



**ICML**  
International Conference  
On Machine Learning

Tutorial @ ICML 2024

# Challenges in LM Evaluation

Lehar 1-4

3.30 - 5.30pm CEST



**Lintang Sutawika**  
@lintangsutawika



**Hailey Schoelkopf**  
@haileysch\_

# Experience from the Trenches

The screenshot shows a GitHub repository page for 'lm-evaluation-harness'. The repository is marked as 'Public'. At the top, there are buttons for 'Edit Pins', 'Unwatch' (36), 'Fork' (1.6k), and 'Starred' (5.9k). Below this, there are navigation links for 'main' (selected), '72 Branches', and '7 Tags'. A search bar contains 'Go to file' and a dropdown for 'Add file'. A green button labeled 'Code' with a dropdown arrow is also present. To the right, a description reads: 'A framework for few-shot evaluation of language models.'

- Unified library for prompted LM evaluations
- Frequently used by LM trainers and researchers
- Backend for [Open LLM Leaderboard](#)'s evaluation tasks
- Experience reproducible evaluation, and *seeing what can go horribly wrong*

# Outline

- **Fundamentals of LM Evaluation**

Evaluation background, Measurement methods, Metrics, ...

- **LM-specific complications**

Unique reproducibility difficulties, Non-robustness, Data contamination, ...

- **General benchmarking complications**

General evaluation pitfalls: Measurement validity, Benchmark saturation, ..

- **Addressing Pitfalls**

Publishing evaluation code, Better reporting, ...

- **Future Directions**

Dynamic eval sets, Evaluating more complex capabilities, Multimodality, Agents, ...

# Goals

You should leave with understanding/knowledge of

- **How LM Evaluation is currently performed**
- **What issues are often faced in evaluating LMs**
- **Best practices for reliable, reproducible LM evaluation**
- **Areas that are open for future research**

# Scope

- Primary focus: evaluation of *base* and *instruction-tuned* LMs
- On zero- and few-shot prompted tasks
- What won't be the focus:
  - Agent Evaluation
  - Tool Use + Function Calling
  - Retrieval-Augmented Generation ("RAG")

# A Key Challenge in LM Evaluation

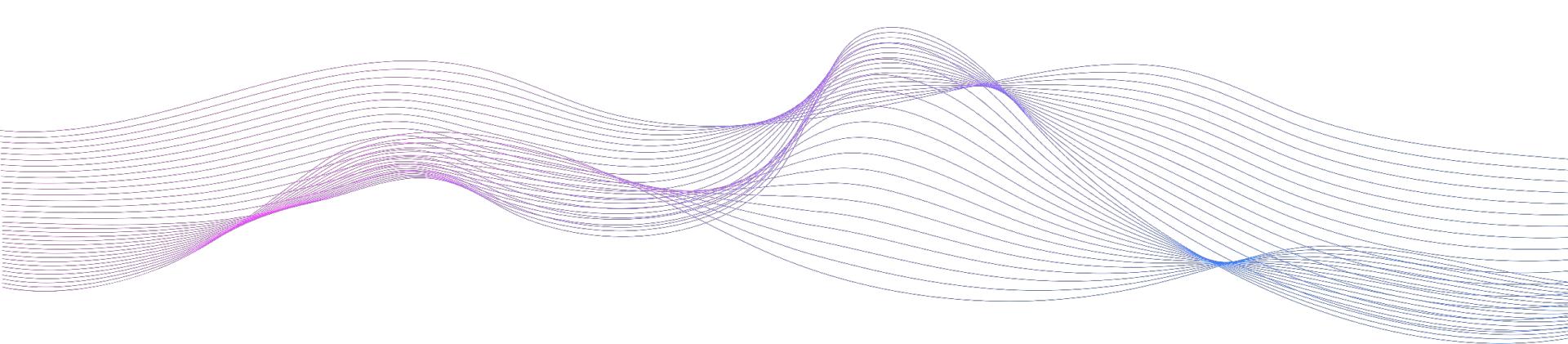
Source	The Bells of St. Martin's Fall Silent as Churches in Harlem Struggle .
Translation	Die Glocken von St. Martin verstummen , da Kirchen in Harlem Probleme haben .
Paraphrase	Die Probleme in Harlems Kirchen lassen die Glocken von St. Martin verstummen .
Paraphrase	Die Kirchen in Harlem kämpfen mit Problemen , und so läuten die Glocken von St. Martin nicht mehr .

[Freitag et al. \(2020\). BLEU might be Guilty but References are not Innocent](#)

There can be many semantically equivalent but syntactically different ways of expressing the same idea.

However, the best tools are the very models we are seeking to evaluate.

There are no perfect ways to evaluate the correctness of arbitrary natural language responses



# LM Evaluation Fundamentals

# Why Evaluate?

- Track progress in the field
- Compare and rank models
- Evaluate progress during training / finetuning
- Measure “intrinsic capabilities”
- Prevent regressions

# Why Evaluate?

- **Tracking progress**
  - Are models getting stronger?
- **Quantitative measures**
  - Able to objectively, reproducibly argue for improvement



# Why Evaluate?

- **Making Comparisons**

- Is method X better than the baseline method Y?
- In what situations is X better?
- Which model should I use for my task?

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	<b>82.0</b>	81.9	79.0
GPQA 0-shot	39.5	<b>41.5</b> CoT	<b>38.5</b> CoT
HumanEval 0-shot	<b>81.7</b>	71.9	73.0
GSM-8K 8-shot, CoT	<b>93.0</b>	91.7 11-shot	<b>92.3</b> 0-shot
MATH 4-shot, CoT	50.4	<b>58.5</b> Minerva prompt	40.5

# Why Evaluate?

- **Assess training runs**
  - Sanity-check training, compare ablations, ...
- **Prevent regressions**
  - During fine-tuning, model compression, ...

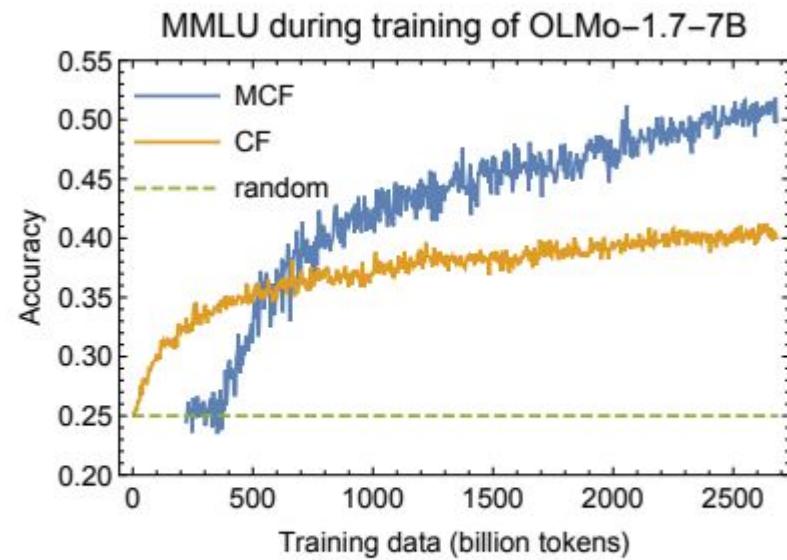
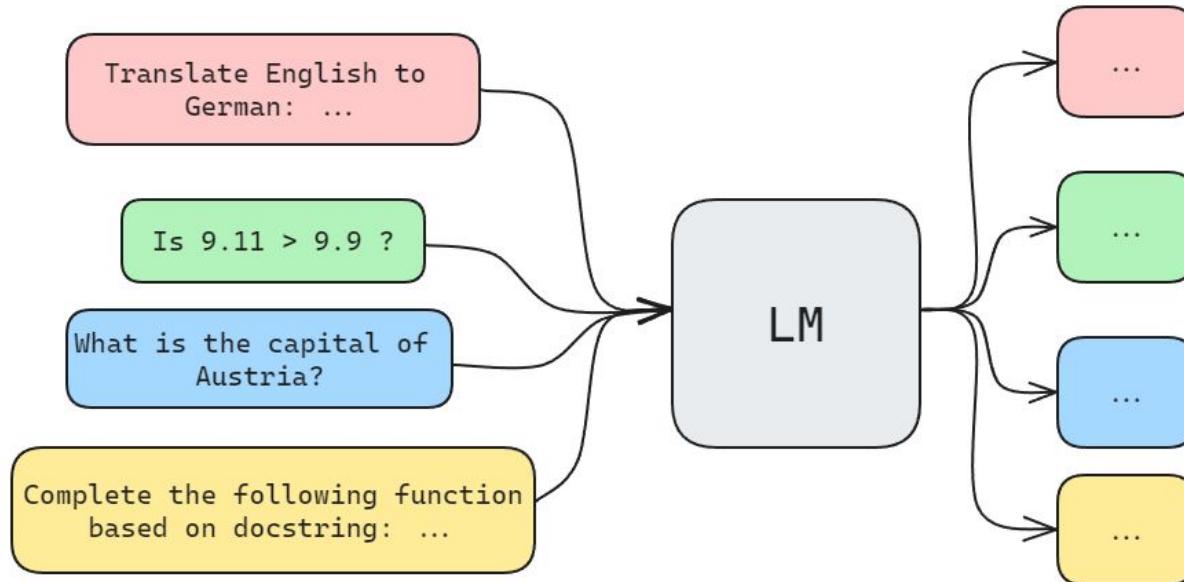


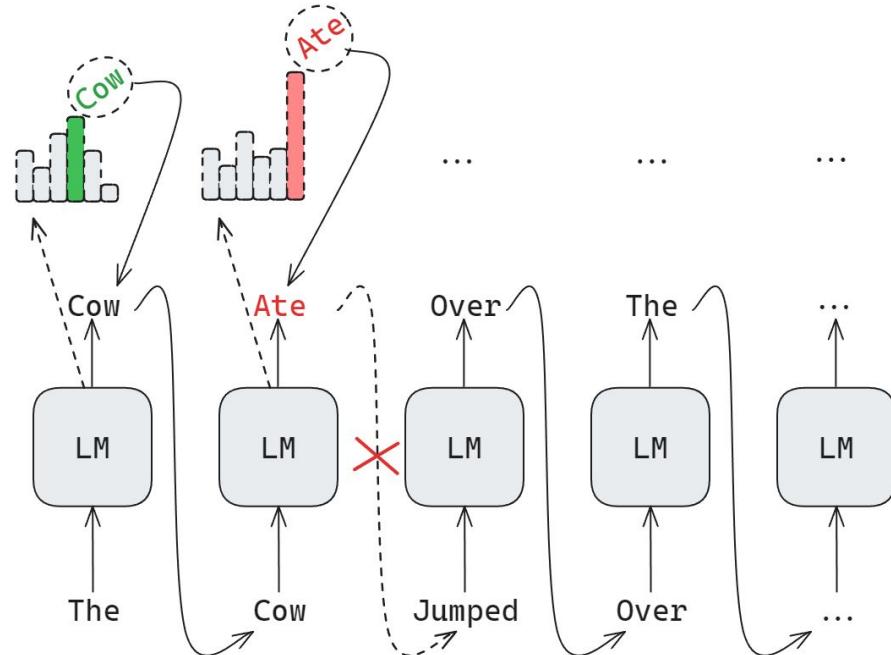
Figure 1: Performance on MMLU validation set during the training of OLMo-1.7-7B model.

# What Do We Want to Evaluate?



# LM Background

- LMs are probabilistic sequence models producing **Logit** distribution over **Vocabulary**
  - $\rightarrow \text{Softmax}(\text{Logits}) = P(x_n | x_{<n})$
  - $\rightarrow \text{Log}(\text{Softmax}(\text{Logits})) = \log P(x_n | x_{<n})$
- **Teacher Forcing:** compute  $P(x_j | x_{<j})$  for every  $j < n$  in parallel
  - Used for efficient autoregressive training



# Measurement Methods

- **How can we interact with an LM?**
  - How will the model be actually used? Chat settings, reranking, classification...
- Obtain an **observation** we can use to score or rank task performance on a given test example
  - Note: limiting to *prompted, training-free* use-cases

# Measurement Methods

- **Perplexity**
- **Conditional Loglikelihoods**
- **Text Generation**

# Perplexity

- A.k.a. “Rolling Loglikelihood”
- (Exponentiated) average *per-token* negative loglikelihood

$$PPL = \exp \left( \frac{-1}{\sum_{j=1}^{|D|} N_j} \sum_{j=1}^{|D|} \sum_{i=1}^{N_j} \log P(y_{j_i} | y_{j_1}, \dots, y_{j_{i-1}}) \right)$$

# Pros and Cons

-  Directly measures language modeling → good for base LMs ; scales smoothly
-  Can be performed using any data distribution–no annotation or labeling required
-  Not as useful for instruction-tuned LMs
-  Does not measure “real-world” freeform generation

# Conditional Loglikelihoods

The cow jumped over *the moon*  
\_\_\_\_\_  
Input                      Target

$\log P(\text{Target} \mid \text{Input})$

# Conditional Loglikelihoods

- To compute  $\log P(y|x)$  in 1 LM call:
  - Feed in  $(x + y)$  to LM, check how likely each token in  $y$  is. *Sum* per-token log probabilities

$$\log P(y|x) = \sum_{i=0}^{m-1} \log p(y_i|x, y_0, \dots, y_{i-1}) = \sum_{i=0}^{m-1} l(n+i, y_i),$$

# Conditional Loglikelihoods

- To compute  $\log P(\mathbf{y}|\mathbf{x})$  in 1 LM call:
  - Feed in  $(\mathbf{x} + \mathbf{y})$  to LM, check how likely each token in  $\mathbf{y}$  is. *Sum* per-token log probabilities

$$\log P(y|x) = \sum_{i=0}^{m-1} \log p(y_i|x, y_0, \dots, y_{i-1}) = \sum_{i=0}^{m-1} l(n+i, y_i),$$

# Conditional Loglikelihoods

- To compute  $\log P(\mathbf{y}|\mathbf{x})$  in 1 LM call:
  - Feed in  $(\mathbf{x} + \mathbf{y})$  to LM, check how likely each token in  $\mathbf{y}$  is. *Sum* per-token log probabilities

$$\log P(y|x) = \sum_{i=0}^{m-1} \log p(y_i|x, y_0, \dots, y_{i-1}) = \sum_{i=0}^{m-1} l(n+i, y_i),$$

# Loglikelihood-based Multiple-Choice

- The cow jumped over *the moon* 
- the earth* 
- the hay* 
- the galaxy* 

# Loglikelihood-based Multiple-Choice

- Compare loglikelihoods  $\log P(y_i|x)$  across a fixed set of answer strings  $y_i$
- Model's answer:  $\text{argmax}_i(\log P(y_i|x))$

A sample question from MMLU

Source: [Hendrycks et al., 2021](#)

Image: Stanford HAI (2024). Artificial Intelligence Index Report  
2024

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.

Answer:

A  $P("A" | c)$

B  $P("B" | c)$

C  $P("C" | c)$

D  $P("D" | c)$

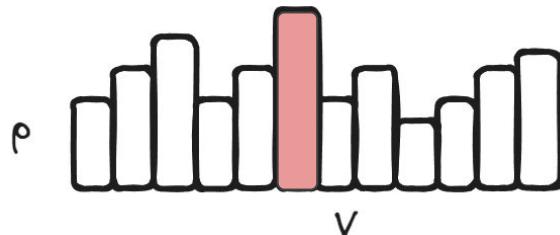
→ Model's answer: A

# Pros and Cons

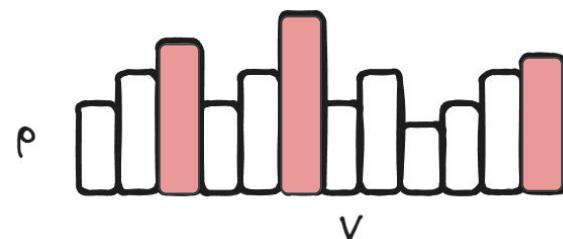
- LM always selects an answer
- Very efficient to evaluate—only need (num. choices) calls to LM
- Closer to training distribution → good for base LMs
- Artificially easy
- “Real-world” usage is not multiple-choice

# Text Generation

- Can probabilistically *sample* from an LM's output probability distribution



**Greedy:** pick most likely token



**Sampling:** pick one of the  $K$  highest, pick randomly, ...

- Sample new token and repeat to generate text
- How most models are used

# Scoring Freeform Generation

- Must extract and *parse*, compare to gold answer
  - Heuristically, using LLM-as-a-Judge, ...

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. ✓

Parse

`find("The answer is {x}")`

Model's answer: 9

# Pros and Cons

-  “Realistic” setting
-  Allows for techniques like Chain-of-Thought
-  Calculating accuracy requires heuristic *parsing, extraction* rules →  
**Scores skewed by parsing failures**
-  Much more expensive computationally
-  Many different decoding hyperparameters to select

# Reproducibility

- All 3 approaches contain hyperparameters that can be varied
- These can strongly affect performance, but often underspecified!

# Tokenization

- Dealing with tokenization properly can be nightmarish
- All 3 approaches implicitly rely on tokenization
- How to establish “fair” comparisons across tokenizers?

# Normalization

- Comparing across tokenizers confounds loglikelihoods
- The tokenizer is part of the system even if not the model!

- **Token-length normalization:** each  $a_i$ 's loglikelihood is divided by  $m_i$ , its length in tokens, to gain the per-token loglikelihood of each answer. This approach requires no additional LM calls, and is used alternately with raw loglikelihoods for most tasks by [Brown et al. \(2020\)](#).
- **Byte-length normalization:** each  $a_i$ 's loglikelihood is divided by its length in bytes, removing the dependence on the model's tokenizer but still normalizing by answer string length. `lm-eval` provides this metric where applicable as `acc_norm`.
- **Mutual Information:** each  $a_i$ 's loglikelihood is defined as  $\log P(a_i|x) - \log P(a_i|null)$ , where *null* is either the empty string, a BOS token, or a placeholder such as "Answer:". This can be thought of as a notion of the *pointwise mutual information* [Shannon, 1948](#); [Askell et al., 2021](#)),  $\log \left( \frac{P(a_i|x)}{P(a_i)} \right)$ , which measures the increase in the likelihood of outputting  $a_i$  when conditioned on the input  $x$ , compared to the likelihood of outputting  $a_i$  unconditionally. Intuitively, this measure of mutual information captures the extent to which introducing  $x$  makes  $a_i$  more likely. Although this approach is nonstandard, it is provided in `lm-eval` under the option `acc_mutual_info`, and used selectively by [Brown et al. \(2020\)](#) and [Askell et al. \(2021\)](#) for certain tasks.

- **Bits per Byte:** This metric measures the average number of bits required to encode each byte of the input text, providing a tokenization-agnostic measure of language modeling performance [Gao et al. \(2020\)](#). Formally:

$$BPB = \frac{-1}{\log(2)} \left( \frac{-1}{\sum_{j=1}^{|D|} B_j} \sum_{j=1}^{|D|} \sum_{i=1}^{N_j} \log P(y_{ji}|y_{j_1}, \dots, y_{j_{i-1}}) \right), \quad (3)$$

where  $\log$  is in base  $e$  and  $B_j$  is the length in bytes of document  $y_j$ . Alternately, bits per byte can be written as

$$BPB = \frac{\sum_{j=1}^{|D|} N_j}{\sum_{j=1}^{|D|} B_j} \log_2(PPL) = \frac{\sum_{j=1}^{|D|} N_j}{\sum_{j=1}^{|D|} B_j} \frac{\log(PPL)}{\log(2)}. \quad (4)$$

That is, taking the base-2 log of perplexity and renormalizing by the number of bytes rather than tokens.

