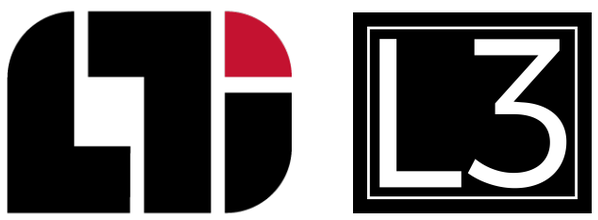
CS11-711 Advanced NLP

Evaluation and Benchmarks

Sean Welleck



https://cmu-l3.github.io/anlp-fall2025/

https://github.com/cmu-l3/anlp-fall2025-code

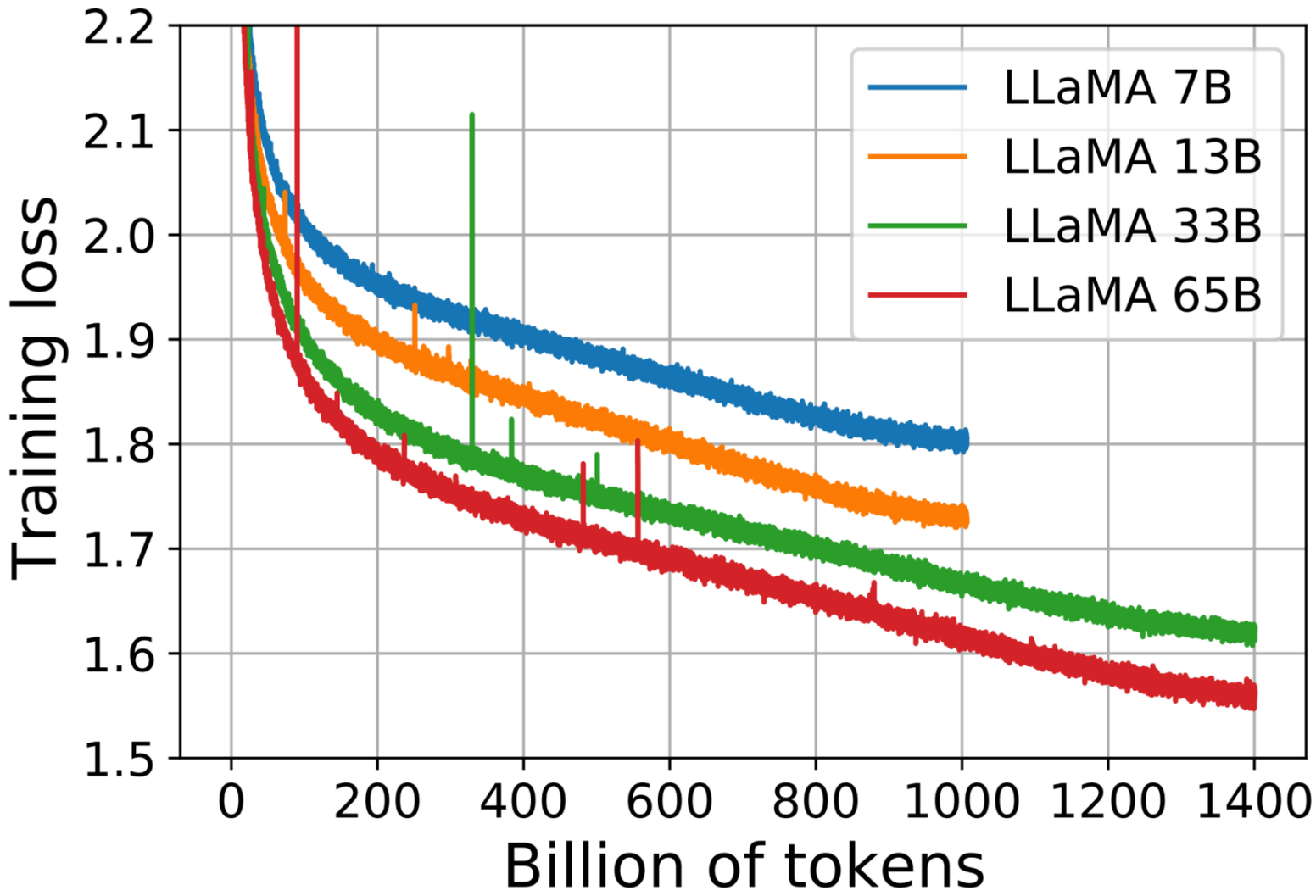
Recap

- Learning: pre-training, fine-tuning
- Inference: decoding algorithms
- Today: How do we evaluate how good our system is?
  - How do we choose between two systems?

Recap: Loss-Based Evaluation

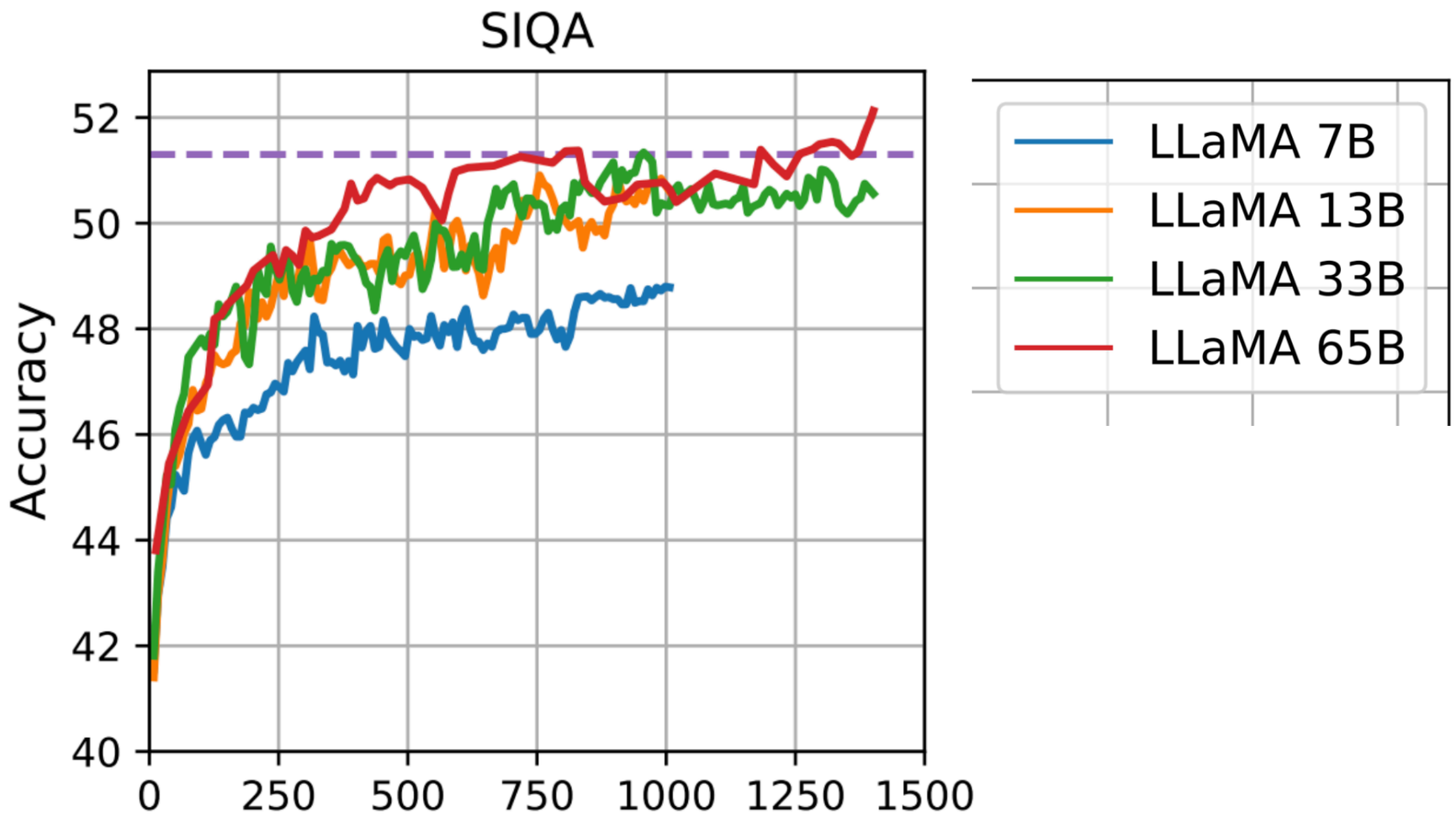
- Let p_θ be a language model
- Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be a dataset
- Let $\mathscr{L}(\theta; \mathscr{D})$ be the loss function
  - Typically $-\log p_\theta(y \mid x)$
- Then we can use the training loss, validation loss, or test loss to evaluate the model

