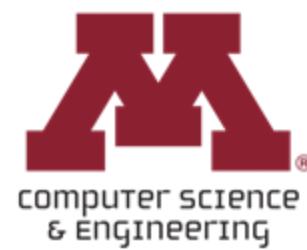


CSCI 5541: Natural Language Processing

Lecture 17: Modern Evaluation



Slides borrowed from Seungone Kim (CMU)

Topics to cover

- Why do we need benchmarks
- Properties of Good Benchmarks
- Widely Used Benchmarks and their Metrics
- Latest Benchmarks

Why do we need Benchmarks?

Why do we need benchmarks

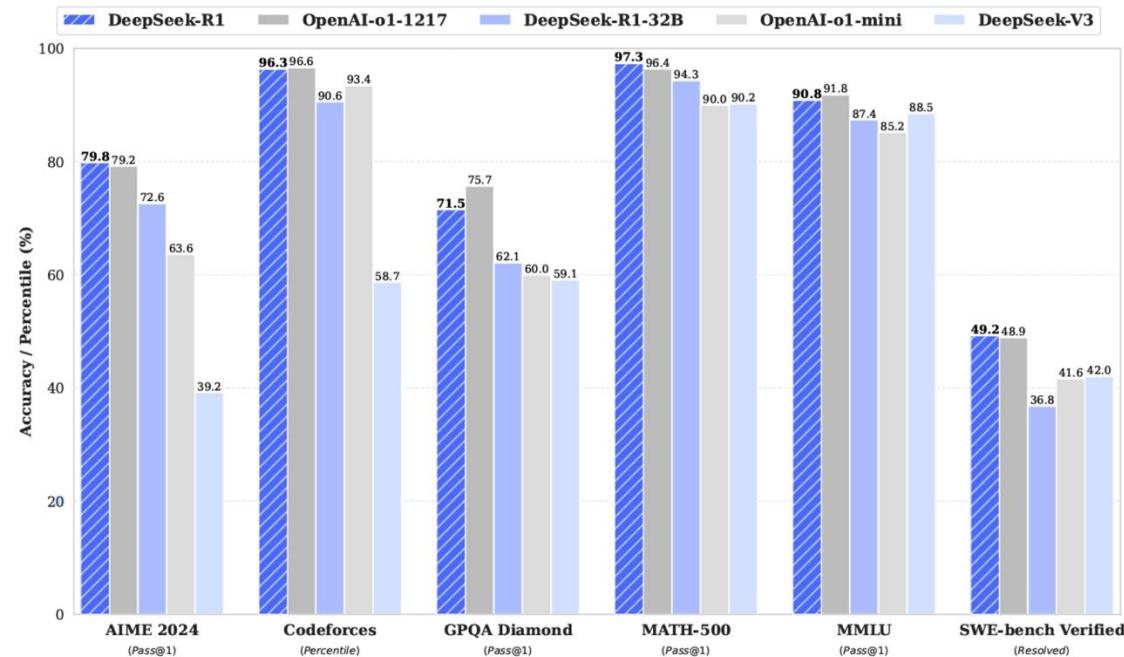
- ❑ Track Progress to Compare Performance

Why do we need benchmarks

- ❑ Track Progress to Compare Performance
 - Given two models (Model A and Model B), which is better?

Why do we need benchmarks

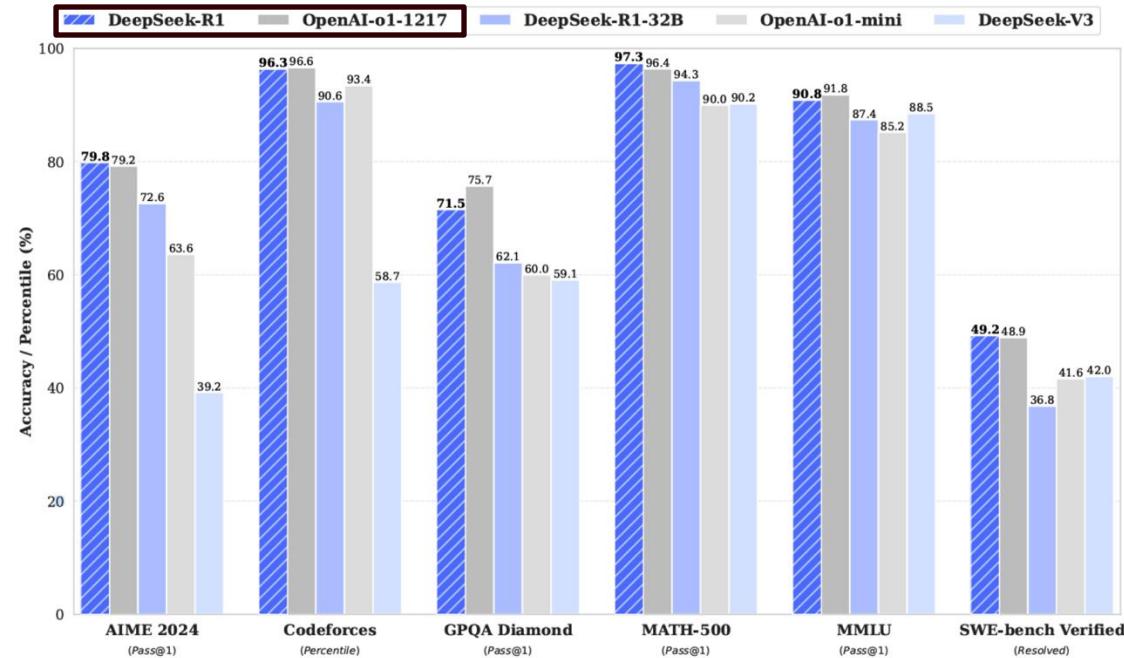
- ❑ Track Progress to Compare Performance
 - Given two models (Model A and Model B), which is better?



Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. and Zhang, X., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Why do we need benchmarks

- ❑ Track Progress to Compare Performance
 - Given two models (Model A and Model B), which is better?



Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. and Zhang, X., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Why do we need benchmarks

- ❑ Track Progress to Compare Performance
 - Given two models (Model A and Model B), which is better?

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	88.4 QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	92.0%* 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Why do we need benchmarks

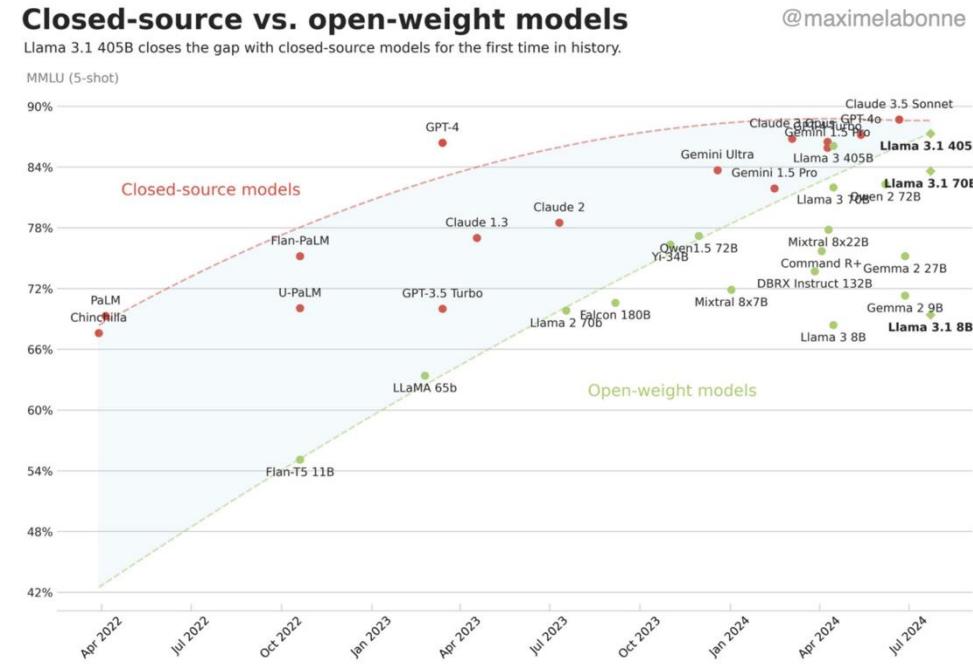
- ❑ Track Progress to Compare Performance
 - Given two models (Model A and Model B), which is better?

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	88.4 QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	92.0%* 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Why do we need benchmarks

- ❑ Track Progress to Compare Performance
 - How good are open source LLMs compared to proprietary LLMs



<https://x.com/maximelabonne/status/1816008591934922915>

Properties of Good Benchmarks

Properties of Good Benchmarks

- ❑ Difficulty
- ❑ Diversity
- ❑ Usefulness
- ❑ Reproducibility
- ❑ Data Contamination

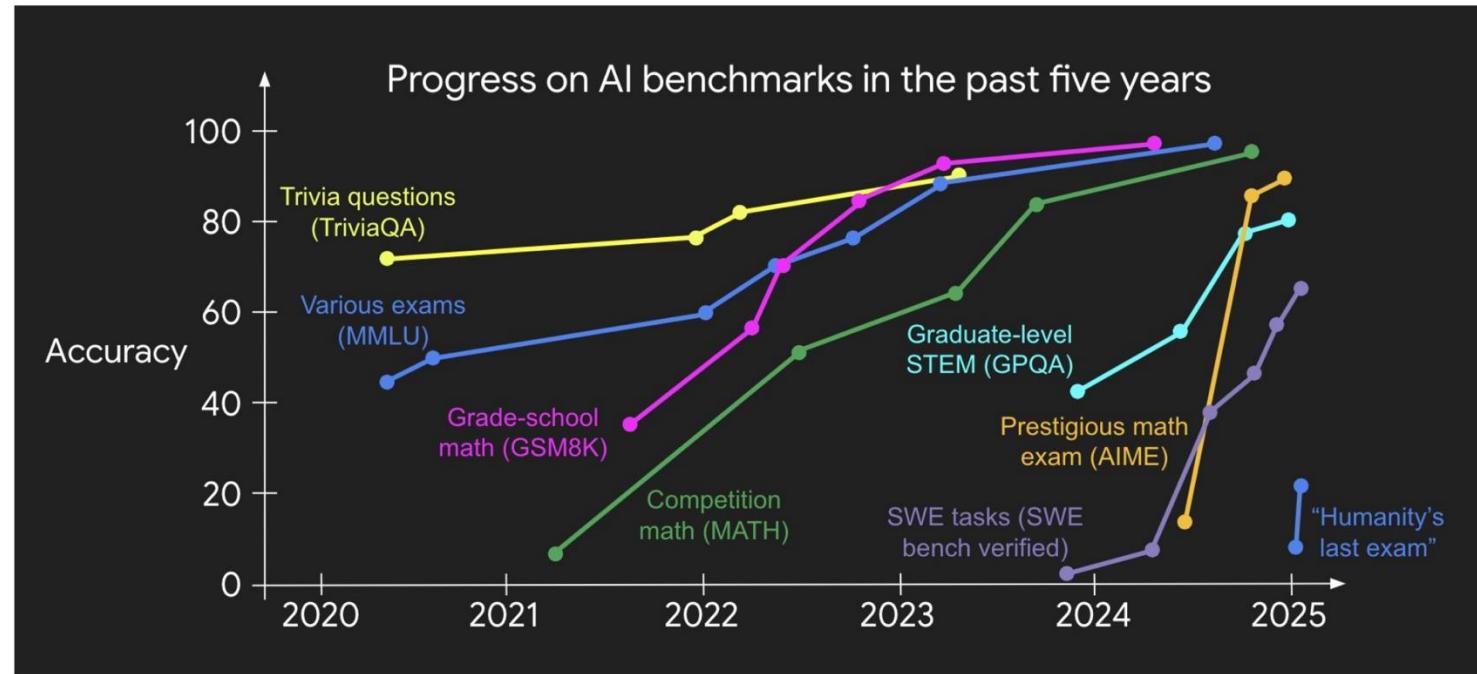
Properties of Good Benchmarks

- Difficulty

Properties of Good Benchmarks

□ Difficulty

- Are the problems sufficiently difficult to distinguish the capable models from those which are less capable?



https://x.com/_jasonwei/status/1889096555254456397

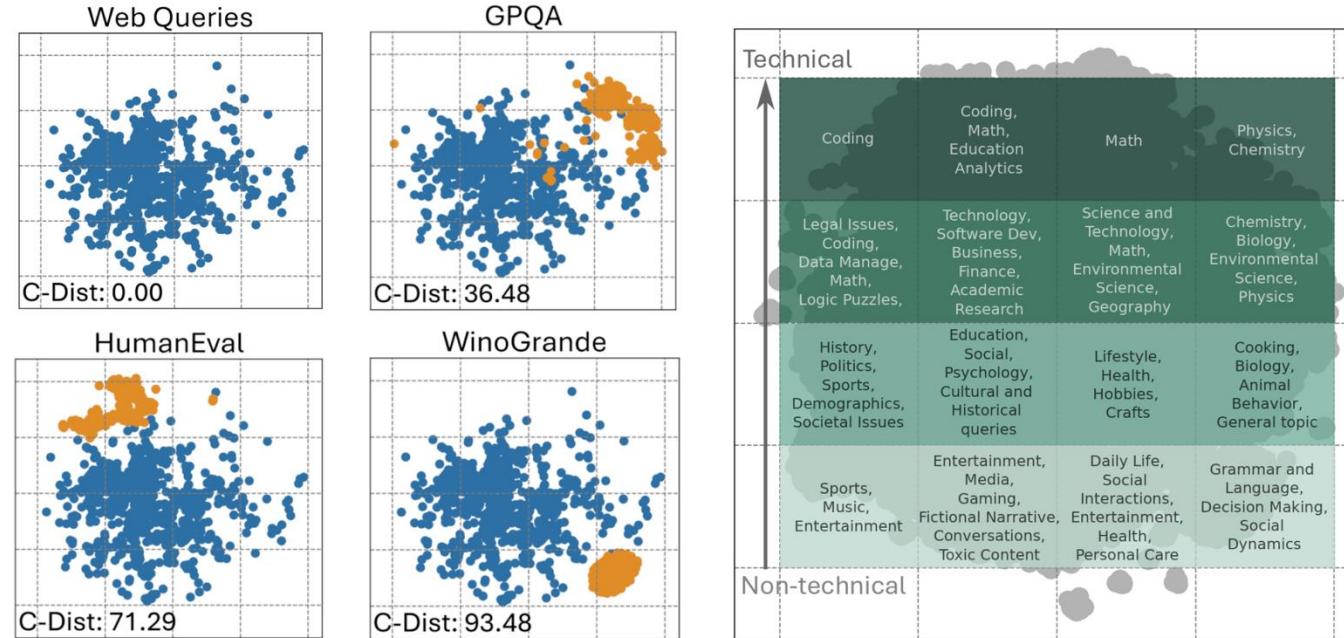
Properties of Good Benchmarks

- ❑ Diversity

Properties of Good Benchmarks

□ Diversity

- Are the problems diverse enough to ensure that the model is effectively being evaluated on its ability to handle a wide range of queries?



Ni, J., Xue, F., Yue, X., Deng, Y., Shah, M., Jain, K., Neubig, G. and You, Y., 2024. MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. arXiv preprint arXiv:2406.06565

Properties of Good Benchmarks

- ❑ Usefulness

Properties of Good Benchmarks

❑ Usefulness

- Does achieving a high score on the benchmark have actual meaning? Is it correlated with something usable in the real world?