- **Word-Level Perplexity:** By tokenizing the input text into words, such as via splitting on whitespace, we can calculate perplexity based on the average loglikelihood per *word* rather than per-token, making the metric comparable across models with different subword tokenizers.
- **Byte-level Perplexity:** Similarly, calculating perplexity averaged over the number of *bytes* instead allows for a different tokenization-independent perplexity calculation, as the number of bytes in each document's string remains constant regardless of the tokenizer used.

# Tokenization Boundaries

- Generation is not free of painful implementation details...
  - Switching the following prompts changes HumanEval scores significantly

1.

```
from typing import List

def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """Check if in given list of numbers, are any two elements closer than the given threshold.

    Parameters
    ----------
    numbers : List[float]
        The list of numbers to check.
    threshold : float
        The threshold distance between elements.

    Returns
    -------
    bool
        True if there are any two elements whose absolute difference is less than or equal to the threshold, False otherwise.
    """
    for i in range(len(numbers)):
        for j in range(i + 1, len(numbers)):
            if abs(numbers[i] - numbers[j]) <= threshold:
                return True
    return False

if __name__ == "__main__":
    print(has_close_elements([1.0, 2.0, 3.0], 0.5))  # False
    print(has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0], 0.3))  # True
```

2.

```
from typing import List

def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """Check if in given list of numbers, are any two elements
    closer than the given threshold.
    """
    for i in range(len(numbers)):
        for j in range(i + 1, len(numbers)):
            if abs(numbers[i] - numbers[j]) < threshold:
                return True
    return False

print(has_close_elements([1.0, 2.0, 3.0], 0.5))
print(has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0, 3.5], 0.3))
print(has_close_elements([1.0, 2.0, 3.0, 4.0, 5.0, 2.0, 3.5], 0.1))
```

# Sliding Window Perplexity

- **Early Token Curse:** initial tokens in document are more difficult to predict
- How to measure perplexity on docs longer than model's context length?

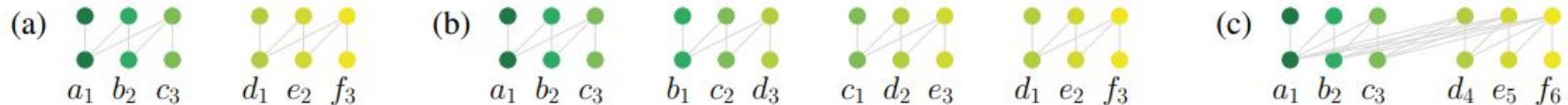
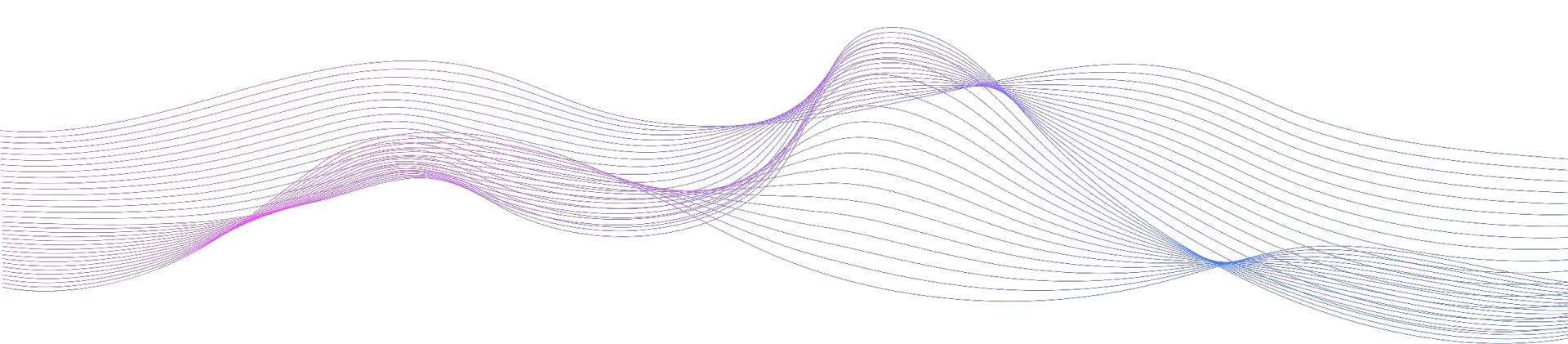


Figure 1: Language model modes for generating or evaluating 6 tokens ( $a, b, \dots, f$ ) when subsequence length  $L = 3$ . The numbers denote the position embeddings (P.E.). (a) Nonoverlapping (§2). (b) Sliding window, stride  $S = 1$ . Here, after the first inference pass we ignore all outputs other than the last (§2). (c) Caching (§5.2) where each subsequence attends to representations of the previous one. (In the next iteration, tokens  $d, e$  and  $f$  become the cache, with P.E. 1, 2 and 3, the three new tokens get P.E. 4, 5, and 6.)

# Underdocumentation and Tacit Knowledge

- Many papers underspecify their evaluation setups / measurement methods at a fundamental level!
- No one “correct” set of implementation details
- Knowing all these details requires *tacit knowledge* and *field experience*  
→ hence this tutorial



# LM-Specific Complications

# LM-Specific Complications

What are the reasons evaluation *of LMs* in particular is so challenging?

# Reproducibility

- Could you calculate these precise numbers yourself?
- How fair are these comparisons?

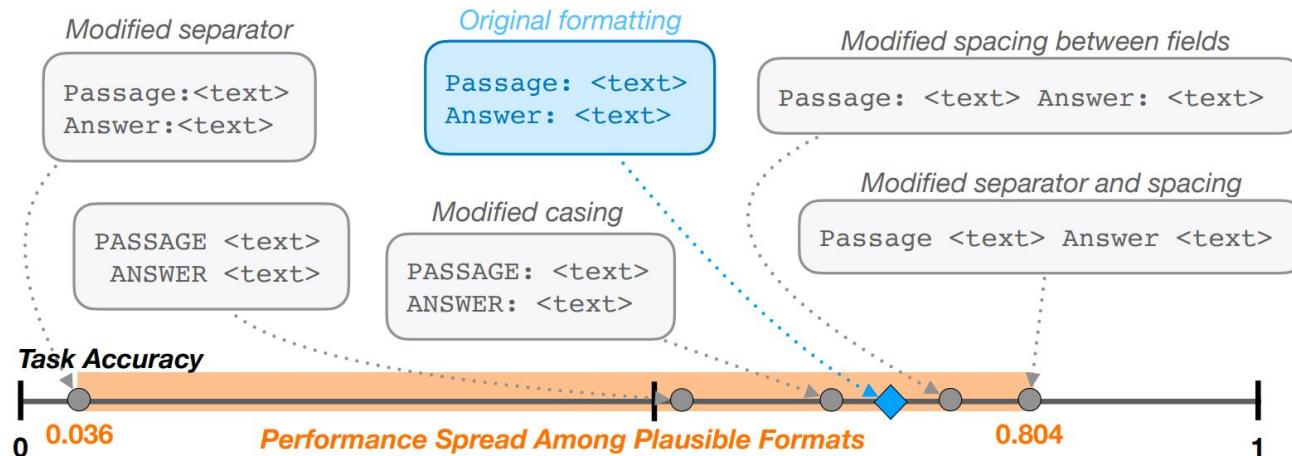
	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	<b>86.8%</b> 5 shot	<b>79.0%</b> 5-shot	<b>75.2%</b> 5-shot	<b>86.4%</b> 5-shot	<b>70.0%</b> 5-shot	<b>83.7%</b> 5-shot	<b>71.8%</b> 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	<b>50.4%</b> 0-shot CoT	<b>40.4%</b> 0-shot CoT	<b>33.3%</b> 0-shot CoT	<b>35.7%</b> 0-shot CoT	<b>28.1%</b> 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	<b>95.0%</b> 0-shot CoT	<b>92.3%</b> 0-shot CoT	<b>88.9%</b> 0-shot CoT	<b>92.0%</b> 5-shot CoT	<b>57.1%</b> 5-shot	<b>94.4%</b> Maj1@32	<b>86.5%</b> Maj1@32
Math problem-solving <i>MATH</i>	<b>60.1%</b> 0-shot CoT	<b>43.1%</b> 0-shot CoT	<b>38.9%</b> 0-shot CoT	<b>52.9%</b> 4-shot	<b>34.1%</b> 4-shot	<b>53.2%</b> 4-shot	<b>32.6%</b> 4-shot
Multilingual math <i>MGSM</i>	<b>90.7%</b> 0-shot	<b>83.5%</b> 0-shot	<b>75.1%</b> 0-shot	<b>74.5%</b> 8-shot	—	<b>79.0%</b> 8-shot	<b>63.5%</b> 8-shot
Code <i>HumanEval</i>	<b>84.9%</b> 0-shot	<b>73.0%</b> 0-shot	<b>75.9%</b> 0-shot	<b>67.0%</b> 0-shot	<b>48.1%</b> 0-shot	<b>74.4%</b> 0-shot	<b>67.7%</b> 0-shot
Reasoning over text <i>DROP, Fi score</i>	<b>83.1</b> 3-shot	<b>78.9</b> 3-shot	<b>78.4</b> 3-shot	<b>80.9</b> 3-shot	<b>64.1</b> 3-shot	<b>82.4</b> Variable shots	<b>74.1</b> Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	<b>86.8%</b> 3-shot CoT	<b>82.9%</b> 3-shot CoT	<b>73.7%</b> 3-shot CoT	<b>83.1%</b> 3-shot CoT	<b>66.6%</b> 3-shot CoT	<b>83.6%</b> 3-shot CoT	<b>75.0%</b> 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	<b>96.4%</b> 25-shot	<b>93.2%</b> 25-shot	<b>89.2%</b> 25-shot	<b>96.3%</b> 25-shot	<b>85.2%</b> 25-shot	—	—
Common Knowledge <i>HellaSwag</i>	<b>95.4%</b> 10-shot	<b>89.0%</b> 10-shot	<b>85.9%</b> 10-shot	<b>95.3%</b> 10-shot	<b>85.5%</b> 10-shot	<b>87.8%</b> 10-shot	<b>84.7%</b> 10-shot

# Reproducibility

- LMs in particular are often non-robust *in counterintuitive ways*

# Prompt Sensitivity

- Choice of the prompt to use for evaluation can be make-or-break



# Prompt Sensitivity

	ARC Challenge		MMLU	
	Cloze	MMLU-style	Hybrid	MMLU-style
GPT-NeoX-20B	<b>38.0 ± 2.78 %</b>	26.6 ± 2.53%	<b>27.6 ± 0.74%</b>	24.5 ± 0.71%
Llama-2-7B	<b>43.5 ± 2.84%</b>	42.8 ± 2.83%	39.8 ± 0.79%	<b>41.3 ± 0.80%</b>
Falcon-7B	<b>40.2 ± 2.81%</b>	25.9 ± 2.51%	<b>29.1 ± 0.75%</b>	25.4 ± 0.72%
Mistral-7B	50.1 ± 2.86%	<b>72.4 ± 2.56%</b>	48.3 ± 0.80%	<b>58.6 ± 0.77%</b>
Mixtral-8x7B	56.7 ± 2.84%	<b>81.3 ± 2.23%</b>	59.7 ± 0.77%	<b>67.1 ± 0.72%</b>

## Multiple-Choice (“MMLU-style”) Formulation

Question: Earth’s core is primarily composed of which of the following materials?

- (A) basalt (B) iron (C) magma (D) quartz

Answer: (B)

## Cloze Formulation

Question: Earth’s core is primarily composed of which of the following materials?

Answer: <answer>, where each answer choice is separately substituted in for <answer>.

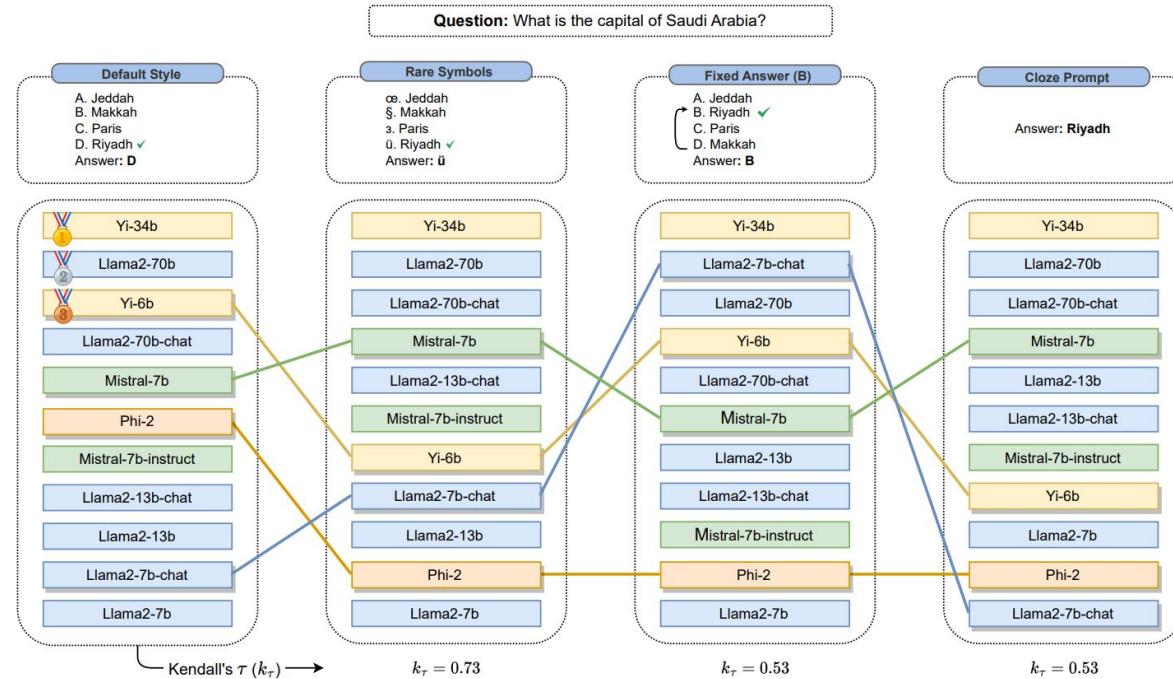
# Prompt Sensitivity

- “Preferred” prompt and output format differs across models
- → Rankings and experimental conclusions are changed by prompt choice!

Table 1: Comparison of 0-shot model performance (acc) for several pretrained LMs (Black et al., 2022; Touvron et al., 2023b; Penedo et al., 2023; Jiang et al., 2023, 2024) on ARC (Challenge subset) and MMLU across two commonly used prompt styles.

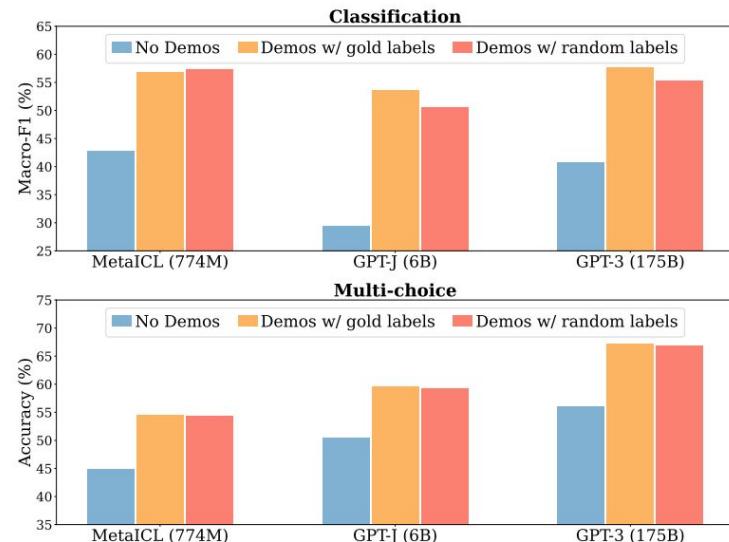
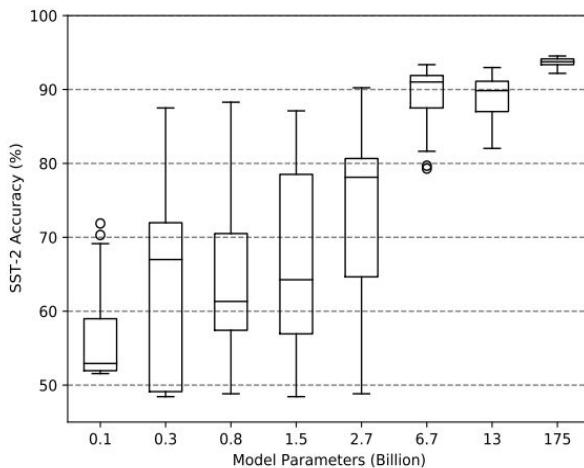
	ARC Challenge		MMLU	
	Cloze	MMLU-style	Hybrid	MMLU-style
GPT-NeoX-20B	<b>38.0 ± 2.78 %</b>	26.6 ± 2.53%	<b>27.6 ± 0.74%</b>	24.5 ± 0.71%
Llama-2-7B	<b>43.5 ± 2.84%</b>	42.8 ± 2.83%	39.8 ± 0.79%	<b>41.3 ± 0.80%</b>
Falcon-7B	<b>40.2 ± 2.81%</b>	25.9 ± 2.51%	<b>29.1 ± 0.75%</b>	25.4 ± 0.72%
Mistral-7B	50.1 ± 2.86%	<b>72.4 ± 2.56%</b>	48.3 ± 0.80%	<b>58.6 ± 0.77%</b>
Mixtral-8x7B	56.7 ± 2.84%	<b>81.3 ± 2.23%</b>	59.7 ± 0.77%	<b>67.1 ± 0.72%</b>

# Prompt Sensitivity



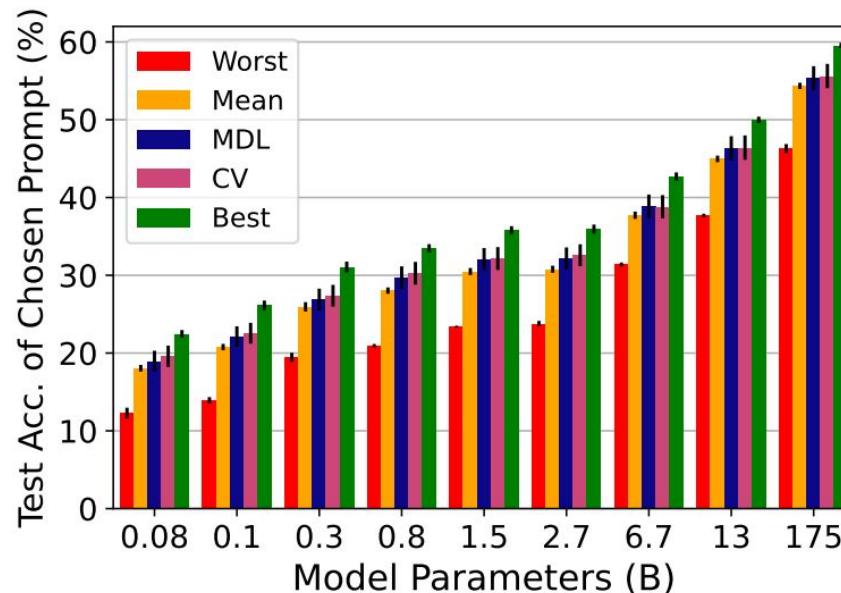
# Few-shot Example Sensitivity

- Choices and orderings of few-shot examples can significantly impact performance



# To Prompt Engineer or Not To Prompt Engineer

- Engineering and taking the best prompt can overestimate performance in *real* few-shot settings



# Details Matter

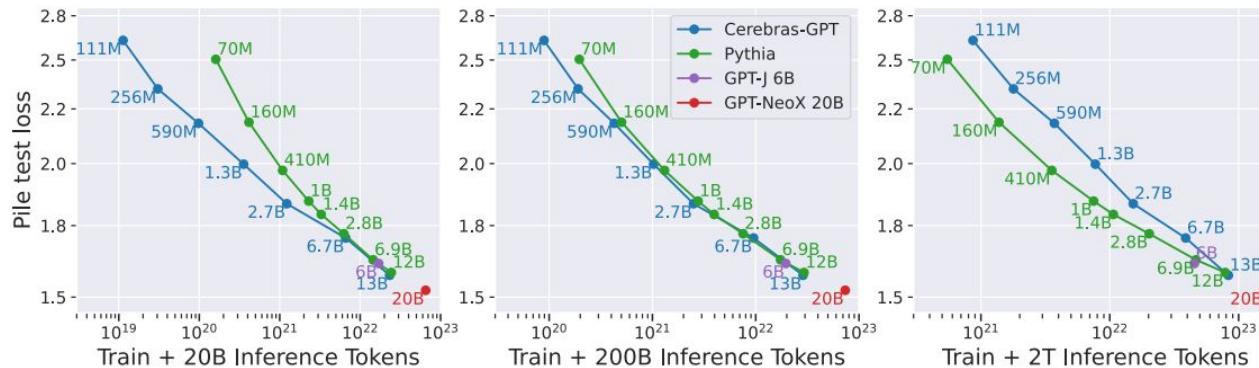


# Fair Comparisons

- What constitutes a 1-to-1 or “apples-to-apples” comparison of two models?
- Should we...
  - Pick the “best” prompt per model? “Worst” prompt per model? Hold the prompt constant?
  - ...