Recap: Loss-Based Evaluation



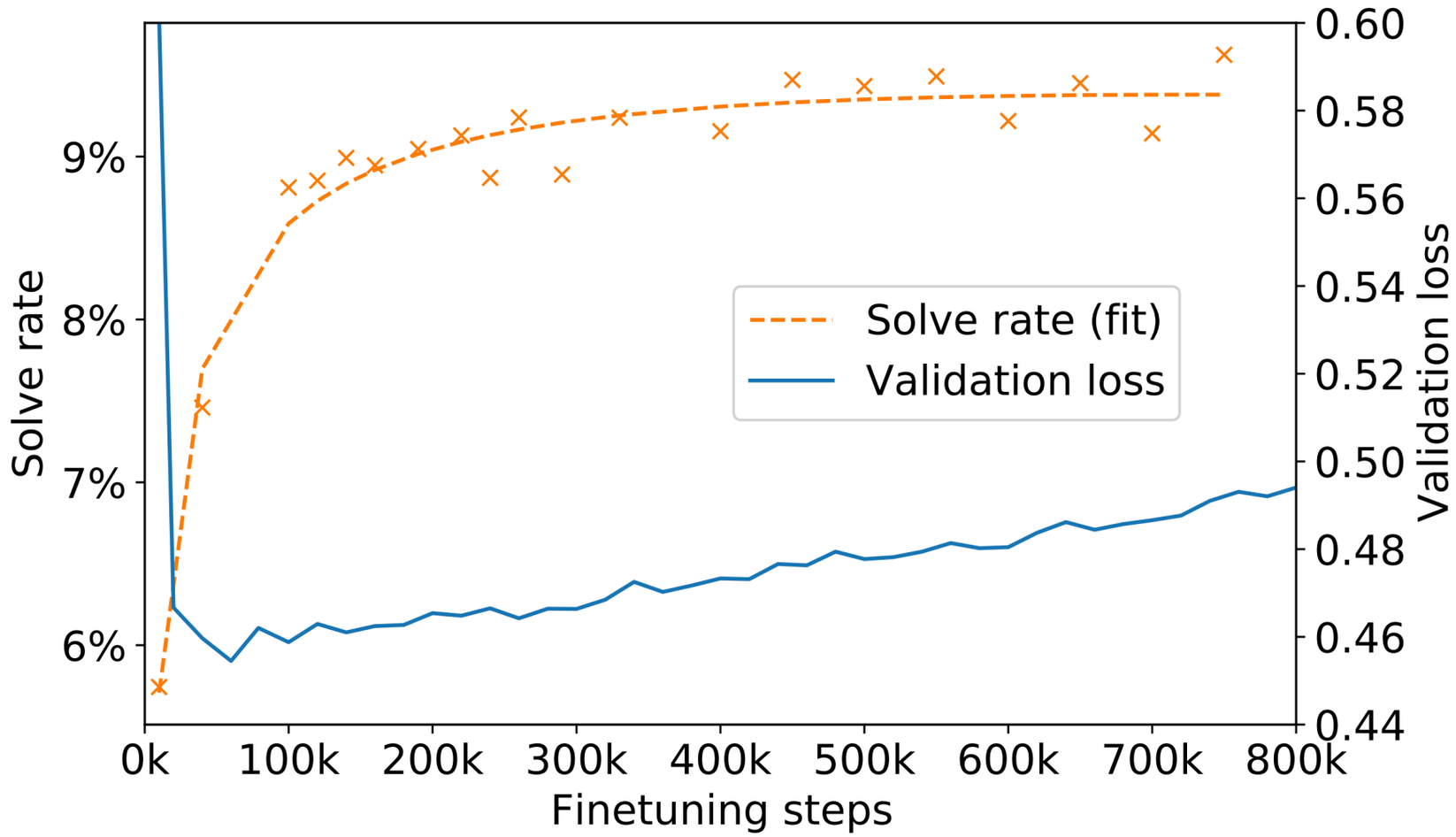
Claim: “Llama 33B is better than llama 13B”

Loss isn't always enough



Claim: "Llama 33B is better than llama 13B" (?)

Loss isn't always enough



Li et al. 2022 (DeepMind): *Competition-Level Code Generation with AlphaCode*

Loss isn’t always enough

- Loss measures how well the model fits the data distribution
- It does not directly measure how well a model performs tasks

Task metrics and downstream evaluation

- Let $g(p_\theta, x)$ denote running an inference algorithm
- Let $m(y, \hat{y}) \rightarrow \mathbb{R}$ be a task metric (e.g., solve rate)
- Let $p(D)$ be a distribution over task datasets
- We want the model to have high expected task performance

$$\mathbb{E}_{D \sim p(D)} \mathbb{E}_{x,y \sim D} \mathbb{E}_{\hat{y} \sim g(p_\theta, x)} [m(y, \hat{y})]$$

- Each dataset D represents a “downstream task”, i.e. a task that we use the model for downstream of its training

How do we evaluate?