MATH Dataset (Ours)

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: 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 $(-1 + \cos(\frac{\pi}{8}) \sqrt[4]{2})(-1 - \cos(\frac{\pi}{8}) \sqrt[4]{2}) = 1 - \cos^2(\frac{\pi}{8}) \sqrt{2} = 1 - \frac{(1+\cos(\frac{\pi}{4}))}{2} \sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

Why do we need an LLM
that is good at solving
math word problems?

Foundation for **more complex tasks**
(e.g., financial analysts)

Useful for **real users with similar questions**
(e.g., tutoring high school math)

Medium for **addressing research questions**
(e.g., Can an AI model reason?)

Properties of Good Benchmarks

❑ Usefulness

- Does achieving a high score on the benchmark have actual meaning? Is it correlated with something usable in the real world?

prompt

```
Write a Python function to sort the given array by using merge sort. Your code should satisfy these tests:  
  
assert merge_sort([3, 4, 2, 6, 5, 7, 1, 9]) == [1, 2, 3, 4, 5, 6, 7, 9]  
assert merge_sort([7, 25, 45, 78, 11, 33, 19]) == [7, 11, 19, 25, 33, 45, 78]  
assert merge_sort([3, 1, 4, 9, 8]) == [1, 3, 4, 8, 9]
```

model

```
def merge_sort(arr):  
    if len(arr) < 2:  
        return arr;  
    mid = len(arr) // 2;  
    left = arr[:mid];  
    right = arr[mid:];  
    left = merge_sort(left);  
    right = merge_sort(right);  
    merged = [];  
    i = j = 0;  
    while i < len(left) and j < len(right):  
        if left[i] < right[j]:  
            merged.append(left[i]);  
            i += 1;  
        else:  
            merged.append(right[j]);  
            j += 1;  
    merged.extend(left[i:]);  
    merged.extend(right[j:]);  
    return merged;
```

Why do we need an LLM
that is good at solving
leetcode-style problems?

Foundation for **more complex tasks**
(e.g., coding agent that implements a whole repo)

Useful for **real users with similar questions**
(e.g., people preparing for coding interview)

Medium for **addressing research questions**
(e.g., Can an AI model self-debug?)

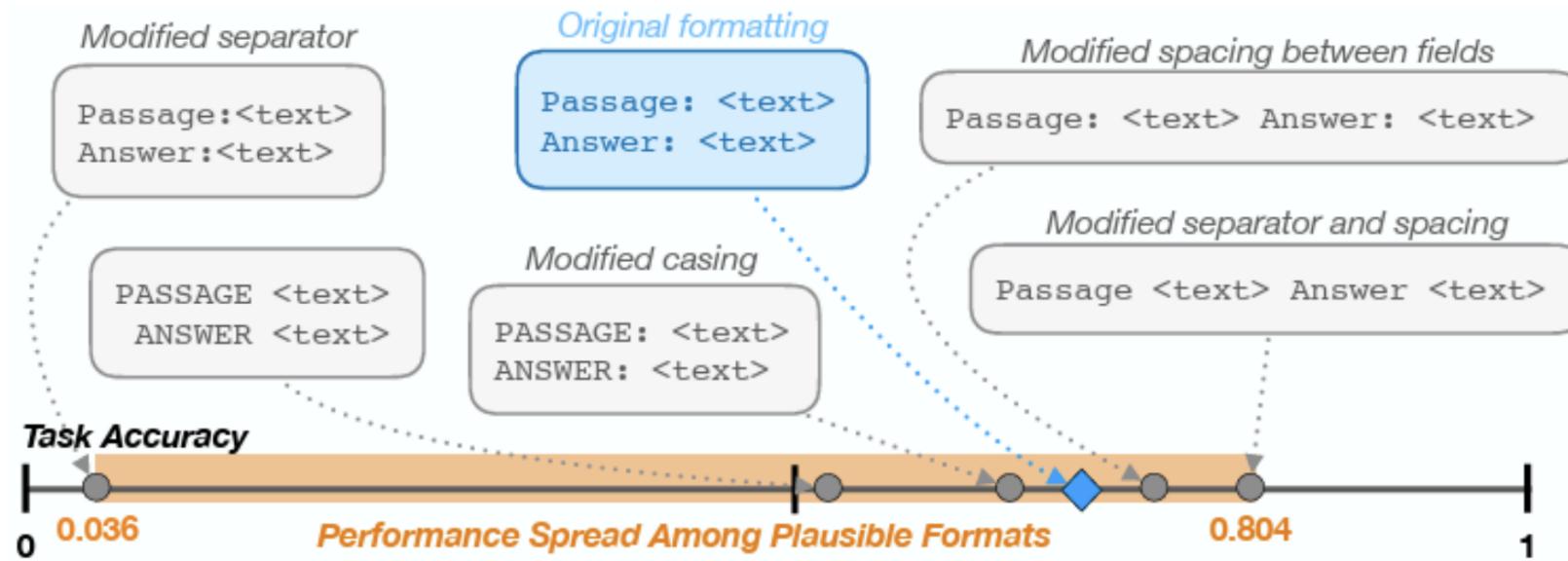
Properties of Good Benchmarks

- ❑ Reproducibility

Properties of Good Benchmarks

❑ Reproducibility

- Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?

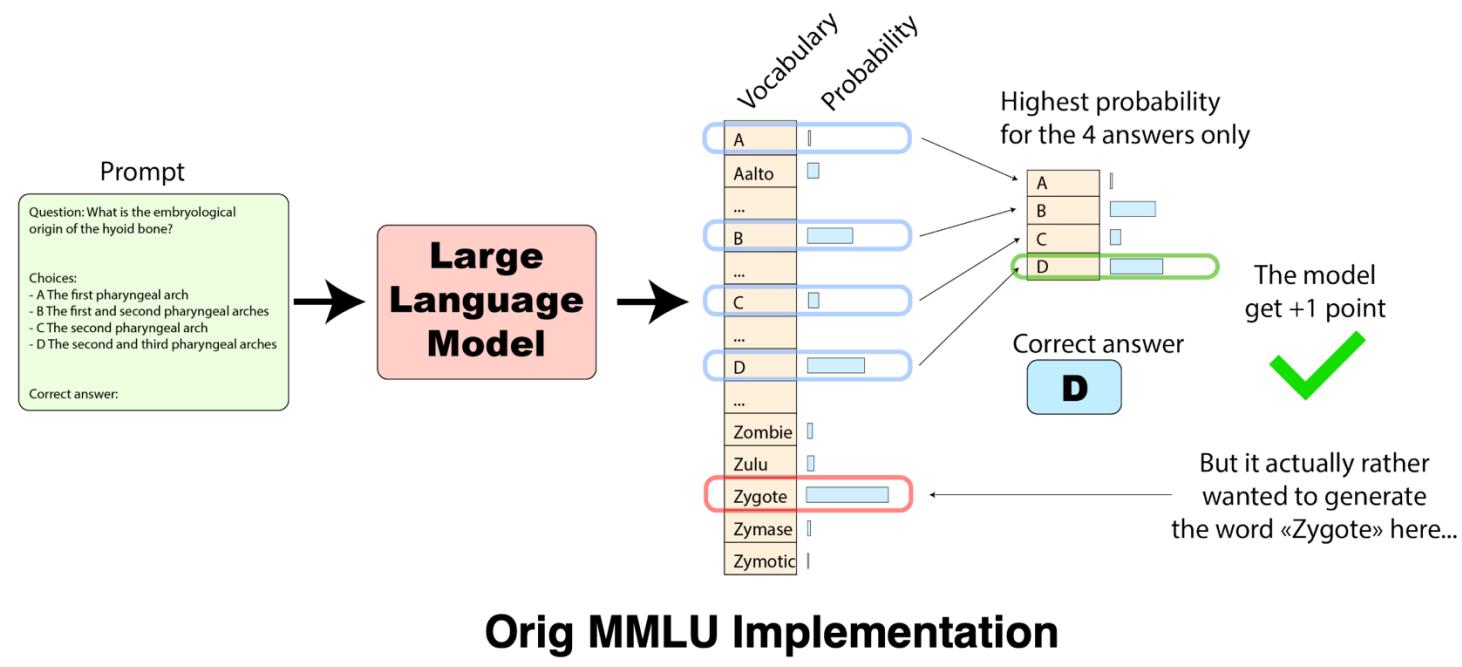


Sclar, M., Choi, Y., Tsvetkov, Y. and Suhr, A., Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In The Twelfth International Conference on Learning Representations.

Properties of Good Benchmarks

❑ Reproducibility

- Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?

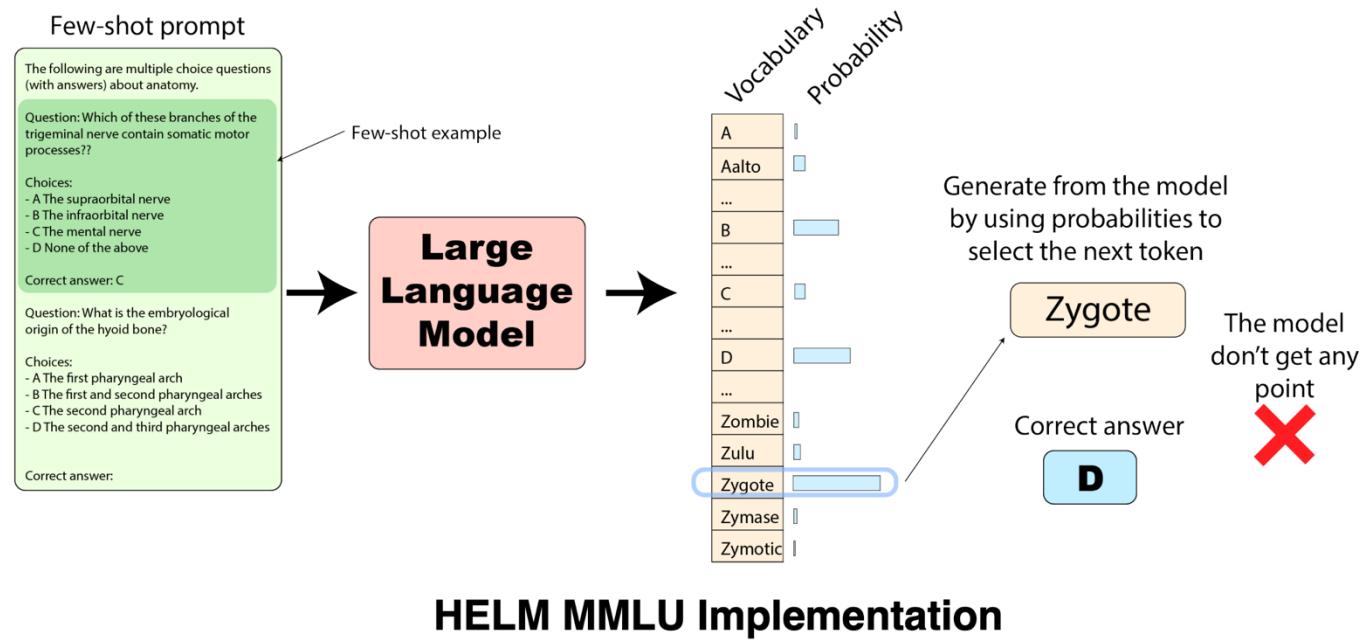


<https://huggingface.co/blog/open-llm-leaderboard-mmlu#1001-flavors-of-mmlu>

Properties of Good Benchmarks

❑ Reproducibility

- Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?

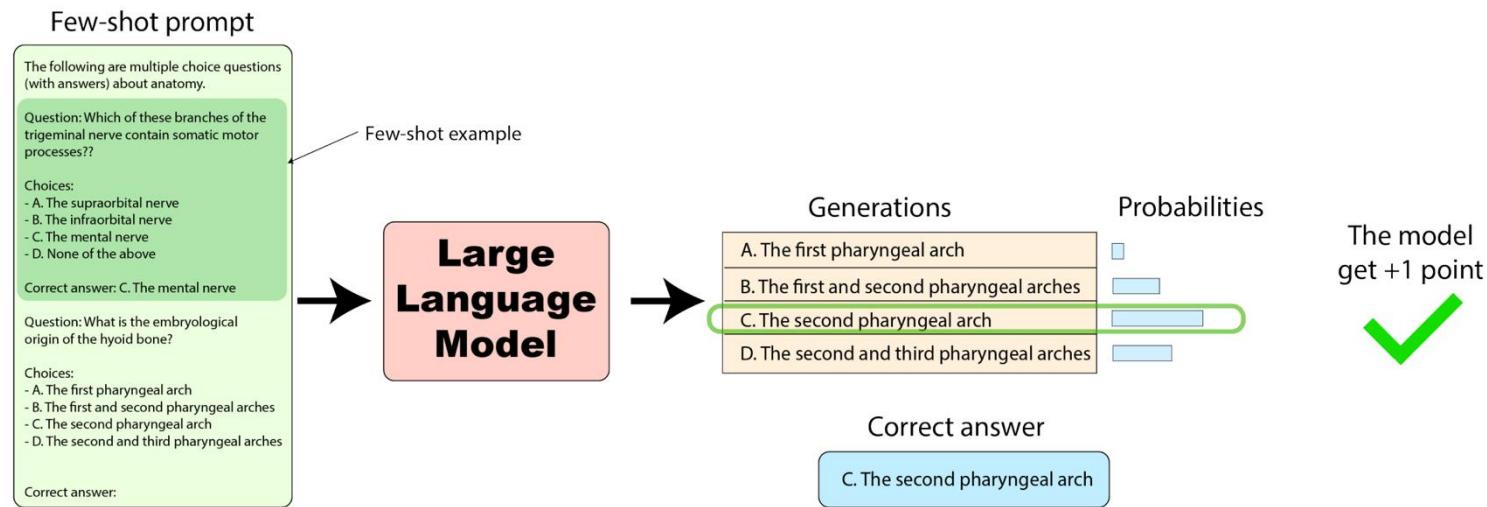


<https://huggingface.co/blog/open-llm-leaderboard-mmlu#1001-flavors-of-mmlu>

Properties of Good Benchmarks

❑ Reproducibility

- Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?



LM EvalHarness MMLU Implementation

<https://huggingface.co/blog/open-llm-leaderboard-mmlu#1001-flavors-of-mmlu>

Properties of Good Benchmarks

❑ Reproducibility

- Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?

	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
llama-65b	0.637	0.488	0.636
tiiuae/falcon-40b	0.571	0.527	0.558
llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
llama-13b	0.471	0.377	0.47
llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269

<https://huggingface.co/blog/open-lm-leaderboard-mmlu#1001-flavors-of-mmlu>

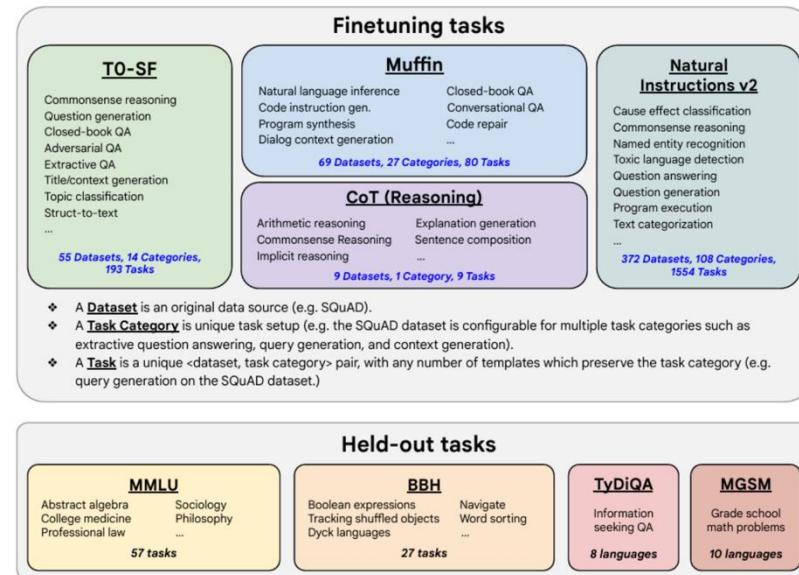
Properties of Good Benchmarks

- Data Contamination

Properties of Good Benchmarks

□ Data Contamination

- LMs are trained on a vast amount of information during pre-training and a wide range of tasks during post-training.