# Fair Comparisons

- “Fairness” will often be context-dependent!
  - Research question matters: minimizing VRAM? Training FLOP? Data efficiency?



# Evaluating Models Vs. Systems

When using GPT-4o, ChatGPT Free users will now have access to features such as:

- Experience GPT-4 level intelligence
- Get responses from both the model and the web
- Analyze data and create charts
- Chat about photos you take
- Upload files for assistance summarizing, writing or analyzing
- Discover and use GPTs and the GPT Store
- Build a more helpful experience with Memory

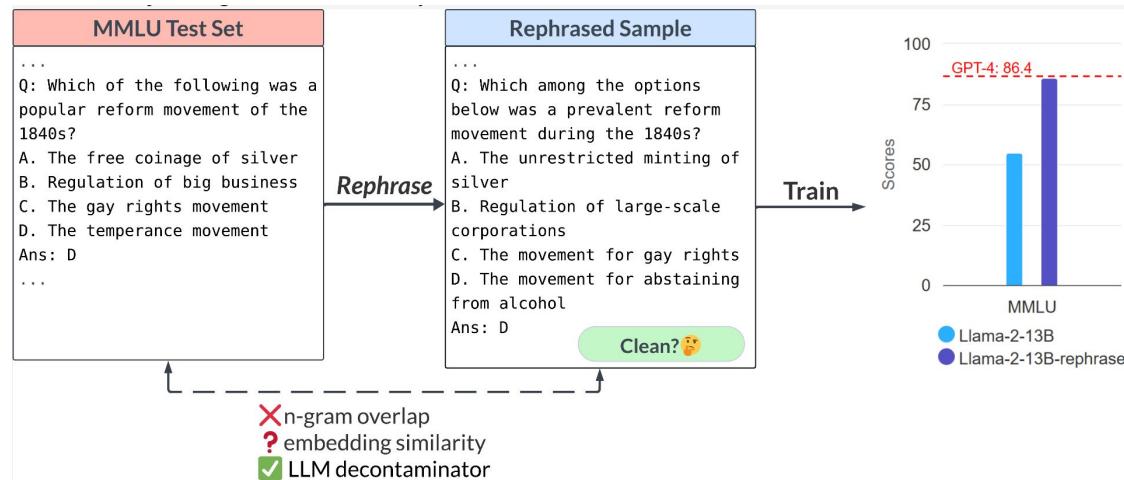
# Dataset Contamination

- Benchmarks are built assuming novelty, generalization
  - Often using internet data as a source
- But LMs are trained on massive internet-scale datasets
  - Easy for test set contents to leak into pretraining data
  - Assumptions during construction may not hold (“validity”)

We create a massive multitask test consisting of multiple-choice questions from various branches of knowledge. The test spans subjects in the humanities, social sciences, hard sciences, and other areas that are important for some people to learn. There are 57 tasks in total, which is also the number of Atari games (Bellemare et al., 2013), all of which are listed in Appendix B. The questions in the dataset were manually collected by graduate and undergraduate students from freely available sources online. These include practice questions for tests such as the Graduate Record Examination

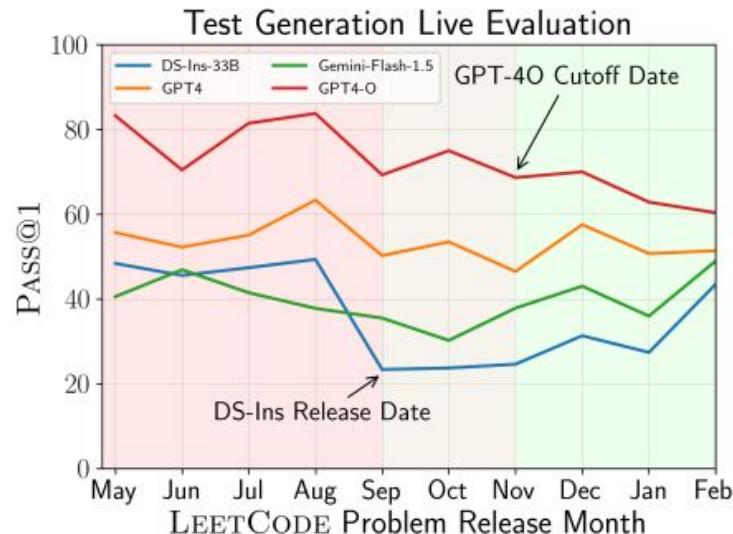
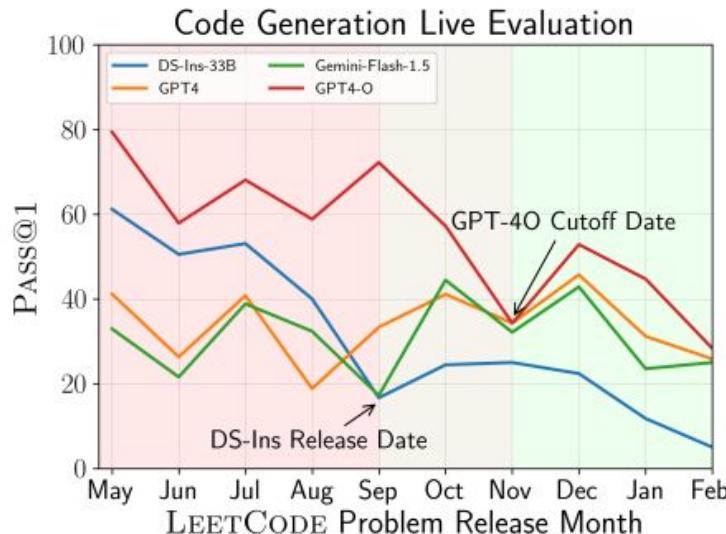
# Dataset Contamination

- Contamination may not always be verbatim
- Very difficult to detect and prove!



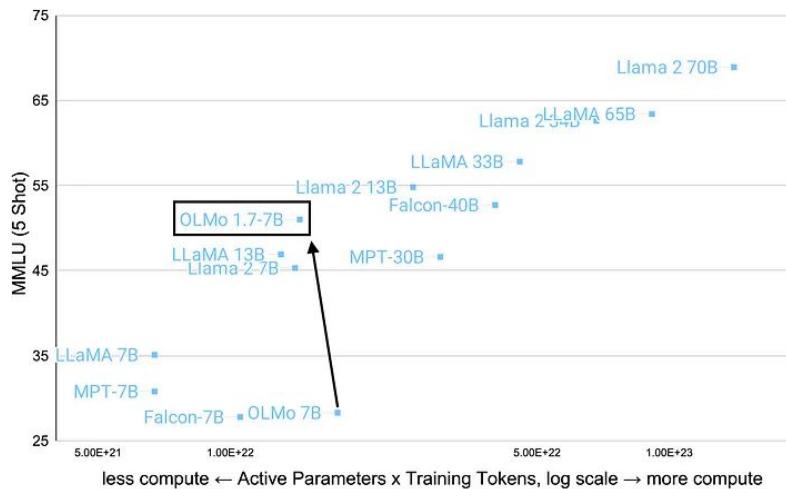
# Dataset Contamination

- *What even “counts” as contamination?*
- *How can we design contamination-proof evals?*



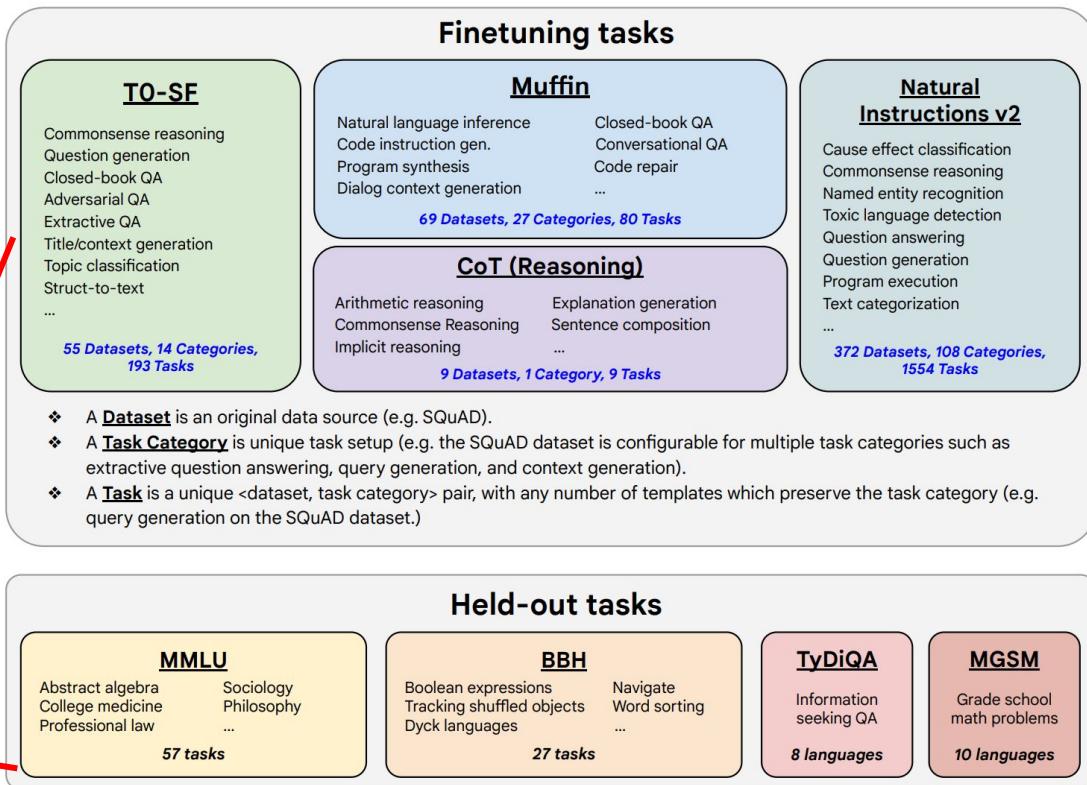
# Task Contamination

- OLMo-1.7-7B: one way to get a good MMLU score is to pretrain with instruction data included
- Is training on instruction-following data “cheating”? No, but violates assumptions



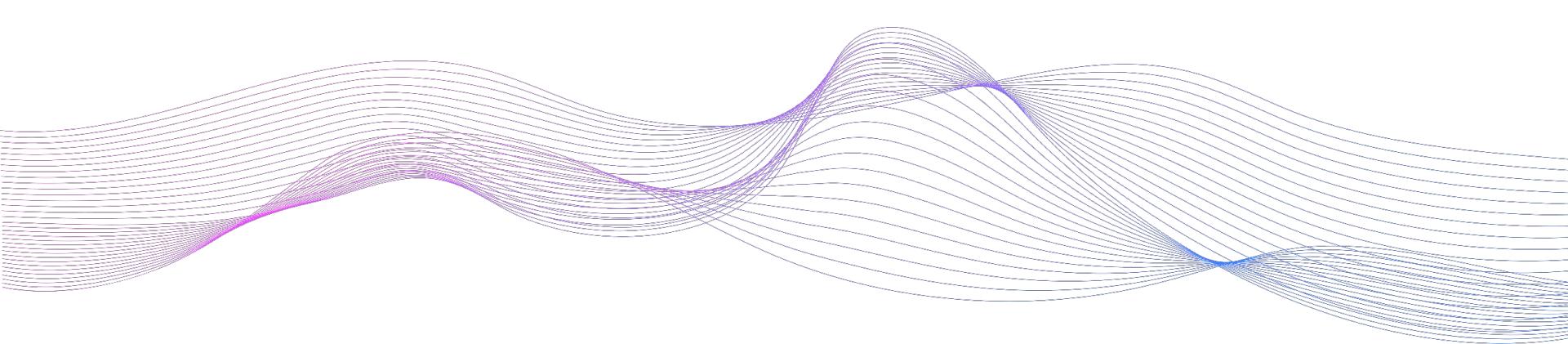
# Are LMs “zero-shot”?

Subset	Provenance	New in Dolma?
Dolma CC	Common Crawl via Dolma 1.6	Updated
Refined Web	Refined Web dataset	Yes
StarCoder	StarCoder dataset	Yes
C4	C4 dataset via Dolma 1.6	Updated
Dolma Reddit	Pushshift API via Dolma 1.6	Updated
Semantic Scholar	S2ORC/Pes2o via Dolma 1.6	No
ArXiv	RedPajama v1	Yes
Stack Exchange	RedPajama v1	Yes
Flan	Flan Collection, reproduced following the original code, as performed by Dettmers et al. (2023)	Yes



# LMs Introduce New Benchmarking Challenges

- Doing reproducible evaluation on LMs is difficult—details matter
- The “right” evaluation choice is not universal
  - Some choices (e.g. drawing comparisons) must be contextual
- Novel validity challenges are introduced by scale
  - LMs at times work well due to *everything being within-distribution*. How can we truly test their generalization?
  - Need to move beyond simple knowledge tests



# General Benchmarking Complications

# General Benchmarking Complications

Why is evaluation difficult *in general*?

What are the challenges in constructing useful datasets for LM evaluation?

# Where do benchmarks come from?

## SQuAD: 100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang  
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu  
Computer Science Department  
Stanford University

Reading Comprehension seen as a useful task

## MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks  
UC Berkeley

Collin Burns  
Columbia University

Steven Basart  
UChicago

Andy Zou  
UC Berkeley

Mantas Mazeika  
UIUC

Dawn Song  
UC Berkeley

Jacob Steinhardt  
UC Berkeley

From observed model multitask capabilities

## Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models

Maribeth Rauh\* John Mellor Jonathan Uesato Po-Sen Huang Johannes Welbl

Laura Weidinger Sumanth Dathathri Amelia Glaese Geoffrey Irving

Iason Gabriel William Isaac Lisa Anne Hendricks

DeepMind

Models are observed to produce toxic content

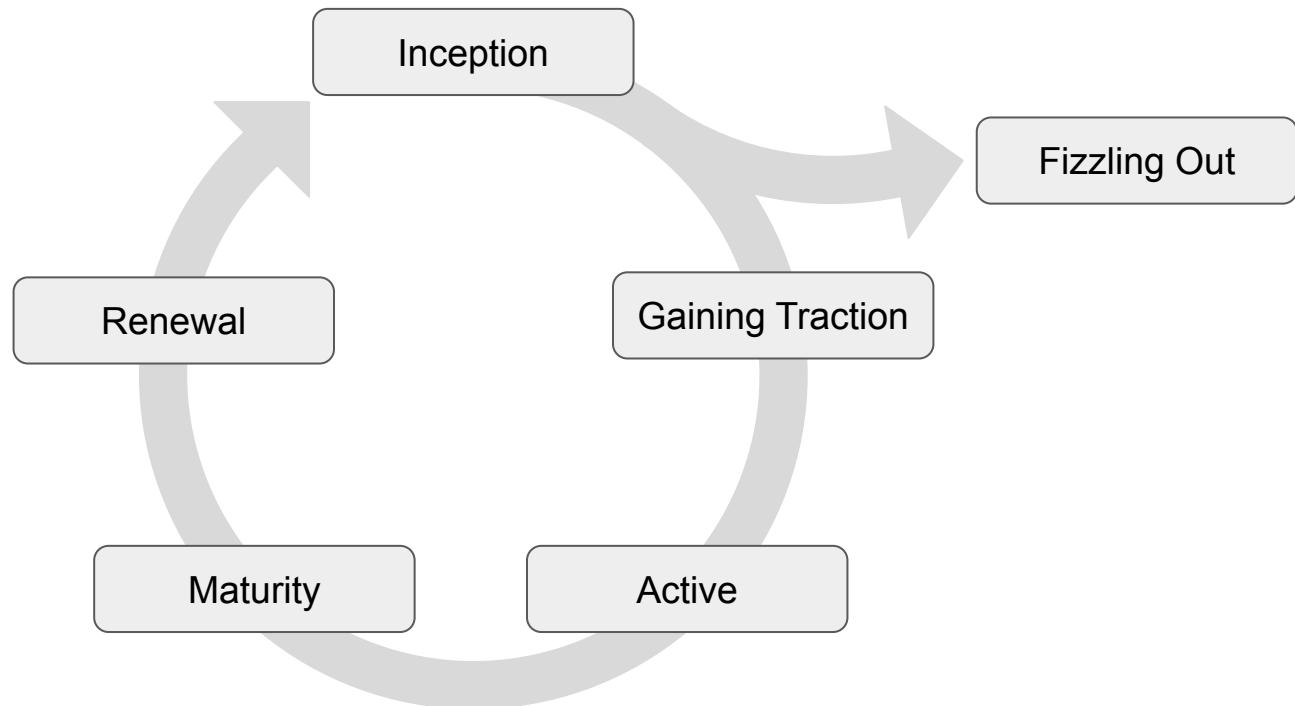
## FELM: Benchmarking Factuality Evaluation of Large Language Models

Shiqi Chen<sup>1\*</sup> Yiran Zhao<sup>3</sup> Jinghan Zhang<sup>2</sup> I-Chun Chern<sup>4</sup>  
Siyang Gao<sup>1</sup> Pengfei Liu<sup>5</sup> Junxian He<sup>2</sup>