- In practice, we use a finite number of *evaluation benchmarks.* Each benchmark typically consists of:
  - A dataset D (sometimes only inputs x)
  - A task metric m (typically $m(y, \hat{y})$)
- We want the benchmarks to cover the breadth and depth of functionality that we want in our system.
  - For a machine translation system, we want various machine translation benchmarks
  - For a language model API, we want benchmarks related to the many capabilities that customers want from the API

This lecture

- Intro/evaluation setup
- Properties of good benchmarks
  - Example benchmarks
  - Example task metrics
- Can I trust the benchmark score?

Example: Llama 3 paper

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**

Table 2 Performance of finetuned Llama 3 models on key benchmark evaluations. The table compares the performance of

- "Key benchmarks" in the Llama 3 paper

Example: MMLU

- Multiple-choice questions in 57 subjects
  - Model generates A, B, C, or D
  - Metric: exact match; is the answer the same as the one in the dataset
    - $m(y, \hat{y}) = 1[y = \hat{y}]$

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. ✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(D) Cannot say unless the speed of throw is given. ✗

Example: HumanEval

- HumanEval: LeetCode-style Python problems
  - Model generates code
  - Metric: execute the code and check whether it passes all test cases
    - $m(y, \hat{y}) = \prod_{j=1}^{numtests} 1[\hat{y} \text{ passes test } j]$
- $\text{pass@}K(y, \{y_1, \ldots, y_K\}) = \max\left(m(y, y_1), \ldots, m(y, y_K)\right)$

```python
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)
```

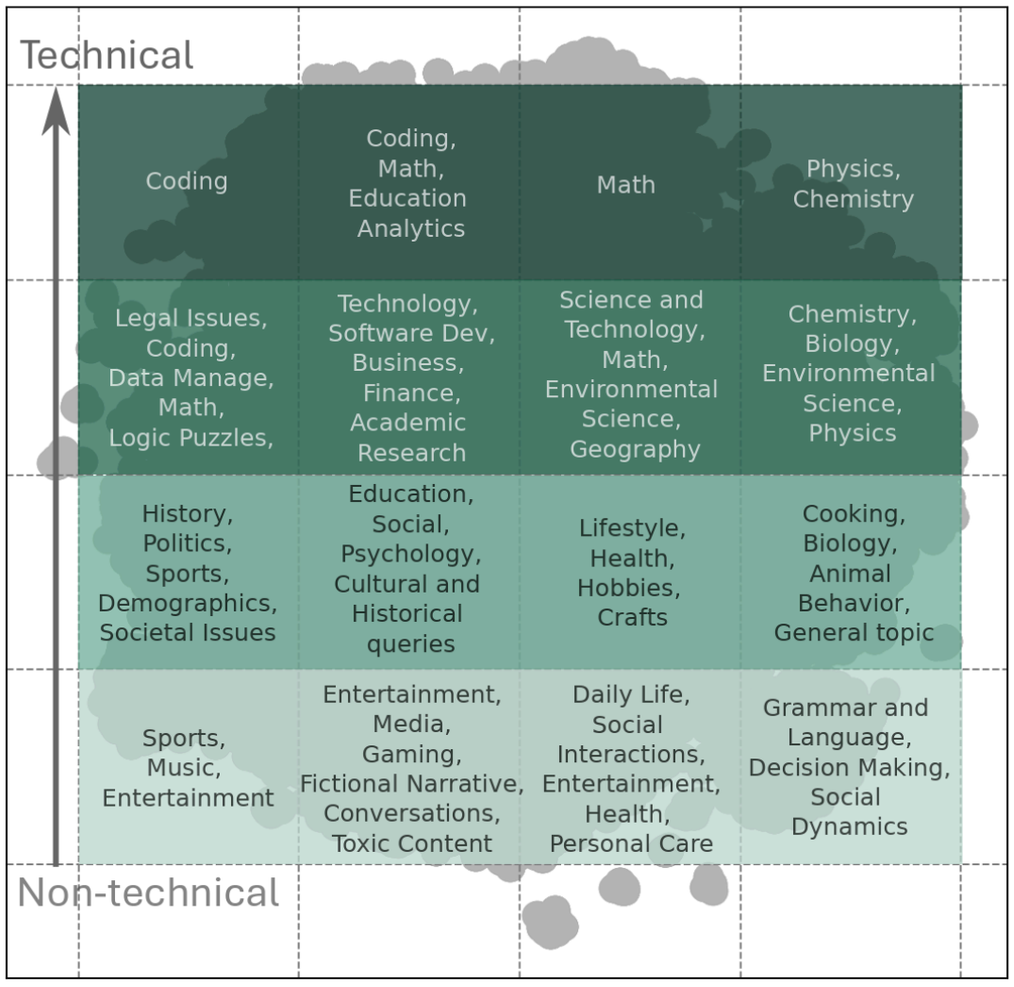
Example: GSM8k

- Grade school mathematics questions
  - Model generates a chain-of-thought and then an answer
  - Metric: check whether the answer is the same as the one in the dataset

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*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

Discussion: properties of good benchmarks?

