2$.

 We are given a monic real polynomial $p(x)$ of degree $2n$, which has the form

$$p(x) = x^{2n} + a_{2n-1}x^{2n-1} + \dots + a_1x + a_0$$

and we know that for integers k such that $1 \leq |k| \leq n$, the polynomial satisfies the condition


$$p\left(\frac{1}{k}\right) = k^2.$$

We are tasked with finding all other real

OpenAI o1-preview

such that $1 \leq |k| \leq n$.

Find all other real numbers x for which $p\left(\frac{1}{x}\right) = x^2$.

 Thought for 5 seconds ^

We are told that for all integer values of k satisfying $1 \leq |k| \leq n$,

$$p\left(\frac{1}{k}\right) = k^2$$

Our problem is: Given that p is monic of degree $2n$, and that $p\left(\frac{1}{k}\right) = k^2$ for $k = \pm 1, \pm 2, \dots, \pm n$, find all other real numbers x such that $p\left(\frac{1}{x}\right) = x^2$.

In other words, we are to find all real numbers x such that $p\left(\frac{1}{x}\right) = x^2$.

~2000 Tokens

OpenAI o1 generates **longer CoTs** for reasoning

RQ2: What training factors influence reasoning ability?

1. How do **long CoTs** compared to **short CoTs**
2. Where does reasoning ability come from?
3. How does model size affect reasoning ability?

Training Setups

Supervised
Fine-tuning (SFT)

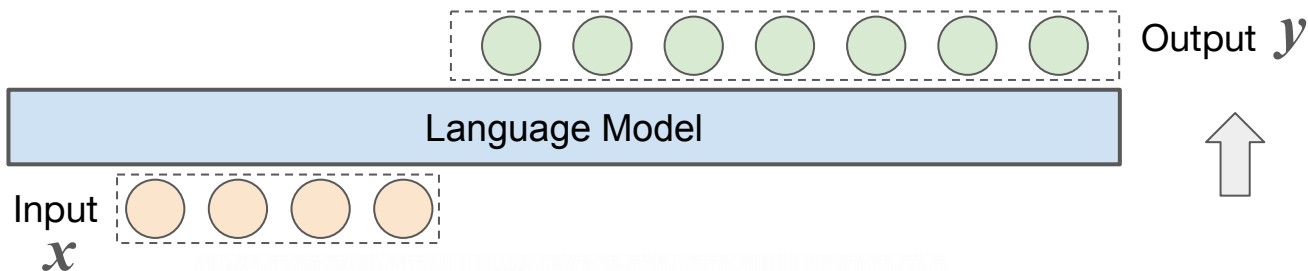
Reinforcement
Learning (RL)

Post-training

Supervised Fine-tuning (SFT)

Question: Lily has 48 apples. She gives 12 to her friend and then divides the rest equally into 6 baskets. How many apples are in each basket?

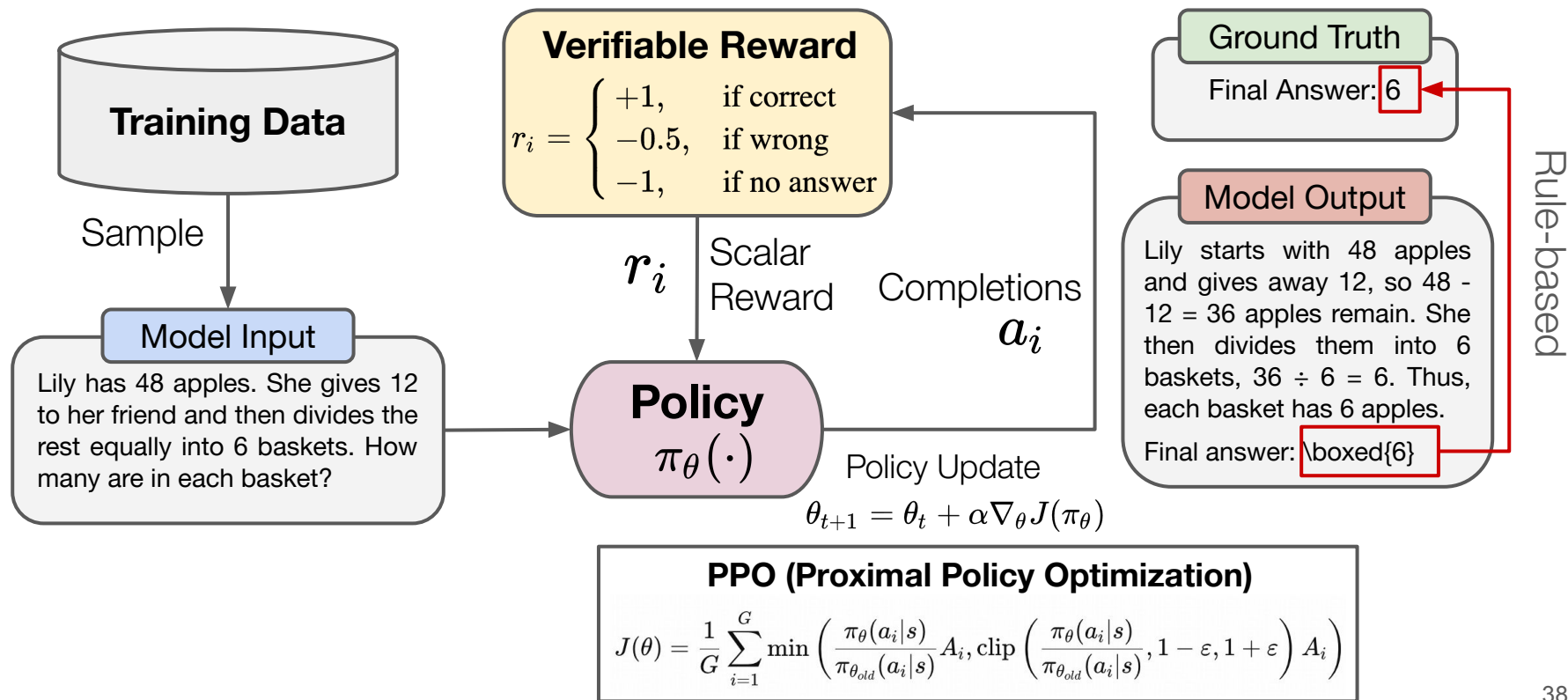
Answer: Lily starts with 48 apples and gives away 12, so $48 - 12 = 36$ apples remain. **Wait, let me verify it again:** 48 minus 12 indeed equals 36. She then divides them into 6 baskets, $36 \div 6 = 6$. **Let me double-check:** ... Thus, each basket contains 6 apples. The final answer: **\boxed{6}**



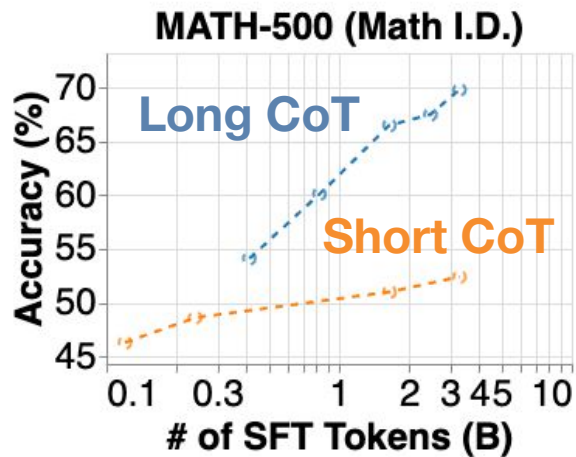
$$\mathcal{L}(\theta) = - \sum_{t=1}^T \log p_{\theta}(y^t | x, y^{<t})$$

Only calculate
the loss over y

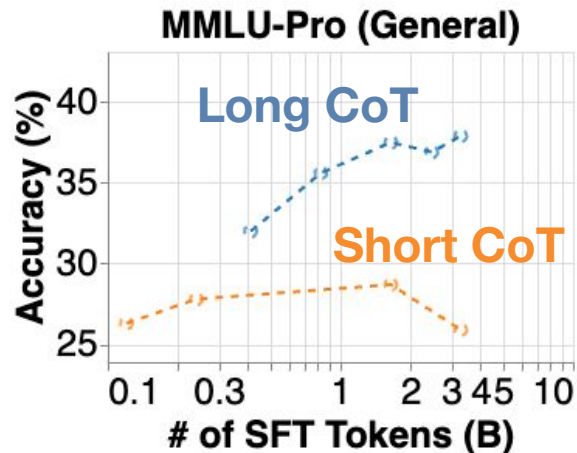
Reinforcement Learning (RL)



Effectiveness of Scaling Inference Compute



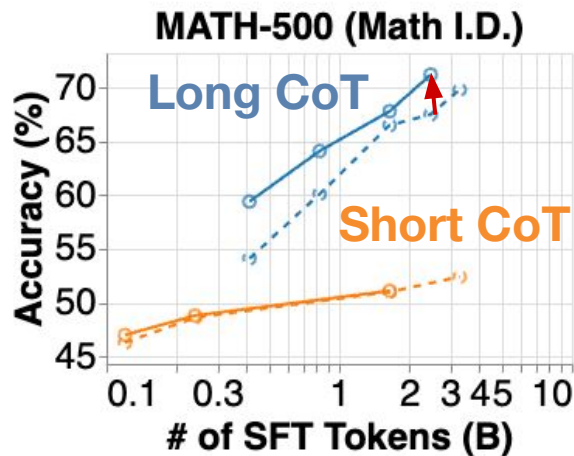
In-domain Math



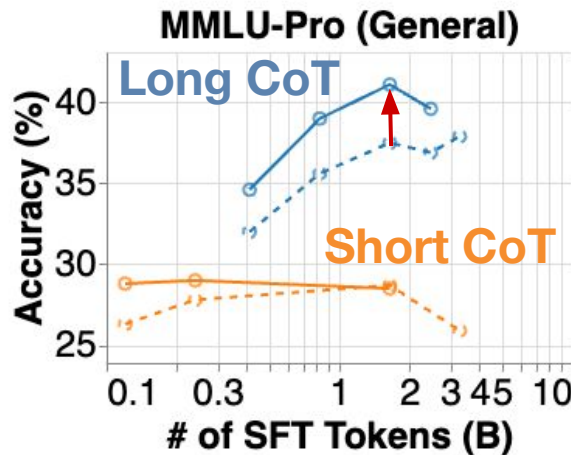
OOD General

Scaling up *inference* compute is effective:
Long CoT can scale up to a much higher upper limit than **short CoT**

RL Enables Better Accuracy and Generalization



In-domain Math



○ SFT + RL
○ SFT

OOD General

RL can often provide significant improvements beyond SFT, particularly in **long CoT scenarios for OOD evaluation tasks**.

Why long CoTs work better than short CoTs?

Long CoT Contains Behaviors Like:

So the user is requesting a bash script that can take a string representing a matrix, such as '[1,2],[3,4],[5,6]' and output its transpose, in the same format.

Clarify intents and break down into small steps

Alternatively, perhaps combine the numbers in some way.

Alternatively, think about their positions in the alphabet.

Alternatively, perhaps the letters are encrypted via a code.

Trying multiple strategies

Let's list them properly.

Wait, earlier I missed some letters there.

Let's re-express the sixth word letters:

mynznvaatzacdfoulxxz

so $48 - 12 = 36$ apples remain.

Wait, let me verify it again: 48 minus 12 indeed equals 36. She then divides them into 6 baskets, $36 \div 6 = 6$.

Let me double-check: ...

Error validation and correction

Reinforcement Learning from Pre-trained LLMs

An interesting “*aha moment*” of an intermediate version of **DeepSeek-R1-Zero**

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...



RQ2: What training factors influence reasoning ability?

1. How do long CoTs compared to short CoTs
- 2. Where does reasoning ability come from?**
3. How does model size affect reasoning ability?

Our hypothesis: the base model might already acquire such an ability during **pre-training**

Search “Aha” Phrases in Pre-training Corpus

“Aha” Phrases

- "Let's think step by step."
- **"Alternatively, ..."**
- "Breaking it down step by step..."
- "Thinking about it logically, first..."
- "Step 1: Let's figure out the starting point."
- "If we follow the steps carefully, we get..."
- "To solve this, let's analyze it piece by piece."
- "Going through this systematically, we have..."
- "Okay, let's solve this gradually."
- **"Does that make sense?"**
- **"Is this correct?"**
- "Wait, does that check out?"
- "Wait, actually..."
- "Oh, hold on..."
- **"Wait a second..."**
- **"Actually, let me rethink that."**
- "Hmm, let me go back for a moment."