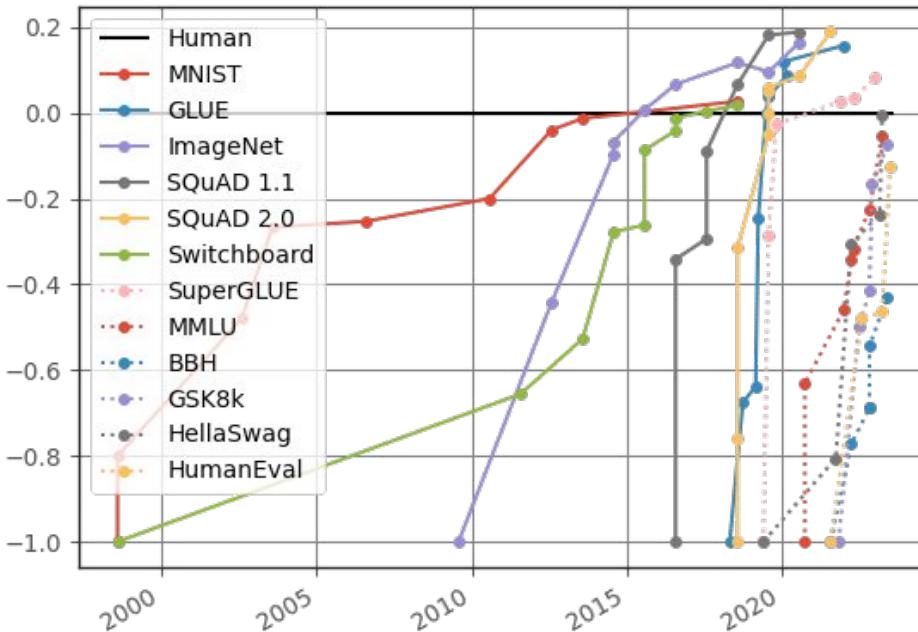
<sup>1</sup>City University of Hong Kong <sup>2</sup>The Hong Kong University of Science and Technology  
<sup>3</sup>National University of Singapore <sup>4</sup>Carnegie Mellon University <sup>5</sup>Shanghai Jiao Tong University  
schen438-c@my.cityu.edu.hk, junxianh@cse.ust.hk

Models are observed to hallucinate

# Life of a Benchmark

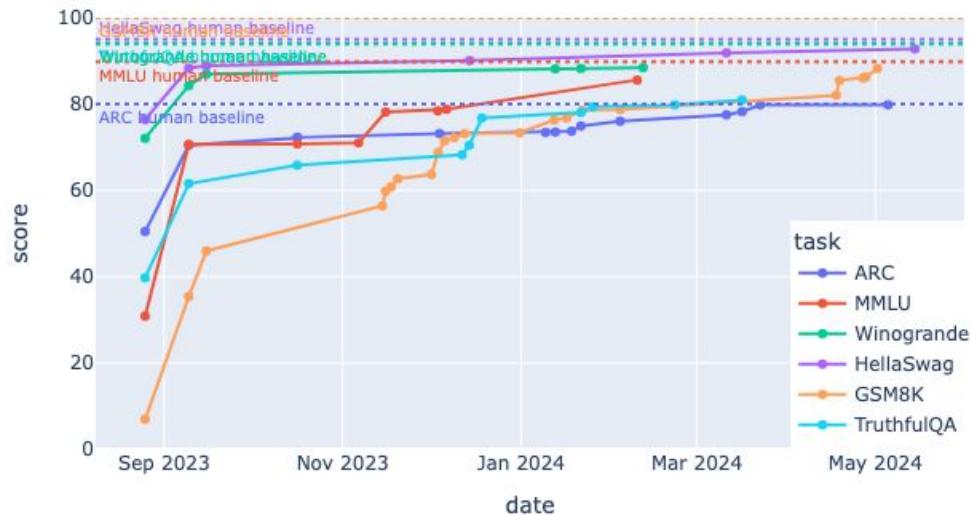


# Benchmarks are Saturating Fast



# OpenLLM Leaderboard Through Time

Top Scores and Human Baseline Over Time (from last update)



# Benchmarks Influence Progress

## Task Selection Bias

- We optimize for what we can measure
- Benchmarks determine what can be measurable

Tasks	Top-5 Performing Models (In Order)
H	Universal, Switch, Adaptive Softmax, Weighted, Vanilla
G	MoE, Switch, Vanilla, Funnel, Universal
A, B	Adaptive Softmax, Vanilla, MoE, Switch, Weighted
A, C	MoE, Switch, Adaptive Softmax, Vanilla, Universal
D, H	Switch, Universal, Adaptive Softmax, MoE, Weighted
B, E, H	Adaptive Softmax, Switch, MoE, Vanilla, Weighted
F, G, H	Switch, MoE, Adaptive Softmax, Universal, Vanilla
A, F, G	MoE, Switch, Vanilla, Adaptive Softmax, Vanilla
C, F, G, H	Switch, MoE, Adaptive Softmax, Vanilla, Universal
A, C, D, G	MoE, Switch, Adaptive Softmax, Vanilla, Universal
All	Switch, MoE, Adaptive Softmax, Vanilla, Universal

A=BoolQ, B=CB, C=CoPA, D=MultiRC, E=ReCoRD, F=RTE, G=WiC, H=WSC

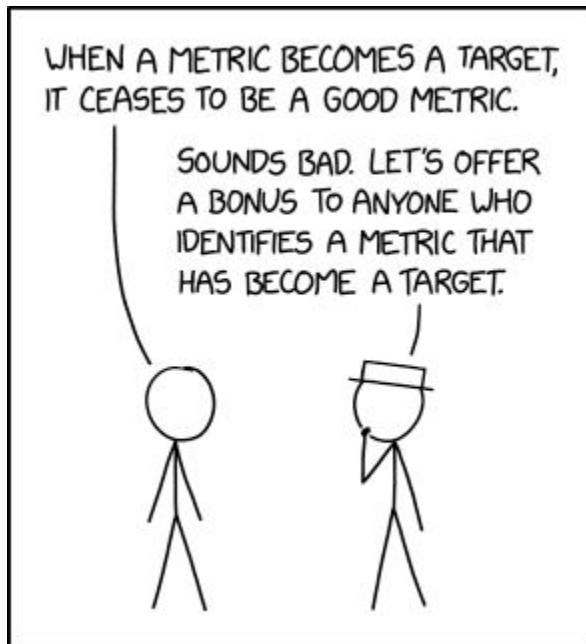
## Community Bias

- Specific benchmarks gain outsized popularity and influence

*“the method was not evaluated on X or Y dataset” or “the method’s performance is not SOTA on dataset Z”.*

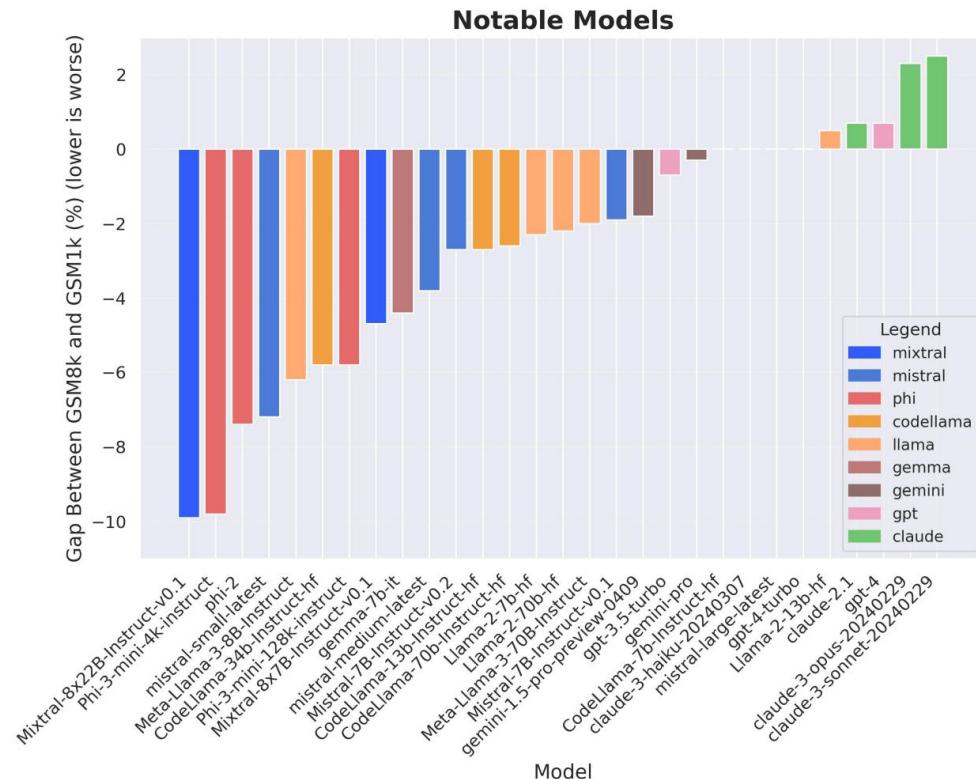
# Overfitting

Goodhart's Law

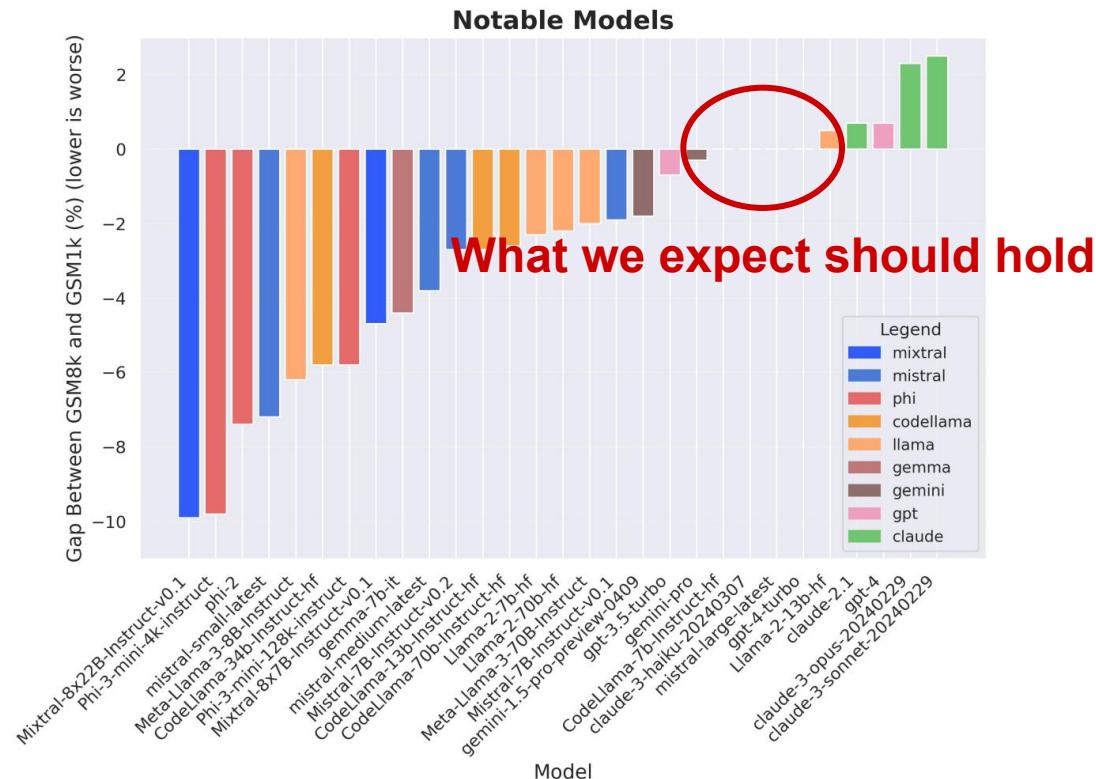


<https://xkcd.com/2899/>

# Benchmarks Get Overfitted

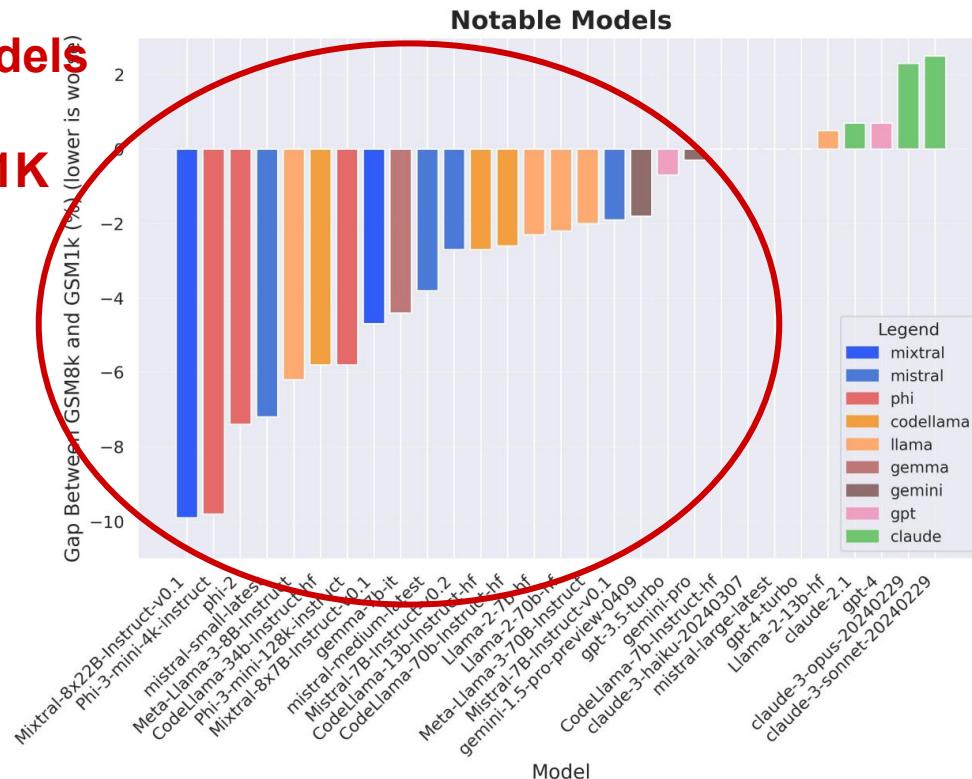


# Benchmarks Get Overfitted



# Benchmarks Get Overfitted

Majority of models tested perform worst on GSM1K



# Evaluation Validity

- Benchmarks are frequently *proxies* for “real” performance
  - Certain benchmarks may not be a good proxy! (Saphra et al., 2023)
- “Measurement Validity”
  - Are our benchmarks measuring “true” improvements / capabilities?
  - Are improvements on benchmarks “real”?
- “Measurement Reliability”
  - Are our benchmarks reproducible?
  - Able to produce consistent results?
  - A measurement can be *reliable* but not *valid*

# Spurious Heuristics

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	<b>The doctor was paid by the actor.</b> → The doctor paid the actor. <b>WRONG</b>
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near <b>the actor danced</b> . → The actor danced. <b>WRONG</b>
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If <b>the artist slept</b> , the actor ran. → The artist slept. <b>WRONG</b>

Benchmarks can be faulty or be solved by models by relying on something other than what was intended to be measured.

McCoy et al proposed a NLI task that sought to measure this.

# Spurious Heuristics

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	<b>The doctor</b> was <b>paid</b> by <b>the actor</b> . → The doctor paid the actor. WRONG
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near <b>the actor</b> danced. → The actor danced. WRONG
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If <b>the artist</b> slept, the actor ran. → The artist slept. WRONG

# Spurious Heuristics

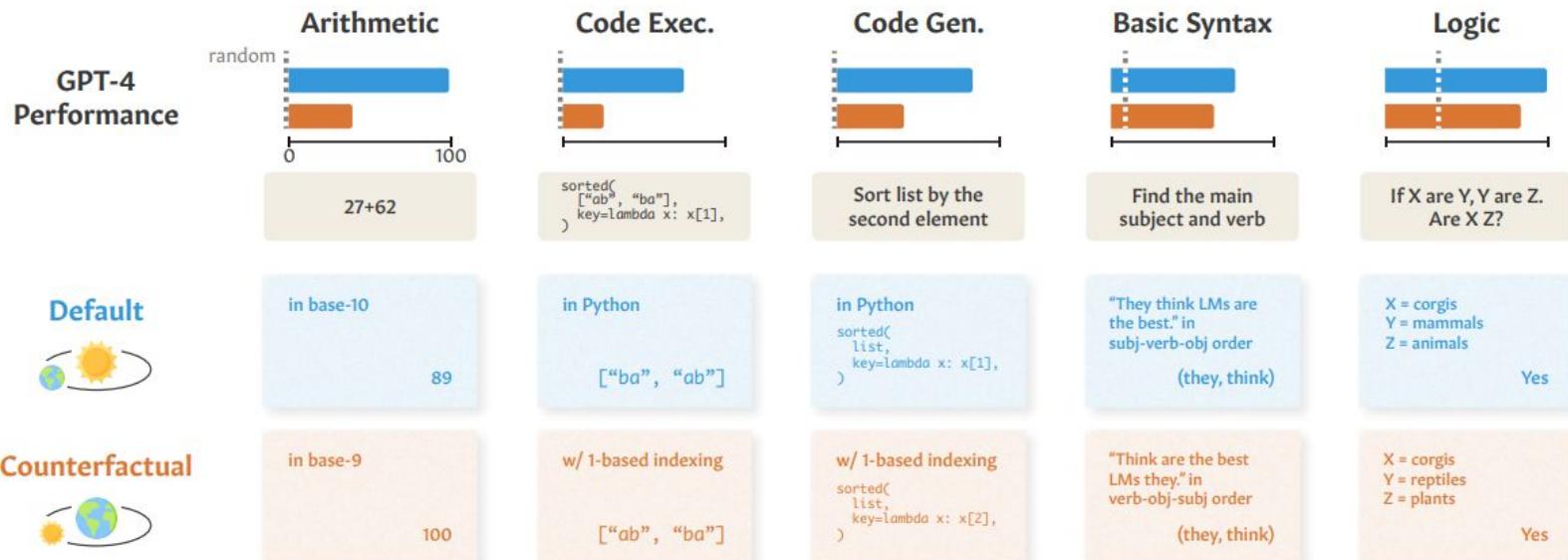
Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	<b>The doctor was paid by the actor.</b> → The doctor paid the actor. <b>WRONG</b>
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near <b>the actor danced.</b> → <b>The actor danced.</b> <b>WRONG</b>
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If <b>the artist slept</b> , the actor ran. → <b>The artist slept.</b> <b>WRONG</b>

# Spurious Heuristics

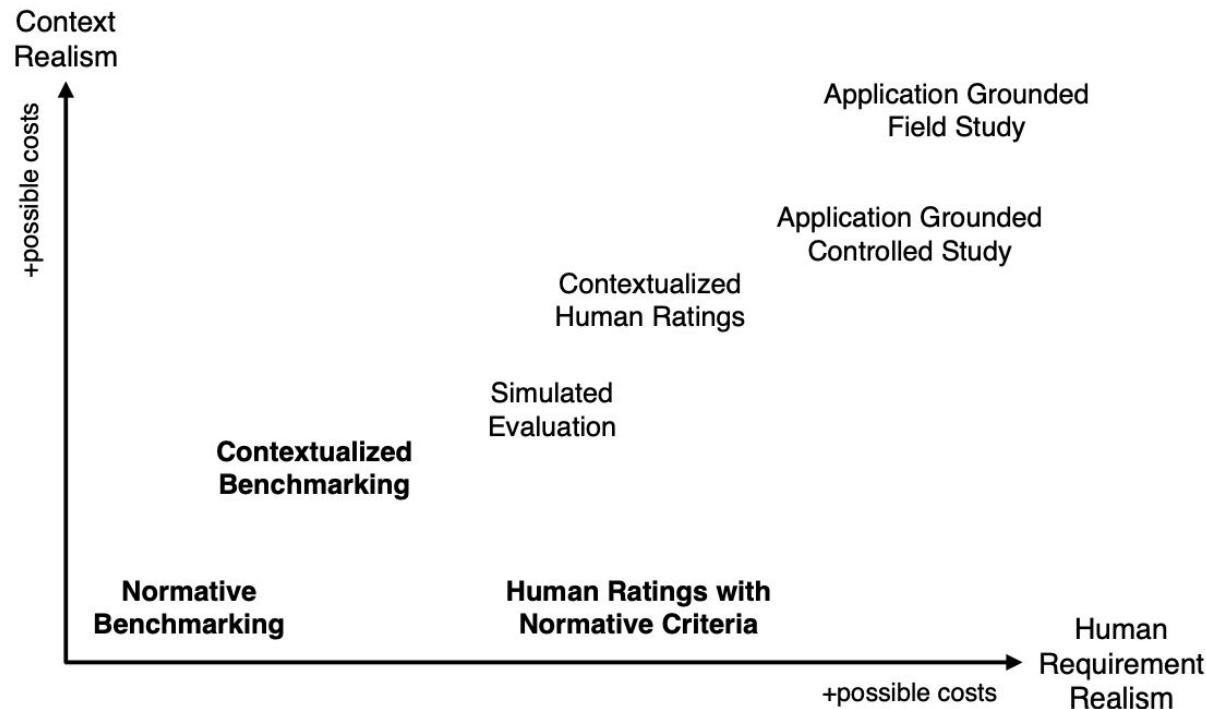
Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	<b>The doctor was paid by the actor.</b> → The doctor paid the actor. <b>WRONG</b>
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near <b>the actor danced</b> . → The actor danced. <b>WRONG</b>
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If <b>the artist slept</b> , the actor ran. → <b>The artist slept</b> . <b>WRONG</b>

# Challenging Intuitions on Generalization

- Testing models on purely “natural” tasks overestimates performance on truly-unseen data
- Performance on task A may not intuitively translate to task B!