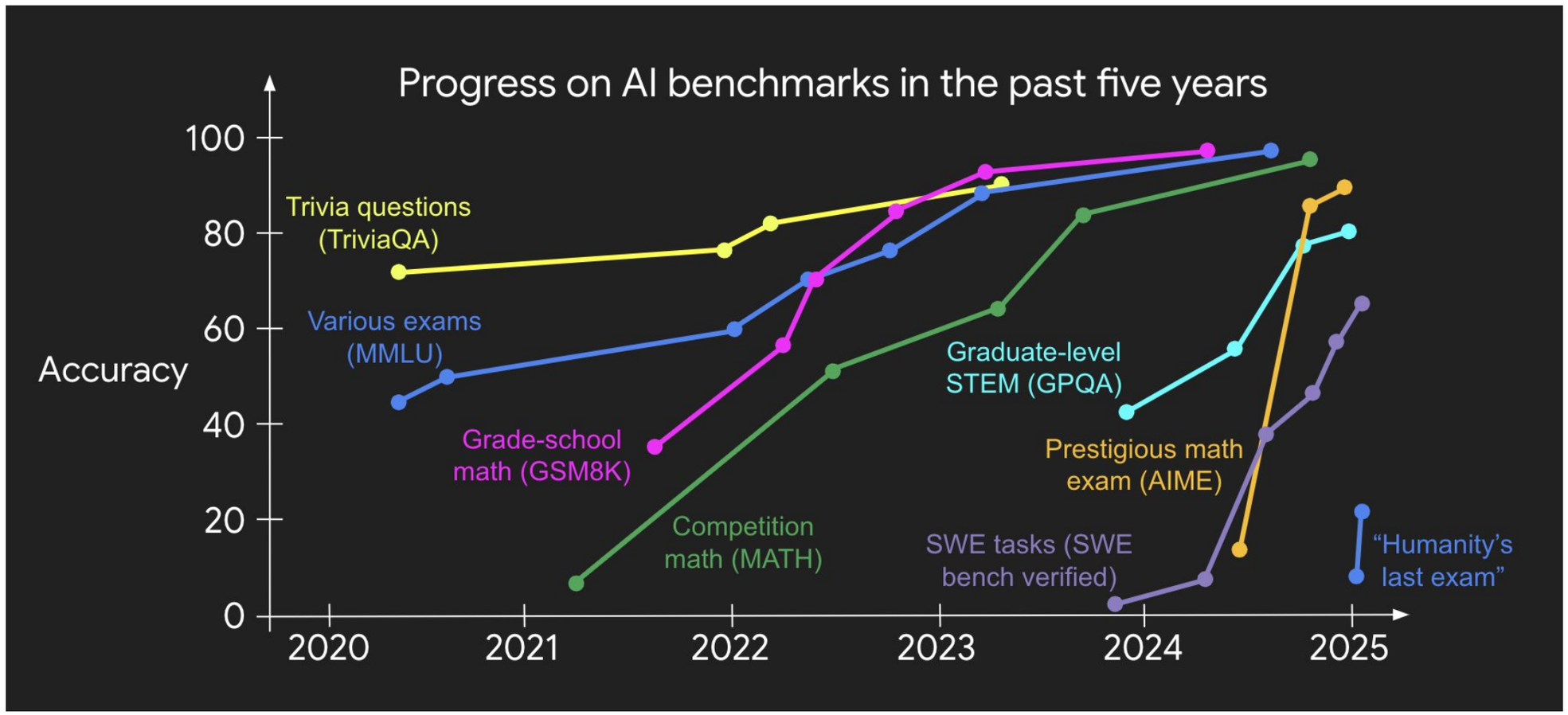
- Breadth/diversity
  - Evaluates all desired input-output functionality



Prompt types from active users [Ni et al 2024]

Discussion: properties of good benchmarks?

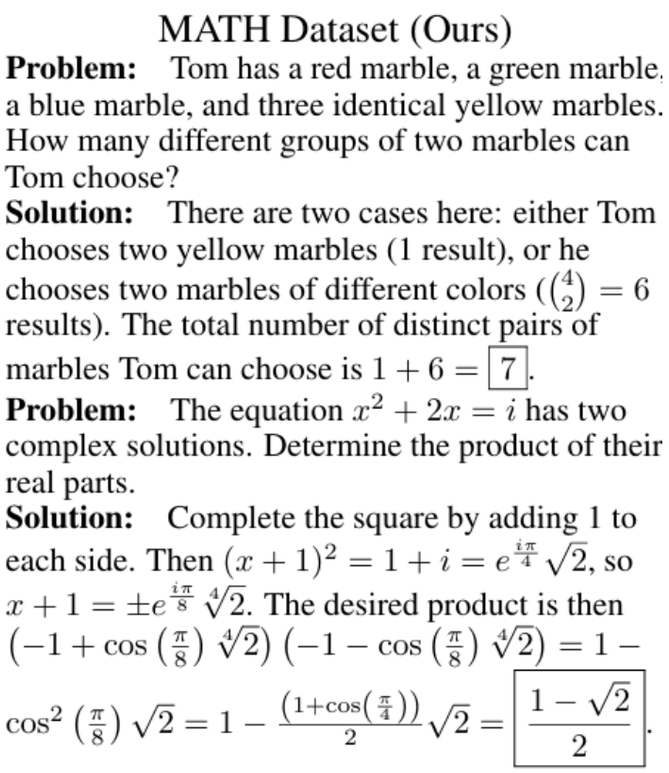
- Depth/difficulty
  - Examples are difficult enough to distinguish good and bad models.



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

Discussion: properties of good benchmarks?

- Utility
  - Task is a proxy for desired functionality



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

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

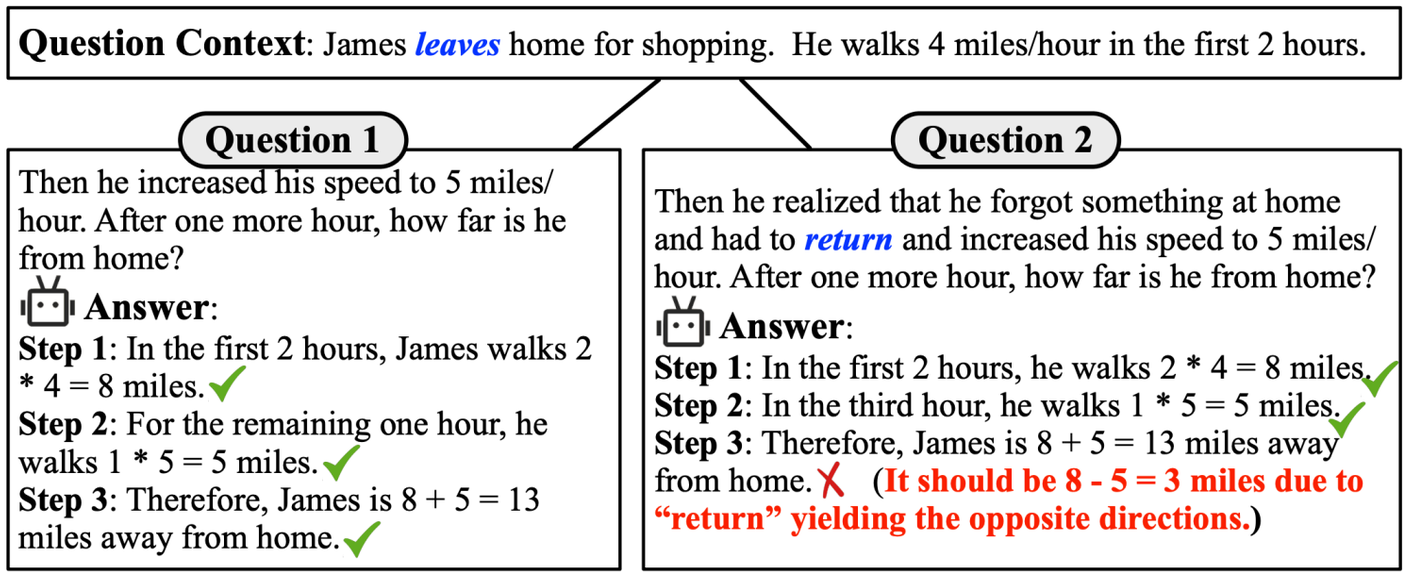
Foundation for **more complex tasks** (e.g., financial analysts)