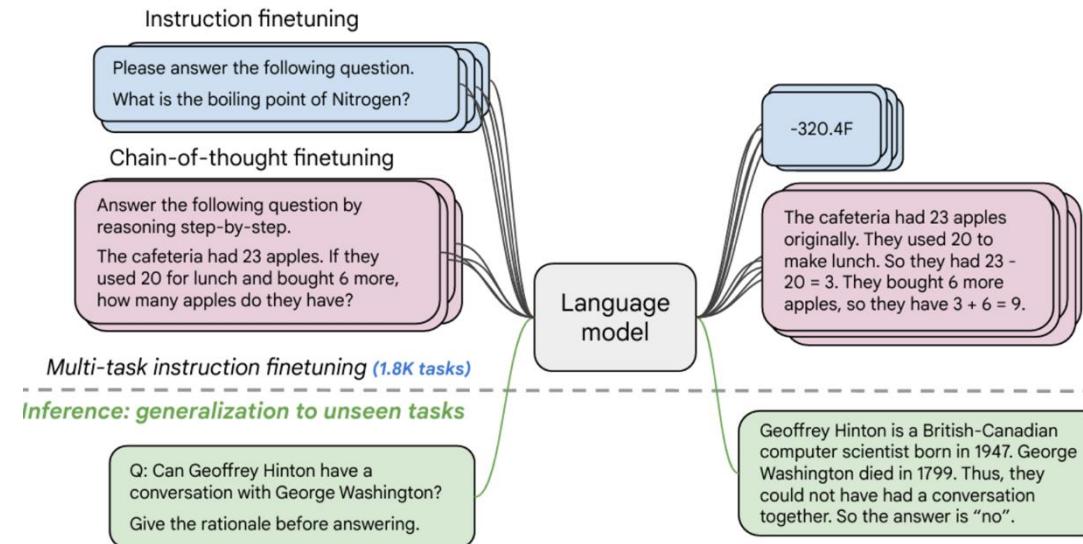


Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. and Webson, A., 2024. Scaling instruction finetuned language models. *Journal of Machine Learning Research*, 25(70), pp.1-53.

Properties of Good Benchmarks

❑ Data Contamination

- How can we ensure that we are evaluating their ability to generalize to unseen novel tasks rather than memorizing similar problem encountered during training?

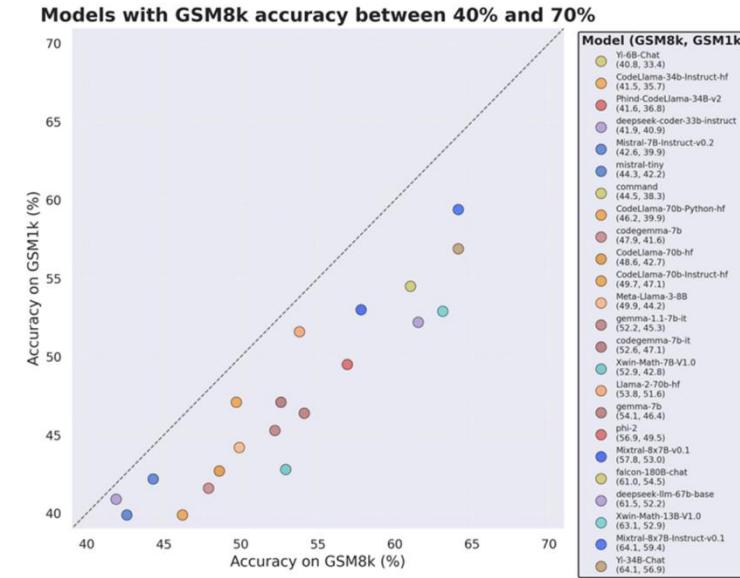
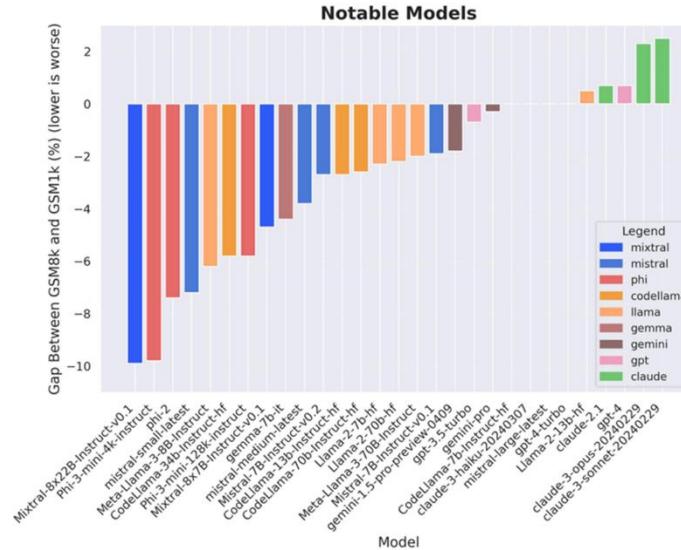


Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. and Webson, A., 2024. Scaling instructionfinetuned language models. *Journal of Machine Learning Research*, 25(70), pp.1-53.

Properties of Good Benchmarks

Data Contamination

- How can we ensure that we are evaluating their ability to generalize to unseen novel tasks rather than memorizing similar problem encountered during training?



Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Slack, D., Lyu, Q. and Hendryx, S., 2024. A careful examination of large language model performance on grade school arithmetic. arXiv preprint arXiv:2405.00332.

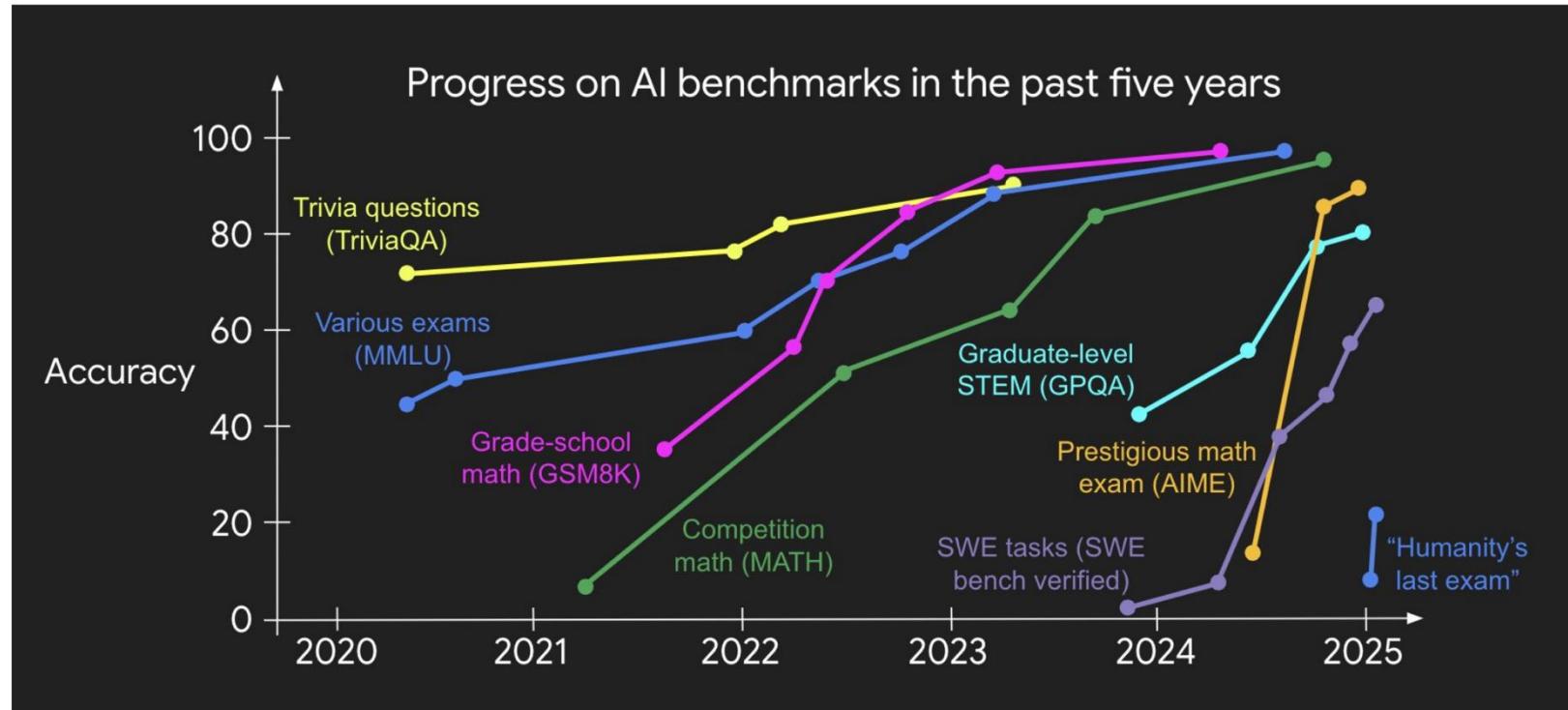
Widely Used Benchmarks and their Metrics

Widely Used Benchmarks & their Metrics

- ❑ Multiple-Choice QA (Classification)
- ❑ Generation

Widely Used Benchmarks & their Metrics

- ❑ Multiple-Choice QA (Classification)
 - (Recap) Benchmarks saturate over time



Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- HellaSwag: Commonsense reasoning around everyday events

 ACTIVITYNET A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

+


A. rinses the bucket off with soap and blow dry the dog's head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.

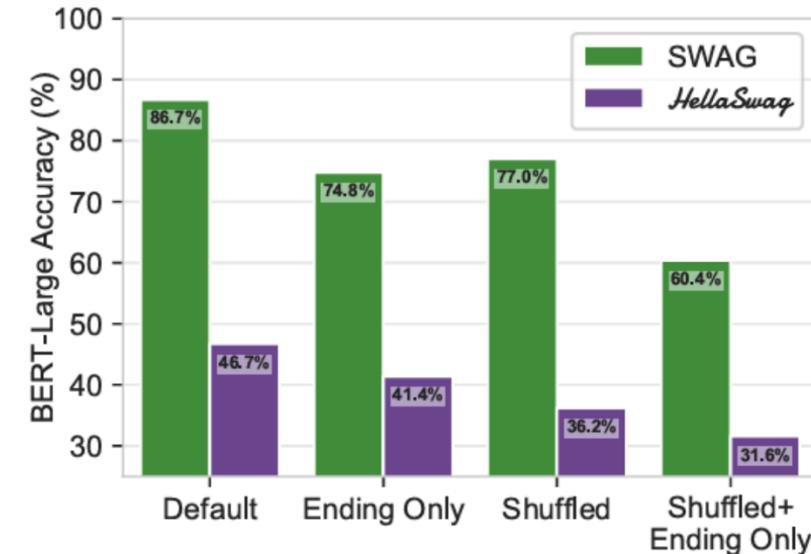
 wikiHow How to do anything
How to determine who has right of way.

+


Come to a complete halt at a stop sign or red light. At a stop sign, come to a complete halt for about 2 seconds or until vehicles that arrived before you clear the intersection. If you're stopped at a red light, proceed when the light has turned green ...

A. Stop for no more than two seconds, or until the light turns yellow. A red light in front of you indicates that you should stop.
B. After you come to a complete stop, turn off your turn signal. Allow vehicles to move in different directions before moving onto the sidewalk.
C. Stay out of the oncoming traffic. People coming in from behind may elect to stay left or right.
D. If the intersection has a white stripe in your lane, stop before this line. Wait until all traffic has cleared before crossing the intersection.

  **easy!**     ???



Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y., 2019, July. HellaSwag: Can a Machine Really Finish Your Sentence?. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4791-4800).

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- SuperGlue: Linguistic Understanding Benchmark

BoolQ **Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*

Question: *is barq's root beer a pepsi product* **Answer:** No

CB **Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?*

Hypothesis: *they are setting a trend* **Entailment:** Unknown

COPA **Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*

Alternative 1: *The sun was rising.* **Alternative 2:** *The grass was cut.*

Correct Alternative: 1

RTE **Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*

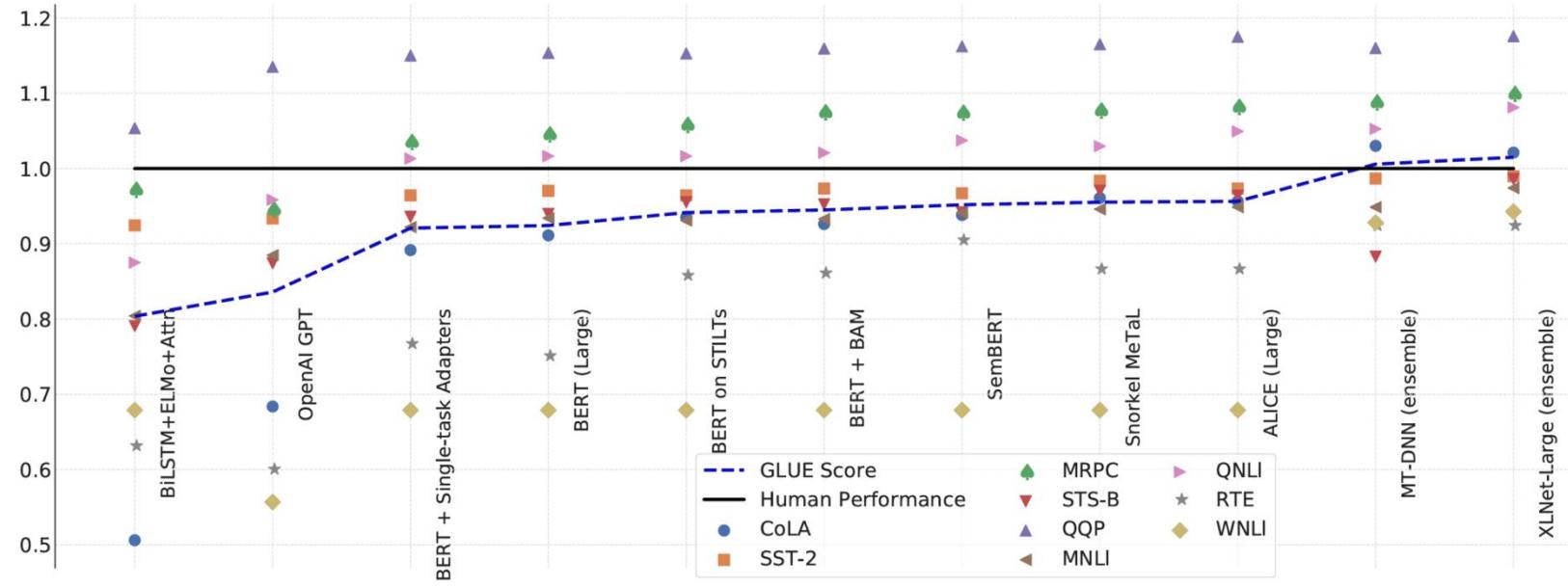
Hypothesis: *Christopher Reeve had an accident.* **Entailment:** False