# Ecological Validity of Benchmarking



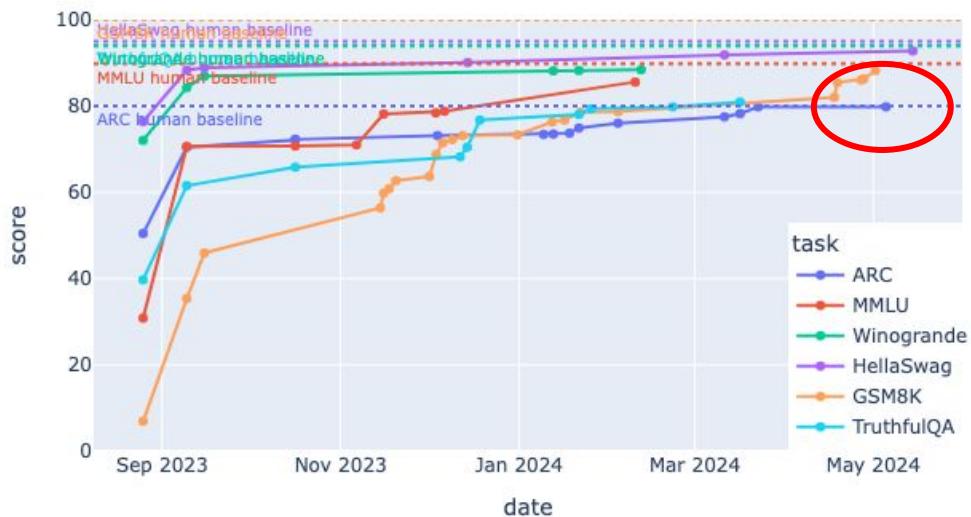
# Alternatives: Extrinsic Evaluations

Model output quality evaluated based on utility towards a specific downstream application.

- Evaluate MT models based on how many manual corrections had to be made ([Snover et al., 2006](#)).
- Evaluate models translation or summaries by answering reading comprehension questions based on those artifacts ([Jones et al., 2005](#); [Callison-Burch, 2009](#); [Scarton and Specia, 2016](#); [Wang et al., 2020](#))

# When do scores become less meaningful?

Top Scores and Human Baseline Over Time (from last update)



# Observed Errors in MMLU Samples

This becomes an even bigger deal as benchmarks saturate!

## Erroneous Instances in MMLU

What is the current best option for preventing future outbreaks of Ebola?

- A. Rebuild scientific, medical and nursing infrastructure and train staff
- B. Early and accurate diagnosis with molecular kits
- C. Develop effective vaccines
- D. Arrange rapid intervention into West Africa with EU and USA army teams

Correct answer, from a Human Virology 5e quiz

Incorrect answer, from MMLU Virology

The number of energy levels for the  $^{55}\text{Mn}$  nuclide are:

- A. 3
- B. 5
- C. 8
- D. 4

Incorrect answer, from MMLU College Chemistry

The woman who conducted a longitudinal study on herself and found increased retrieval difficulty as she got older was named

- A. Clark
- B. Smith
- C. Whitebear
- D. Ebbinghaus

Ambiguous question, from MMLU Human Aging

# Error by Category

## OK

What is the capital city of Indonesia?

- A. Berlin
- B. Paris
- C. Rome
- D. Jakarta

Ground Truth Answer: D  
Correct Answer: D

## Bad Question Clarity

Where is the headquarter of the company mentioned in question 21?

- A. Edinburgh
- B. Madrid
- C. London
- D. Paris

Ground Truth Answer: D  
Correct Answer: ?

## Bad Options Clarity

What is the largest ocean on Earth?

- A. Atlantic
- B. Ocean
- C. Pacific Ocean
- D. Arctic Ocean

Ground Truth Answer: C  
Correct Answer: C

## Multiple Correct Answers

Which of the following countries are located in both Europe and Asia?

- A. Russia
- B. Turkey
- C. Kazakhstan
- D. Georgia

Ground Truth Answer: B  
Correct Answer: A, B

## No Correct Answer

Who won the Champions League in the 2020-2021 session?

- A. Manchester
- B. Real Madrid
- C. Liverpool
- D. Barcelona

Ground Truth Answer: A  
Correct Answer: Chelsea

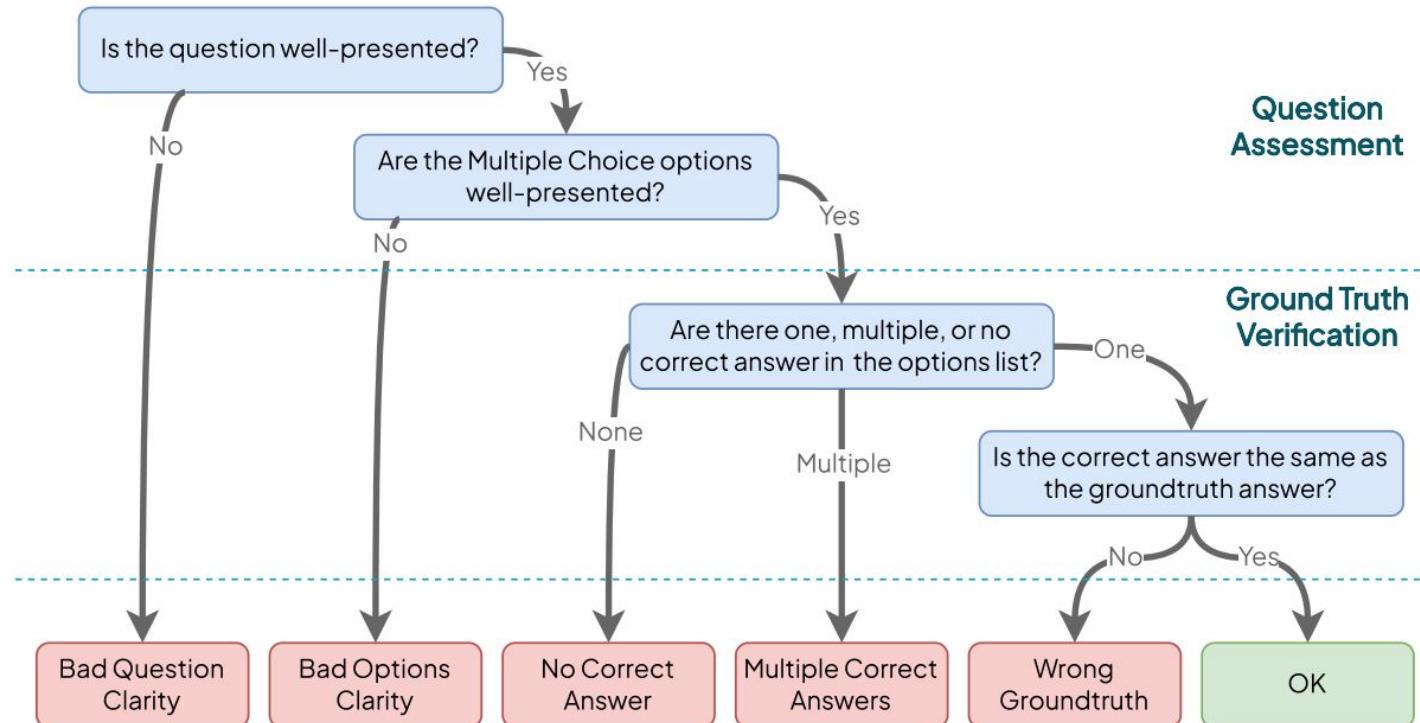
## Wrong Groundtruth

A virus such as influenza which emerges suddenly and spreads globally is called:

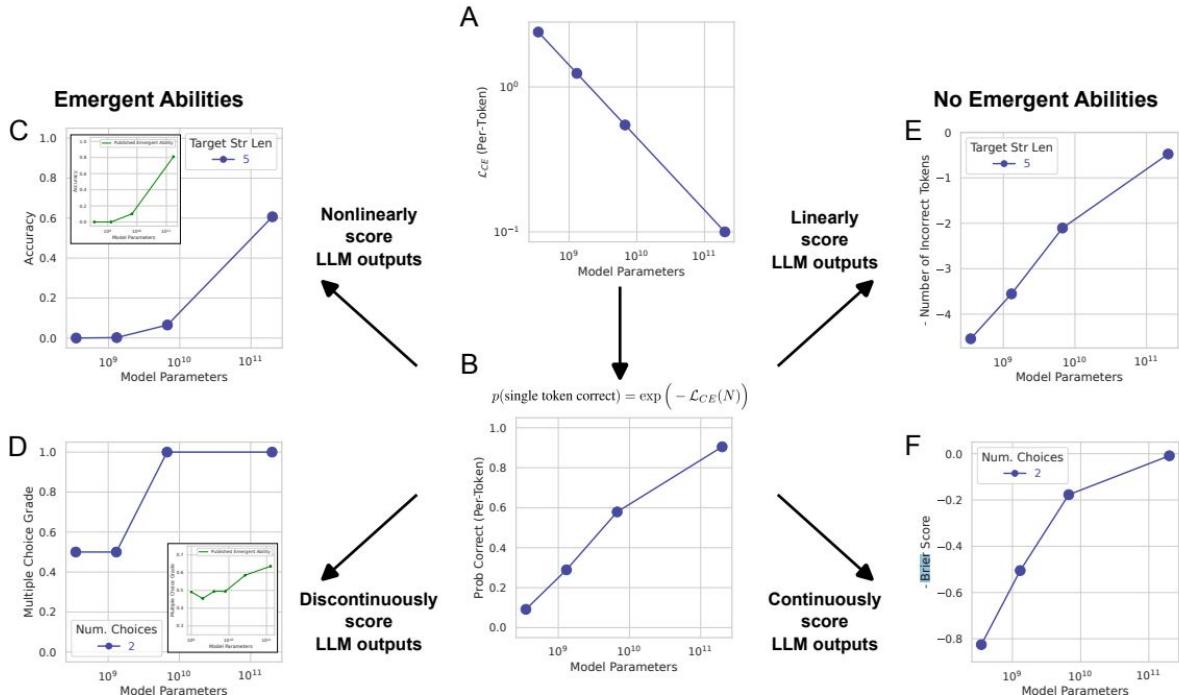
- A. Epidemic
- B. Endemic
- C. Pandemic
- D. Zoonotic

Ground Truth Answer: B  
Correct Answer: C

# Error Analysis Heuristic

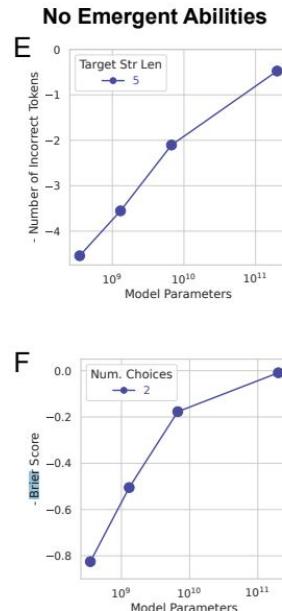
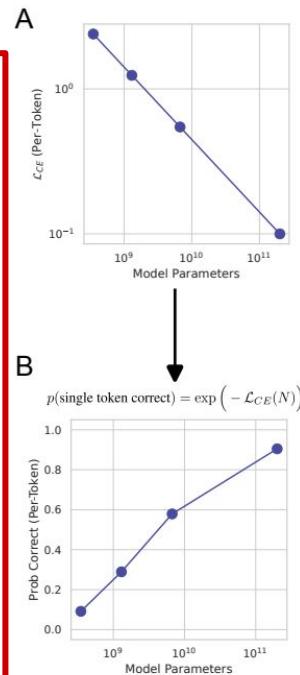
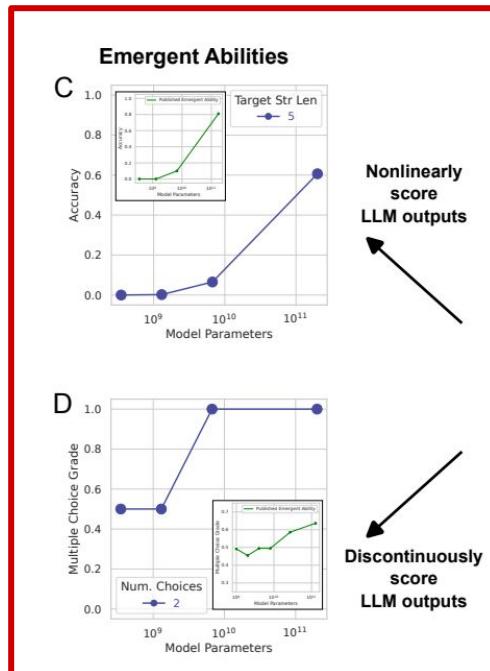


# Are we Using the Right Metrics?



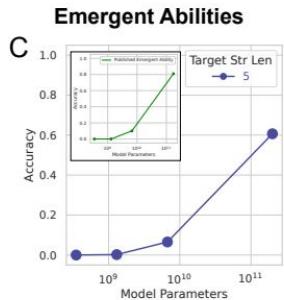
# Are we Using the Right Metrics?

Using Non-linear or discontinuous scores,  
can observe “emergent capabilities”

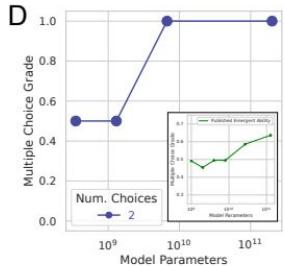


# Are we Using the Right Metrics?

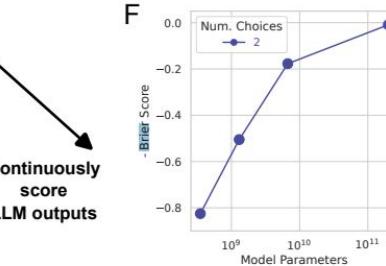
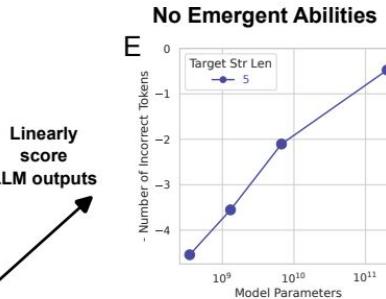
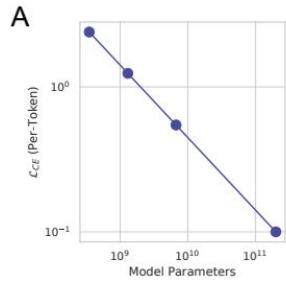
Using linear or continuous scores, can observe performance scales predictably



Nonlinearly score LLM outputs



Discontinuously score LLM outputs



Linearly score LLM outputs

Continuously score LLM outputs

# Automatic Metrics May Not Lead to Best Results

BLEU may have impeded progress in MT

## To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation

Tom Kocmi Christian Federmann Roman Grundkiewicz Marcin Junczys-Dowmunt Hitokazu Matsushita Arul Menezes  
Microsoft  
1 Microsoft Way  
Redmond, WA 98052, USA  
`{tomkocmi,chrife,rogrundk,marcind,himatsus,arulm}@microsoft.com`

BLEU might be Guilty but References are not Innocent

Markus Freitag, David Grangier, Isaac Caswell  
Google Research  
`{freitag,grangier,icaswell}@google.com`

Rouge favors systems that produce longer summaries

## How to Compare Summarizers without Target Length? Pitfalls, Solutions and Re-Examination of the Neural Summarization Literature

Simeng Sun<sup>1</sup> Ori Shapira<sup>2</sup> Ido Dagan<sup>2</sup> Ani Nenkova<sup>1</sup>

<sup>1</sup>Department of Computer and Information Science, University of Pennsylvania

<sup>2</sup>Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel

`{simsun, nenkova}@seas.upenn.edu`  
`obspp18@gmail.com, dagan@cs.biu.ac.il`

# Human Evaluations are Diverse

Paper	Criterion Name in Paper	quality-criterion properties			Evaluation Mode		
		Type of Quality	Form/Content	Frame of Reference (FoR)	obj. / subj.	abs. / rel.	extr. / intr.
<i>Group 1 – Same name, different quality-criterion properties, same evaluation modes (2 example sets):</i>							
Yu et al. (2020)	Fluency	goodness	form	none	subj.	abs.	intr.
Van de Cruys (2020)	Fluency	correctness	form	none	subj.	abs.	intr.
Pan et al. (2020)	Fluency	correctness	(a) form (b) content (c) content	(a) none (b) none (c) external FoR	subj.	abs.	intr.
Van de Cruys (2020)	Coherence	goodness	content	none	subj.	abs.	intr.
Juraska et al. (2019)	Coherence	(a) correctness (b) goodness	form	none	subj.	abs.	intr.
Chai and Wan (2020)	Coherence	goodness	content	external FoR	subj.	abs.	intr.
Barros et al. (2017)	Coherence	correctness	content	none	subj.	abs.	intr.
<i>Group 2 – Different names, same quality-criterion properties, same evaluation modes:</i>							
Wang et al. (2020)	Faithfulness	correctness	content	FoR = input	obj.	abs.	intr.
Cao et al. (2020)	Content Similarity	correctness	content	FoR = input	obj.	abs.	intr.
Zhou et al. (2020)	Content Preservation	correctness	content	FoR = input	obj.	abs.	intr.
<i>Group 3 – Different names, same quality-criterion properties, different evaluation modes (2 example sets):</i>							
Gatt and Belz (2008)	Reading Time	goodness	both	none	obj	abs	extr
Forrest et al. (2018)	Ease of Reading	goodness	both	none	subj.	abs.	intr.
Miliaev et al. (2003)	Usefulness	goodness	both	external FoR	subj.	abs.	intr.
Qu and Green (2002)	Task success	goodness	both	external FoR	obj.	abs.	extr.
<i>Group 4 – Equivalent names, same quality-criterion properties, different evaluation modes:</i>							
Moraes et al. (2016)	Text Complexity	feature	both	none	subj.	rel.	intr.
Narayan and Gardent (2016)	Simplicity	feature	both	none	subj.	abs.	intr.
<i>Group 5 – Different names, different quality-criterion properties, different evaluation modes, related definitions:</i>							
Chai and Wan (2020)	Coreference	correctness	both	none	subj.	abs.	intr.
Funakoshi et al. (2004)	Accuracy	correctness	both	external FoR	obj.	abs.	extr.
Gatt and Belz (2008)	Identification Time	goodness	both	external FoR	obj	abs	extr