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

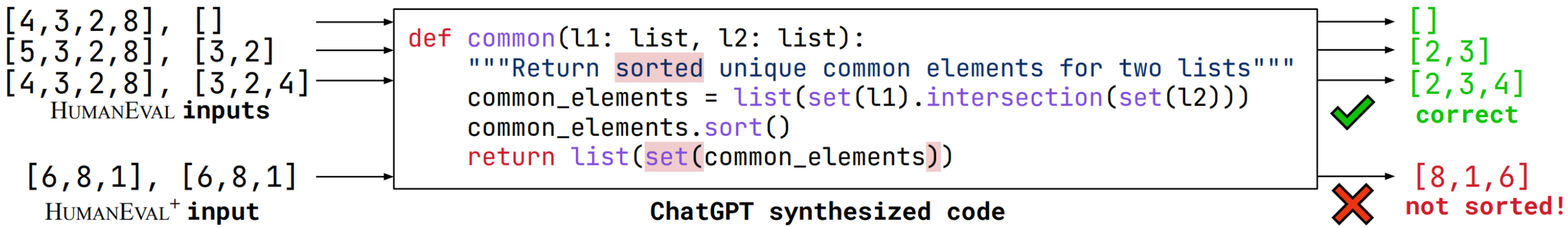
Medium for **addressing research questions** (e.g., Can an AI model reason?)

Discussion: properties of good benchmarks?

- Robustness



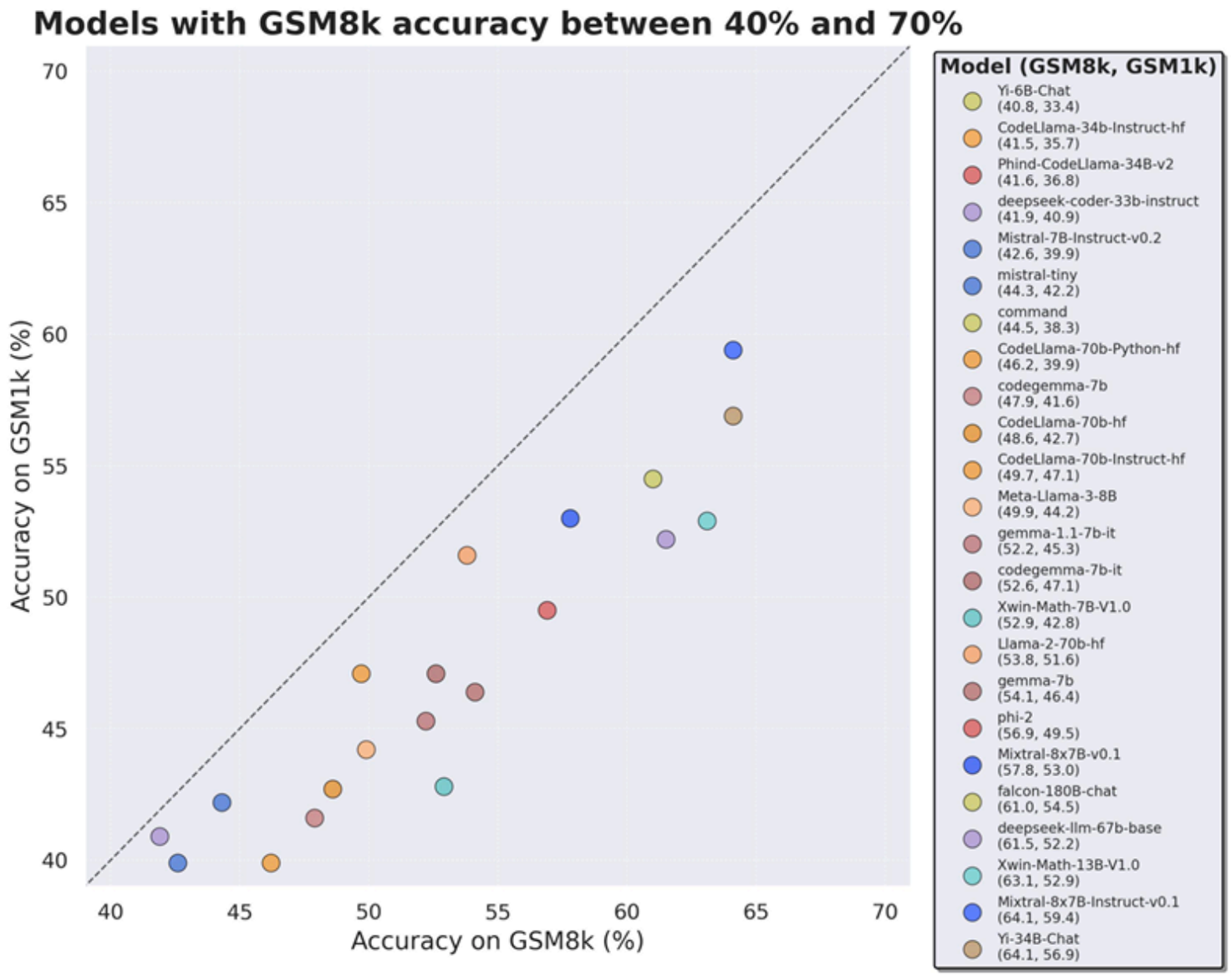
GSM-Plus: Robustness to input variations



HumanEval-Plus: a more robust metric

Discussion: properties of good benchmarks?

- Data contamination



Discussion: properties of good benchmarks?