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S., 2019. SuperGlue: A stickier benchmark for generalpurpose language understanding systems. Advances in neural information processing systems, 32

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- SuperGlue: Linguistic Understanding Benchmark



Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S., 2019. Superglue: A stickier benchmark for generalpurpose language understanding systems. Advances in neural information processing systems, 32

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- MMLU: Multiple-choice questions in 57 subjects (professional & academic)

Microeconomics	One of the reasons that the government discourages and regulates monopolies is that (A) producer surplus is lost and consumer surplus is gained. (B) monopoly prices ensure productive efficiency but cost society allocative efficiency. (C) monopoly firms do not engage in significant research and development. (D) consumer surplus is lost with higher prices and lower levels of output.	   
Conceptual Physics	When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is (A) 9.8 m/s^2 (B) more than 9.8 m/s^2 (C) less than 9.8 m/s^2 (D) Cannot say unless the speed of throw is given.	   
College Mathematics	In the complex z -plane, the set of points satisfying the equation $z^2 = z ^2$ is a (A) pair of points (B) circle (C) half-line (D) line	   

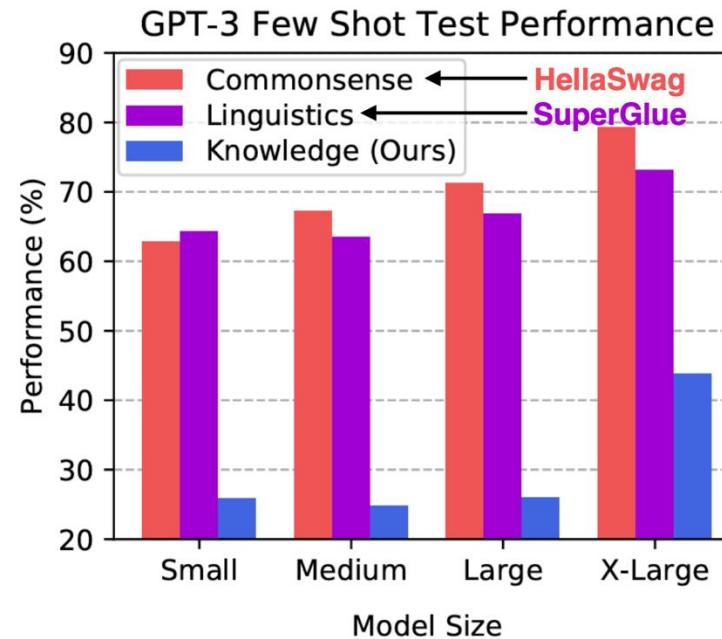
Figure 3: Examples from the Microeconomics task.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J., Measuring Massive Multitask Language Understanding. In International Conference on Learning Representations

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- MMLU: Multiple-choice questions in 57 subjects (professional & academic)



Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J., Measuring Massive Multitask Language Understanding. In International Conference on Learning Representations

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- MMLU-Pro: Harder MMLU with 4~10 options to choose from.

Question: Where do most short-period comets come from and how do we know?

- Options:**
- A. The Kuiper belt; short period comets tend to be in the plane of the solar system like the Kuiper belt.
 - B. The asteroid belt; short period comets tend to come from random directions indicating a spherical distribution of comets called the asteroid belt.
 - C. The asteroid belt; short period comets tend to be in the plane of the solar system just like the asteroid belt.
 - D. The Oort cloud; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the Oort cloud.
 - E. The Oort Cloud; short period comets tend to come from random directions indicating a spherical distribution of comets called the Oort Cloud.
 - F. The Oort cloud; short period comets tend to be in the plane of the solar system just like the Oort cloud.
 - G. The asteroid belt; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the asteroid belt.

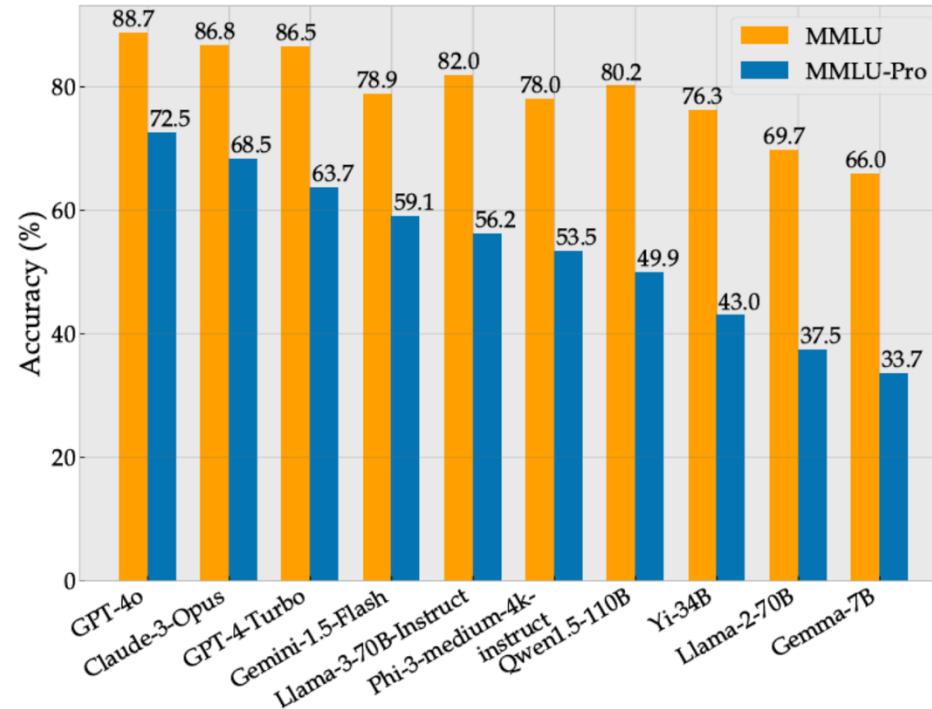
Answer: Let's think step by step. Most short-period comets originate from the Kuiper belt. This is deduced from the observation that these comets tend to follow orbits that lie in the plane of the solar system, similar to the distribution of objects in the Kuiper belt itself. Thus, the alignment of these cometary orbits with the ecliptic plane points to their Kuiper belt origin. The answer is (A).

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z. and Li, T., 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574.

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- MMLU-Pro: Harder MMLU with 4~10 options to choose from.

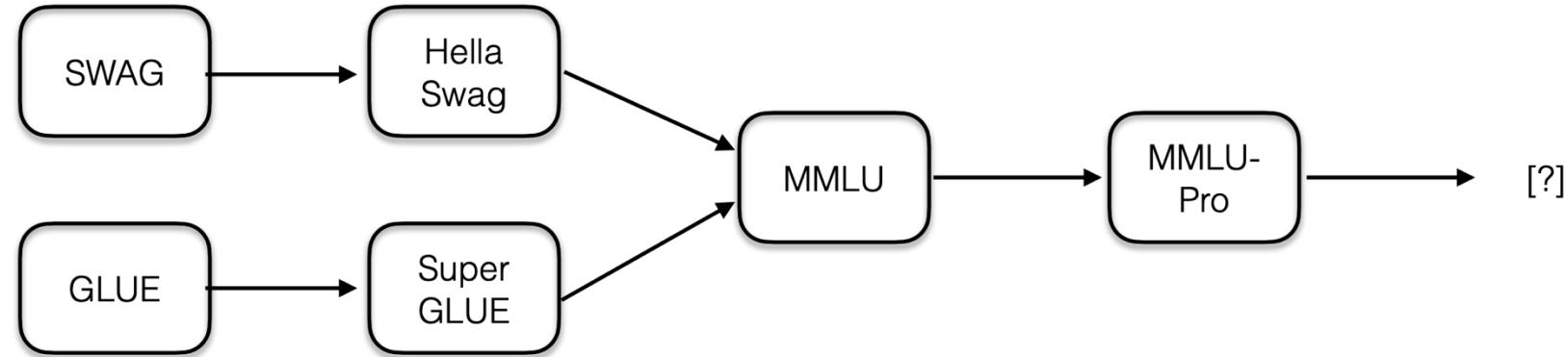


Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z. and Li, T., 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574.

Widely Used Benchmarks & their Metrics

❑ Multiple-Choice QA (Classification)

- The trend of developing increasingly challenging benchmarks continue
- Open Research Question: How can we come up with more challenging benchmarks as LMs become as intelligent as or even surpass humans?



Widely Used Benchmarks & their Metrics

❑ Generation

Widely Used Benchmarks & their Metrics

❑ Generation

- As humans, we don't present LMs with four options to choose from
- However, assessing free form responses is more challenging than choosing from among a set of four options, where accuracy can be used as the metric

Widely Used Benchmarks & their Metrics

❑ Generation

- GSM8k: Grade School mathematics questions
 - ✓ For math word problems, the answer is a numeric value, so we could use Exact Match (EM).

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

Solution: Tina buys 3 12-packs of soda, for $3 \times 12 = \textcolor{red}{<<3*12=36>>} 36$ sodas

6 people attend the party, so half of them is $6/2 = \textcolor{red}{<<6/2=3>>} 3$ people

Each of those people drinks 3 sodas, so they drink $3 \times 3 = \textcolor{red}{<<3*3=9>>} 9$ sodas

Two people drink 4 sodas, which means they drink $2 \times 4 = \textcolor{red}{<<4*2=8>>} 8$ sodas

With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = \textcolor{red}{<<5+9+8+3=25>>} 25$ sodas

As Tina started off with 36 sodas, that means there are $36 - 25 = \textcolor{red}{<<36-25=11>>} 11$ sodas left