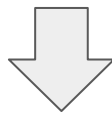
Pre-training
Corpus



COMMON
CRAWL





OpenWebMath



Search over the web
pre-training corpus

Where Does the “aha” Ability Come From?

 Interpretation of multilevel parameters
General ■ brms

 martinmodrak Stan Developer Feb 2021

Tiny:

So, are you basically stating that I can “forget” this dimensionality property in the first instance as a sort of analysis phase to see how parameters behave, and come back to the right dimensionality once it is decided how to use the results?

I am not sure I follow your thought here, but maybe that's just because I would have worded it differently? I definitely don't think “forget” is the right word. It is good to be aware of what your parameters mean. My point was more like: “OK, so we have interdependent parameters, so maybe focus less on each parameter separately and rather look what they all together imply about the world”. Kind of like if you modelled speed and capacity of a vehicle, but were really interested in how long does it take to transport a pile of stuff - looking at each parameter separately tells you something, but it is really hard to interpret without the other parameter. The model as a whole however contains enough information to answer your question.

An alternative approach would be to try to find a different parametrization of the model where the parameters are interpretable separately but that might be hard. And in the end, you will not get any new information, just a different rephrasing, so if you are not having any problems with fitting, I think it is unlikely to be very helpful.

Also, if this is the parametrization of the process used by many in the field, than maybe people would expect you to report as $(\frac{L}{n\omega})^{n-1}s^{-1}$, because that's what everybody has been doing (although possibly with fixed n)?

Does that make sense?

<https://discourse.mc-stan.org/t/interpretation-of-multilevel-parameters/20846>

So the question is then to find the right prediction task, looking at your setup, those may include:

...

... I am not sure I follow your thought here, but **maybe that's just because I would have worded it differently?**

... **An alternative approach** would be to try to find a different parametrization of the model where the parameters are interpretable separately, **but that might be hard.**

Does that make sense?

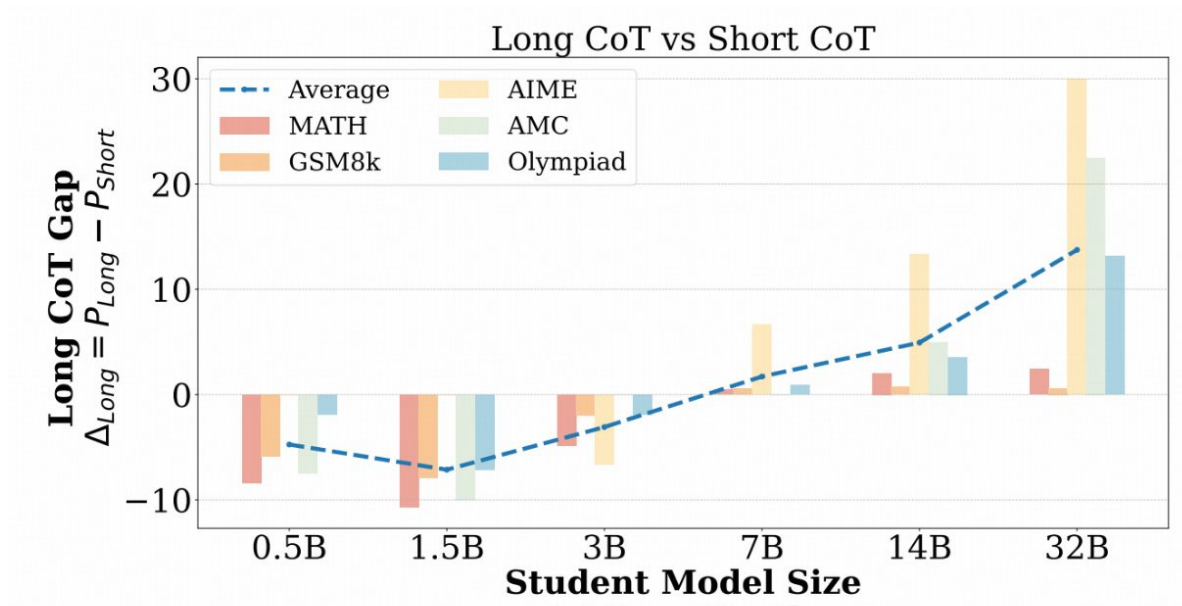
The base model might already acquire such skills during **pre-training. RL reinforces and increases the frequency of these patterns.**



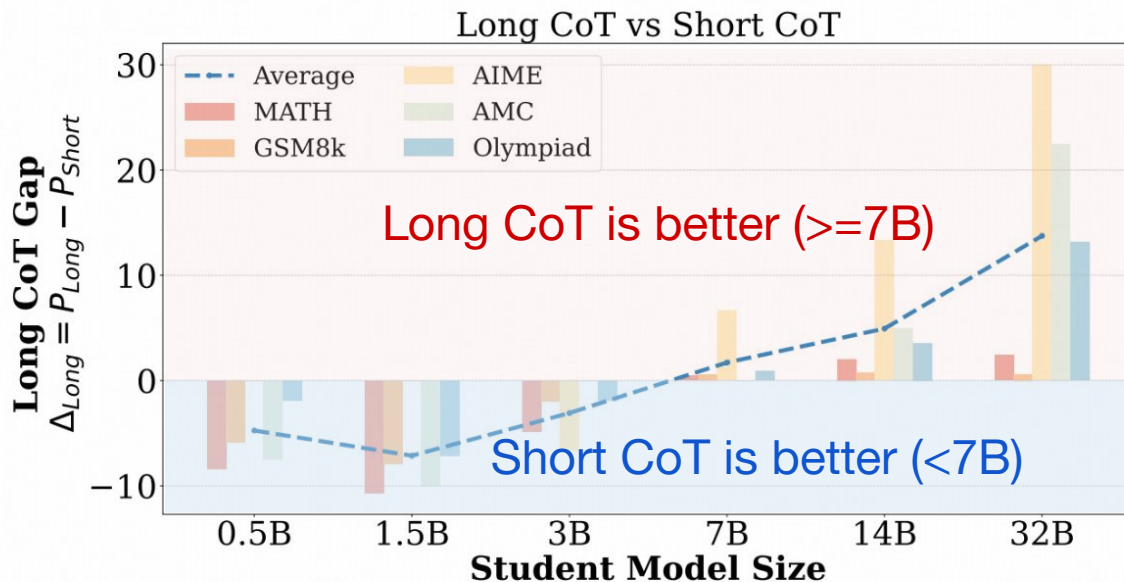
RQ2: What training factors influence reasoning ability?

1. How do long CoTs compared to short CoTs
2. Where does reasoning ability come from?
- 3. How does model size affect reasoning ability?**

Model Size Impacts the Reasoning Learning Ability



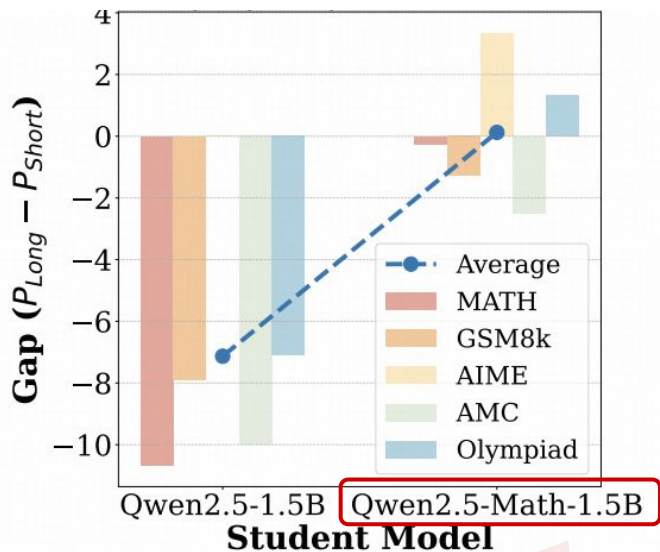
Model Size Impacts the Reasoning Learning Ability



Small student models tend to benefit more from **short CoT**, while **large student** models gain greater advantages from **long CoT**

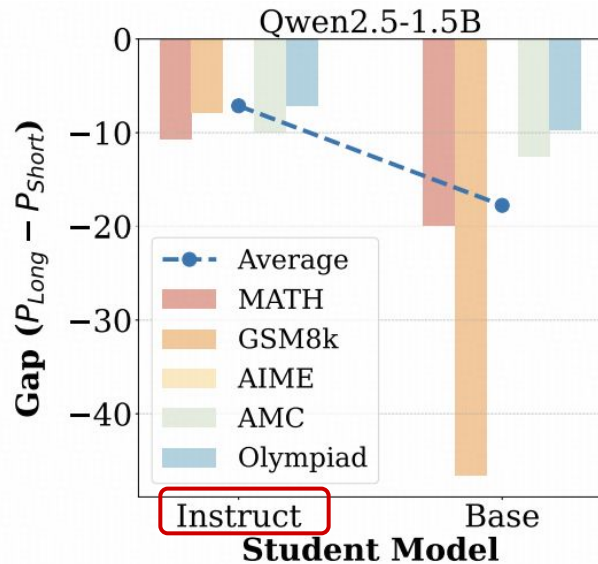
What Factors Lead to Learning Ability Gap?

Domain-specific Knowledge



Pre-trained on large-scale **math** corpus (>100B tokens)

Instruction Following Ability



Fine-tuned on high-quality general and math **instructions**



Summary: Understanding Reasoning

Building rigorous **benchmarks** to expose models' limitations

Controlled experiments in **ablation study** and **probing analysis**

hourmovers

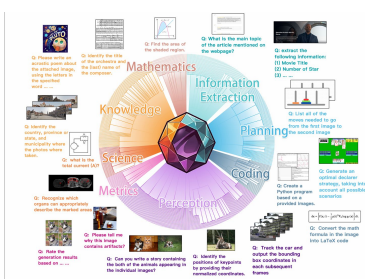
Table 23: Each of the following situations relates to a different company. (Image 1) For company D, find the missing amounts.

| | Company A | Company B | Company C | Company D |
|---------------------------------------|-------------|-----------|-------------|-----------|
| 1. Revenue | \$1,400,000 | \$100,000 | \$1,004,114 | |
| 2. Expenses | \$400,400 | 1,130,000 | 76,320 | ? |
| 3. Costs | 0 | ? | 4,120 | \$205 |
| 4. Profit (or loss) as a % of revenue | 32.70% | 3.00% | 30.35% | 20.35% |

Options:

| | |
|-----------------|-----------------|
| (A) \$1,001,204 | (F) \$1,130,594 |
| (C) \$1,004,320 | (H) \$1,200,320 |
| (E) \$1,130,320 | (I) \$1,130,732 |
| (G) \$1,130,732 | (J) \$1,004,630 |
| (B) \$1,070,732 | (K) \$1,000,630 |

View Available Computer Science Tutors



Engage with LLM in Debate

Let's have a conversation over the provided question and try to decide the correct answer together.

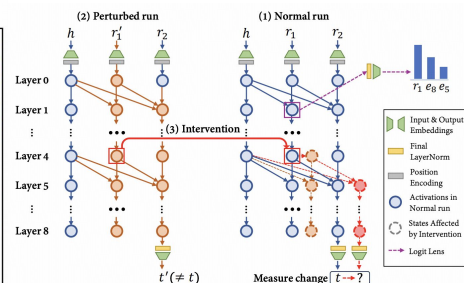
User: Henry and 3 of his friends order 7 pizzas for lunch. Each pizza is cut into 8 slices. If Henry and his friends want to share the pizzas equally, how many slices can each of them have?

LLM: There are 4 people sharing 7 pizzas, so there are $7 \times 8 = 56$ slices in total. If they want to share the pizzas equally, they can divide the slices by the number of people: $56 / 4 = 14$. Each of them can have 14 slices. The answer is 14.

User: Since there are 7 pizzas and each pizza is cut into 8 slices, the total number of pizza slices is 14. Henry and his 3 friends make a group of 4 people. So, each of them can have 4 slices. The answer is 4.

LLM: You are correct! I apologize for my mistake. Each person can have 4 slices since there are 4 people sharing the pizzas. Thank you for correcting me.

User: No problem at all! I'm glad that we reached an agreement on the final answer.