- Data contamination

OpenWebMath document

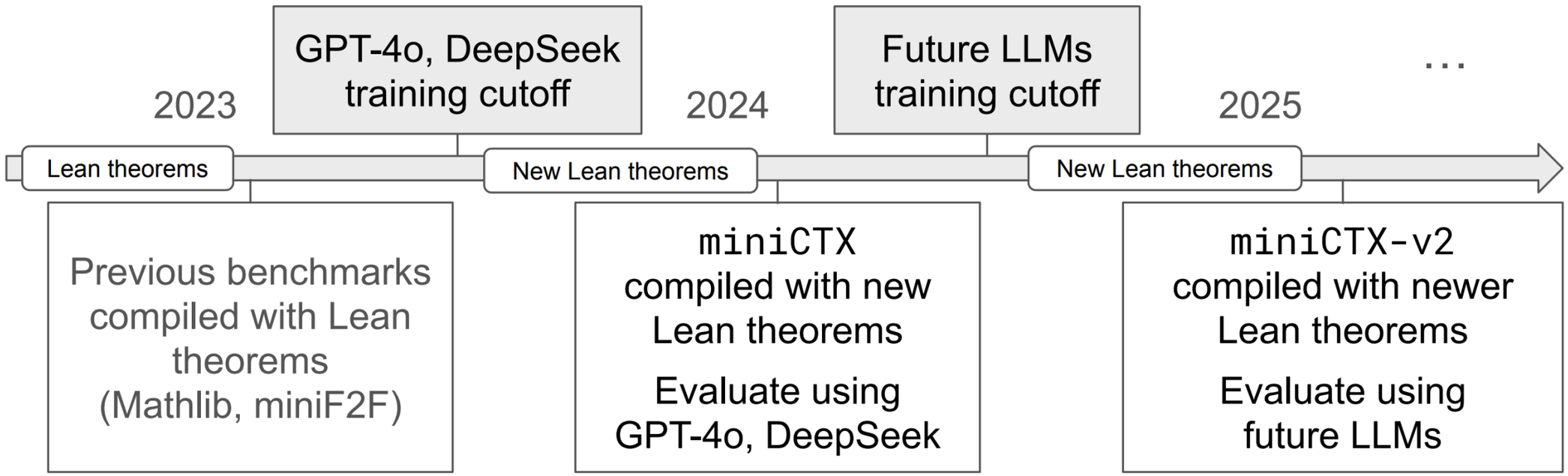
```
A triangle is formed with edges along the line $y=\frac{2}{3}x+5$,
the $x$-axis, and the line $x=k$. If the area of the triangle is less than $20$,
find the sum of all possible integral values of $k$.

Feb 28, 2018

Look at the graph, here...two triangles are possible :
https://www.desmos.com/calculator/m6wnjpgldq
The  height of the triangles at any point will be formed by

[ (2/3)x + 5 ]

And the bases will be  [ x - (- 7.5)]   = [ x + 7.5]
So....we want to solve this
(1/2) [ (2/3)x + 5 ] [ x + 7.5 ]  =  20
[ (2/3)x + 5 ] [ x + 7.5]  =  40
(2/3)x^2 + 5x + 5x  + 37.5  = 0
(2/3)x^2 + 10x - 2.5  =  0

Using a little technology.....the max x  value for the
triangle formed above the x axis will be = .246

And the min x value for the triangle formed below the x axis will be = -15.246

With the given boundaries, the  integer sums  of all possible x values of k giving
triangles with an area < 20 units^2  =

[ (-15) + (-14 ) + (-13) + ...+ ( -2) + ( - 1 ) + 0 ]  =
-  (15) (16) / 2   =
-120

Feb 28, 2018
```

MATH Problem: A triangle is formed with edges along the line $y = \frac{2}{3}x + 5$, the x-axis, and the line $x = k$. If the area of the triangle is less than 20, find the sum of all possible integral values of k.
Answer: -120

Source: Llemma paper [Azerbayev et al]

Discussion: properties of good benchmarks?

- Data contamination



"Temporal split" in miniCTX [Hu et al]

Discussion: properties of good benchmarks?

- Efficient to evaluate
    - Hire an expert human to read a long chat and decide if it's correct
        - Expensive, slow, error prone
- …
- Check whether a generated integer answer matches the answer in the dataset
    - Cheap, fast, exact, but simplifies the task

Automatically evaluating complex text

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.

Automatically evaluating complex text

- Option 1: word overlap
  - **Rouge**: Count the word overlap between the prediction & ref. answer

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 this desirable?

  - **Rouge-L**: [“exercise”, “mood”] and [“exercise”, “and”] are common subsequences
-> Precision: 2/5, Recall: 2/6, F1: 0.364

Discuss: Pros? Cons?

Automatically evaluating complex text

- Option 2: use a LLM ("LLM as a judge")
  - Example: $z \sim \boldsymbol{LLM}(x, y_A, y_B;$ prompt$)$
    - Prompt: "Which output is better, A or B? Think carefully."
  - Then we can compute the *win-rate* of model A versus model B.

Discuss: Pros? Cons?

Automatically evaluating complex text

- Option 2: use a LLM (“LLM as a judge”)

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.

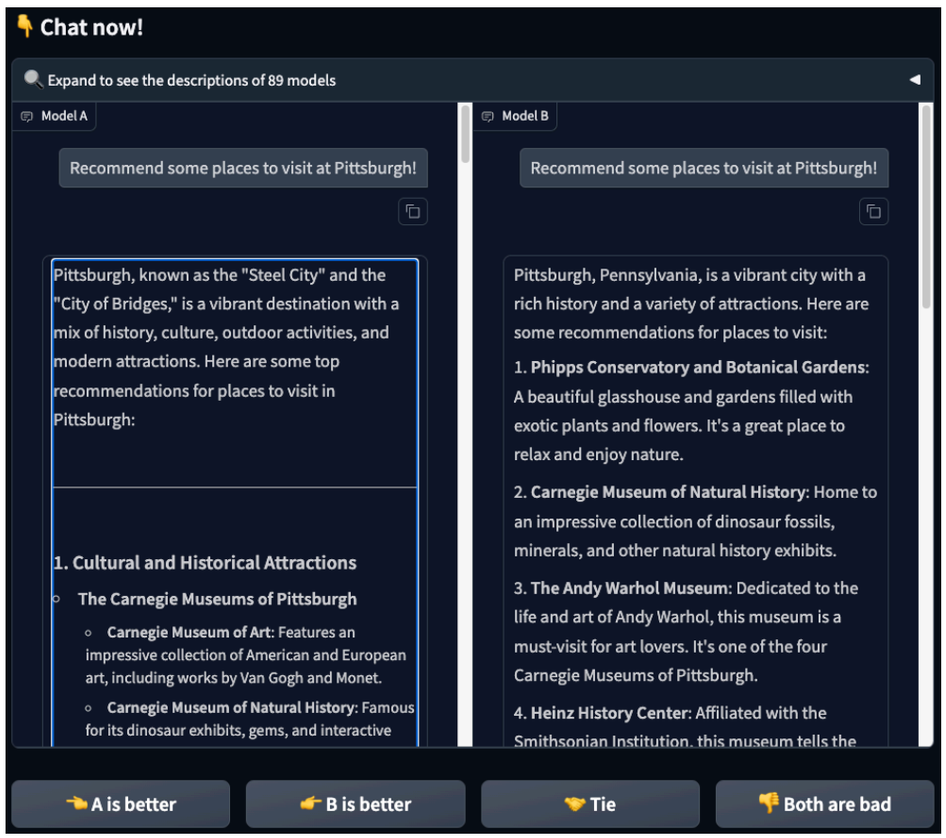
GPT-4 Judgment (when B is placed in the first position):