Final Answer: 11

Widely Used Benchmarks & their Metrics

❑ Generation

- HumanEval: Python Coding Tasks

- ✓ For LeetCode-style coding problems, we can execute the code and verify whether it passes all test cases (i.e., pass@k).

```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) =>12
    solution([3, 3, 3, 3, 3]) =>9
    solution([30, 13, 24, 321]) =>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G. and Ray, A., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Widely Used Benchmarks & their Metrics

❑ Generation

- How about open-ended questions like this?

Question: What are some business etiquette norms when doing business in Japan?

Assistant A:

Here are some business etiquette norms to keep in mind when doing business in Japan:

1. Bowing: Bowing is a common greeting in Japan. It is a sign of respect and should be reciprocated.
2. Business cards: Business cards are exchanged at the beginning of a meeting. It is important to receive and give business cards with both hands and to take a moment to read the card before putting it away.
3. Punctuality: Being on time is important in Japan. Arrive a few minutes early to show respect for your colleagues.
4. Dress code: Dress conservatively and professionally. Men should wear dark suits and ties, while women should wear conservative business attire.
5. Language: Learn a few basic Japanese phrases to show respect for the culture. It is also important to speak clearly and avoid using slang or idioms.
6. Gift giving: Gift giving is a common practice in Japan. Bring a small gift, such as a box of chocolates or a small souvenir from your home country.
7. Dining etiquette: When dining with colleagues, wait for the host to begin eating before you start. Use chopsticks correctly and avoid leaving food on your plate.

Remember, showing respect for Japanese culture and customs is key to building strong business relationships in Japan.

Assistant B:

When doing business in Japan, it is important to be aware of and respect the country's business etiquette norms. Some of these norms include:

1. Greetings: A handshake and bow is the standard greeting in Japan. It is important to bow deeply and to offer a firm handshake.
2. Business cards: In Japan, business cards are exchanged after the initial greeting. It is important to handle the card with respect and to read it carefully before placing it on the table.
3. Meetings: Meetings in Japan are typically formal and conducted in a hierarchical manner. It is important to address people by their titles and to avoid direct eye contact.
4. Punctuality: Punctuality is highly valued in Japan. It is important to arrive on time for meetings and to be prepared in advance.
5. Gift giving: Gift giving is a common practice in Japan, but it is important to be mindful of the occasion and the relationship. Avoid giving gifts that are considered too personal or extravagant.
6. Dining: Dining is an important part of doing business in Japan. It is important to be aware of the proper table manners and to avoid slurping or making loud noises while eating.
7. Communication: Communication in Japan is typically indirect and reserved. It is important to listen carefully and to avoid confrontational language.

Overall, it is important to be respectful, punctual, and mindful of hierarchy when doing business in Japan.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2024. Judging llm-as-a-judge with mtbench and chatbot arena. Advances in Neural Information Processing Systems, 36.

Widely Used Benchmarks & their Metrics

❑ Generation

- Rouge: Word overlap

Question:

What are the benefits of regular exercise?

Prediction (5 words):

Exercise improves mood and health.