- MMMU [CVPR'24 Best Paper Finalist]
- MMLU-Pro [NeurIPS'24]
- MegaBench [ICLR'25]
- MixEval/MixEval-X [NeurIPS'24; ICLR'25]

- Evaluate via Debate [EMNLP'23 Findings]
- Grokked Transformers [NeurIPS'24]
- Small LMs Learning Ability Gap [arXiv'25]
- Demystify Long CoT [arXiv'25]

This Talk: Learning to Reason with LLMs



Understanding Reasoning

(~20min)



Improving Reasoning

(~20min)



Future Work

(~5min)



Inform better modeling



Understanding Reasoning

Benchmark and Ablation Study:

- MMMU [🏆CVPR'24 Best Paper Finalist]
- Demystify Long CoT [ICML'25 Submission]



Improving Reasoning

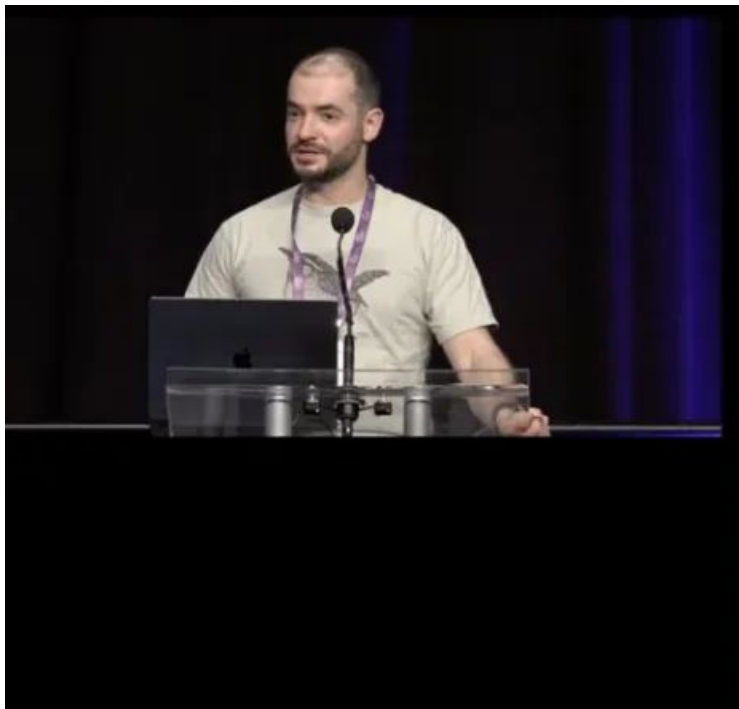
Math and STEM Reasoning:

- MAmmoTH-2 [NeurIPS'24d]
- Demystify Long CoT [ICML'25 Submission]

How does our understanding of reasoning contribute to the improvement?



Challenges in Pre-training Scaling



Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

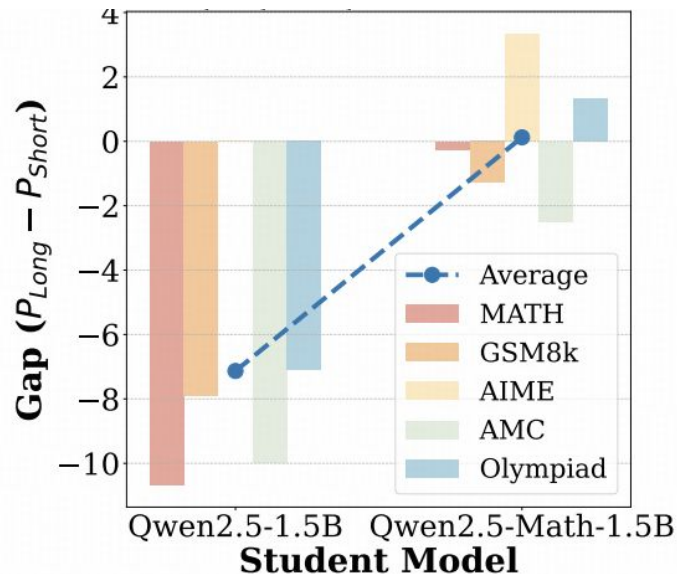
Ilya Sutskever (former chief scientist at OpenAI)'s Keynote at NeurIPS 2024



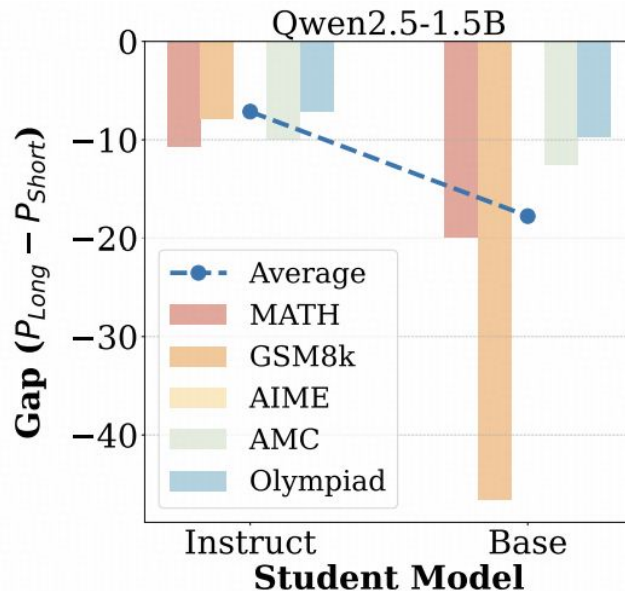
Improving LLM Reasoning

- **RQ1:** How to develop efficient and scalable post-training methods beyond pre-training?
- **RQ2:** How can RL reward shaping control LLM output (e.g., length) for better reasoning?

Recap: Factors Lead to Learning Ability Gap



Domain-specific Knowledge



Instruction Following Ability



Recap: Factors Lead to Learning Ability Gap

Domain-specific
Knowledge

Instruction
Following Ability



Two key factors for scaling in post-training

Solution 1: Human-annotated Data are Costly

Training Data

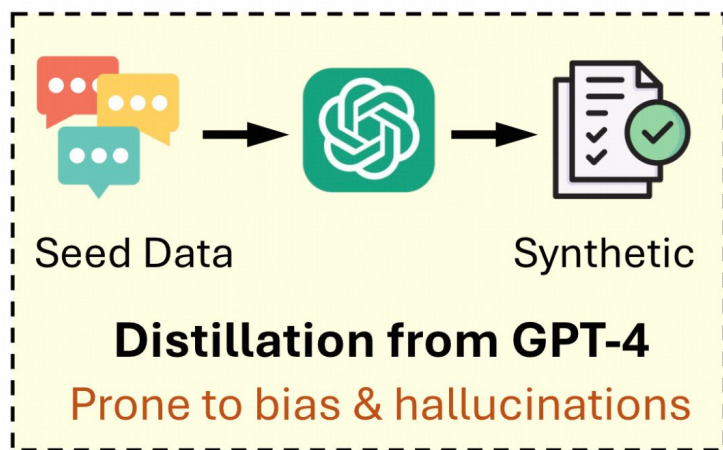


Overview Llama 3 was pretrained on over 15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 10M human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data.

| | |
|-----------------------------|-----------------|
| Reward per assignment | \$10.000 |
| Commission per assignment | \$4.000 |
| Total cost per assignment | \$14.000 |
| Total number of assignments | 10000000 |
| Total cost | \$140000000.000 |



Solution 2: Distilling from Proprietary Models is Error-Prone and Lacks Diversity



Instruction 1: A farmer is planting trees in a 24m×18m rectangular orchard. The trees are spaced 2m apart in both directions. How many trees can be planted?

Response 1: Calculate the number of trees along the length: $24/2=12$

✗ (Mistake: Forgot to add 1, should be 13.)

Calculate the number of trees along the width: $18/2=9$

✗ (Mistake: Forgot to add 1, should be 10.)

Multiply to find the total number of trees: $12 \times 9 = 108$.

Correctness of examples is not guaranteed

Instruction 2: A farmer is planting trees in a 30m × 20m...

Instruction 3: A gardener is planting bushes in a 30m×20m...

Instruction 4: A park manager is setting up benches in a 40m × 25m...

Instruction 5: A landscaper is placing flower pots in a 36m × 16m...

The generated queries lack diversity

Our Solution: *WebInstruct*

10M Synthetic Instruction Data from the Web

Physics Forums

Health & Medicine / Diseases in Health & Medicine / Inflammation

Are there infinite combination equilibrium constant?

zenterix · Wednesday, 3:2

Wednesday, 3:26 PM

zenterix

WOLFRAM COMMUNITY

Homework Statement: The chem

Balancing mi

PO: CON-6 (EU), CON-6.A (LO), C

In 1890, Louisiana enacted a black people and for white p only* train. He refused to mc was arrested. The case went ruled to uphold the Louisian

Which statement accurately Ferguson (1896) decision?

Do 4 proble

Outline the interventions in practice suppress inflammatory response c

Is there a way to show the factors of a Perfect Number expressed as a sum rather than a product

Bob Freeman, Retired

Posted 21 hours ago

HL

Normally, FactorInteger[] function outputs all the prime factors of a number as a list composed of the factors and their exponents, but I would like to see the output expressed as the sum of all the proper factors (all factors less than the value). Here are 3 examples that should help illustrate what I'm looking for:

1. FactorInteger[24] output would show (2, 3), (3, 1) but I would like to see it expressed as:

$24 = 2 + 2 + 4 + 7 + 24$

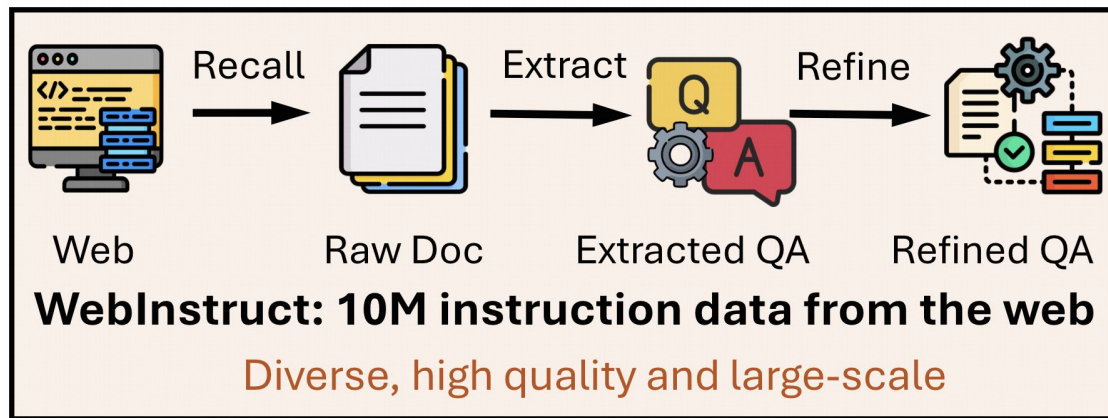
2. FactorInteger[240] output would show (2, 4), (3, 1), (5, 1), but I would like to see it expressed as:

$480 = 1 + 2 + 4 + 8 + 16 + 31 + 62 + 124 + 248$


3. FactorInteger[1225] output would show (2, 6), (127, 1), but I would like to see it expressed as:

$8128 = 2 + 2 + 4 + 8 + 16 + 32 + 64 + 127 + 254 + 508 + 1016 + 2032 + 4064$

I also just came across the Divisors[] function that creates a list of all possible divisors of a given value, which is quite close, but would really like to see the values expressed as a sum rather than a list of values. I also found the Divisors[] function that would remove the last element so that it's easy to see the entire list of proper divisors, but really would like to express it as a sum rather than the list. E.G.