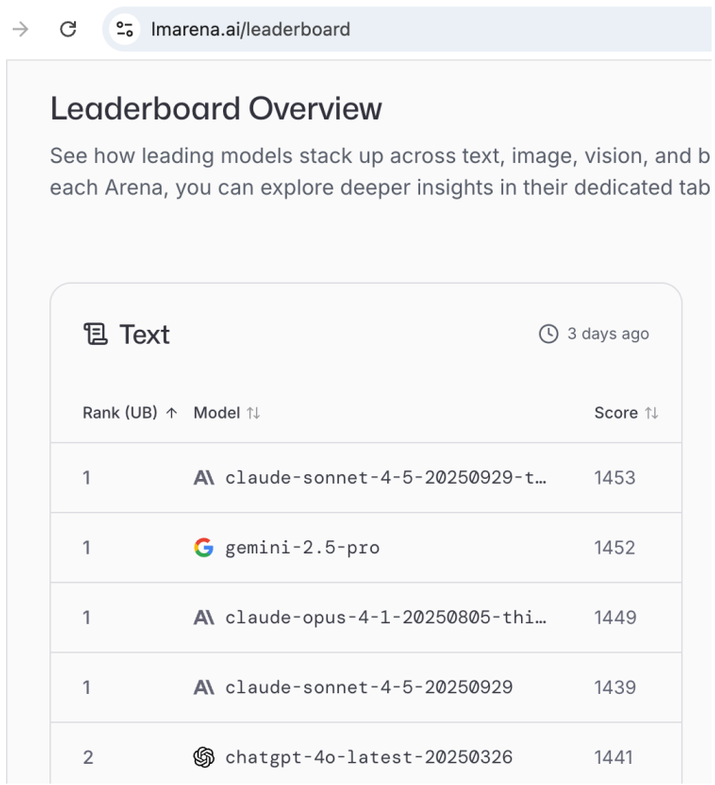
Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B’s answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication 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.

Discuss: Pros? Cons?

Automatically evaluating complex text

- Option 3: use humans
  - LMSys Chatbot Arena: ask humans to vote which response is better





Discuss: Pros? Cons?

Discussion: properties of good benchmarks?

- Breadth/diversity
- Depth/difficulty
- Utility
- Robustness
- Data (un-)contamination
- Efficient evaluation (automatic, low cost)

This lecture

- Intro/evaluation setup
- Properties of good benchmarks
  - Example benchmarks
  - Example task metrics
- **Next: What can I conclude from the benchmark score?**

Eval \ Model	"Galleon"	"Dreadnought"	Difference
MATH	65.5%	63.0%	+2.5%
HumanEval	83.6%	87.7%	−3.1%
MGSM	75.3%	78.0%	−2.7%

- What can we conclude?

Statistical background on evaluation

- Suppose an eval consists of *N* independently drawn questions, $q_1, \ldots, q_N$
- Let $\bar{s} = \frac{1}{n}\sum_i s_i$ be the average of observed model scores s_i
- Let μ be the unobserved true underlying score, $\mu = \mathbb{E}[s]$

Statistical background on evaluation

- By the law of large numbers, we can estimate $\mu \approx \bar{s}$
- By the central limit theorem, the standard error of the estimator can be estimated as:
- $$SE_{CLT} = \sqrt{Var(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_i - \bar{s})^2\right)/n}$$
- $$SE_{Bernoulli} = \sqrt{\bar{s}(1-\bar{s})/n}$$

Confidence interval

- $CI_{95\%} = \bar{s} \pm 1.96 \times SE$
- We can report:
  - Number of questions N

- The standard error or a confidence interval

	# Questions	"Galleon"	"Dreadnought"
MATH	5,000	65.5% (0.7%)	63.0% (0.7%)
HumanEval	164	83.6% (3.2%)	86.7% (3.0%)
MGSM	2,500	75.3% (0.9%)	78.0% (0.9%)

Code example

Clustered questions

- We assumed that questions are drawn independently, but often they are not
- For instance, we may have a single math problem translated into multiple languages (MGSM)
- We can account for such “clustering” of question using a different standard error estimator:

$$\mathrm{SE}_{\text{clustered}} = \left(\mathrm{SE}^2_{\text{C.L.T.}} + \frac{1}{n^2} \sum_c \sum_i \sum_{j \neq i} (s_{i,c} - \bar{s})(s_{j,c} - \bar{s}) \right)^{1/2}$$

Clustered questions

	# Questions	# Clusters	"Galleon"	"Dreadnought"
DROP	9,622	588	87.1 (0.8)	83.1 (0.9)
RACE-H	3,498	1,045	91.5% (0.5%)	82.9% (0.7%)
MGSM	2,500	250	75.3% (1.6%)	78.0% (1.5%)

	$SE_{clustered}$	$SE_{C.L.T.}$	Ratio
DROP	(1.34)	(0.44)	3.05
RACE-H	(0.51%)	(0.46%)	1.10
MGSM	(1.62%)	(0.86%)	1.88

Comparing models: unpaired

- Difference of means: $\hat{\mu}_{A-B} = \hat{\mu}_A - \hat{\mu}_B$
    - Null hypothesis: difference of means is 0
- Standard error: $SE_{A-B} = \sqrt{SE_A^2 + SE_B^2}$
- Confidence interval: $CI_{A-B,95\%} = \hat{\mu}_{A-B} \pm 1.96 \times SE_{A-B}$
    - If this doesn't include 0, the result is statistically significant
- Compute z score: $z_{A-B} = \hat{\mu}_{A-B} / SE_{A-B}$
    - Standardizes the difference
- Get associated *p*-value
    - Probability of observing this difference under the null hypothesis
- If below a threshold (e.g., $p < 0.01$), reject the null hypothesis

Code example

Comparing models: paired

- Evaluate both systems on the same examples
  - Suppose we have access to all of the evaluations, (x, y_A, y_B)
  - Then we can use a “paired” test that typically has reduced variance.

$$\mathrm{SE}_{A-B,\mathrm{paired}} = \sqrt{\mathrm{Var}(s_{A-B})/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_{A-B,i} - \bar{s}_{A-B})^2\right)/n}$$