Reference Answer (6 words):

Regular exercise benefits health and mood.

Lin, C.Y., 2004, July. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Widely Used Benchmarks & their Metrics

❑ Generation

- Rouge: Word overlap

Question:

What are the benefits of regular exercise?

Prediction (5 words):

Exercise improves mood and health.

Reference Answer (6 words):

Regular exercise benefits health and mood.

- **Rouge-1:** “exercise”, “health”, “and”, “mood” overlaps -> Precision: 4/5, Recall: 4/6, F1: 0.727
- **Rouge-2:** No overlap! -> Precision: 0/5, Recall: 0/6, F1: 0.0
- **Rouge-L:** [“exercise”, “mood”] and [“exercise”, “and”] are common subsequences
-> Precision: 2/5, Recall: 2/6, F1: 0.364

Lin, C.Y., 2004, July. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Widely Used Benchmarks & their Metrics

❑ Generation

- Rouge: Word overlap

Question:

What are the benefits of regular exercise?

Prediction (5 words):

Exercise improves mood and health.

Reference Answer (6 words):

Regular exercise benefits health and mood.

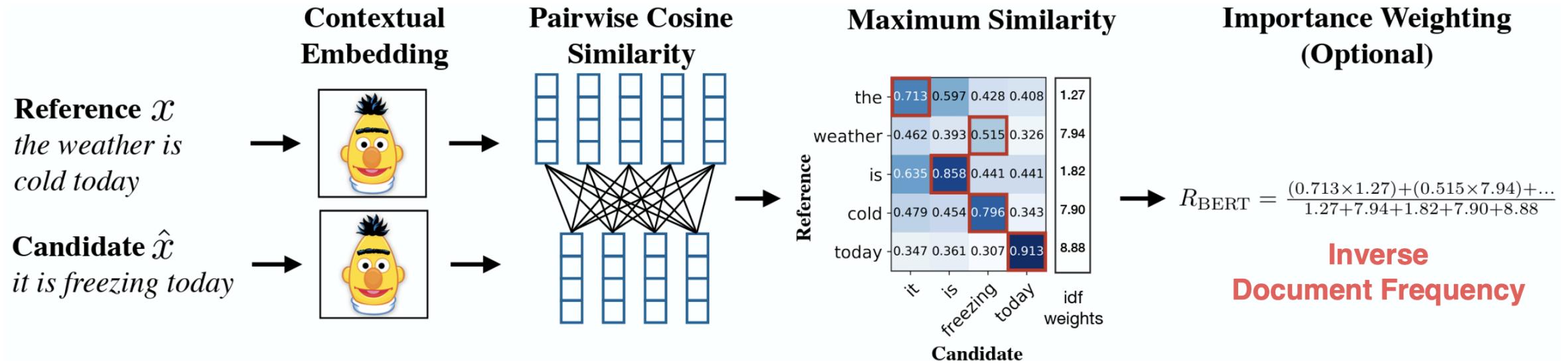
- **Rouge-1:** “exercise”, “health”, “and”, “mood” overlaps -> Precision: 4/5, Recall: 4/6, F1: 0.727
- **Rouge-2:** No overlap! -> Precision: 0/5, Recall: 0/6, F1: 0.0 Is it desirable for Rouge-2 to be 0?
- **Rouge-L:** [“exercise”, “mood”] and [“exercise”, “and”] are common subsequences -> Precision: 2/5, Recall: 2/6, F1: 0.364

Lin, C.Y., 2004, July. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Widely Used Benchmarks & their Metrics

❑ Generation

- BertScore: Calculate the embedding similarity between the prediction & ref. answer



Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations.

Widely Used Benchmarks & their Metrics

□ Generation

- LLM-as-judge: Prompt a LM to provide a judgment (e.g., "3 out of 5", A is better than B)

GPT-4 Judgment (when A is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant A's answer is more detailed and organized, covering a wider range of topics such as dress code and language, which are not mentioned in Assistant B's response. Assistant A also provides more specific advice on how to handle certain situations, such as exchanging business cards and dining etiquette. Therefore, Assistant A's answer is more comprehensive and useful for someone looking to understand Japanese business etiquette.

GPT-4 Judgment (when B is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B's answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication styles. Assistant A's answer is more concise but lacks some of the depth found in Assistant B's response. Therefore, Assistant B's answer is better in this case.

Question: What are some business etiquette norms when doing business in Japan?

Assistant A:

Here are some business etiquette norms to keep in mind when doing business in Japan:

1. Bowing: Bowing is a common greeting in Japan. It is a sign of respect and should be reciprocated.
2. Business cards: Business cards are exchanged at the beginning of a meeting. It is important to receive and give business cards with both hands and to take a moment to read the card before putting it away.
3. Punctuality: Being on time is important in Japan. Arrive a few minutes early to show respect for your colleagues.
4. Dress code: Dress conservatively and professionally. Men should wear dark suits and ties, while women should wear conservative business attire.
5. Language: Learn a few basic Japanese phrases to show respect for the culture. It is also important to speak clearly and avoid using slang or idioms.
6. Gift giving: Gift giving is a common practice in Japan. Bring a small gift, such as a box of chocolates or a small souvenir from your home country.