WebInstruct: 10M Synthetic Instruction Data

 **Raw Docs** *Unformatted Text, Site Information, Ads*

Topics Science\nAnatomy&Physiology\nAstronomy\nAstrophysics
\nBiology\nChemistry \n...Socratic Meta...Featured Answers
How do you simplify $\frac{(u^4 v^3)}{(u^2 v^{-1})^4}^0$ 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\nWhat is the quotient of powers property?
\n\n\nHow 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](#)

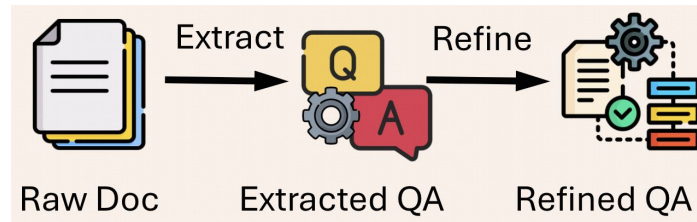
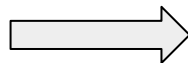



Extract

 **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.

Rewrite



 **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



10M Synthetic Data vs 10M Human Annotated Data

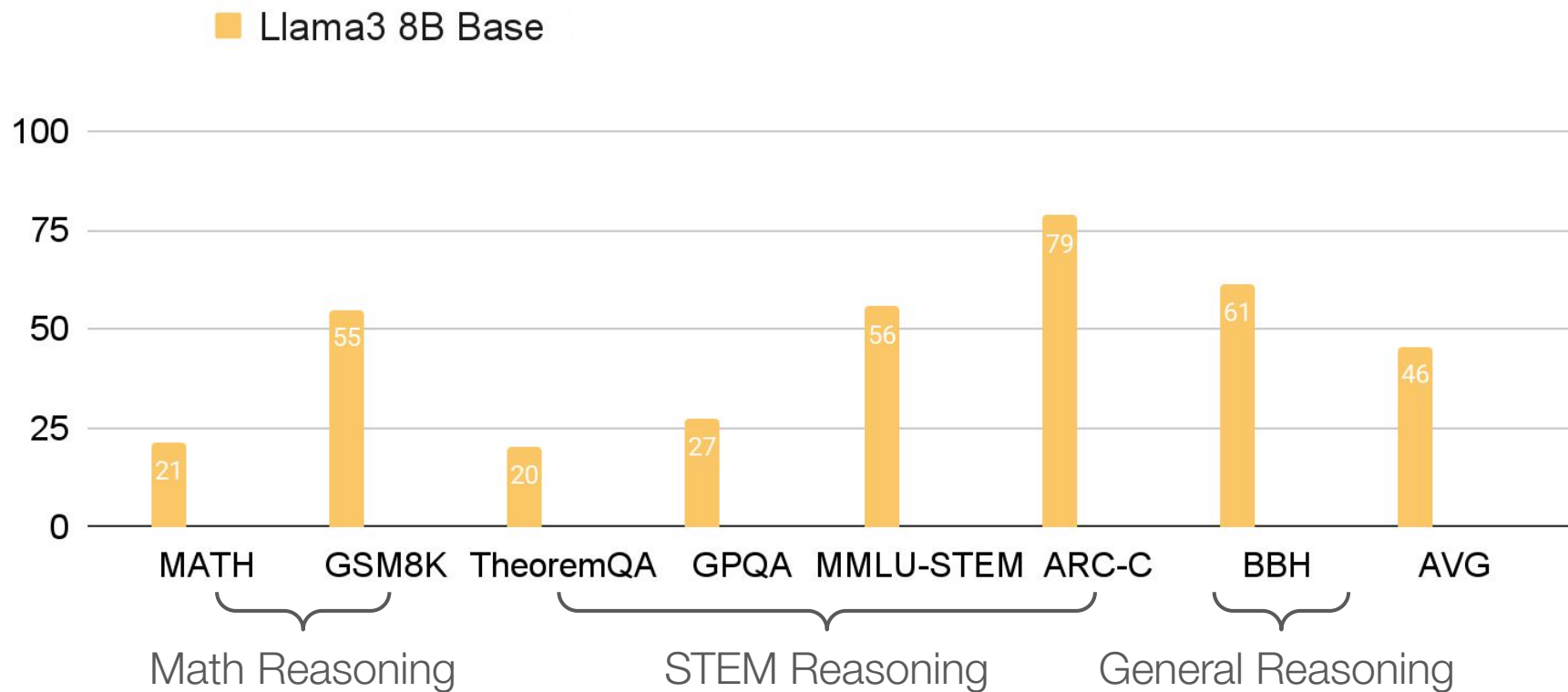
MATH GSM8K TheoremQA GPQA MMLU-STEM ARC-C BBH AVG

Math Reasoning STEM Reasoning General Reasoning

Yue, X., Zheng, T., Zhang, G., & Chen, W. (2024). MAMmoTH2: Scaling instructions from the web. **NeurIPS 2024**



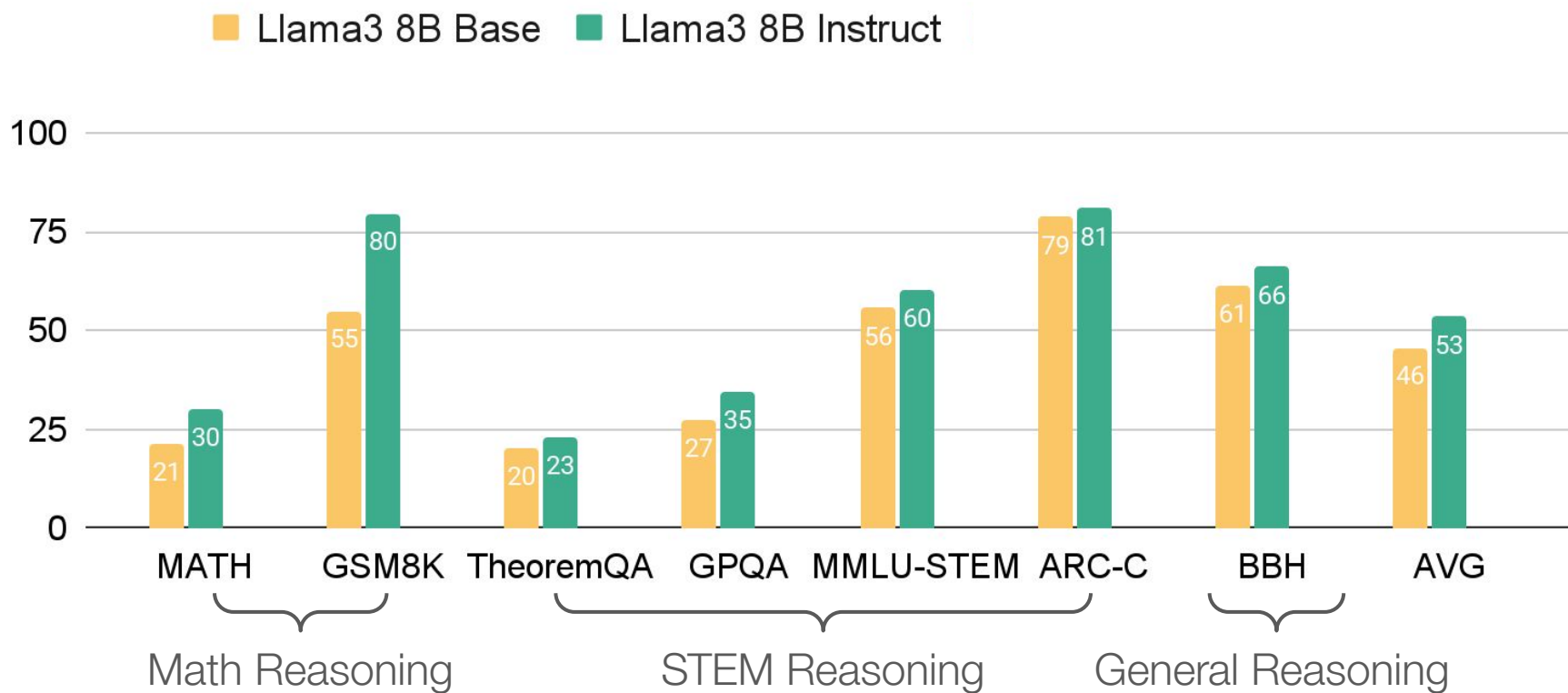
10M Synthetic Data vs 10M Human Annotated Data



Yue, X., Zheng, T., Zhang, G., & Chen, W. (2024). MAMmoTH2: Scaling instructions from the web. **NeurIPS 2024**



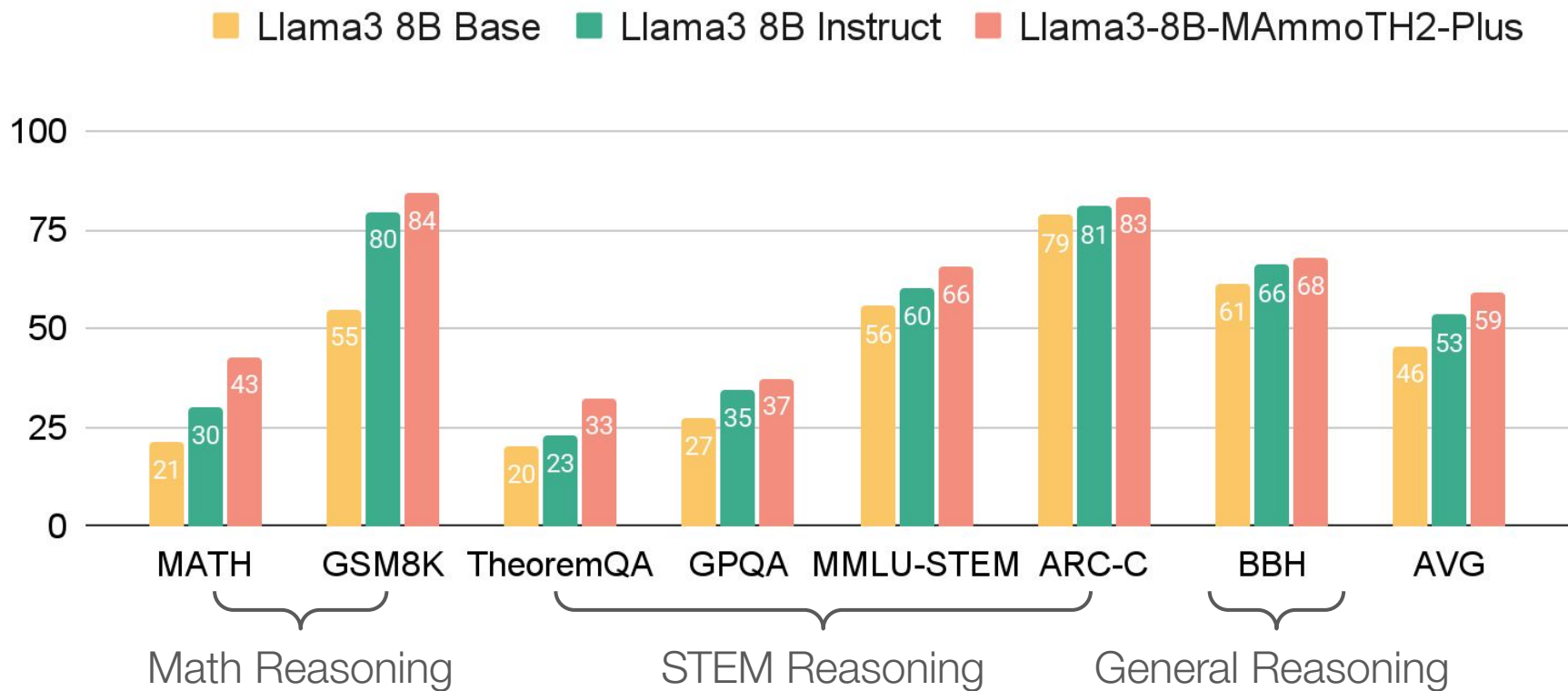
10M Synthetic Data vs 10M Human Annotated Data



Yue, X., Zheng, T., Zhang, G., & Chen, W. (2024). MAMmoTH2: Scaling instructions from the web. **NeurIPS 2024**



10M Synthetic Data vs 10M Human Annotated Data



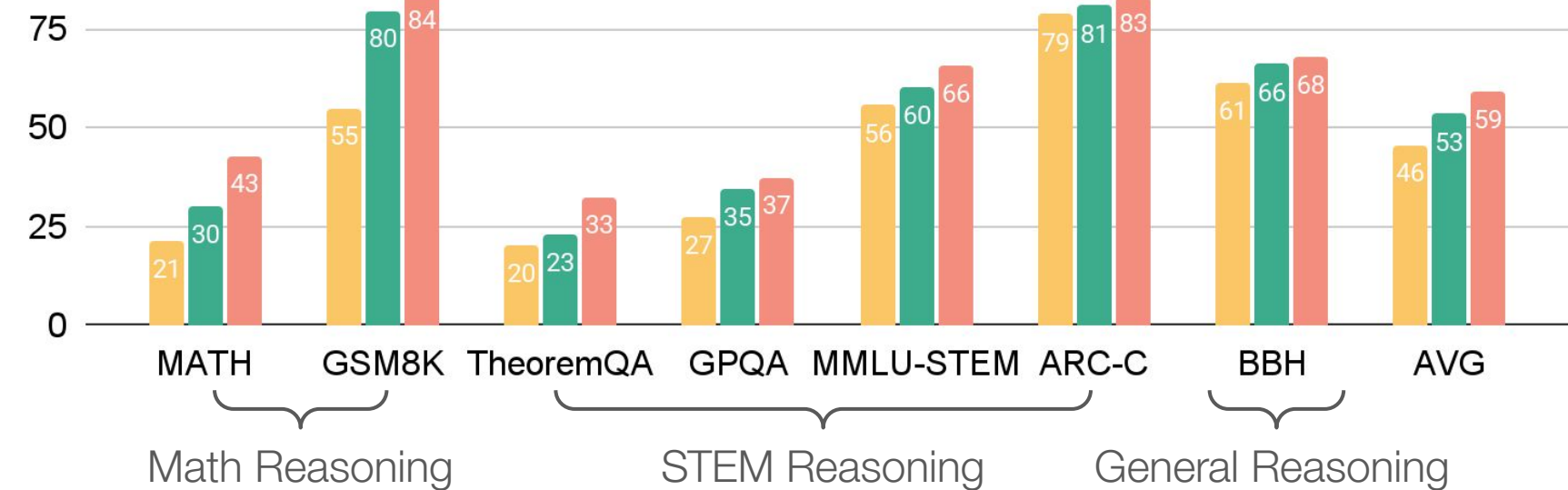
Yue, X., Zheng, T., Zhang, G., & Chen, W. (2024). MAMmoTH2: Scaling instructions from the web. **NeurIPS 2024**