# Human Evaluations are Diverse

Paper	Criterion Name in Paper	quality-criterion properties			Evaluation Mode		
		Type of Quality	Form/Content	Frame of Reference (FoR)	obj. / subj.	abs. / rel.	extr. / intr.
<i>Group 1 – Same name, different quality-criterion properties, same evaluation modes (2 example sets):</i>							
Yu et al. (2020)	Fluency	goodness	form	none	subj.	abs.	intr.
Van de Cruys (2020)	Fluency	correctness	form	none	subj.	abs.	intr.
Pan et al. (2020)	Fluency	correctness	(a) form (b) content (c) content	(a) none (b) none (c) external FoR	subj.	abs.	intr.
Van de Cruys (2020)	Coherence	goodness	content	none	subj.	abs.	intr.
Juraska et al. (2019)	Coherence	(a) correctness (b) goodness	form	none	subj.	abs.	intr.
Chai and Wan (2020)	Coherence	goodness	content	external FoR	subj.	abs.	intr.
Barros et al. (2017)	Coherence	correctness	content	none	subj.	abs.	intr.
<i>Group 2 – Different names, same quality-criterion properties, same evaluation modes:</i>							
Wang et al. (2020)	Faithfulness	correctness	content	FoR = input	obj.	abs.	intr.
Cao et al. (2020)	Content Similarity	correctness	content	FoR = input	obj.	abs.	intr.
Zhou et al. (2020)	Content Preservation	correctness	content	FoR = input	obj.	abs.	intr.
<i>Group 3 – Different names, same quality-criterion properties, different evaluation modes (2 example sets):</i>							
Gatt and Belz (2008)	Reading Time	goodness	both	none	obj.	abs.	extr.
Forrest et al. (2018)	Ease of Reading	goodness	both	none	subj.	abs.	intr.
Miliaev et al. (2003)	Usefulness	goodness	both	external FoR	subj.	abs.	intr.
Qu and Green (2002)	Task success	goodness	both	external FoR	obj.	abs.	extr.
<i>Group 4 – Equivalent names, same quality-criterion properties, different evaluation modes:</i>							
Moraes et al. (2016)	Text Complexity	feature	both	none	subj.	rel.	intr.
Narayan and Gardent (2016)	Simplicity	feature	both	none	subj.	abs.	intr.
<i>Group 5 – Different names, different quality-criterion properties, different evaluation modes, related definitions:</i>							
Chai and Wan (2020)	Coreference	correctness	both	none	subj.	abs.	intr.
Funakoshi et al. (2004)	Accuracy	correctness	both	external FoR	obj.	abs.	extr.
Gatt and Belz (2008)	Identification Time	goodness	both	external FoR	obj.	abs	extr

# Human Evaluations are Diverse

Paper	Criterion Name in Paper	quality-criterion properties			Evaluation Mode		
		Type of Quality	Form/Content	Frame of Reference (FoR)	obj. / subj.	abs. / rel.	extr. / intr.
<i>Group 1 – Same name, different quality-criterion properties, same evaluation modes (2 example sets):</i>							
Yu et al. (2020)	Fluency	goodness	form	none	subj.	abs.	intr.
Van de Cruys (2020)	Fluency	correctness	form	none	subj.	abs.	intr.
Pan et al. (2020)	Fluency	correctness	(a) form (b) content (c) content	(a) none (b) none (c) external FoR	subj.	abs.	intr.
Van de Cruys (2020)	Coherence	goodness	content	none	subj.	abs.	intr.
Juraska et al. (2019)	Coherence	(a) correctness (b) goodness	form	none	subj.	abs.	intr.
Chai and Wan (2020)	Coherence	goodness	content	external FoR	subj.	abs.	intr.
Barros et al. (2017)	Coherence	correctness	content	none	subj.	abs.	intr.
<i>Group 2 – Different names, same quality-criterion properties, same evaluation modes:</i>							
Wang et al. (2020)	Faithfulness	correctness	content	FoR = input	obj.	abs.	intr.
Cao et al. (2020)	Content Similarity	correctness	content	FoR = input	obj.	abs.	intr.
Zhou et al. (2020)	Content Preservation	correctness	content	FoR = input	obj.	abs.	intr.
<i>Group 3 – Different names, same quality-criterion properties, different evaluation modes (2 example sets):</i>							
Gatt and Belz (2008)	Reading Time	goodness	both	none	obj.	abs.	extr.
Forrest et al. (2018)	Ease of Reading	goodness	both	none	subj.	abs.	intr.
Miliaev et al. (2003)	Usefulness	goodness	both	external FoR	subj.	abs.	intr.
Qu and Green (2002)	Task success	goodness	both	external FoR	obj.	abs.	extr.
<i>Group 4 – Equivalent names, same quality-criterion properties, different evaluation modes:</i>							
Moraes et al. (2016)	Text Complexity	feature	both	none	subj.	rel.	intr.
Narayan and Gardent (2016)	Simplicity	feature	both	none	subj.	abs.	intr.
<i>Group 5 – Different names, different quality-criterion properties, different evaluation modes, related definitions:</i>							
Chai and Wan (2020)	Coreference	correctness	both	none	subj.	abs.	intr.
Funakoshi et al. (2004)	Accuracy	correctness	both	external FoR	obj.	abs.	extr.
Gatt and Belz (2008)	Identification Time	goodness	both	external FoR	obj.	abs	extr

# Human Evaluations are Diverse

Paper	Criterion Name in Paper	quality-criterion properties			Evaluation Mode		
		Type of Quality	Form/Content	Frame of Reference (FoR)	obj. / subj.	abs. / rel.	extr. / intr.
<i>Group 1 – Same name, different quality-criterion properties, same evaluation modes (2 example sets):</i>							
Yu et al. (2020)	Fluency	goodness	form	none	subj.	abs.	intr.
Van de Cruys (2020)	Fluency	correctness	form	none	subj.	abs.	intr.
Pan et al. (2020)	Fluency	correctness	(a) form (b) content (c) content	(a) none (b) none (c) external FoR	subj.	abs.	intr.
Van de Cruys (2020)	Coherence	goodness	content	none	subj.	abs.	intr.
Juraska et al. (2019)	Coherence	(a) correctness (b) goodness	form	none	subj.	abs.	intr.
Chai and Wan (2020)	Coherence	goodness	content	external FoR	subj.	abs.	intr.
Barros et al. (2017)	Coherence	correctness	content	none	subj.	abs.	intr.
<i>Group 2 – Different names, same quality-criterion properties, same evaluation modes:</i>							
Wang et al. (2020)	Faithfulness	correctness	content	FoR = input	obj.	abs.	intr.
Cao et al. (2020)	Content Similarity	correctness	content	FoR = input	obj.	abs.	intr.
Zhou et al. (2020)	Content Preservation	correctness	content	FoR = input	obj.	abs.	intr.
<i>Group 3 – Different names, same quality-criterion properties, different evaluation modes (2 example sets):</i>							
Gatt and Belz (2008)	Reading Time	goodness	both	none	obj.	abs.	extr.
Forrest et al. (2018)	Ease of Reading	goodness	both	none	subj.	abs.	intr.
Miliaev et al. (2003)	Usefulness	goodness	both	external FoR	subj.	abs.	intr.
Qu and Green (2002)	Task success	goodness	both	external FoR	obj.	abs.	extr.
<i>Group 4 – Equivalent names, same quality-criterion properties, different evaluation modes:</i>							
Moraes et al. (2016)	Text Complexity	feature	both	none	subj.	rel.	intr.
Narayan and Gardent (2016)	Simplicity	feature	both	none	subj.	abs.	intr.
<i>Group 5 – Different names, different quality-criterion properties, different evaluation modes, related definitions:</i>							
Chai and Wan (2020)	Coreference	correctness	both	none	subj.	abs.	intr.
Funakoshi et al. (2004)	Accuracy	correctness	both	external FoR	obj.	abs.	extr.
Gatt and Belz (2008)	Identification Time	goodness	both	external FoR	obj.	abs	extr

# Human Evaluation is not Gold Standard

Could lead to divergence due to :

- (a) background knowledge,
- (b) preconceptions about language,
- (c) general educational level.

## *How Do Cultural Differences Impact the Quality of Sarcasm Annotation?: A Case Study of Indian Annotators and American Text*

Aditya Joshi<sup>1,2,3</sup> Pushpak Bhattacharyya<sup>1</sup> Mark Carman<sup>2</sup>  
Jaya Saraswati<sup>1</sup> Rajita Shukla<sup>1</sup>

<sup>1</sup>IIT Bombay, India

<sup>2</sup>Monash University, Australia

<sup>3</sup>IITB-Monash Research Academy, India

{adityaj, pb}@cse.iitb.ac.in, mark.carman@monash.edu

## HUMAN FEEDBACK IS NOT GOLD STANDARD

Tom Hosking  
University of Edinburgh  
tom.hosking@ed.ac.uk

Phil Blunsom  
Cohere  
phil@cohere.com

Max Bartolo  
Cohere, UCL  
max@cohere.com

## Assessing Inter-Annotator Agreement for Translation Error Annotation

Arle Lommel, Maja Popović, Aljoscha Burchardt,

DFKI

Alt-Moabit 91c, 10559 Berlin, Germany  
E-mail: arle.lommel@dfki.de, maja.popovic@dfki.de, aljoscha.burchardt@dfki.de

## Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions

David M. Howcroft<sup>1</sup>✉, Anya Belz<sup>2</sup>, Miruna Clinciu<sup>1</sup>, Dimitra Gkatzia<sup>3</sup>,  
Sadid A. Hasan<sup>4</sup>, Saad Mahamood<sup>5</sup>, Simon Mille<sup>6</sup>,  
Emiel van Miltenburg<sup>7</sup>, Sashank Santhanam<sup>8</sup>, and Verena Rieser<sup>1</sup>

<sup>1</sup>The Interaction Lab, MACS, Heriot-Watt University, Edinburgh, Scotland, UK

<sup>2</sup>University of Brighton, Brighton, England, UK

<sup>3</sup>Edinburgh Napier University, Edinburgh, Scotland, UK

<sup>4</sup>CVS Health, Wellesley, MA, USA

<sup>5</sup>trivago N.V., Düsseldorf, Germany

<sup>6</sup>Universitat Pompeu Fabra, Barcelona, Spain

<sup>7</sup>Tilburg Center for Cognition & Communication, Tilburg University, Tilburg, Netherlands

<sup>8</sup>Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

✉ Corresponding author: D. Howcroft@hw.ac.uk

# Things to Consider when Utilizing Human Evaluations

Topic	Best practice
General	Always conduct a human evaluation (if possible).
Criteria	Use separate criteria rather than an overall quality assessment. Properly define the criteria that are used in the evaluation.
Sampling	Preferably use a (large-scale) reader-focused design rather than a (small-scale) expert-focused design. Always recruit sufficiently many participants. Report (and motivate) the sample size and the demographics.
Annotation	For a qualitative analysis, recruit multiple annotators (at least 2, more is better) Report the Inter-Annotator Agreement score with confidence intervals, plus a percentage agreement.
Measurement	For a quantitative study, use multiple item 7-point (preferably) Likert scales, or (continuous) ranking.
Design	Reduce order- and learning effects by counterbalancing/random ordering, and properly report this.
Statistics	If the evaluation study is exploratory, only report exploratory data analysis. If the study is confirmatory, consider preregistering and conduct appropriate statistical analyses.

# Things to Consider when Utilizing Human Evaluations

Topic	Best practice
General	Always conduct a human evaluation (if possible).
Criteria	Use separate criteria rather than an overall quality assessment. Properly define the criteria that are used in the evaluation.
Sampling	Preferably use a (large-scale) reader-focused design rather than a (small-scale) expert-focused design. Always recruit sufficiently many participants. Report (and motivate) the sample size and the demographics.
Annotation	For a qualitative analysis, recruit multiple annotators (at least 2, more is better) Report the Inter-Annotator Agreement score with confidence intervals, plus a percentage agreement.
Measurement	For a quantitative study, use multiple item 7-point (preferably) Likert scales, or (continuous) ranking.
Design	Reduce order- and learning effects by counterbalancing/random ordering, and properly report this.
Statistics	If the evaluation study is exploratory, only report exploratory data analysis. If the study is confirmatory, consider preregistering and conduct appropriate statistical analyses.

# Things to Consider when Utilizing Human Evaluations

Topic	Best practice
General	Always conduct a human evaluation (if possible).
Criteria	Use separate criteria rather than an overall quality assessment. Properly define the criteria that are used in the evaluation.
Sampling	Preferably use a (large-scale) reader-focused design rather than a (small-scale) expert-focused design. Always recruit sufficiently many participants. Report (and motivate) the sample size and the demographics.
Annotation	For a qualitative analysis, recruit multiple annotators (at least 2, more is better) Report the Inter-Annotator Agreement score with confidence intervals, plus a percentage agreement.
Measurement	For a quantitative study, use multiple item 7-point (preferably) Likert scales, or (continuous) ranking.
Design	Reduce order- and learning effects by counterbalancing/random ordering, and properly report this.
Statistics	If the evaluation study is exploratory, only report exploratory data analysis. If the study is confirmatory, consider preregistering and conduct appropriate statistical analyses.

# Things to Consider when Utilizing Human Evaluations

Topic	Best practice
General	Always conduct a human evaluation (if possible).
Criteria	Use separate criteria rather than an overall quality assessment. Properly define the criteria that are used in the evaluation.
Sampling	Preferably use a (large-scale) reader-focused design rather than a (small-scale) expert-focused design. Always recruit sufficiently many participants. Report (and motivate) the sample size and the demographics.
Annotation	For a qualitative analysis, recruit multiple annotators (at least 2, more is better) Report the Inter-Annotator Agreement score with confidence intervals, plus a percentage agreement.
Measurement	For a quantitative study, use multiple item 7-point (preferably) Likert scales, or (continuous) ranking.
Design	Reduce order- and learning effects by counterbalancing/random ordering, and properly report this.
Statistics	If the evaluation study is exploratory, only report exploratory data analysis. If the study is confirmatory, consider preregistering and conduct appropriate statistical analyses.

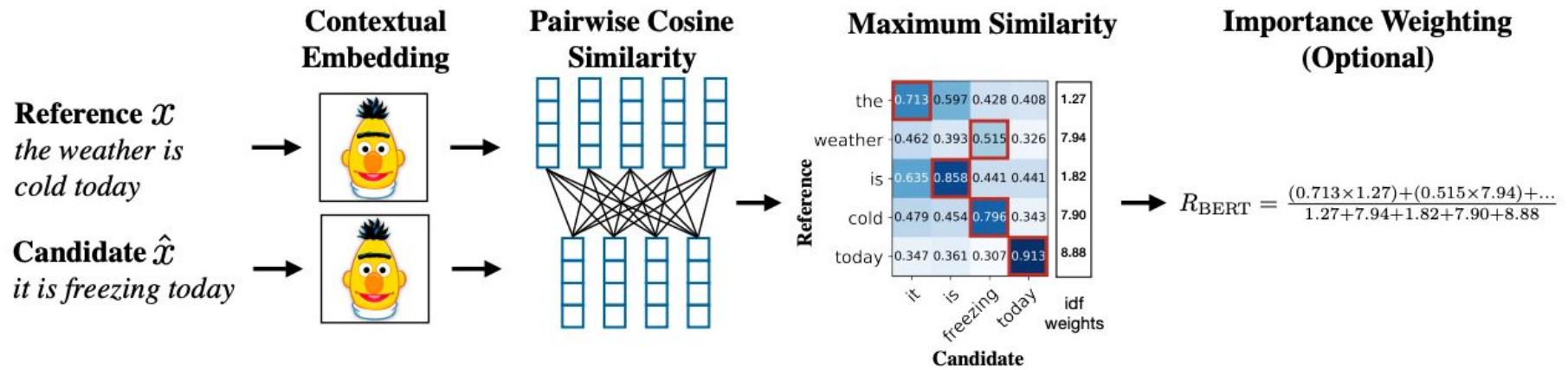
# Things to Consider when Utilizing Human Evaluations

Topic	Best practice
General	Always conduct a human evaluation (if possible).
Criteria	Use separate criteria rather than an overall quality assessment. Properly define the criteria that are used in the evaluation.
Sampling	Preferably use a (large-scale) reader-focused design rather than a (small-scale) expert-focused design. Always recruit sufficiently many participants. Report (and motivate) the sample size and the demographics.
Annotation	For a qualitative analysis, recruit multiple annotators (at least 2, more is better) Report the Inter-Annotator Agreement score with confidence intervals, plus a percentage agreement.
Measurement	For a quantitative study, use multiple item 7-point (preferably) Likert scales, or (continuous) ranking.
Design	Reduce order- and learning effects by counterbalancing/random ordering, and properly report this.
Statistics	If the evaluation study is exploratory, only report exploratory data analysis. If the study is confirmatory, consider preregistering and conduct appropriate statistical analyses.

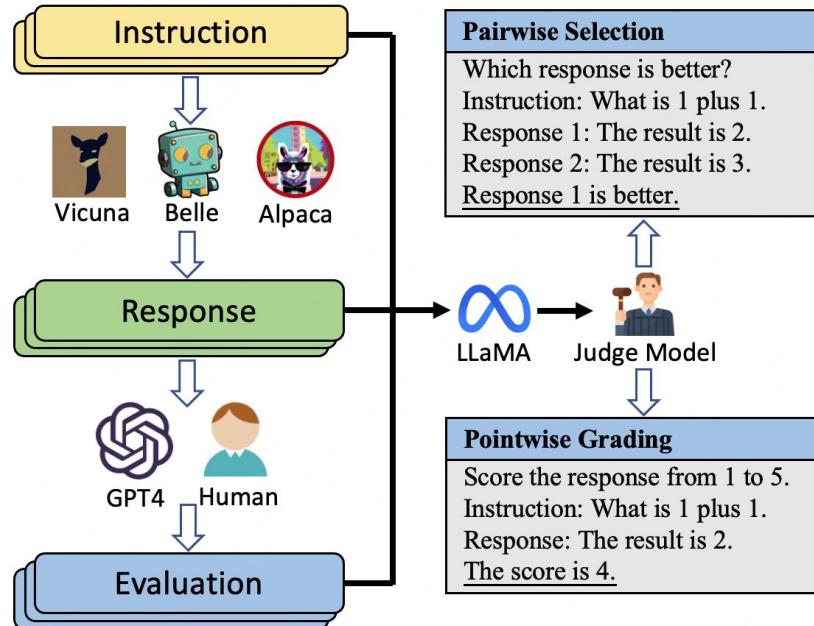
# Things to Consider when Utilizing Human Evaluations

SYSTEM	
task	<b>What problem are you solving (e.g. data-to-text)?</b> How does it relate to other NLG (sub)tasks?
input/output	<b>What do you feed in and get out of your system?</b> Show examples of inputs and outputs of your system. Additionally, if you include pre and post-processing steps in your pipeline, clarify whether your input is to the preprocessing, and your output is from the post-processing, step, or what you consider to be the ‘core’ NLG system. In general, make it easy for readers to determine what form the data is in as it flows through your system.
EVALUATION CRITERIA	
name	<b>What is the name for the quality criterion you are measuring (e.g. grammaticality)?</b>
definition	<b>How do you define that quality criterion?</b> Provide a definition for your criterion. It is okay to cite another paper for the definition; however, it should be easy for your readers to figure out what aspects of the text you wanted to evaluate.
OPERATIONALISATION	
instrument type	<b>How are you collecting responses?</b> Direct ratings, post-edits, surveys, observation? Rankings or rating scales with numbers or verbal descriptors? Provide the full prompt or question with the set of possible response values where applicable, e.g. when using Likert scales.
instructions, prompts, and questions	<b>What are your participants responding to?</b> Following instructions, answering a question, agreeing with a statement? <i>The exact text you give your participants is important for anyone trying to replicate your experiments.</i> In addition to the immediate task instructions, question or prompt, provide the full set of instructions as part of your experimental design materials in an appendix.