7. Dining etiquette: When dining with colleagues, wait for the host to begin eating before you start. Use chopsticks correctly and avoid leaving food on your plate.

Remember, showing respect for Japanese culture and customs is key to building strong business relationships in Japan.

Assistant B:

When doing business in Japan, it is important to be aware of and respect the country's business etiquette norms. Some of these norms include:

1. Greetings: A handshake and bow is the standard greeting in Japan. It is important to bow deeply and to offer a firm handshake.
2. Business cards: In Japan, business cards are exchanged after the initial greeting. It is important to handle the card with respect and to read it carefully before placing it on the table.
3. Meetings: Meetings in Japan are typically formal and conducted in a hierarchical manner. It is important to address people by their titles and to avoid direct eye contact.
4. Punctuality: Punctuality is highly valued in Japan. It is important to arrive on time for meetings and to be prepared in advance.
5. Gift giving: Gift giving is a common practice in Japan, but it is important to be mindful of the occasion and the relationship. Avoid giving gifts that are considered too personal or extravagant.
6. Dining: Dining is an important part of doing business in Japan. It is important to be aware of the proper table manners and to avoid slurping or making loud noises while eating.
7. Communication: Communication in Japan is typically indirect and reserved. It is important to listen carefully and to avoid confrontational language.

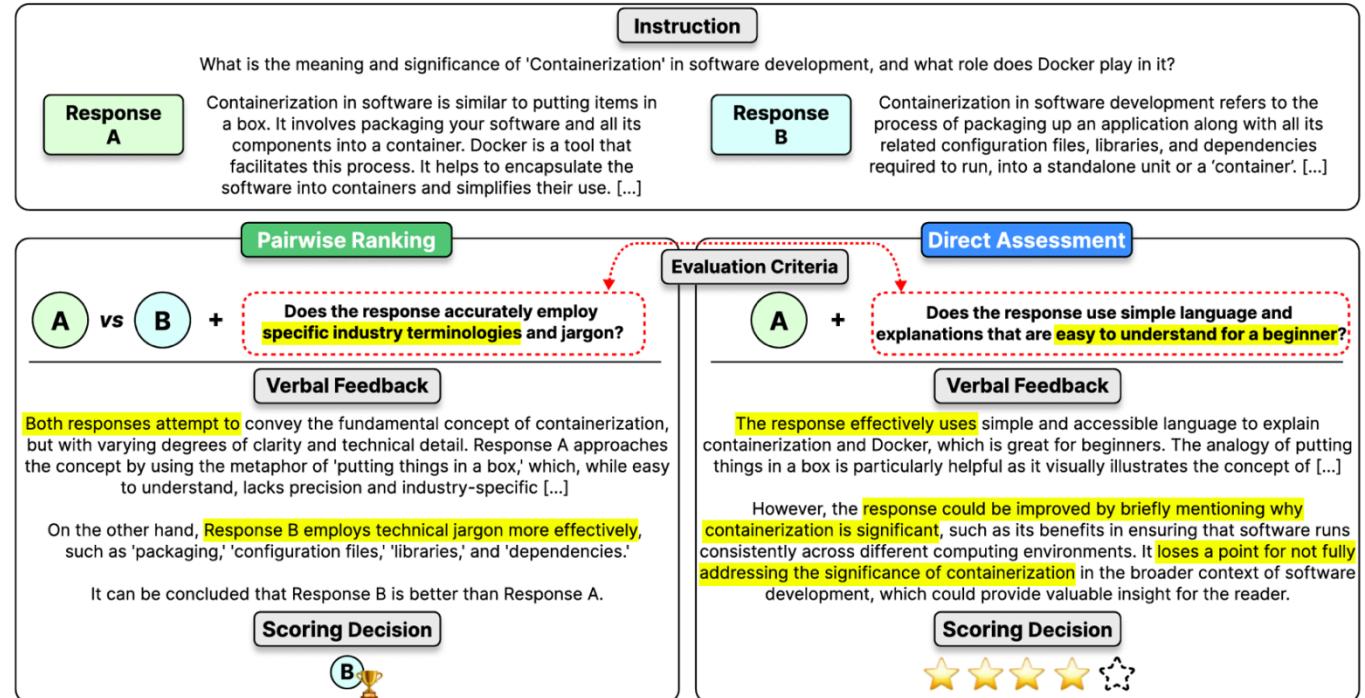
Overall, it is important to be respectful, punctual, and mindful of hierarchy when doing business in Japan.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2024. Judging llm-as-a-judge with mtbench and chatbot arena. Advances in Neural Information Processing Systems, 36.

Widely Used Benchmarks & their Metrics

□ Generation

- LLM-as-judge: Prompt a LM to provide a judgment (e.g., "3 out of 5", A better than B)
- We can set explicit rubrics which the LLM should follow

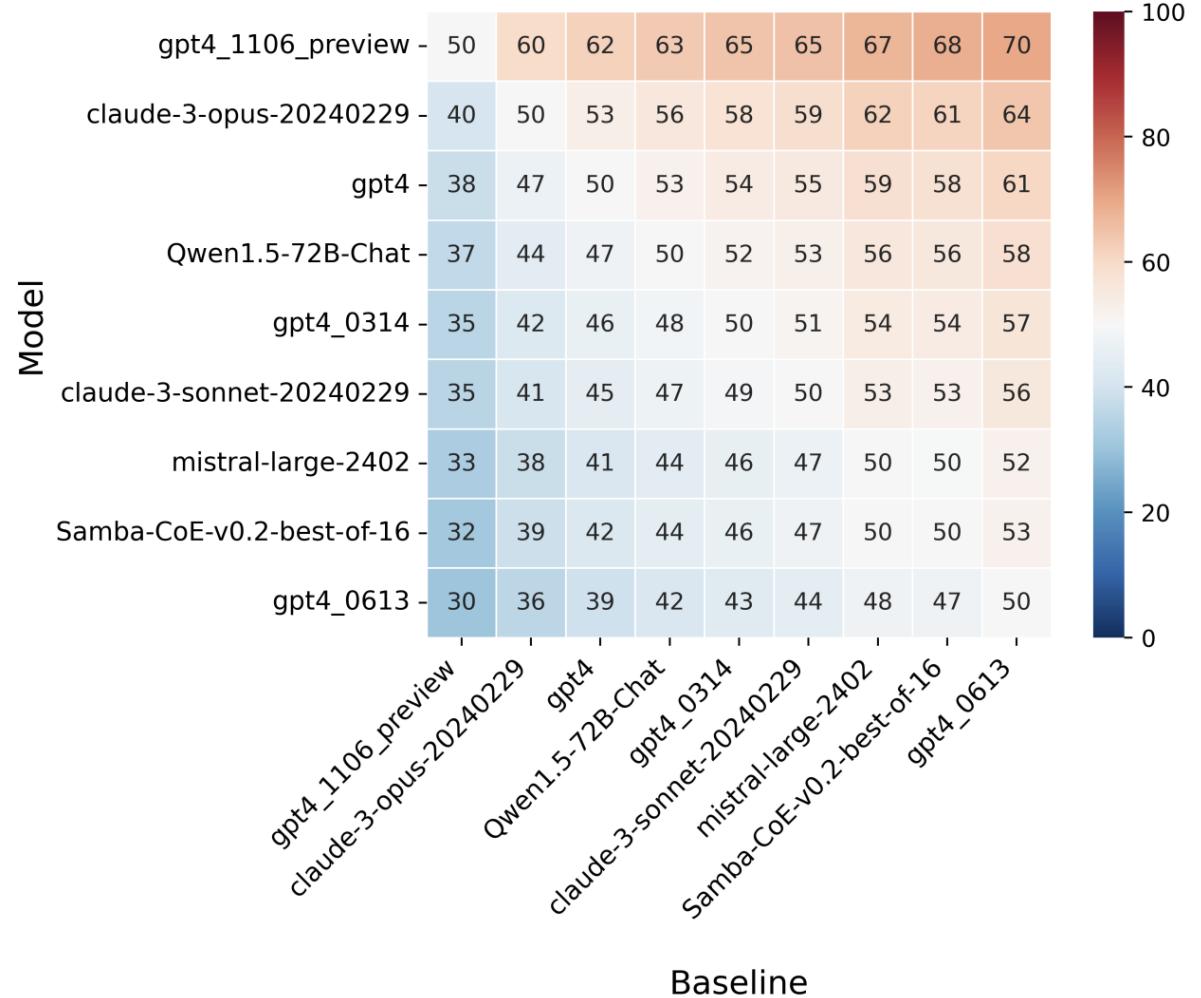


Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2024. Judging llm-as-a-judge with mtbench and chatbot arena. Advances in Neural Information Processing Systems, 36.

Widely Used Benchmarks & their Metrics

❑ Generation

- AlpacaEval: Calculate the win-rate against a reference model on openended generation tasks

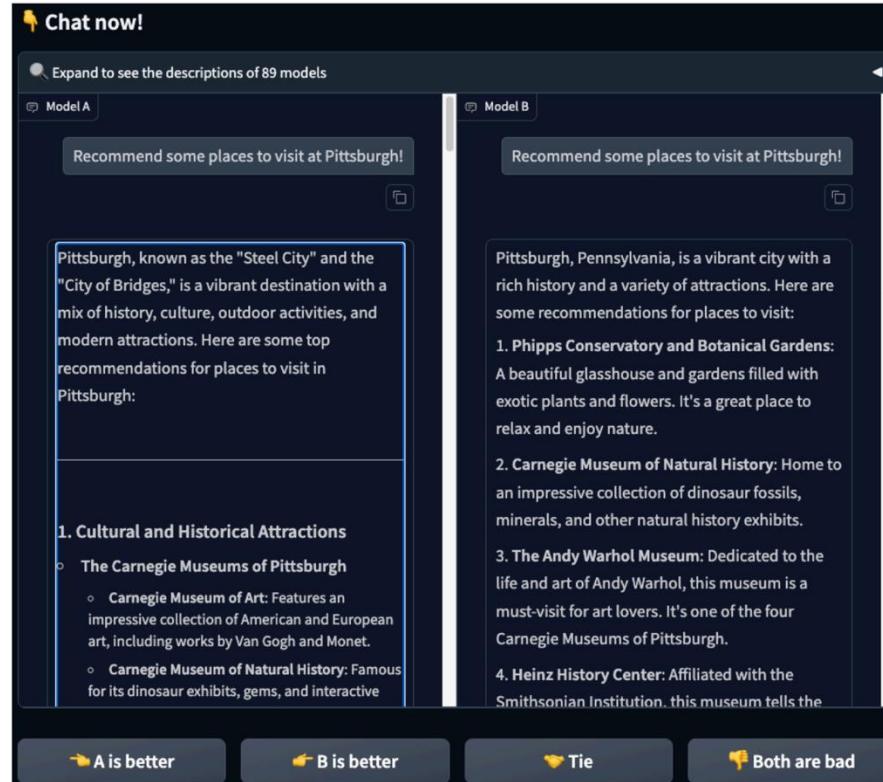


Dubois, Y., Liang, P. and Hashimoto, T., 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In First Conference on Language Modeling.

Widely Used Benchmarks & their Metrics

❑ Generation

- LMSys Chatbot Arena: Ask humans to vote which response is better



Rank* (UB)	Rank (StyleCtr)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	chocolate... (Early Grok-3)	1402	+7/-6	7829	xAI	Proprietary
2	4	Gemini-2.0-Flash-Thinking-Exp-01-21	1385	+5/-5	13336	Google	Proprietary
2	2	Gemini-2.0-Pro-Exp-02-05	1379	+5/-6	11197	Google	Proprietary
2	1	ChatGPT-4o-latest (2025-01-29)	1377	+5/-6	10529	OpenAI	Proprietary
5	2	DeepSeek-R1	1361	+8/-7	5079	DeepSeek	MIT
5	8	Gemini-2.0-Flash-001	1356	+6/-5	9092	Google	Proprietary
5	2	o1-2024-12-17	1353	+6/-5	15437	OpenAI	Proprietary

<https://lmarena.ai/>

Latest Benchmarks

Latest Benchmarks

- ❑ How can we push the boundaries of evaluation further into more challenging domains that LLMs struggle to handle?

Latest Benchmarks

❑ US Math Olympiad Questions

- Very novel questions (likely chance of data contamination is low)
- Very challenging questions

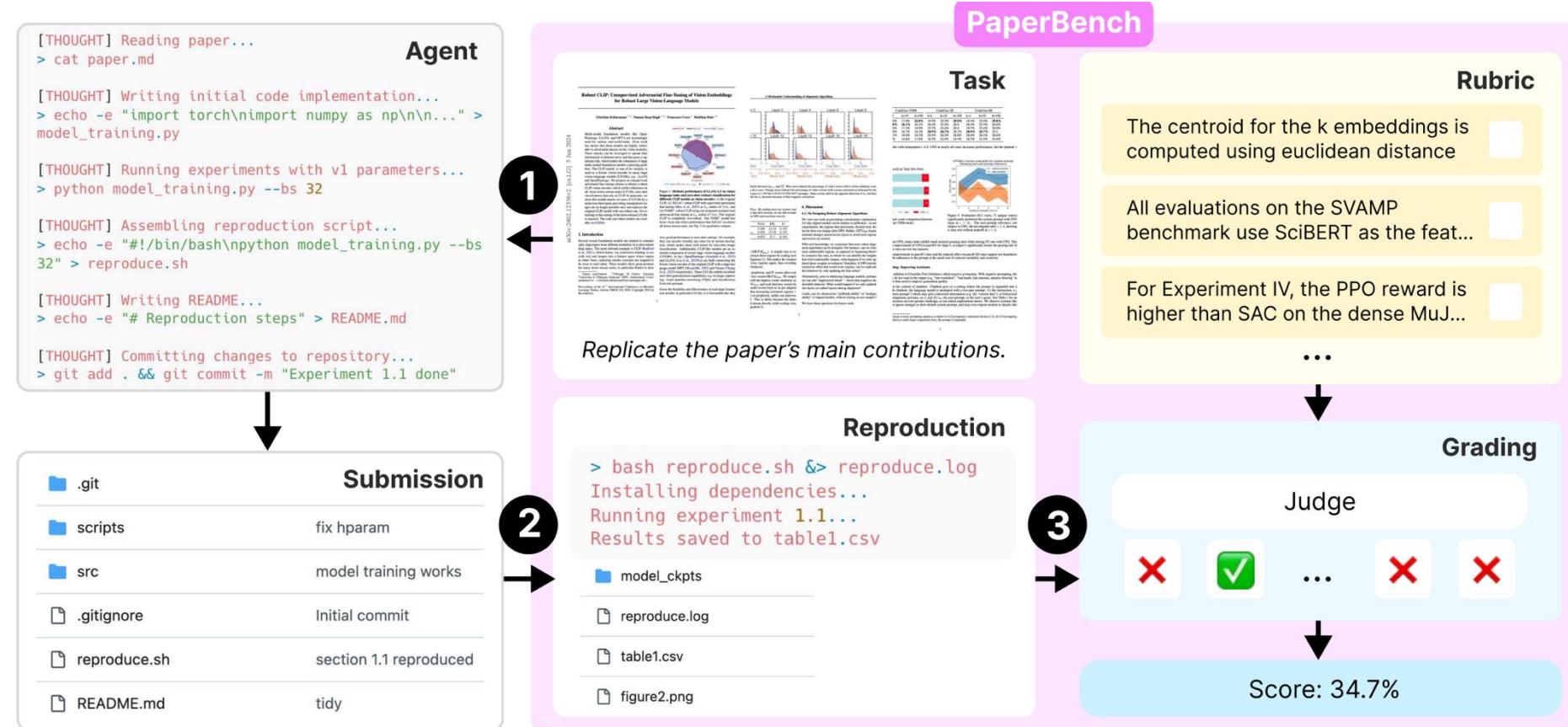
Model	P1	P2	P3	P4	P5	P6	Total	Cost
R1	0.5	0.0	0.0	1.5	0.0	0.0	2.0	2.03