10M Synthetic Data vs 10M Human Annotated Data

■ Llama3 8B Base ■ Llama3 8B Instruct ■ Llama3-8B-MAMmoTH2-Plus

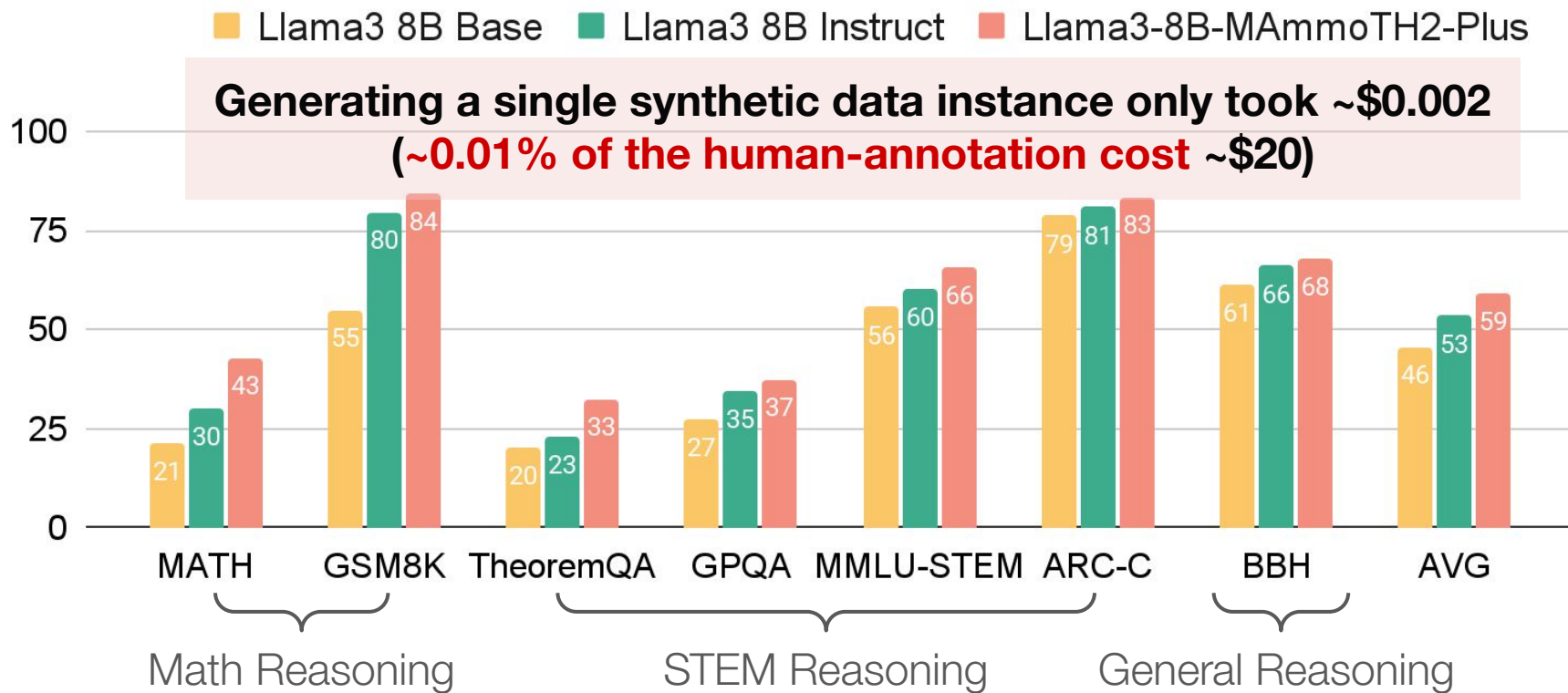
MAMmoTH2 outperforms Llama3-Instruct with the same base model



Yue, X., Zheng, T., Zhang, G., & Chen, W. (2024). MAMmoTH2: Scaling instructions from the web. **NeurIPS 2024**



10M Synthetic Data vs 10M Human Annotated Data



Yue, X., Zheng, T., Zhang, G., & Chen, W. (2024). MAMmoTH2: Scaling instructions from the web. **NeurIPS 2024**



Impact: MAMmoTH Series



The Llama 3 Herd of Models

Llama Team, AI @ Meta¹

¹A detailed contributor list can be found in the appendix of this paper.

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech capabilities into Llama 3 via a compositional approach. We observe this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.

Date: July 23, 2024

Website: <https://llama.meta.com/>

4.3.3 Math and Reasoning

We define reasoning as the ability to perform multi-step computations and arrive at the correct final answer. Several challenges guide our approach to training models that excel in mathematical reasoning:

- **Lack of prompts:** As the complexity of questions increases, the number of valid prompts or questions for Supervised Fine-Tuning (SFT) decreases. This scarcity makes it difficult to create diverse and representative training datasets for teaching models various mathematical skills (Yu et al., 2023; Yue et al., 2023; Luo et al., 2023; Mitra et al., 2024; Shao et al., 2024; Yue et al., 2024b).

MAMmoTH-1

MAMmoTH-2

Improving Reasoning



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao^{1,2,*}, Peiyi Wang^{1,3,*}, Qihao Zhu^{1,3,*}, Runxin Xu¹, Junxiao Song¹, Xiao Bi¹, Haowei Zhang¹, Mingchuan Zhang¹, Y.K. Li¹, Y. Wu¹, Daya Guo^{1*}

¹DeepSeek-AI, ²Tsinghua University, ³Peking University

{zhihongshao, wangpeiyi, zhuqh, guoday}@deepseek.com
<https://github.com/deepseek-ai/DeepSeek-Math>

3.1. SFT Data Curation

We construct a mathematical instruction-tuning dataset covering English and Chinese problems from different mathematical fields and of varying complexity levels: problems are paired with solutions in chain-of-thought (CoT) (Wei et al., 2022), program-of-thought (PoT) (Chen et al., 2022; Gao et al., 2023), and tool-integrated reasoning format (Gou et al., 2023). The total number of training examples is 776K.

MAMmoTH-1

- **English mathematical datasets:** We annotate GSM8K and MATH problems with tool-integrated solutions, and adopt a subset of MathInstruct (Yue et al., 2023) along with the training set of Lila-OOD (Mishra et al., 2022) where problems are solved with CoT or PoT. Our English collection covers diverse fields of mathematics, e.g., algebra, probability, number theory, calculus, and geometry.

Impact: MAMmoTH Series



- Cited and used in the post-training stage of leading industrial LLMs
- Scaling synthetic rationale generation post-training

Granite Code Models: A Family of Open Foundation Models for Code Intelligence

Mayank Mishra* Matt Stallone* Gaoyuan Zhang* Yikang Shen Aditya Prasad
Adriana Meza Soria Michele Merler Parameswaran Selvam Saptha Surendran
Shivdeep Singh Manish Sethi Xuan-Hong Dang Pengyuan Li Kun-Lung Wu
Syed Zawad Andrew Coleman Matthew White Mark Lewis Raju Pavuluri



DotaMath: Decomposition of Thought with Code Assistance and Self-correction for Mathematical Reasoning

Chengpeng Li^{1,2}, Guanting Dong^{2*}, Mingfeng Xue^{2*}, Ru Peng^{2*}, Xiang Wang¹
Dayiheng Liu^{2†}

¹University of Science and Technology of China

²Alibaba Group.

{lichengpeng.lcp, liudayiheng.ldyh}@alibaba-inc.com

ON DESIGNING EFFECTIVE RL REWARD AT TRAINING TIME FOR LLM REASONING

Jiaxuan Gao^{1,2,*} Shusheng Xu^{1,2,*} Wenjie Ye³ Weilin Liu³ Chuyi He³

Wei Fu^{1,2} Zhiyu Mei^{1,2} Guangju Wang² Yi Wu^{1,2,3,†}

¹ Institute for Interdisciplinary Information Sciences, Tsinghua University

² Shanghai Qi Zhi Institute

³OpenPsi Inc.

{samjia2000, xssstory, jxwuyi}@gmail.com

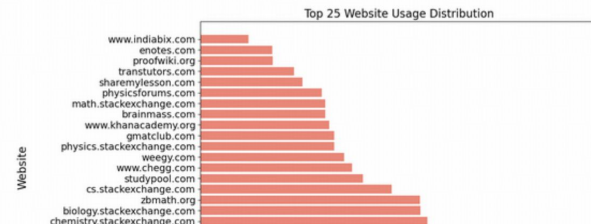
- Inspire future work for reasoning with programming
- Open models and datasets are adopted as baselines for further tuning (e.g., RL)

2023, year of open LLMs

MARKTECHPOST

Home AI Research News New Releases Open Source AI AI P

MAMmoTH2: Scaling Instructions from the Web



Papers Explained 231: MAMmoTH2



Ritvik Rastogi · Follow

4 min read · Oct 14, 2024



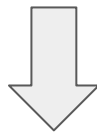
20



- Covered in various blogs



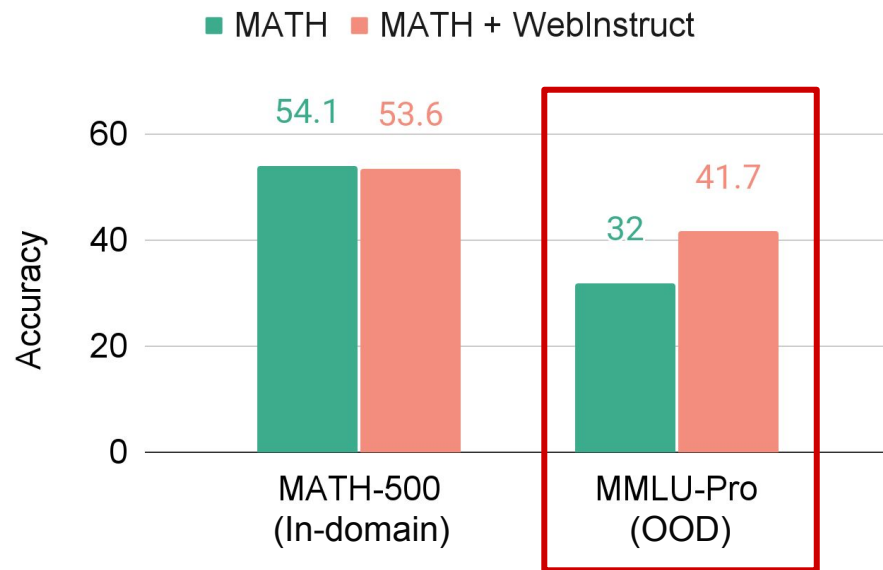
Verifiable reward signal is the key to
improve LLM reasoning with RL



Can we leverage WebInstruct for RL training?