# New Metrics are Increasingly Neural Network-Based

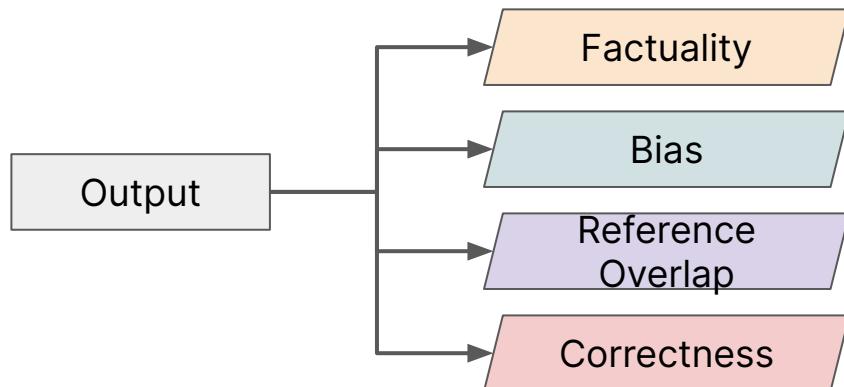


# New Metrics are Increasingly Neural Network-Based



# Must Metrics Reflect Human Evaluation?

While metrics such as BLEU does not correlate to human judgment ([Callison-Burch et al., 2006](#)), it may not necessarily be desirable ([Gehrmann et al., 2022](#)). Opt instead for multidimensional.



# Direct Comparison is Not Straightforward!

Model↓	ARC-CHALLENGE Evaluations:						OPENBOOKQA Evaluations:					
	Ref1	Ref2	Ref3	Ref4	Ref5	Ref6	OLMES	Ref2	Ref4	Ref5	Ref7	Ref8
MPT-7B	47.7	42.6			46.5		45.7	51.4	48.6			52.4
RPJ-INCITE-7B	46.3				42.8		45.3		49.4			49.0
Falcon-7B	47.9	42.4		44.5	47.5		49.7	51.6	44.6	53.0	26.0†	55.2
Mistral-7B	60.0		55.5	54.9			78.6†			52.2	77.6†	80.6†
Llama2-7B	53.1	45.9	43.2	45.9	48.5	53.7†	54.2	58.6	58.6	48.4	58.6	54.4†
Llama2-13B	59.4	49.4	48.8	49.4		67.6†	67.3†	57.0	57.0		57.0	63.4†
Llama3-8B	60.2					78.6†	79.3†				76.6†	77.2†
Num shots	25	0	0	0	0	25	5	0	0	0	0	5
Curated shots	No					No	Yes				No	Yes
Formulation	CF	CF	CF?	CF	CF	MCF	MCF/CF	CF	CF	CF?	CF	MCF
Normalization	char	char	?	char?	pmi	none	none/pmi	pmi	pmi?	pmi	pmi?	none

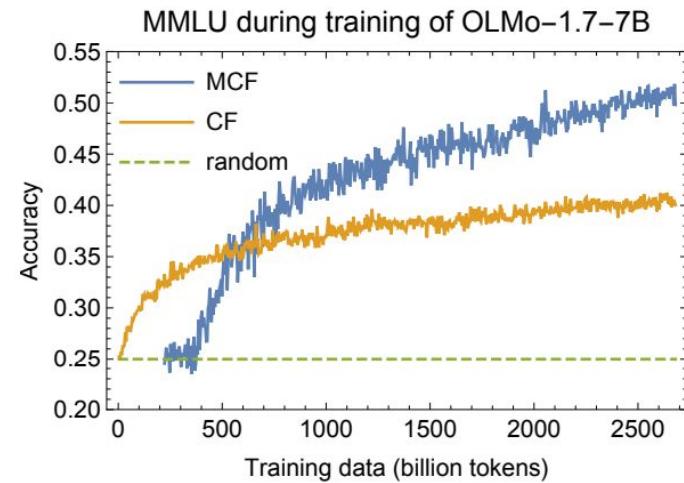
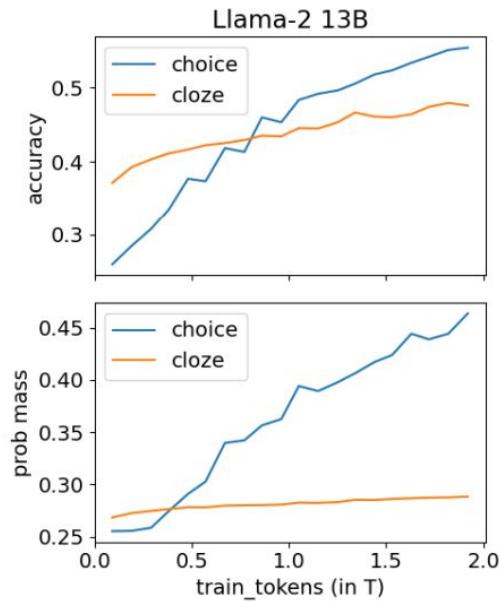
Ref Reference citation

**Ref1** HF Open LLM Leaderboard (Beeching et al., 2023)  
**Ref2** Llama2 paper (Touvron et al., 2023a)  
**Ref3** Mistral 7B (Jiang et al., 2023)  
**Ref4** Falcon paper (Almazrouei et al., 2023)

Ref Reference citation

**Ref5** OLMo paper (Groeneveld et al., 2024)  
**Ref6** Llama3 model card (AI@Meta, 2024)  
**Ref7** Gemma paper (Gemma Team et al., 2024)  
**Ref8** HELM Lite Leaderboard (Liang et al., 2023)

# What Signal do We Want to Measure?



[Gu et al, 2024] OLMES: A Standard for Language Model Evaluations

[Madaan et al, 2024] Quantifying Variance in Evaluation Benchmarks

# Prompt choice affects performance

	ARC Challenge		MMLU	
	Cloze	MMLU-style	Hybrid	MMLU-style
GPT-NeoX-20B	<b>38.0 ± 2.78 %</b>	26.6 ± 2.53%	<b>27.6 ± 0.74%</b>	24.5 ± 0.71%
Llama-2-7B	<b>43.5 ± 2.84%</b>	42.8 ± 2.83%	39.8 ± 0.79%	<b>41.3 ± 0.80%</b>
Falcon-7B	<b>40.2 ± 2.81%</b>	25.9 ± 2.51%	<b>29.1 ± 0.75%</b>	25.4 ± 0.72%
Mistral-7B	50.1 ± 2.86%	<b>72.4 ± 2.56%</b>	48.3 ± 0.80%	<b>58.6 ± 0.77%</b>
Mixtral-8x7B	56.7 ± 2.84%	<b>81.3 ± 2.23%</b>	59.7 ± 0.77%	<b>67.1 ± 0.72%</b>

## Multiple-Choice Formulation

Question: Earth's core is primarily composed of which of the following materials?

- (A) basalt (B) iron (C) magma (D) quartz

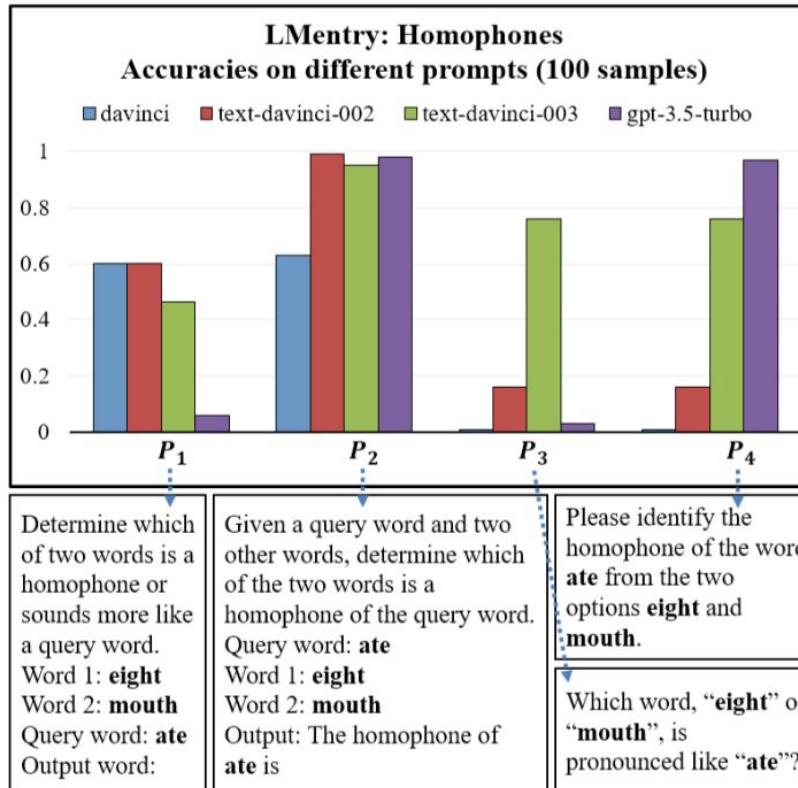
Answer: (B)

## Cloze Formulation

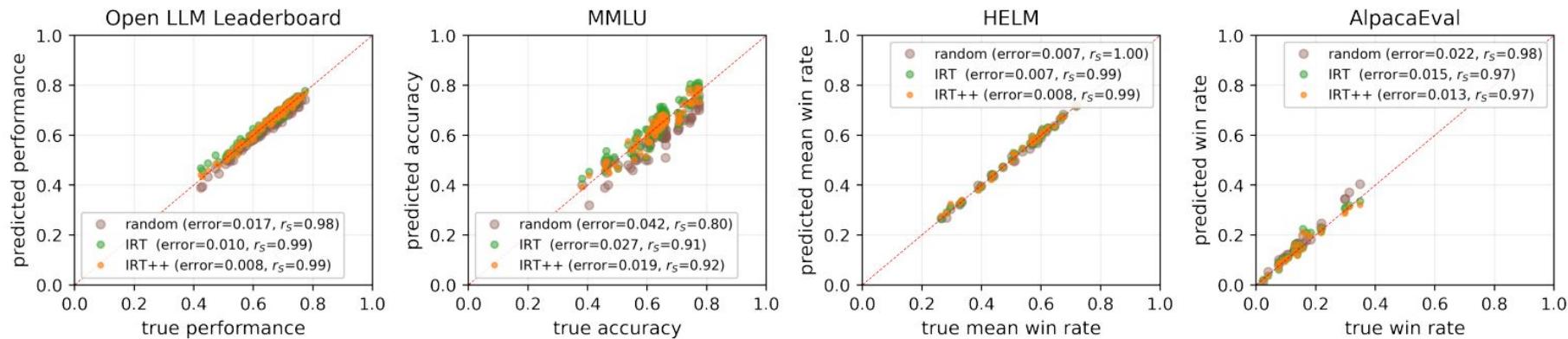
Question: Earth's core is primarily composed of which of the following materials?

Answer: <answer>, where each answer choice is separately substituted in for <answer>.

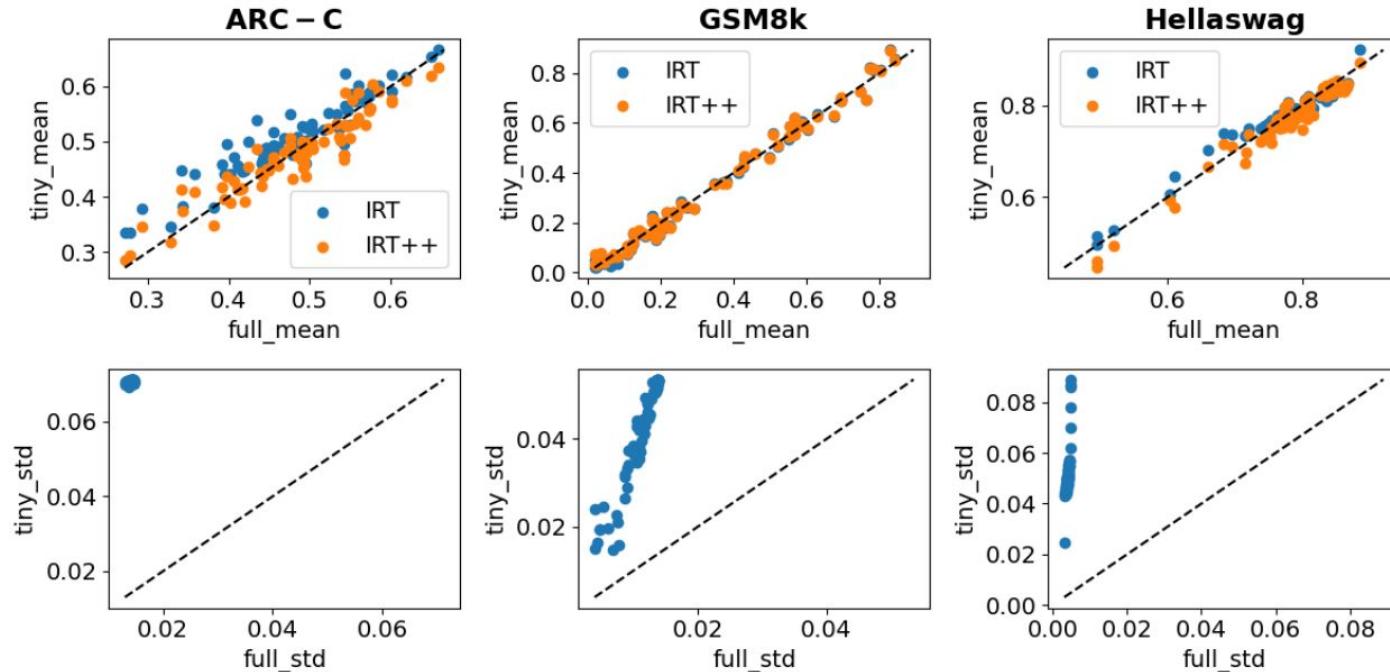
# Set of Prompts Could be Considered Part of the Benchmark



# Making Benchmarks Smaller by Targeted Sampling



# But Shrinking May Not Offer The Full Picture



# Statistical Analysis would Benefit Benchmark Modification

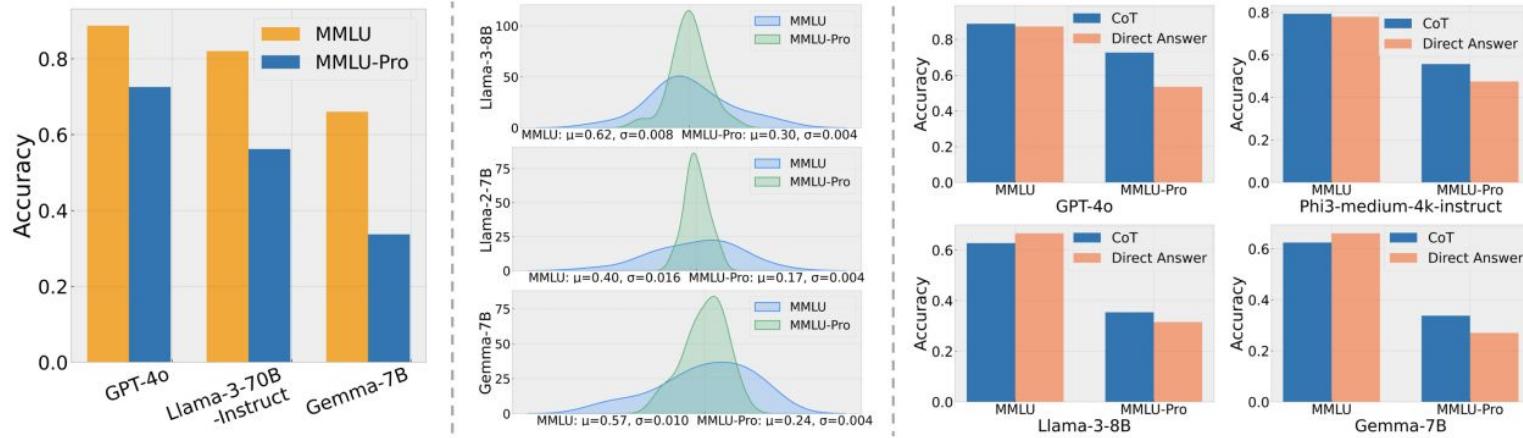


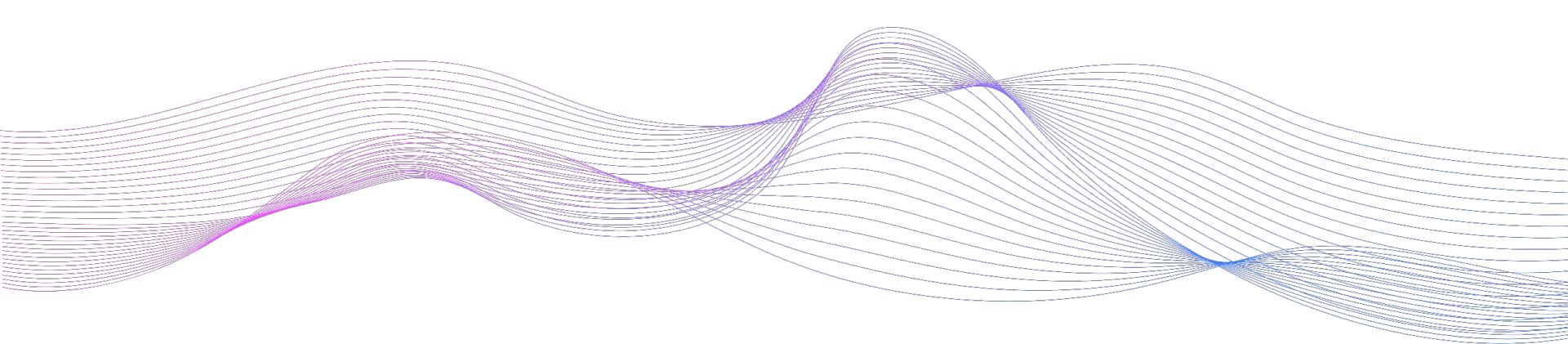
Figure 1: Comparing between MMLU and MMLU-Pro: (Left) Performance gap; (Center) Accuracy distributions affected by 24 prompts, with taller and thinner profiles indicating more stability and shorter and wider profiles indicating greater fluctuations; (Right) Performance using CoT vs. Direct.

# Benchmarks should ...

- Imply robust in-domain performance if good performance is observed
  - **We need more work on dataset design and data collection methods**
- Have examples that are accurately and unambiguously annotated
  - **Test examples should be validated thoroughly enough to remove erroneous examples and to properly handle ambiguous ones**
- Offer adequate statistical power
  - **Much larger and or much harder**
- Reveal plausibly harmful social biases in systems and should not incentivize the creation of biased systems
  - **Encourage the development and use of auxiliary bias evaluation metrics**

# Benchmarking is Difficult

- Benchmarks dictate what we measure → what we end up building
- Must ensure **validity** of our evaluations for findings to be useful
- Careful dataset construction and metric design is crucial



# Addressing Evaluation Pitfalls

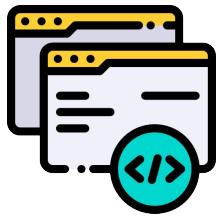
# Addressing Evaluation Pitfalls

What can we do to address these challenges right now?