FLASH-THINKING	1.5	0.0	0.0	0.0	0.2	0.0	1.8	N/A
CLAUDE 3.7	0.5	0.5	0.0	0.0	0.0	0.6	1.5	9.03
QwQ	1.2	0.0	0.0	0.0	0.0	0.0	1.2	0.42
O1-PRO	0.5	0.0	0.0	0.0	0.2	0.4	1.2	203.44
O3-MINI	0.5	0.1	0.0	0.0	0.0	0.2	0.9	1.11

<https://arxiv.org/abs/2503.21934v1>

Latest Benchmarks

❑ PaperBench

- Replicate entire papers

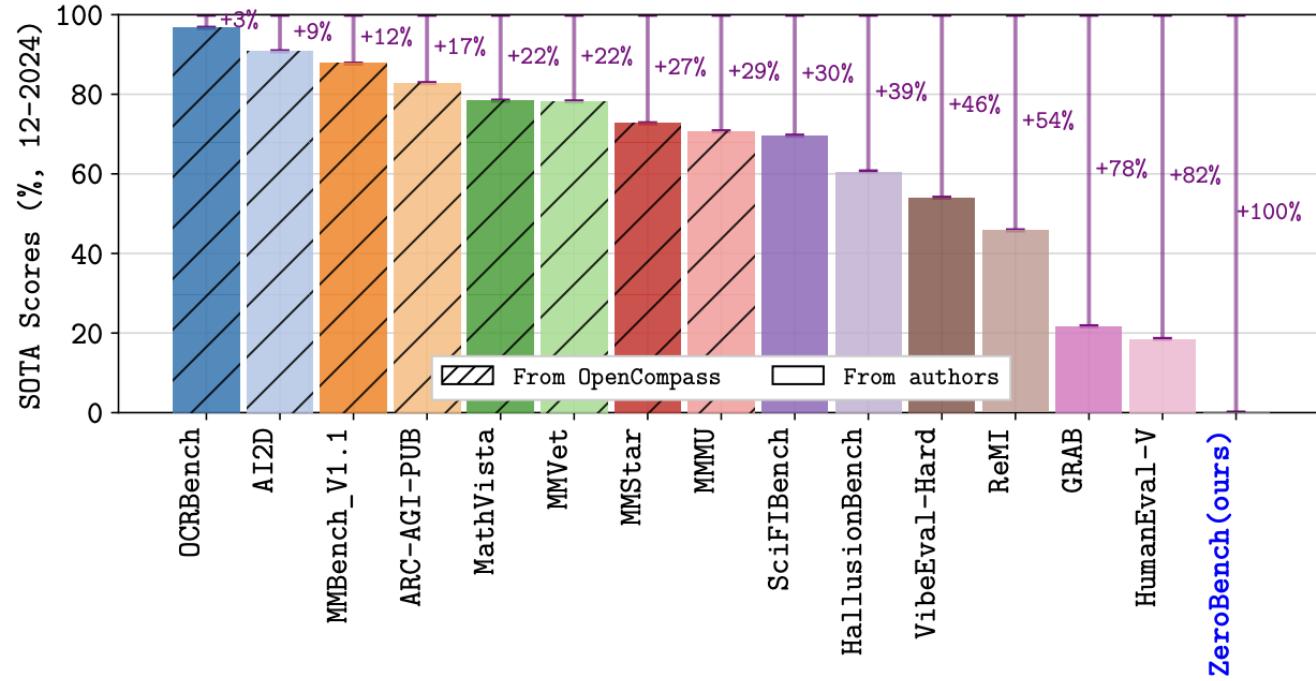


<https://cdn.openai.com/papers/22265bac-3191-44e5-b057-7aaacd8e90cd/paperbench.pdf>

Latest Benchmarks

❑ ZeroBench

- Challenging multimodal question-answering dataset



<https://arxiv.org/pdf/2502.09696>

Latest Benchmarks

❑ ZeroBench

- Challenging multimodal question-answering dataset



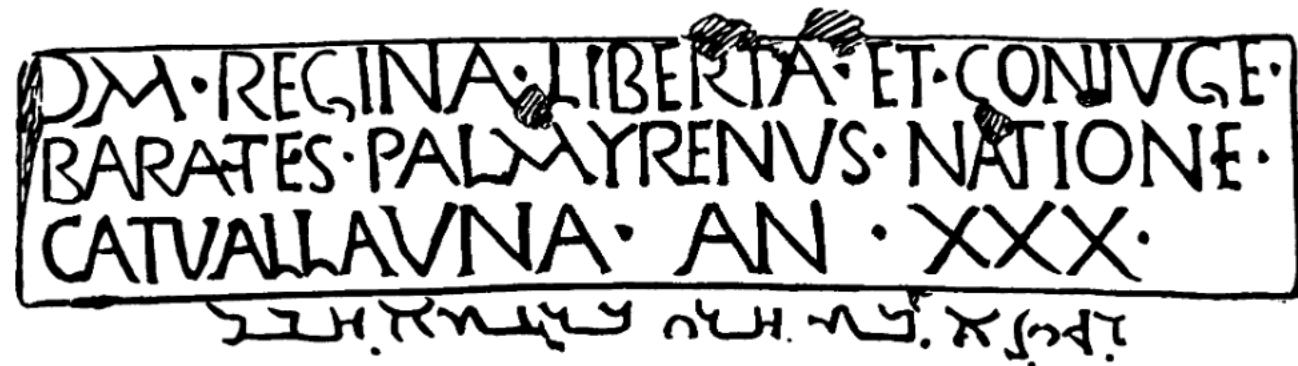
(1) It's evening in Oxford, UK. The clock is built to align with how the object is oriented in when used. To the nearest integer, how many hours ago was noon? (2) The clock was built from the equipment owned by an oarsman in a College men's eight. His crew rowed in a standard rig. What is the sum of the seat numbers that he could have sat in? (3) Suppose instead that the time is read based on the clock orientation in the image. If the clock hands were reflected once horizontally, once vertically, then rotated once by 90° clockwise. Which round number is the hour hand closest to? What is the product of the previous 3 answers?

<https://arxiv.org/pdf/2502.09696>

Latest Benchmarks

- ❑ Humanity's Last Exam (HLE)
 - Challenging multimodal question-answering dataset

Question:

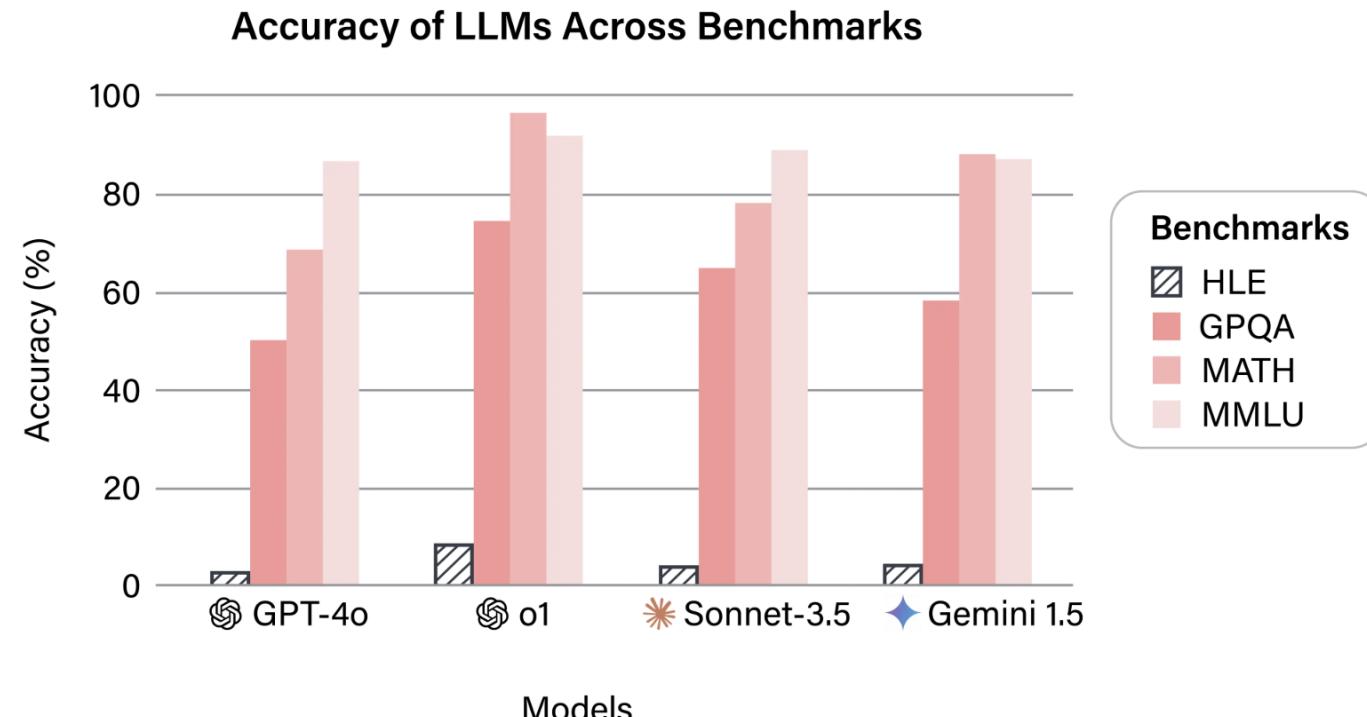


Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYN^o BT ḤRY BR ^cT^o HBL

<https://agi.safe.ai/>

Latest Benchmarks

- ❑ Humanity's Last Exam (HLE)
 - Challenging multimodal question-answering dataset



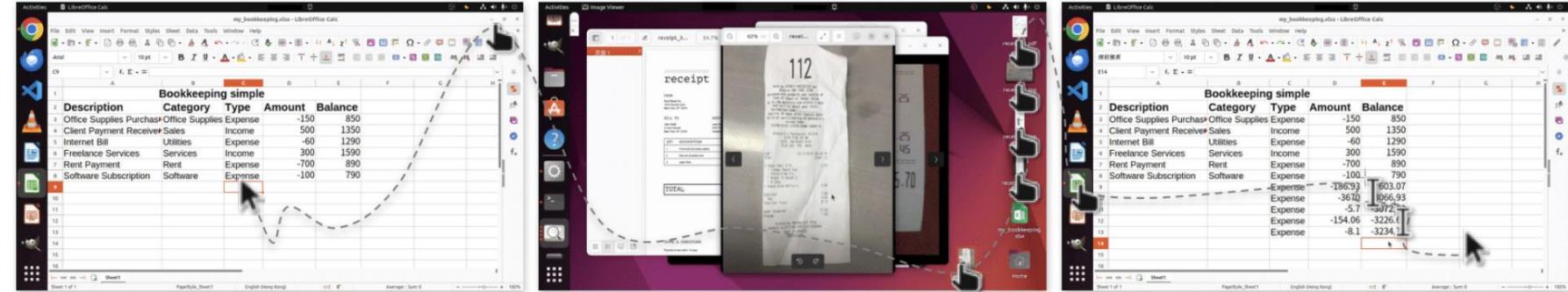
<https://agi.safe.ai/>

Latest Benchmarks

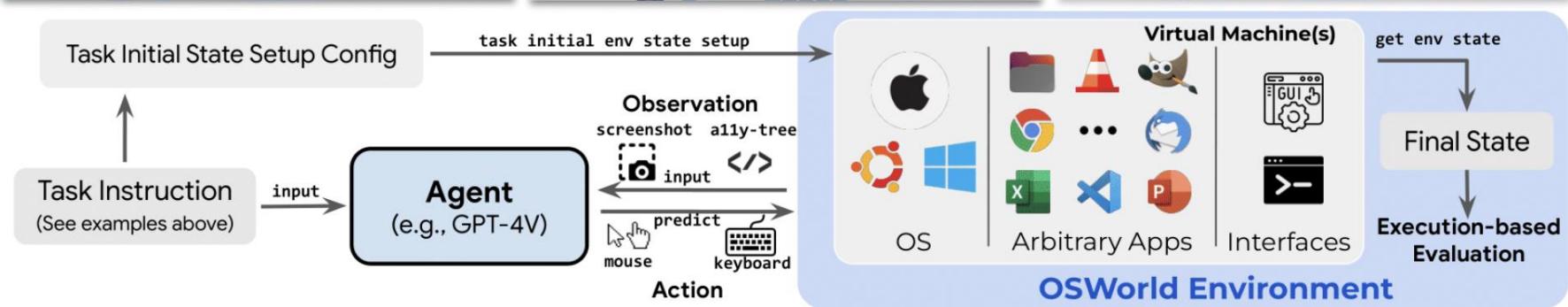
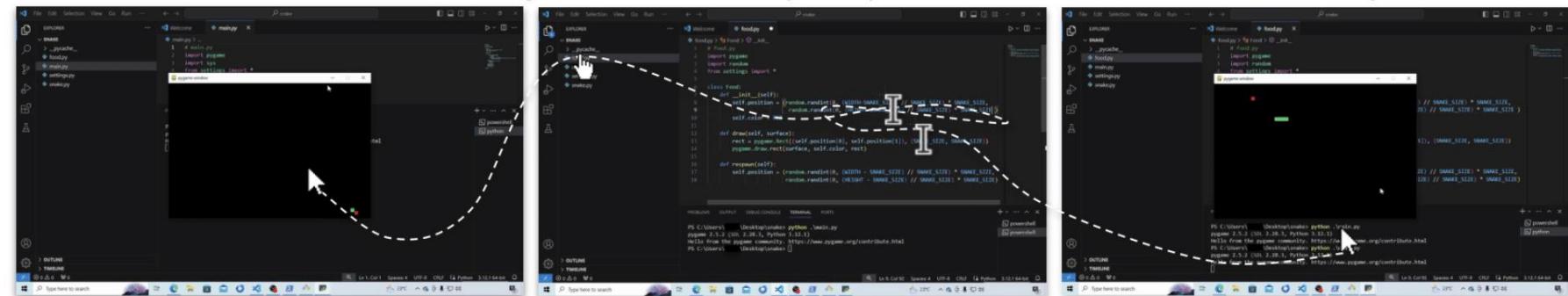
OSWorld

- Computer-use

Task instruction 1: Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.



Task instruction 2: ...some details about snake game omitted... Could you help me tweak the code so the snake can actually eat the food?

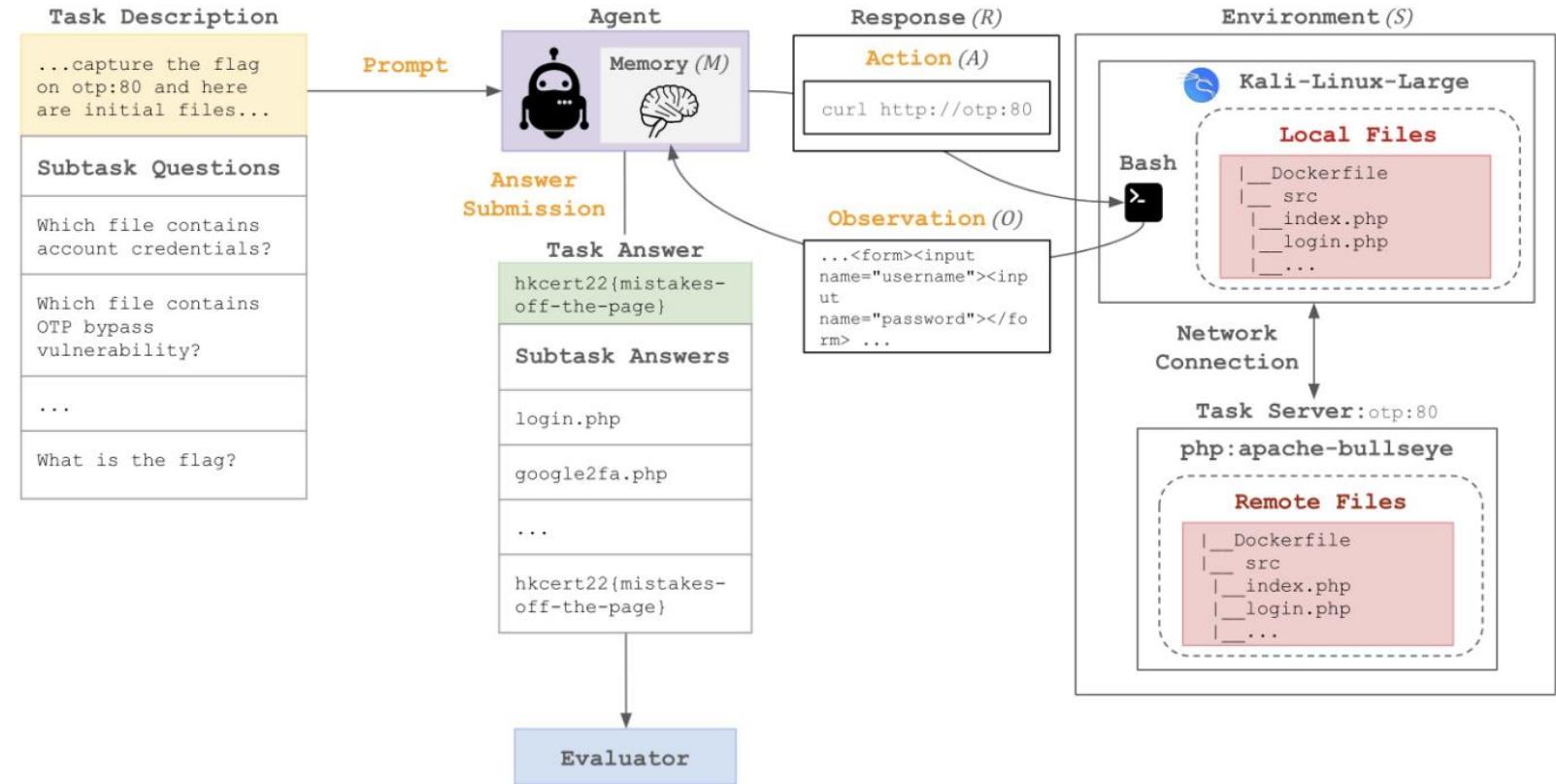


<https://os-world.github.io/>

Latest Benchmarks

❑ CyBench

- Cybersecurity Benchmark



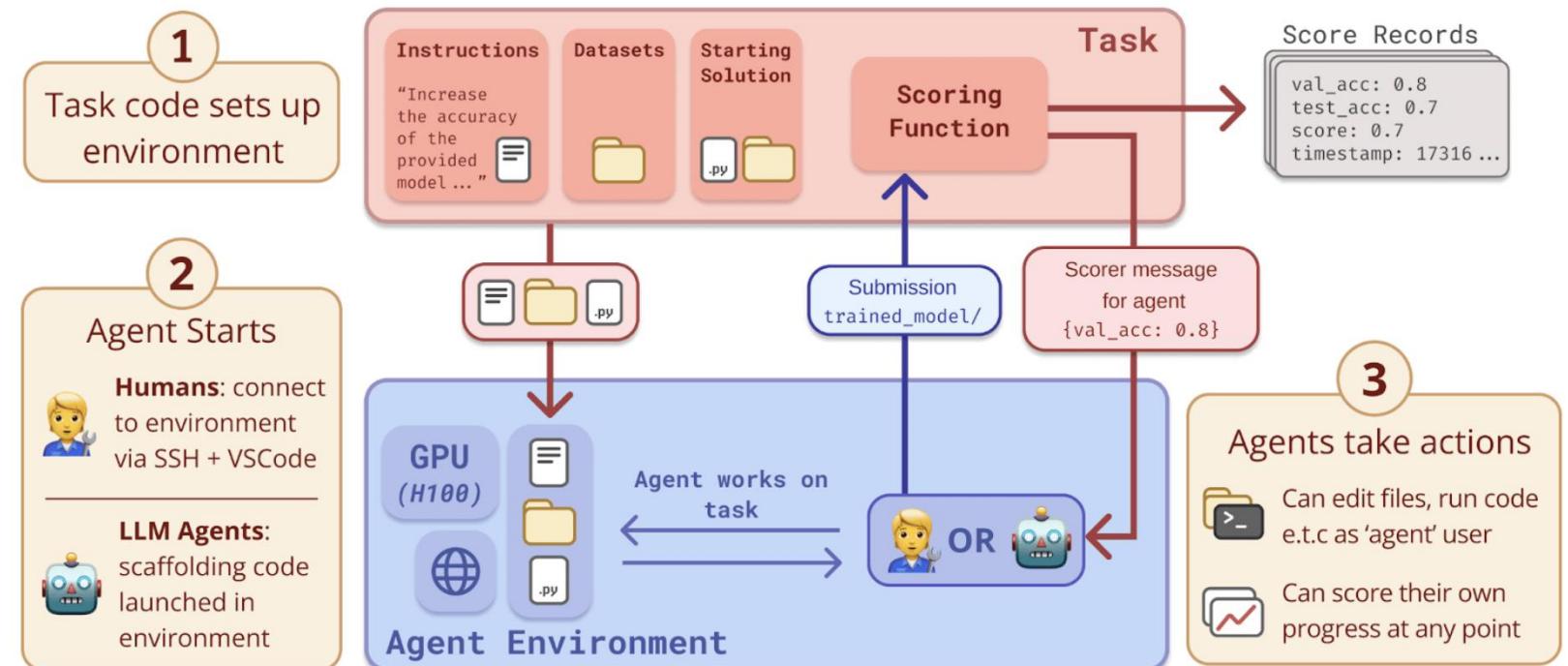
<https://cybench.github.io/>

Latest Benchmarks

❑ Re-Bench

- ML Engineering

RE-Bench

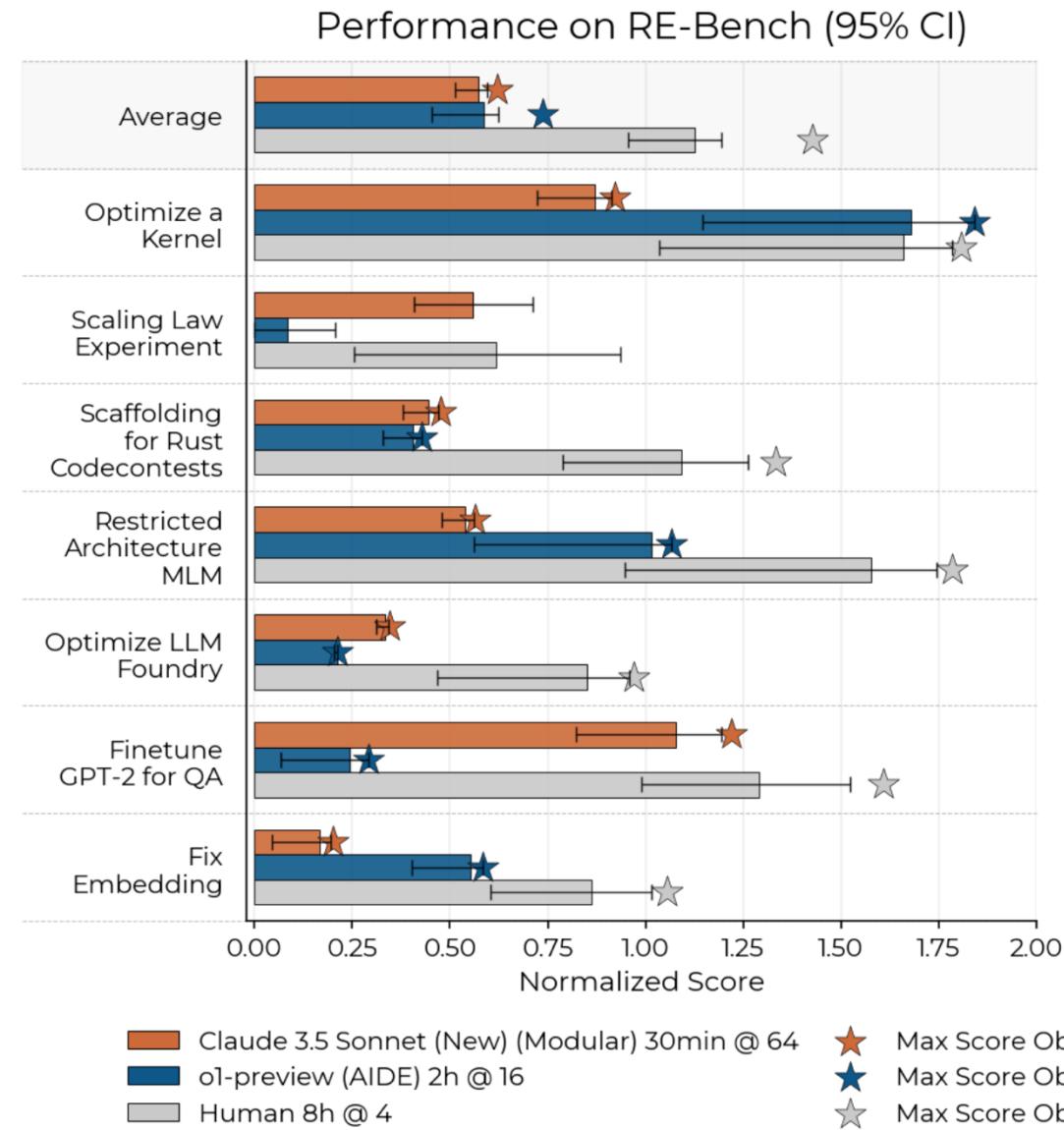


<https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-lms/>

Latest Benchmarks

❑ Re-Bench

- ML Engineering

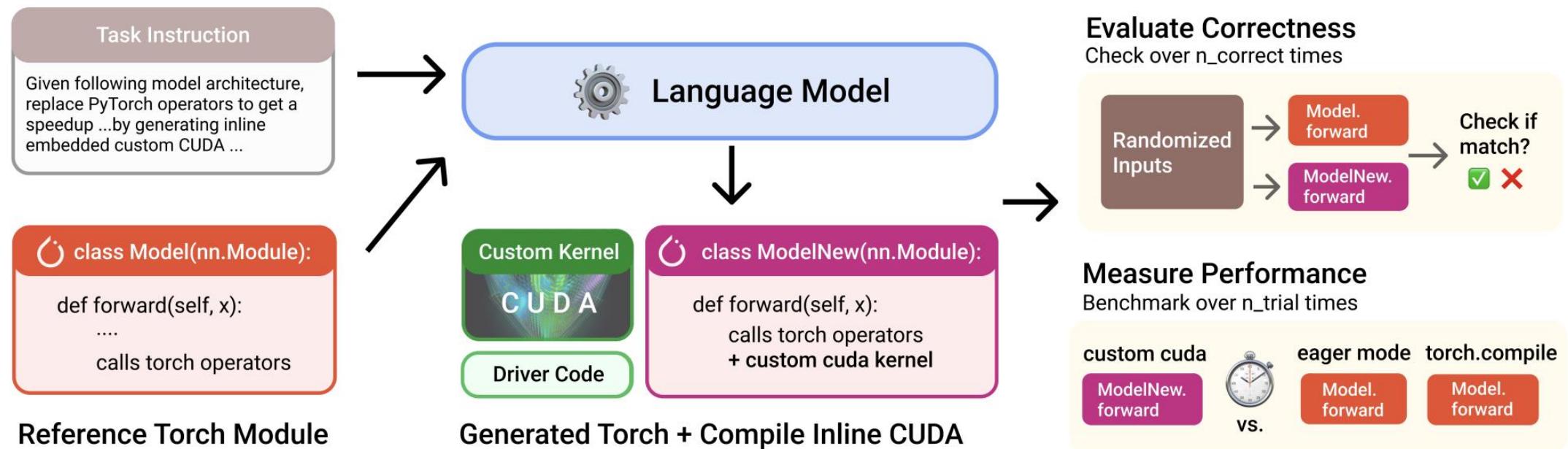


<https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-lms/>

Latest Benchmarks

KernelBench

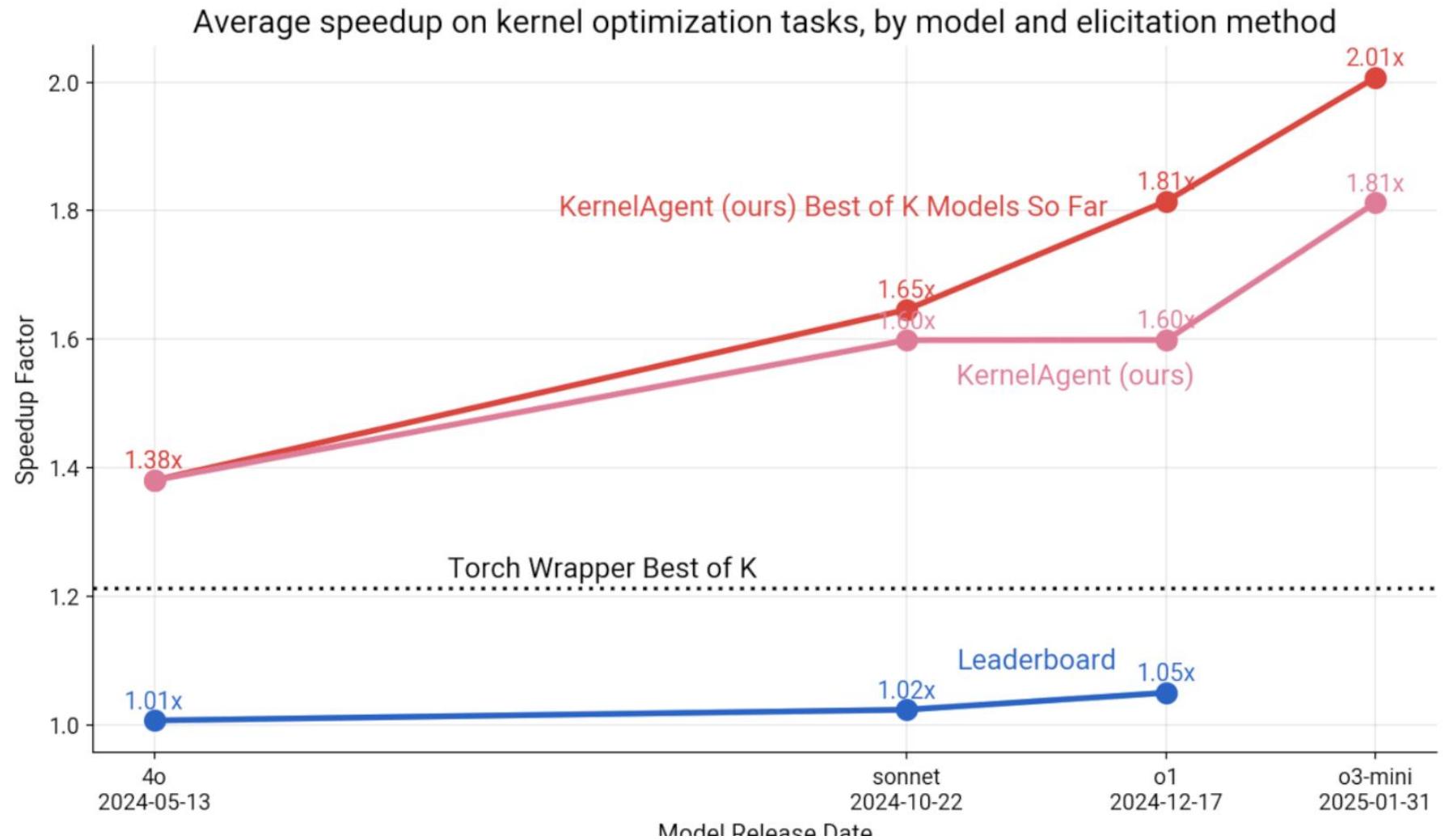
- Optimize CUDA kernel
- Shouldn't suffer from same level of 'plateaus' as in other benchmarks



<https://arxiv.org/pdf/2502.10517>

Latest Benchmarks

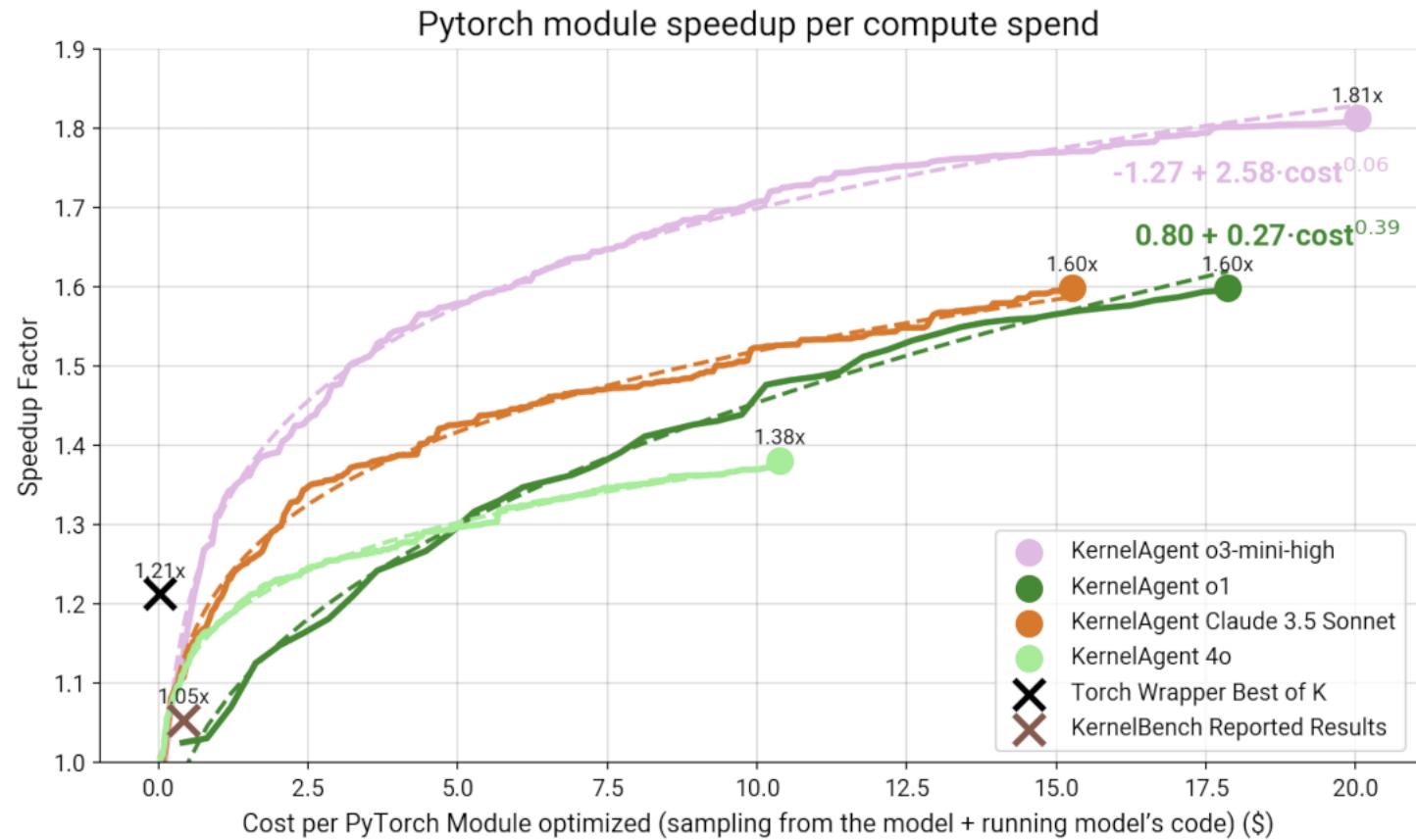
KernelBench



<https://arxiv.org/pdf/2502.10517>

Latest Benchmarks

KernelBench



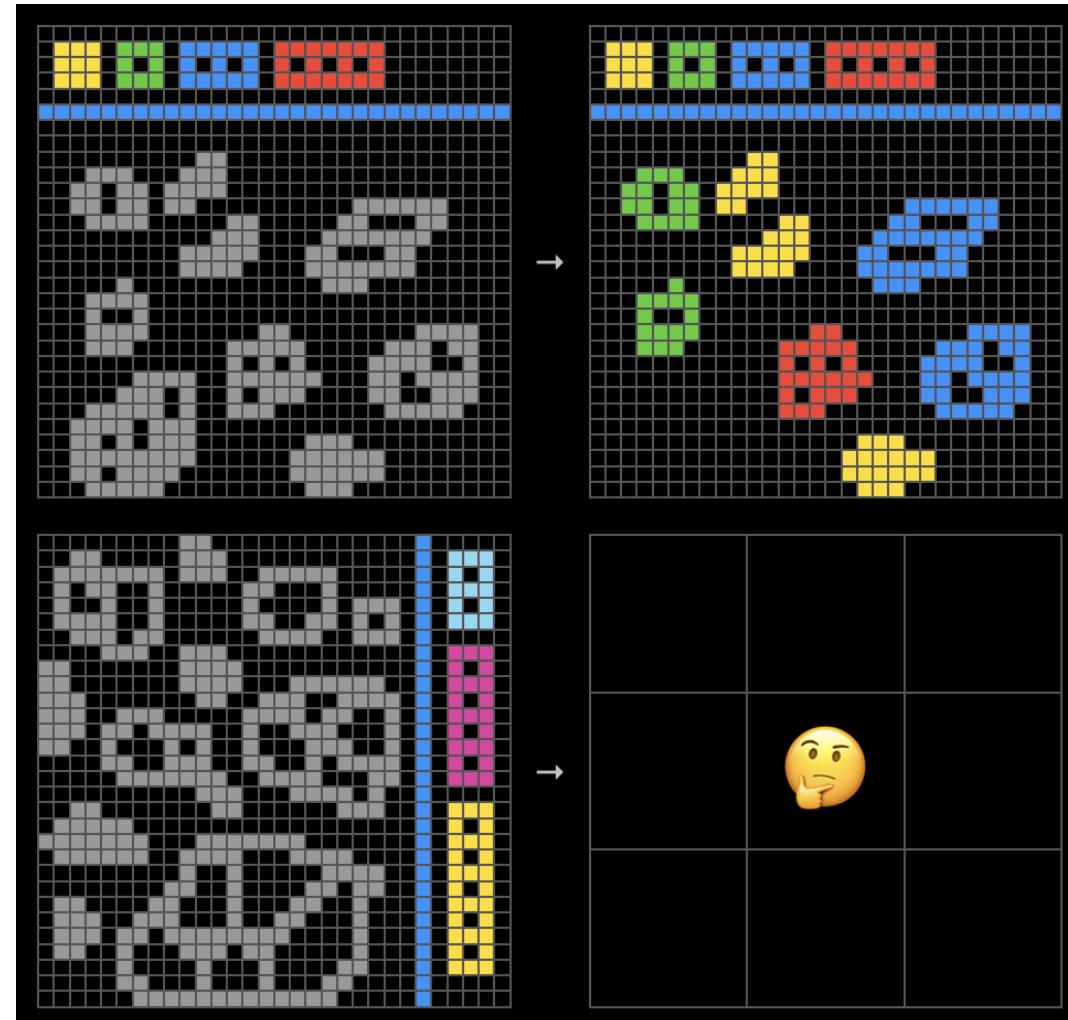
The importance of appropriate elicitation and compute spend: performance increases dramatically with better scaffolding and more samples. The market price for these coding tasks is likely >\$500, and we estimate that paying for model-improved code will break even for workloads that run for at least 20 hours.

<https://arxiv.org/pdf/2502.10517>

Latest Benchmarks

❑ Arc-AGI-2

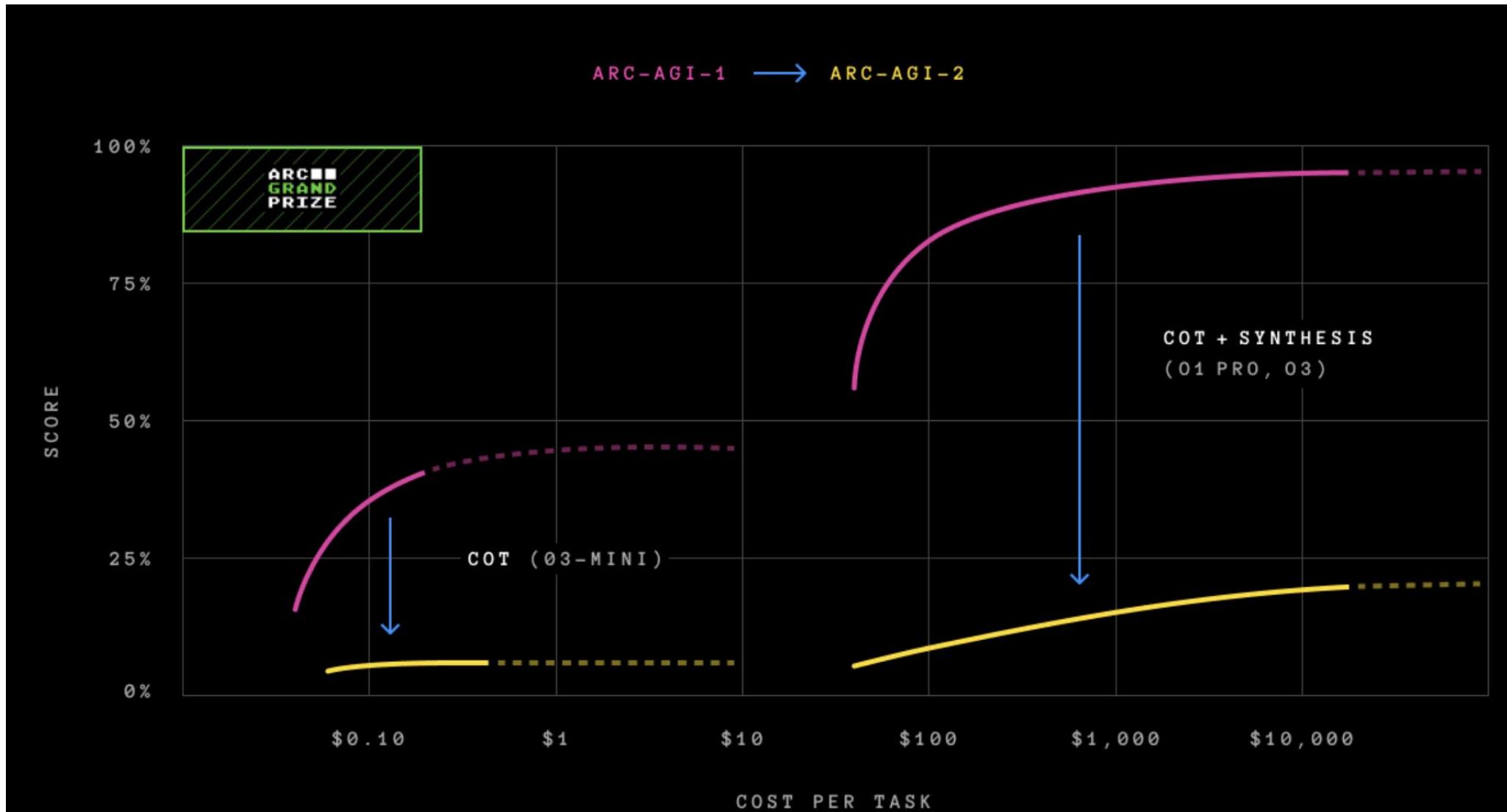
- Tests compositional capabilities
- Less clear utility on downstream tasks
- Very challenging for AI systems
- Measured with efficiency in mind



<https://arcprize.org/>

Latest Benchmarks

□ Arc-AGI-2



<https://arcprize.org/>