Scaling up Verifiable Reward in RL with WebInstruct

WebInstruct achieves better *generalization* in OOD reasoning scenario

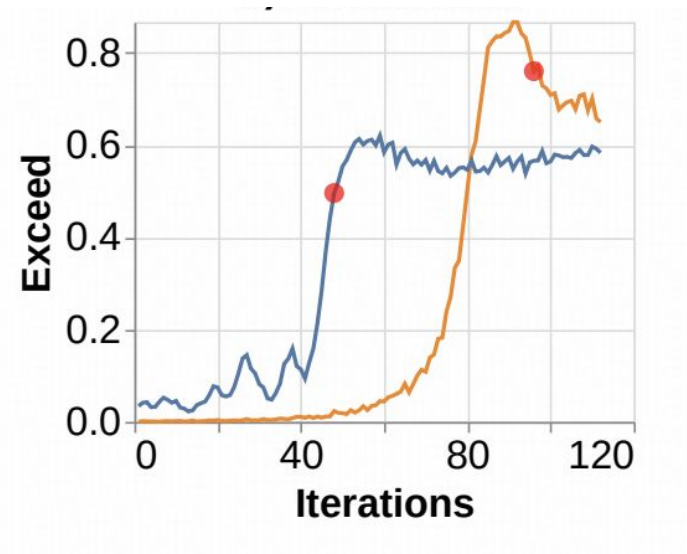
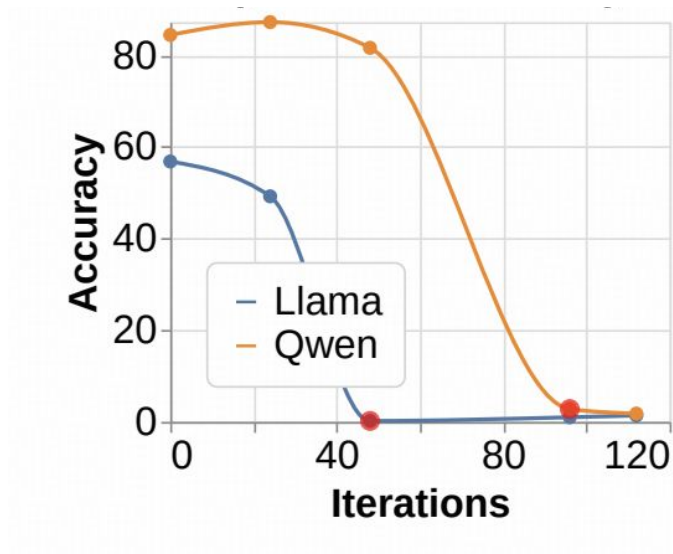


Improving LLM Reasoning

- **RQ1:** How to develop efficient and scalable post-training methods beyond pre-training?
- **RQ2:** How can RL reward shaping control LLM output (e.g., length) for better reasoning?

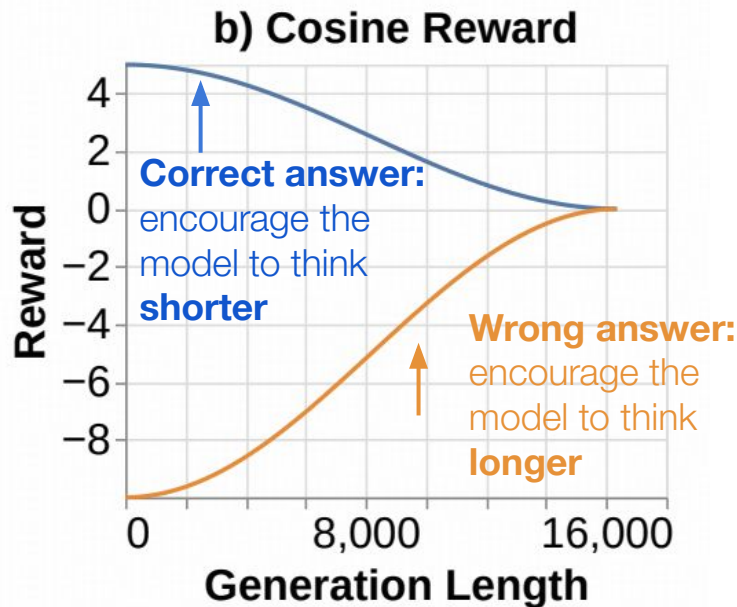
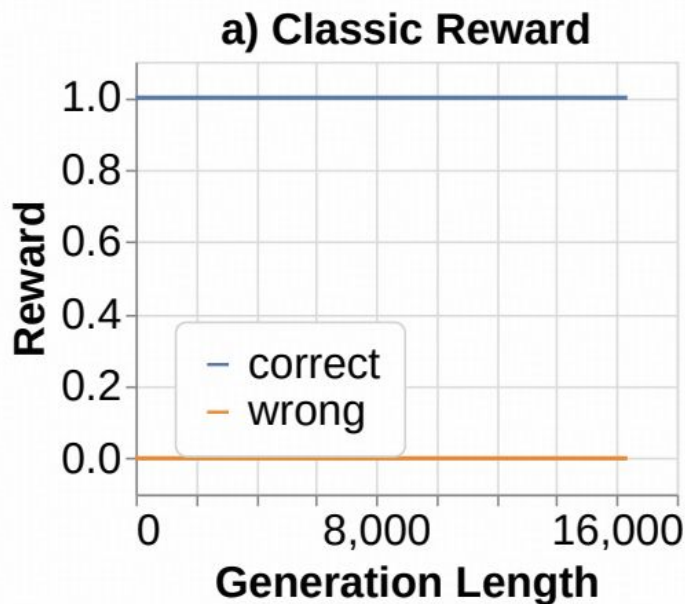


💥 The length of CoT During RL Can Be Exploded!



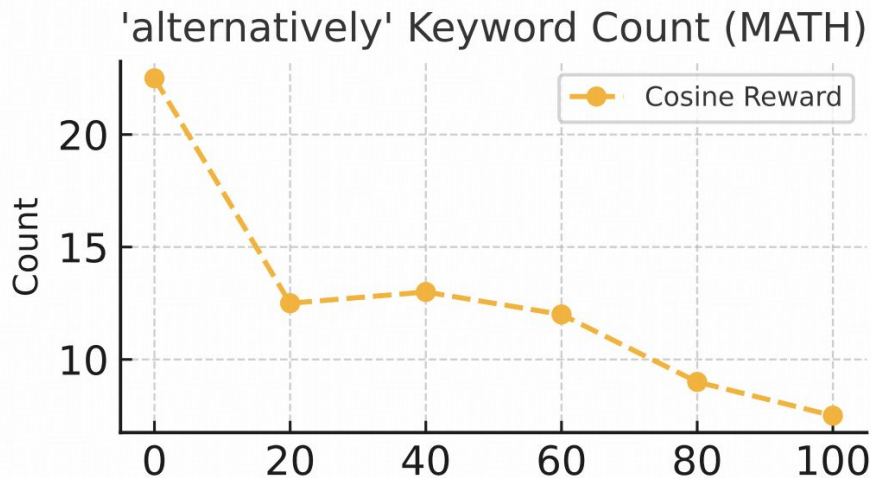
Yeo*, E., Tong*, Y., Niu, M., Neubig, G., & **Yue, X.** (2025). Demystifying Long Chain-of-Thought Reasoning in LLMs. *arXiv 2025*. (*: my advisee)

Impact of Reward Design on Long CoT



We propose a **cosine reward** to control the length scaling!

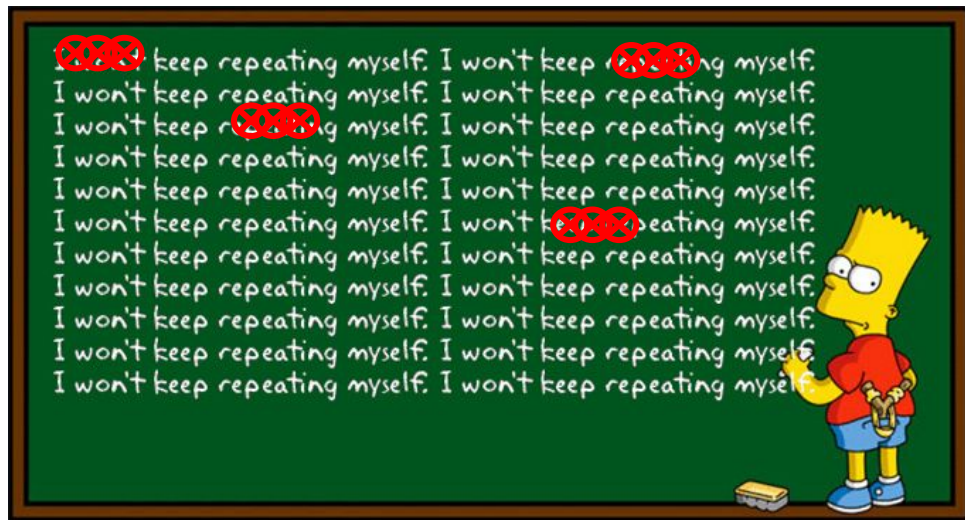
Length Reward Hacking



Length scaling cosine reward can result in **reward hacking** like *repeating* to mitigate penalty instead of exploration which can show as **a drop in CoT branching frequency**



Our Solution: Cosine Reward + Repetition Penalty



Penalize repeating tokens!

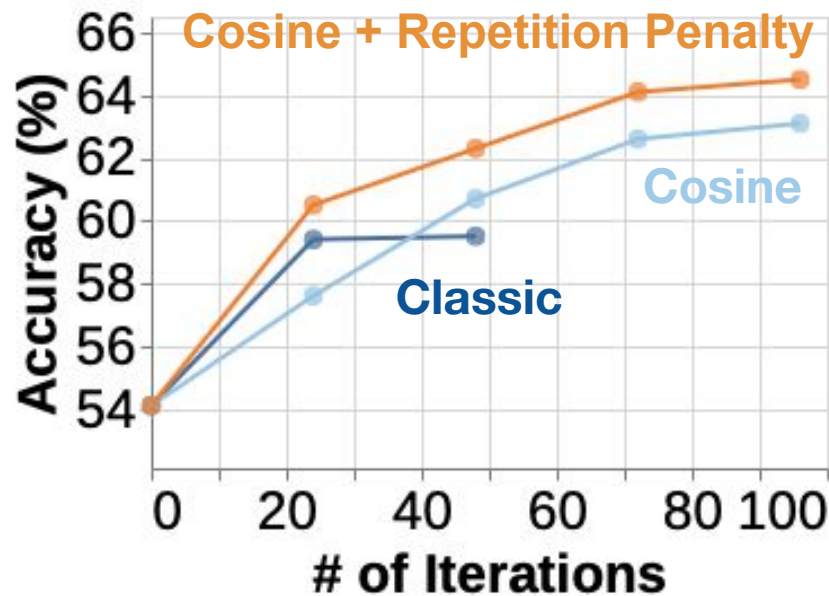
- Considering:

Repeating is usually local behaviors, for which other tokens are seldom responsible.

- Implementation:

Dense rewards with discounting based on N-gram repetition detection

Impact of Reward Design on Long CoT

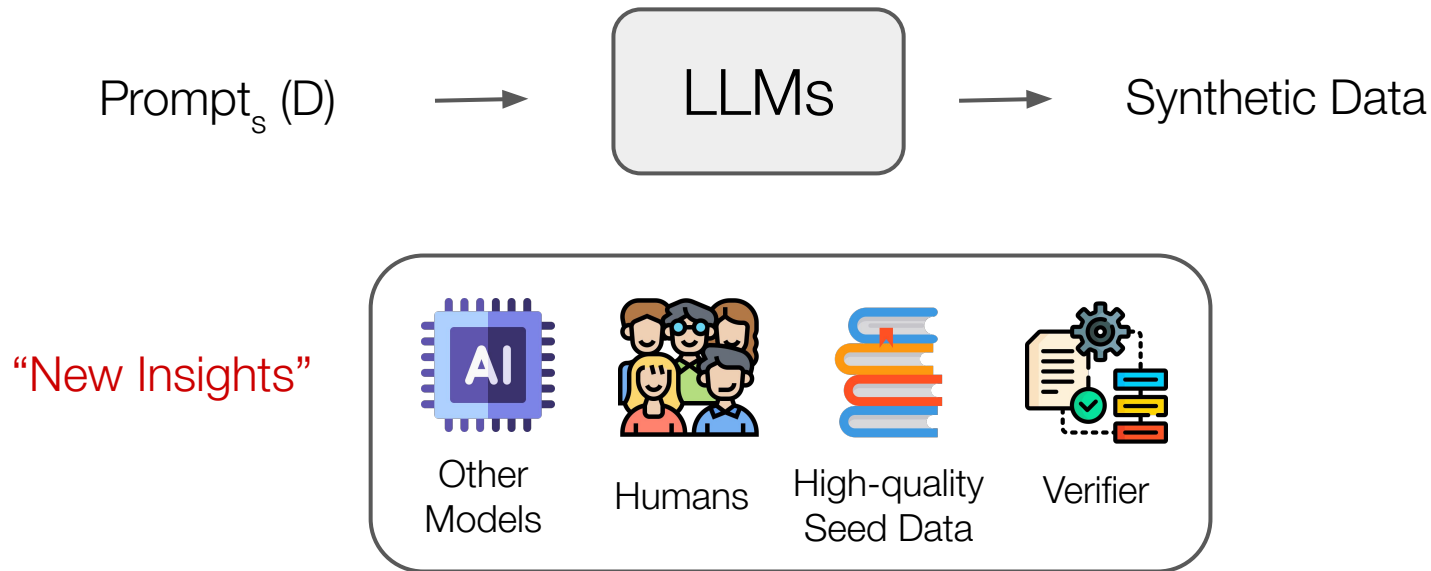


Our **reward shaping** improves model performance on downstream tasks

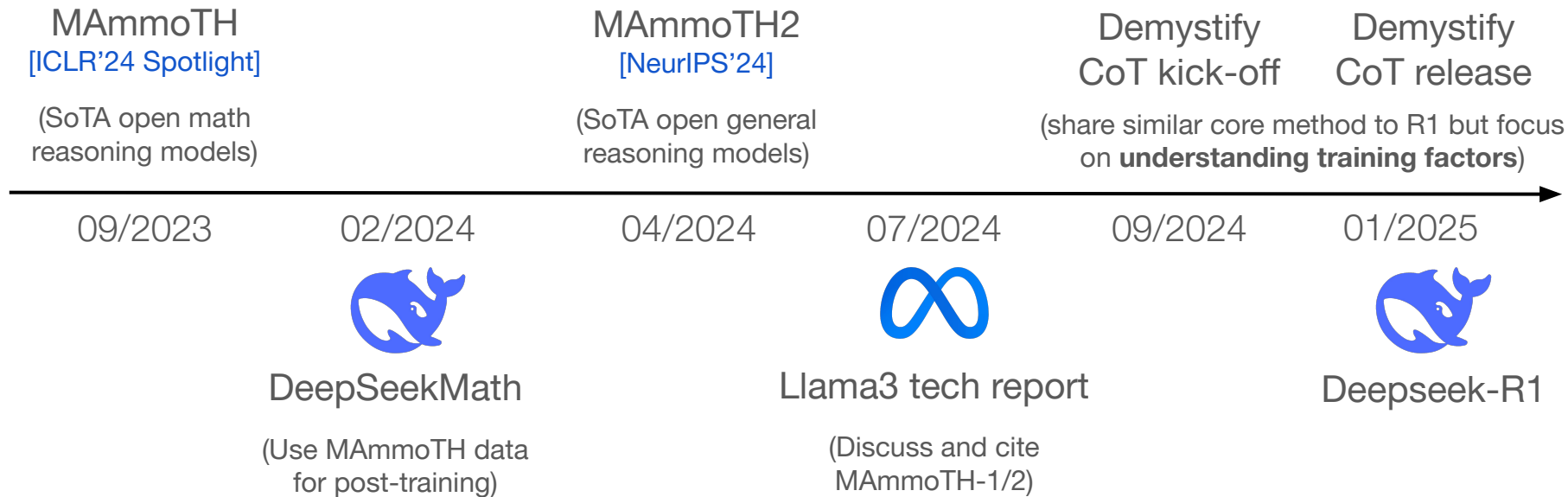


Summary: Improving Reasoning

We see the power of “synthetic” data in improving the reasoning capability of LLMs



Summary: Timeline of Recent Reasoning Projects



My Research



Understanding Reasoning

Benchmarks:

- MMMU [🏆 CVPR'24 Best Paper Finalist]
- MMLU-Pro [NeurIPS'24a]
- MixEval(-X) [ICLR'25 Spotlight]
- MegaBench [ICLR'25b]

Probing & Ablation Study:

- Grokked Transformers [NeurIPS'24c]
- AI Debate [EMNLP'23 Findings]
- Demystify Long CoT [arXiv'25]



Improving Reasoning

Math and STEM Reasoning:

- MAMmoTH [ICLR'24 Spotlight]
- MAMmoTH-2 [NeurIPS'24d]
- Demystify Long CoT [arXiv'25]

Multimodal Reasoning:

- MAMmoTH-VL [arXiv'24]
- MultiUI [ICLR'25c]

Code Reasoning:

- OpenCodeInterpreter [ACL'24 Findings]



Responsible LMs

Multilinguality:

- Pangea [ICLR'25d]
- JMMMU [NAACL'25]

Privacy and Security:

- Machine Unlearning [ACL'24]
- Differential Privacy [ACL'21; CCS'23; 🏆 ACL'23 Best Paper Honorable Mention]

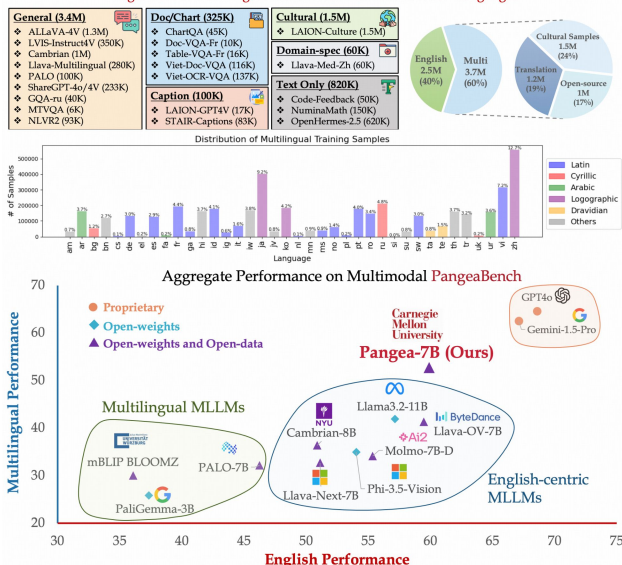
AI for Healthcare:

[ACL'20; 🏆 BIBM'21 Best Paper; arXiv'24]

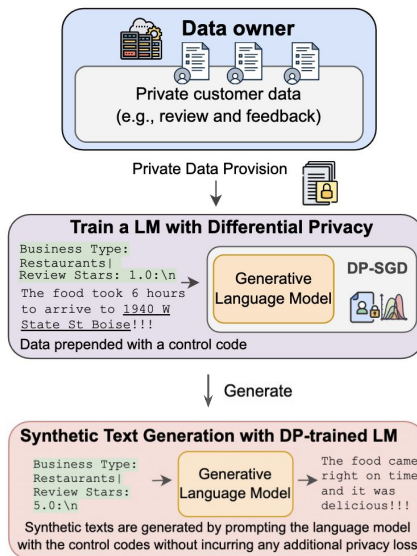
Building Responsible LMs

Multilinguality

PangeaIns: 6M Multilingual Multimodal Instructions for 39 Languages



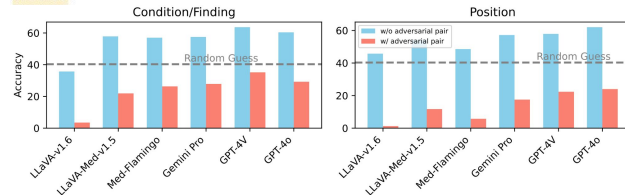
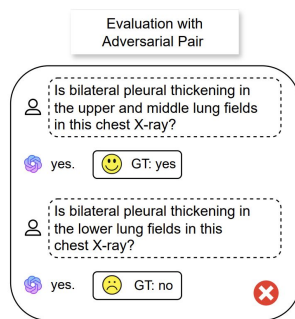
Privacy



AI in Healthcare



Caption: Chest X-ray showing bilateral pleural thickening in the upper and middle lung fields.



Pangea, a fully open-source multimodal LLM achieving state-of-the-art performance across 14 benchmarks in English and multilingual settings.

Xiang Yue et al., Pangea: A fully open multimodal multimodal llm for 39 languages. *ICLR 2025*

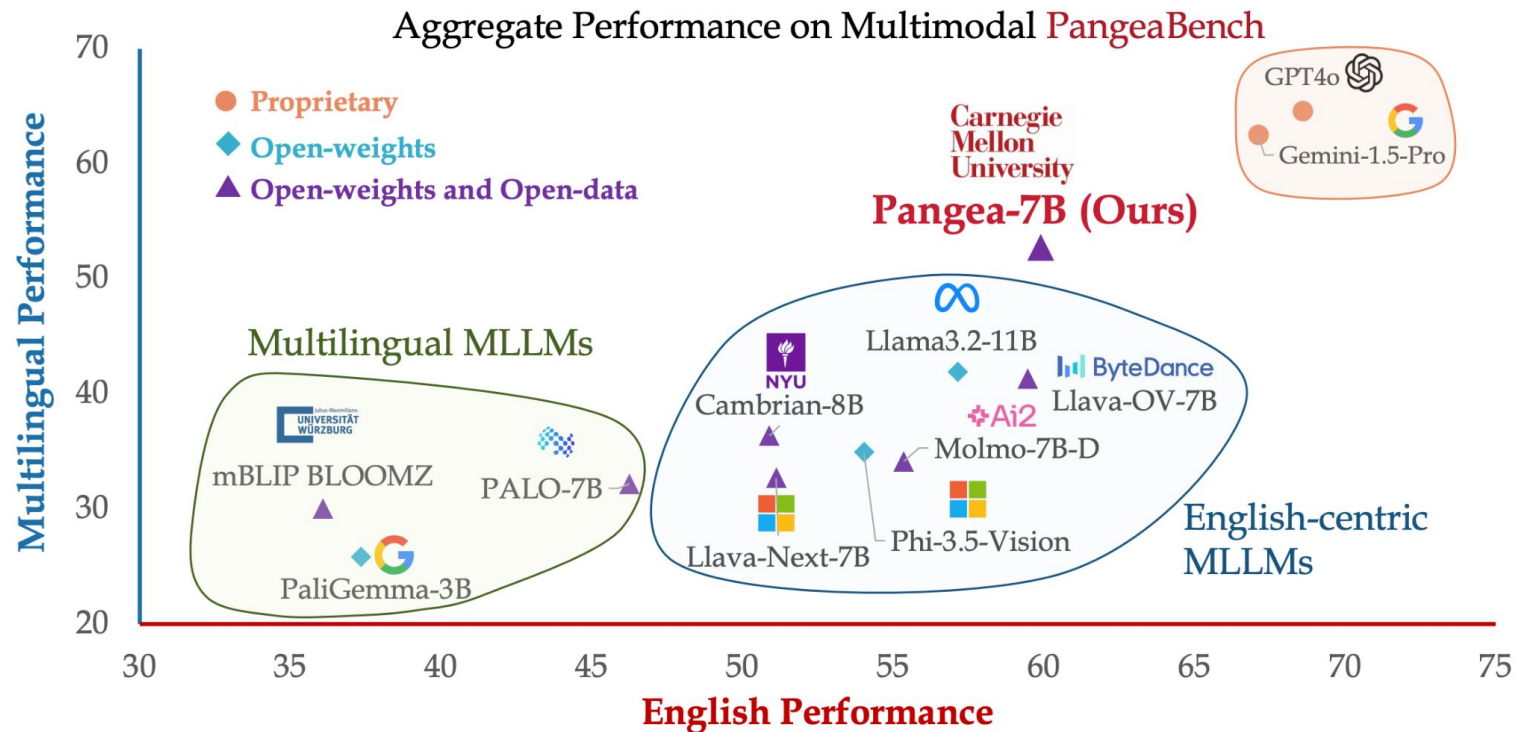
A simple recipe of differentially-private (DP) fine-tuning LMs creates synthetic text with strong privacy and high utility.