Comparing models: paired

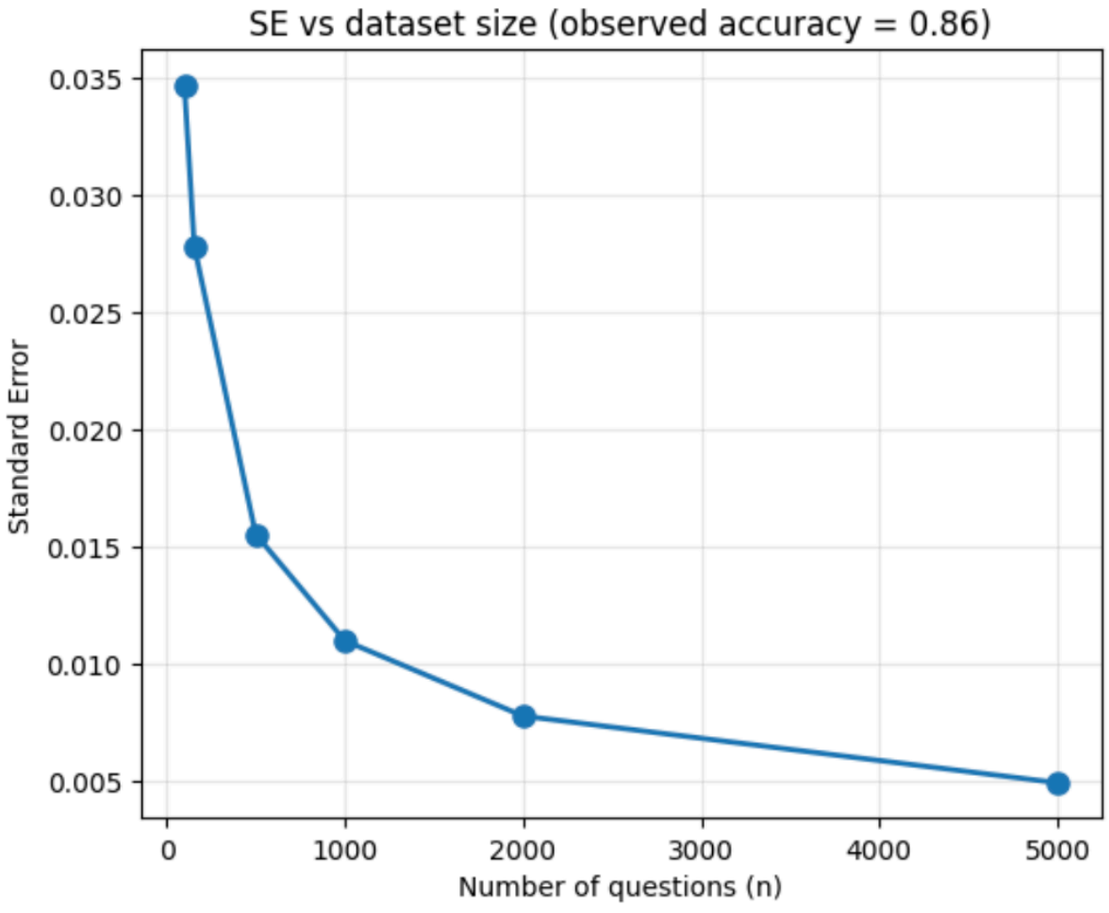
Eval	Model	Baseline	Model – Baseline	95% Conf. Interval	Correlation
MATH	Galleon	Dreadnought	+2.5% (0.7%)	(+1.2%, +3.8%)	0.50
HumanEval	Galleon	Dreadnought	−3.1% (2.1%)	(−7.2%, +1.0%)	0.64
MGSM	Galleon	Dreadnought	−2.7% (1.7%)	(−6.1%, +0.7%)	0.37

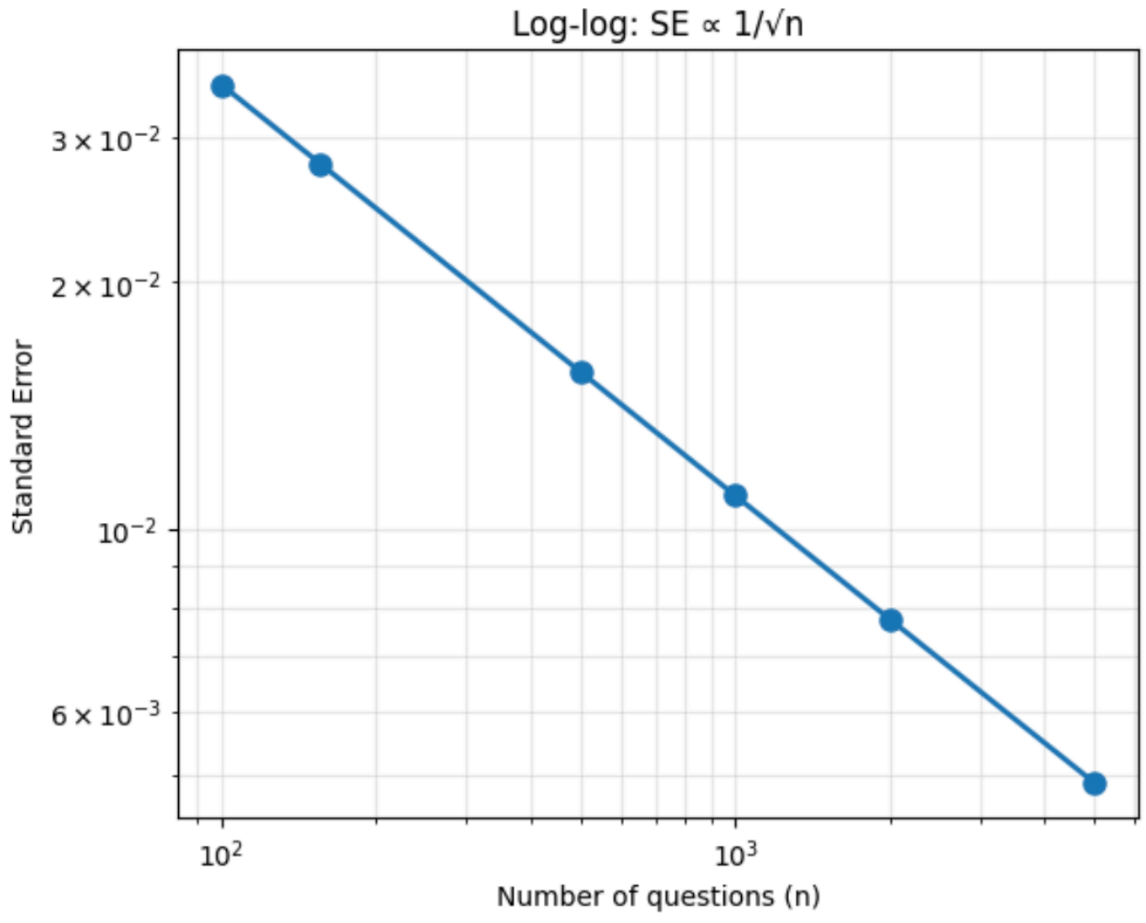
Code example

Variance reduction

- Recall that the estimator is:
  - $\hat{\mu} = \sum_{i=1}^{n} s_i / n$
- Then the variance is $Var(\hat{\mu}) = Var(s)/n$
- To reduce variance:
  - Increase number of questions n
  - If we are using stochastic decoding, sample more outputs and take the average as the score.

Variance reduction






```
n=  100: SE=0.0347
n=  156: SE=0.0278
n=  500: SE=0.0155
n= 1000: SE=0.0110
n= 2000: SE=0.0078
n= 5000: SE=0.0049
```

This lecture

- Intro/evaluation setup
- Properties of good benchmarks
  - Example benchmarks
  - Example task metrics
- What can I conclude from the benchmark score?
  - Basic confidence intervals, hypothesis testing

Thank you