# Reporting Standards for LM Benchmarking

- There are no standards for *sufficient* and *complete* reporting of evaluation details
  - → Many don't report much info on evaluation setup at all (even their prompts!)
  - → For those who do, it's easy to leave key facts out accidentally or through lack of understanding

# Share, share, share!\*



Evaluation Code



Methodology  
Details (Prompts  
included)



Model  
responses

\*Don't overshare: hold out an extra private test set!

# Sharing Code Mitigates Reproducibility Challenges

- Publishing evaluation code used to obtain results can ensure sufficient documentation
- Serving as a “ground truth” reference point for methodology

# Share Model Outputs

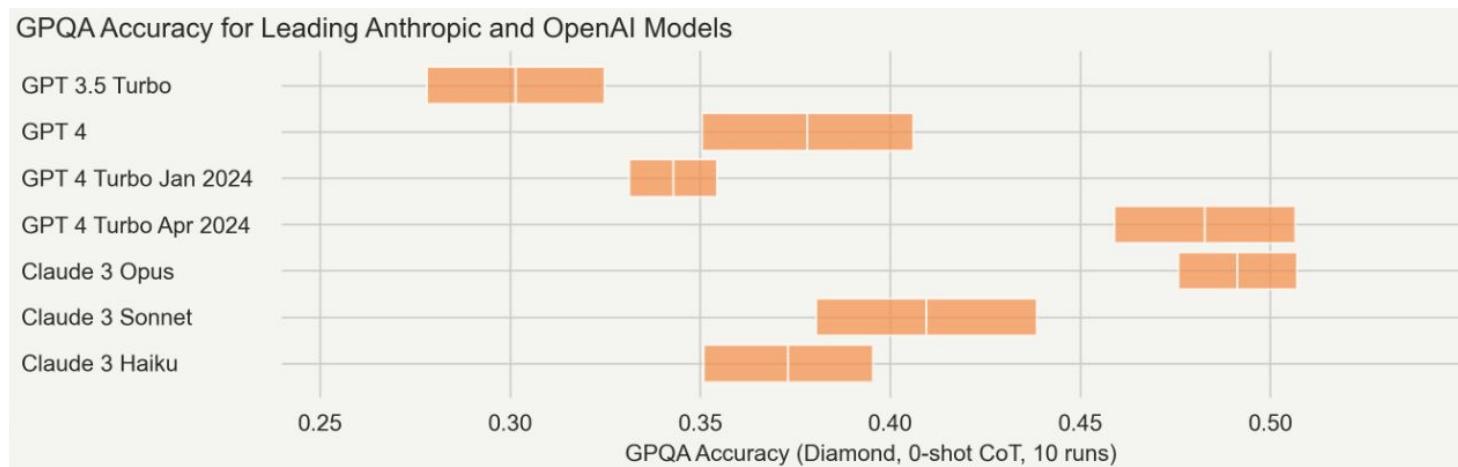
- Allows for reproducibility, even when API models are deprecated
- Allows for future error analysis
- Makes evaluation research possible even without \$\$\$ for evaluating large or expensive models

# Avoid Copying Results Across Publications

- Results across different publications will likely not match in all settings
  - →Comparisons may not be meaningful
- Drawing baseline numbers directly from other work is likely to mislead!

# Improved Statistical Testing

- Most papers do not report error bars *whatsoever*
- Harder benchmarks are getting smaller



# Operationalizing Best Practices

- Reimplementing many evaluation tasks is a lot of work!
  - Hard to account for quirks of every individual benchmark
  - Evaluation code may be entangled with model inference code, etc...
  - → But shareable code goes a long way: <https://github.com/openai/simple-evals>
- You can use existing evaluation libraries as infrastructure to lower the overhead of adopting best practices

# Existing Tooling and Standards



EleutherAI / lm-evaluation-harness



stanford-crfm / helm

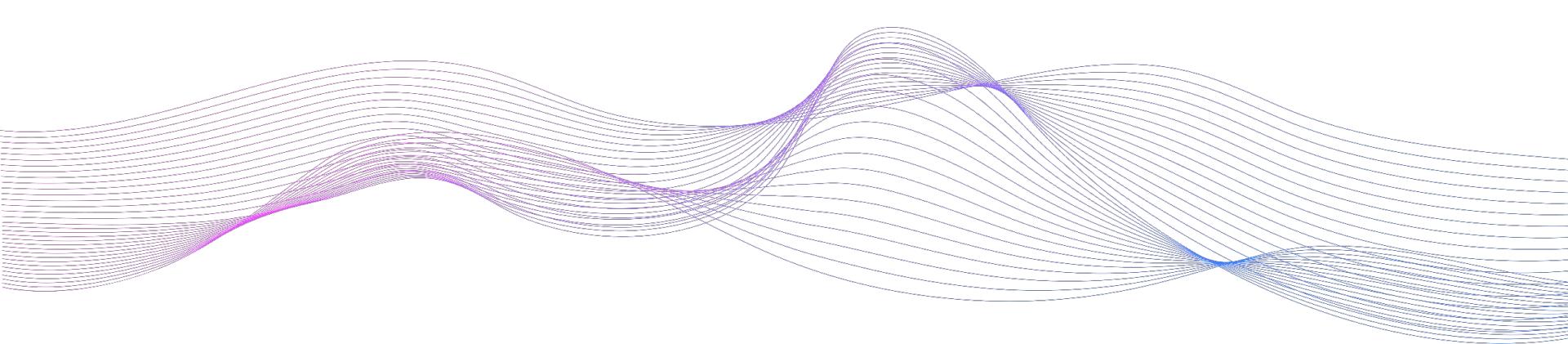


open-compass / opencompass



UKGovernmentBEIS / inspect\_ai

And more...



# Future Directions

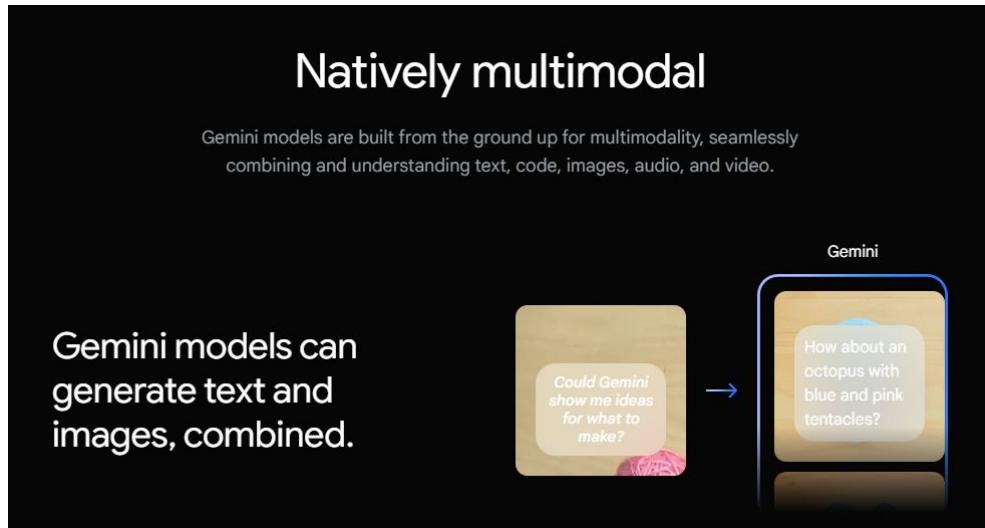
# Future Directions

What are promising future research directions?

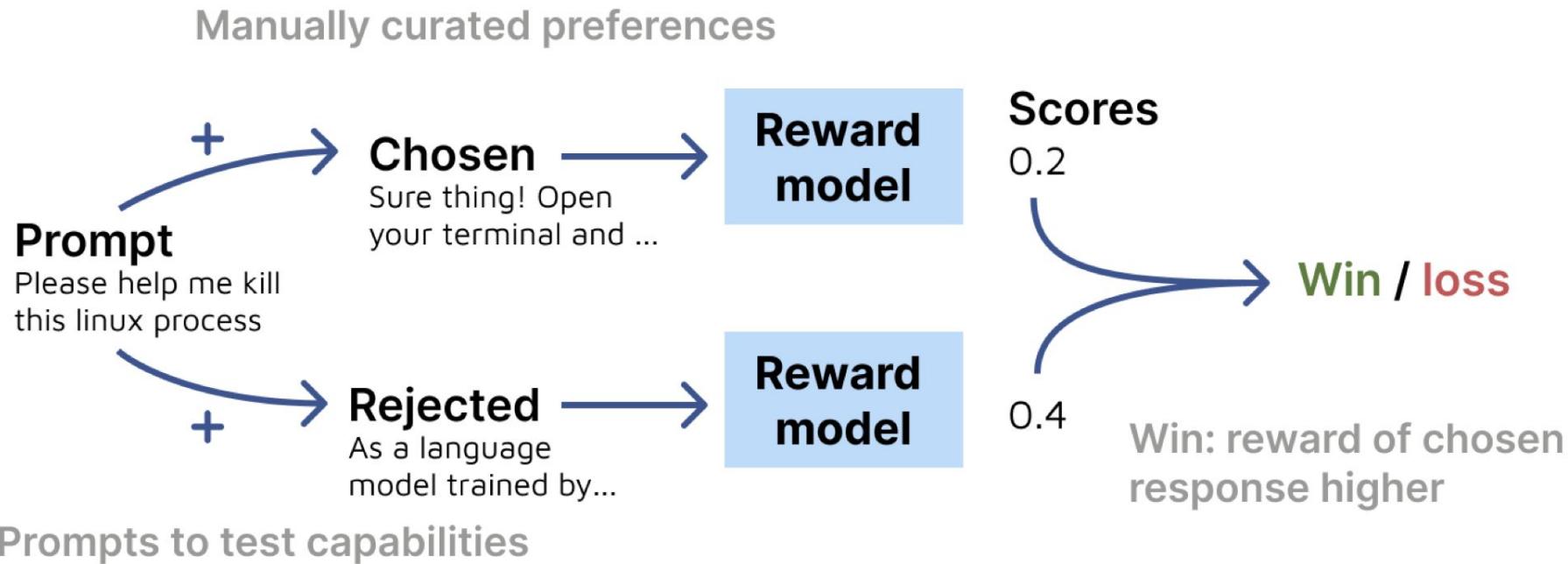
How can future LM evaluations be improved?

# Multimodality

- State-of-the-art “LMs” are no longer text-only
- What are we using multimodal language models for?
- How do we evaluate multimodal understanding and generation well?



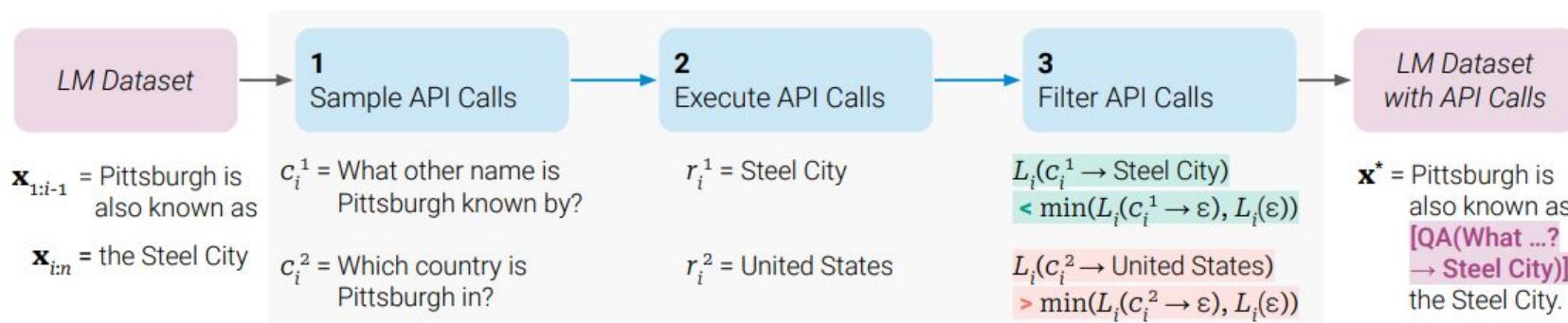
# RewardBench



# Tool Use

- Many LM-based systems are augmented to use external tools
  - Code interpreters
  - Calculators
  - Web search
  - And more!

Out of 1400 participants, 400 (or [\[Calculator\(400 / 1400\)\]](#)  
→ 0.29) 29% passed the test.



# “Agentic” Evaluations

- Evaluating models directly in downstream use cases as part of “agent” loops
- In general, evals need to grow beyond being static, since models are used interactively!

# Agentic Evaluations

## Leaderboard

Model	% Resolved	Date	Logs	Trajs	Site	Verified?	Open?
🟡 Factory Code Droid	19.27	2024-06-17	🔗	-	🔗	✗	✗
🟣 AutoCodeRover (v20240620) + GPT 4o (2024-05-13)	18.83	2024-06-28	🔗	-	🔗	✗	✗
🟤 AppMap Navie + GPT 4o (2024-05-13)	14.60	2024-06-15	🔗	-	🔗	✓	✓
Amazon Q Developer Agent (v20240430-dev)	13.82	2024-05-09	🔗	-	🔗	✗	✗
SWE-agent + GPT 4 (1106)	12.47	2024-04-02	🔗	🔗	🔗	✓	✓
SWE-agent + Claude 3 Opus	10.51	2024-04-02	🔗	🔗	-	✓	✓
RAG + Claude 3 Opus	3.79	2024-04-02	🔗	-	🔗	✓	✓
RAG + Claude 2	1.96	2023-10-10	🔗	-	-	✓	✓
RAG + GPT 4 (1106)	1.31	2024-04-02	🔗	-	-	✓	✓
RAG + SWE-Llama 13B	0.70	2023-10-10	🔗	-	-	✓	✓
RAG + SWE-Llama 7B	0.70	2023-10-10	🔗	-	-	✓	✓
RAG + ChatGPT 3.5	0.17	2023-10-10	🔗	-	-	✓	✓

- The **% Resolved** metric refers to the percentage of SWE-bench instances (2294 total) that were *resolved* by the model.
- "**Verified**" indicates that we, the SWE-bench team, received access to the system and were able to reproduce the patch generations.
- "**Open**" refers to submissions that have open-source code. This does *not* necessarily mean the underlying model is open-source.
- The leaderboard is updated once a week on **Monday**.
- If you would like to submit your model to the leaderboard, please check the [submission](#) page.
- All submissions are Pass@1, do not use `hints_text`, and are in the unassisted setting.

# Red-Teaming

- **Red-teaming:** human annotators attempting to elicit harms
  - Extremely important and useful
  - Very company-specific ; requires care hiring and paying annotator workforce
  - Most “realistic” form of measuring efficacy of guardrails
- How can best practices for red teaming be standardized and improved?
- How can portions be automated?

# Contamination-Proof Benchmarks

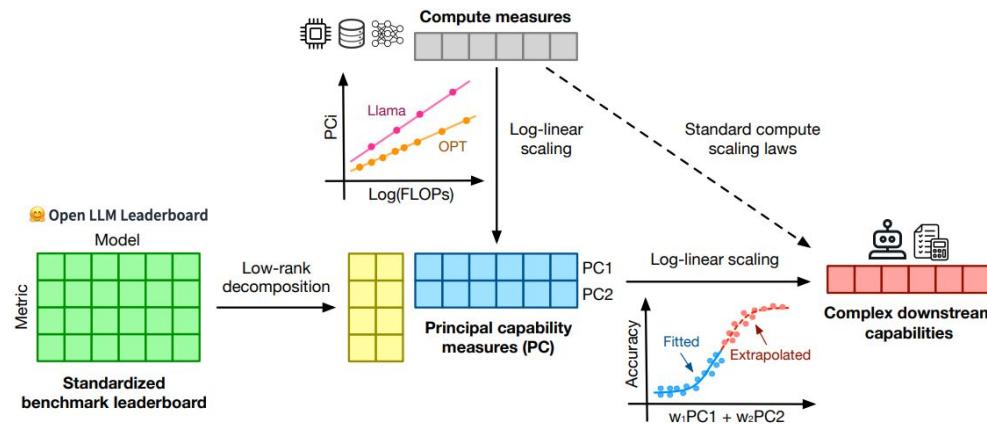
- What eval processes are most reliant to gaming? What ones can be shown to require “true” generalization OOD? How can we measure or ensure this?
- Private test sets can help a bit, but...

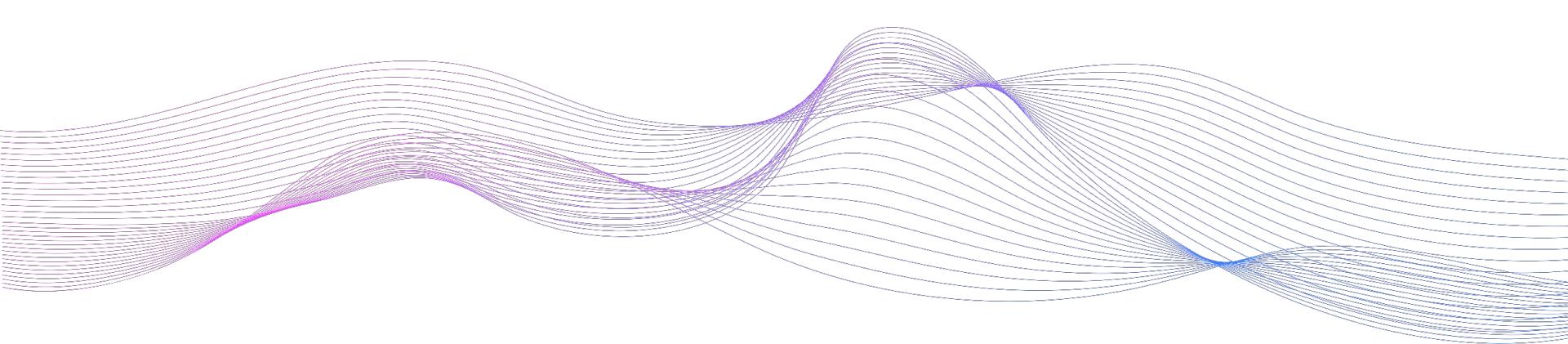
# Dynamic Evaluation Datasets

- Adversarial evolving datasets
  - [Dynabench: Rethinking Benchmarking in NLP](#)
- Targeted model-generated evaluation datasets
  - [Discovering Language Model Behaviors with Model-Written Evaluations](#)
  - [AutoBENCHer: Creating Salient, Novel, Difficult Datasets for Language Models](#)
  - [Task Me Anything](#)

# Predictable Evals + Eval Scaling Laws

- Scaling laws let us derisk model scaling and extrapolate performance estimates for loss.
- Can we do the same for downstream tasks of interest?





# Conclusions

# LM Evaluation is Challenging

- Lack of clear reporting standards and best practices
- Reproducibility is crucial—models are often non-robust to many counterintuitively important factors
- Many exciting areas for future research
  - New application areas
  - Evaluations that are more reflective of how models are used
  - More complex, dynamic evaluation processes
  - Evaluation on more complex capabilities



**ICML**  
International Conference  
On Machine Learning

Tutorial @ ICML 2024

# Challenges in LM Evaluation

Lehar 1-4

3.30 - 5.30pm CEST



**Lintang Sutawika**  
@lintangsutawika



**Hailey Schoelkopf**  
@haileysch\_