Xiang Yue et al., Synthetic text generation with differential privacy: A simple and practical recipe. *ACL 2023 Best Paper Honorable Mention*

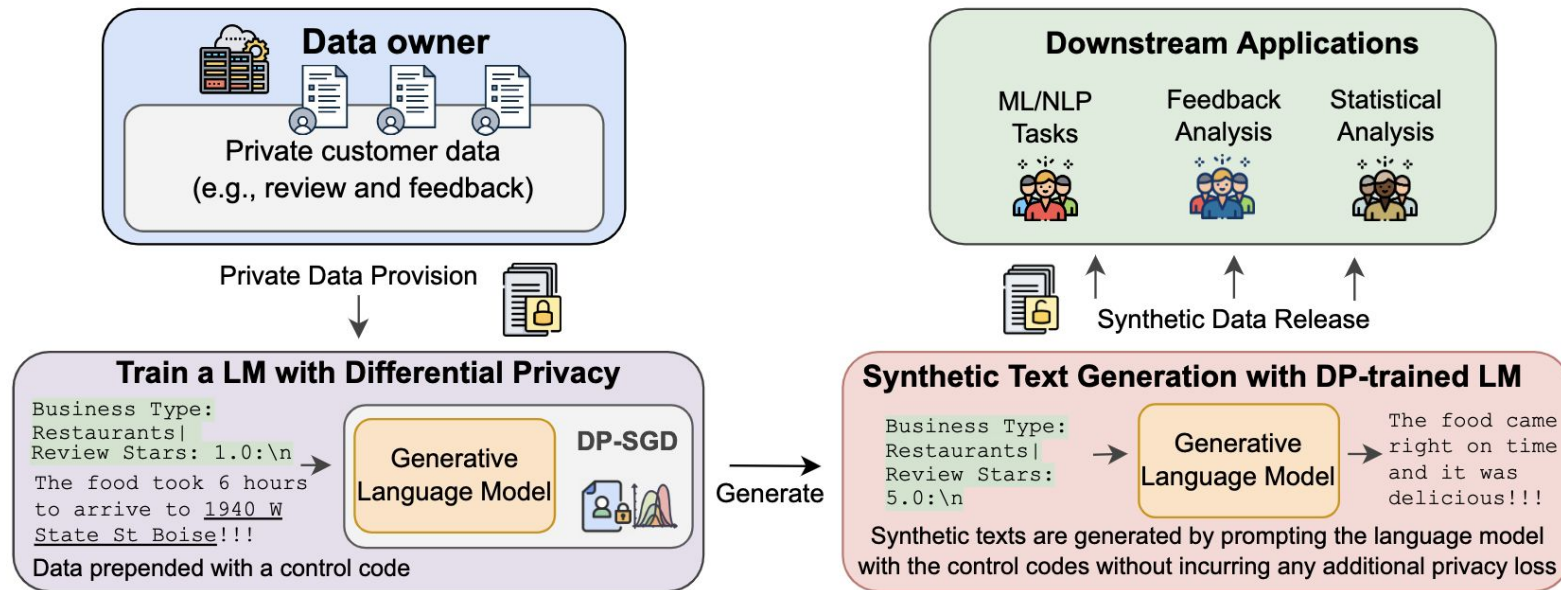
Probing for medical diagnosis: frontier models like GPT-4o perform worse than random guessing on specialized diagnostic questions

Yan, Qianqi, Xuehai He, **Xiang Yue**, and Xin Eric Wang. "Worse than random? An embarrassingly simple probing evaluation of large multimodal models in medical VQA." *arXiv 2024*

Pangea: SOTA fully-open multilingual multimodal LLM



Synthetic Data Generation with Privacy



We *fine-tune language models with Differential Privacy (DP)* and then leverage it for synthetic text generation with prompts





Best Paper Honorable Mention

Open the research direction of differentially-private text generation with LMs

Differentially Private Synthetic Data via Foundation Model APIs 2: Text

Chulin Xie¹ Zinan Lin² Arturs Backurs² Sivakanth Gopi² Da Yu³ Huseyin Inan² Harsha Nori²
Haotian Jiang² Huishuai Zhang² Yin Tat Lee² Bo Li^{1,4} Sergey Yekhanin²

chulinx2@illinois.edu, {zinanlin, arturs.backurs, sivakanth.gopi, huseyin.inan, hanori, haotianjiang, huishuai.zhang, yintatlee, yekhanin}@microsoft.com, yuda3@mail2.sysu.edu.cn, bol@uchicago.edu

Private Synthetic Text Generation with Diffusion Models

Sebastian Ochs¹ and Ivan Habernal²

Trustworthy Human Language Technologies

¹ Department of Computer Science, Technical University of Darmstadt

² Research Center Trustworthy Data Science and Security of the University Alliance Ruhr, Faculty of Computer Science, Ruhr University Bochum

Differentially Private Knowledge Distillation via Synthetic Text Generation

James Flemings Murali Annavaram

University of Southern California

{jamesf17, annavara}@usc.edu

scientific reports

OPEN De-identification is not enough: a comparison between de-identified and synthetic clinical notes

Atiqueur Rahman Sarkar^{1,2}, Yao-Shun Chuang², Noman Mohammed¹ & Xiaoqian Jiang²

PRIVATELY ALIGNING LANGUAGE MODELS WITH REINFORCEMENT LEARNING

Fan Wu^{1*}, Huseyin A. Inan², Arturs Backurs³, Varun Chandrasekaran¹, Janardhan Kulkarni³, Robert Sim²

¹ University of Illinois Urbana-Champaign, ² M365 Research, ³ Microsoft Research
{fanw6, varunc}@illinois.edu, {huseyin.inan, arturs.backurs, jakul, rsim}@microsoft.com

Privacy-Preserving Instructions for Aligning Large Language Models

Da Yu[†] Peter Kairouz[‡] Sewoong Oh[‡] Zheng Xu[‡]

Google Research



This Talk: Learning to Reason with LLMs



Understanding Reasoning

(~20min)



Improving Reasoning

(~20min)



Future Work

(~5min)

Short Term Research Vision



Understanding Reasoning

- How to measure the **transferability** of reasoning in *novel and dynamic environments*?
- How to measure a model's **efficiency** in *learning new skills* beyond final accuracy?

[1] **Yue*, Song*, et al.**, Pangea: A fully open multilingual multimodal llm for 39 languages. *ICLR 2025*

[2] Hu, K., ..., **Yue, X.**, & Liu, Z. Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv 2025*



Improving Reasoning

- How to develop RL approaches for **arbitrary noisy feedback**?
- How to create **human-like memory systems** for **continuous learning**?

[1] Song, Y., Yin, D., **Yue, X.**, Huang, J., Li, S., & Lin, B. Y. (2024). Trial and error: Exploration-based trajectory optimization for llm agents. *ACL 2024*

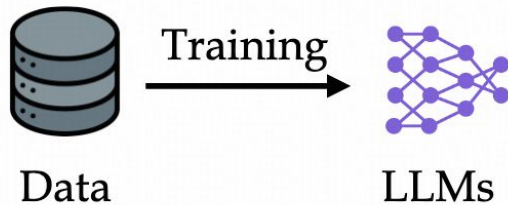
[2] Zheng, T., ... & **Yue, X.** (2024). OpenCodeInterpreter: Integrating code generation with execution and refinement. *Findings of ACL 2024 (1.6K Github Stars; #1 HF Daily paper)*



Mid Term Research Vision

Current

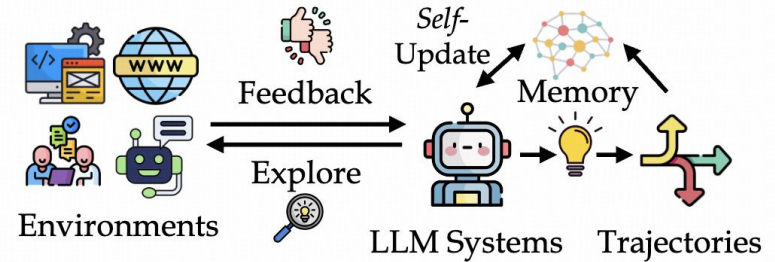
Learning to reason by
mimicking human data



(Static and Passive)

Future

Learning to reason from
exploration and **feedback**



(Dynamic and Active)



Long Term Research Vision

Building intelligent machines capable of **reasoning** across modalities and contexts

- The training **method** of LMs might change
- The **backbone** of LMs might change
- The training **data** of LMs might change

But the need for ***universal understanding*** and **reasoning** across modalities and contexts will remain

Intelligence: A Long Debate

Intelligence as a *collection of task-specific skills*

“Much of the human cognitive function is the result of special-purpose adaptations to solve specific problems.” --Charles Darwin

“AI is the science of making machines capable of performing tasks that would require intelligence if done by humans.”
--Marvin Minsky

Intelligence as a *general learning ability*

“Presumably the child brain is something like a notebook as one buys it from the stationer’s. Rather little mechanism, and lots of blank sheets.” --Alan Turing

“AI is the science and engineering of making machines do tasks they have never seen”
--John McCarthy

Intelligence measures a model’s ability to *efficiently acquire and apply skills to achieve goals* in *novel and dynamic* environments

(My view on “Intelligence”)



Intelligence: A Long Debate

Intelligence as a *collection of task-specific skills*

Intelligence as a *general learning ability*

AI learns efficiently and generalizes like humans

intelligence if done by humans.”
--Marvin Minsky

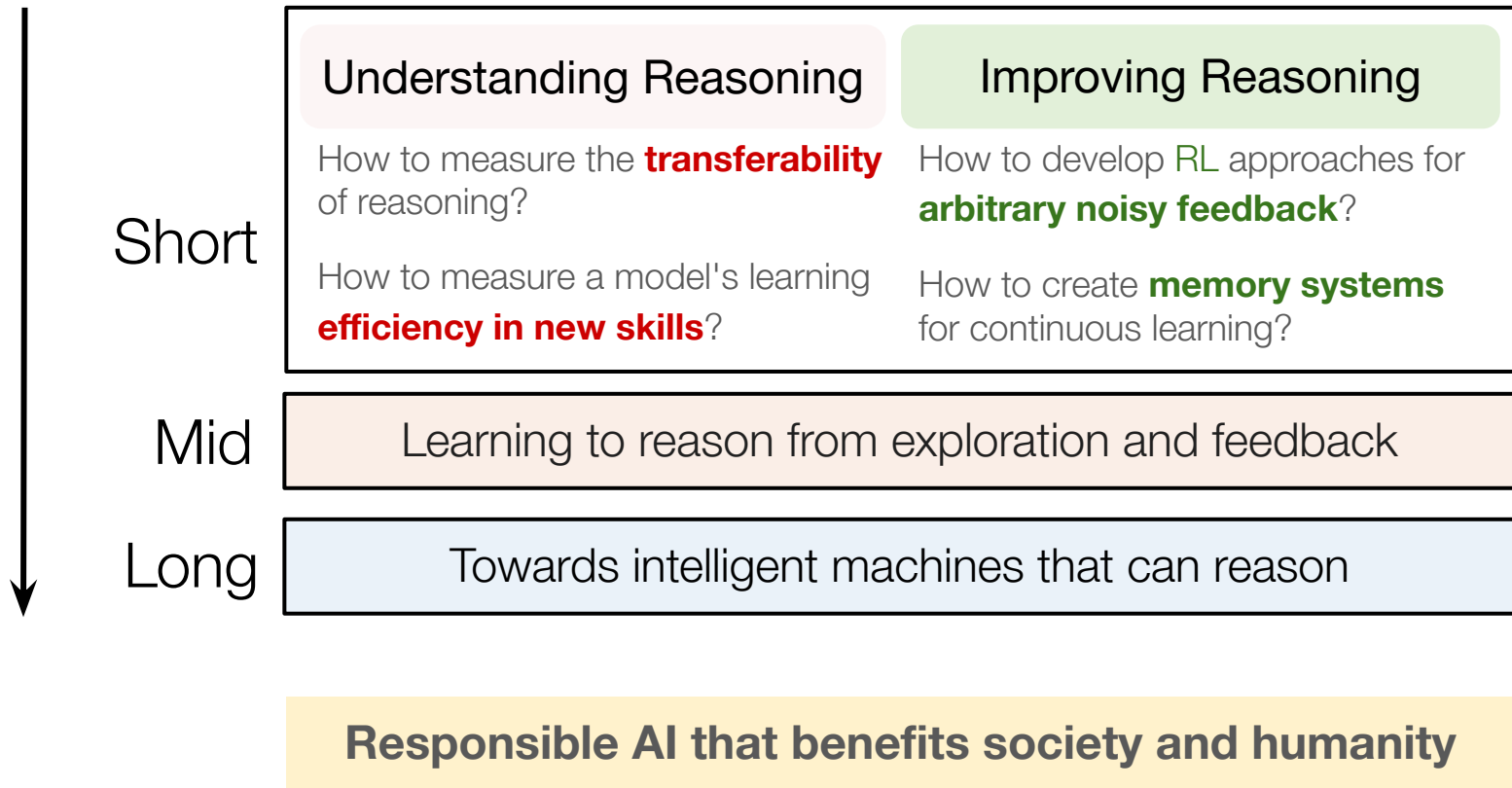
machines do tasks they have never seen”
--John McCarthy

Intelligence measures a model's ability to *efficiently acquire and apply skills to achieve goals* in *novel and dynamic* environments

(My view on “Intelligence”)



Summary: Research Vision



Thank you!



Email: xyue2@andrew.cmu.edu

Homepage: <https://xiangyue9607.github.io/>

Twitter/X: [@xiangyue96](https://twitter.com/xiangyue96)