

# Benchmarking

CSE 5525: Foundations of Speech and Natural Language  
Processing

<https://shocheen.github.io/courses/cse-5525-spring-2025>



**THE OHIO STATE UNIVERSITY**

---

# Logistics

- Homework 3 is due on March 5.
  
- Final project progress report due March 28.

# Overview

- What is a benchmark?
- Quality of good benchmarks
- Benchmark and metrics, evaluation (closed and open-ended evaluation)
- Current evaluations of LLMs
- Issues with benchmarking

# Applications $\Rightarrow$ Tasks

Capabilities the NLP community has been targeting in its sixty-year history:

- Translate text from one language to another
- Summarize one or more documents in a few paragraphs or in a structured table
- Answer a question using information in one or more documents
- Engage in a conversation with a person and follow any instructions they give

A huge number of questions arise, options:

1. Conclude that the desired system is just isn't possible yet or would be very expensive to build with the best available methods
2. Define and tackle **tasks**—versions of the application that abstract away some details while making some simplifying assumptions

# What makes a task?

The term “**task**” is generally used among researchers to refer to a specification of certain components of an NLP system, most notably data and evaluation:

- **Data:** there is a set of realistic demonstrations of possible inputs paired with their desirable outputs
- **Evaluation:** there is a method for measuring, in a quantitative and reproducible way, how well any system’s output matches the desired output

An example of the task you worked on:

- Determine sentiment expressed in text ⇒ Binary sentiment classification
- **Dataset:** The Stanford Sentiment Treebank (SST-2)
  - Inputs are full sentences derived from another dataset of movie reviews by Pang and Lee (2005)
  - Crowdsourced fine-grained assessments of sentiment, then turn them into binary labels
- **Evaluation:** Accuracy (% of correctly predicted)

# What Is Benchmarking?

"Datasets are the telescopes of our field."

- Aravind Joshi

## **Benchmark:**

- \* one or multiple datasets
- \* one or multiple associated metrics
- \* ways to aggregate performance

# Benchmarks Orient AI

Benchmarks set priorities and codify values

Benchmarks are mechanisms for change

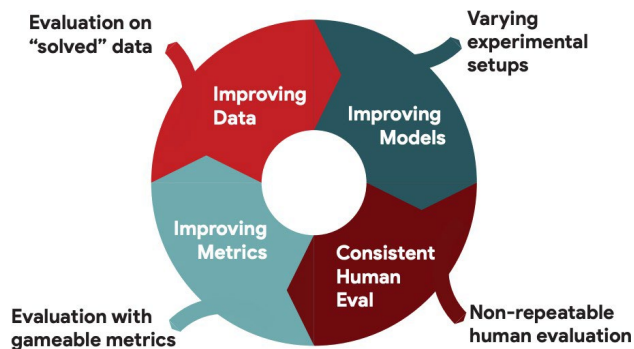


"proper evaluation is a complex and challenging business"

- Karen Spärck Jones (*ACL Lifetime Achievement Award, 2005*)

Spärck Jones and Galliers (1995), Liberman (2010), Ethayarajh and Jurafsky (2020), Bowman and Dahl (2021), Raji et al. (2021), Birhane et al. (2022), Bommasani (2022) *inter alia*

# Benchmarks are useful to track progress



google/**BIG-bench**

Beyond the Imitation Game collaborative benchmark for measuring and extrapolating the capabilities of language models



👤 217  
Contributors

🏠 2  
Used by

★ 2k  
Stars

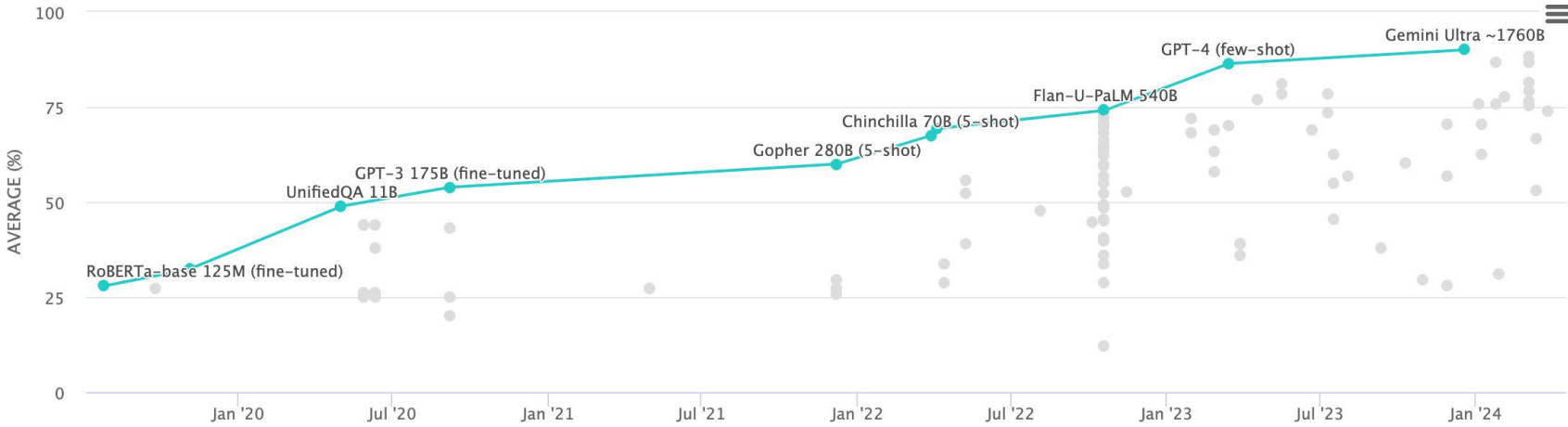
🍴 478  
Forks





# Benchmarks and evaluations drive progress

## MMLU



Benchmarks and how we drive the progress of the field

# A brief history of benchmarking

Benchmarks have a long history of being used to assess the performance of computational systems.

The Standard Performance Evaluation Corporation (SPEC),

Established in 1988 is one of the oldest organizations dedicated to benchmarking the performance of computer hardware

Benchmark sets and performances measured as millions of instructions per second (MIPS).

# Efforts in Machine Learning

MLCommons

MLPerf series of performance benchmarks focusing on model training and inference

DARPA and NIST

TREC workshop in IR

**ML**  
 **Commons**

# Benchmarking Principles

**Relevance:** Benchmarks should measure relatively vital features.

**Representativeness:** Benchmark performance metrics should be broadly accepted by industry and academia.

**Equity:** All systems should be fairly compared.

**Repeatability:** Benchmark results can be verified.

**Cost-effectiveness:** Benchmark tests are economical.

**Scalability:** Benchmark tests should work across systems possessing a range of resources from low to high.

**Transparency:** Benchmark metrics should be easy to understand.

# Two major types of evaluations

Close-ended  
evaluations

## Example

**Text:** Read the book, forget the movie!

**Label:** Negative

Open ended evaluations

**Context (human-written):** In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

**GPT-2:** The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Close-ended  
evaluation

# Close-ended tasks

- Limited number of potential answers
- Often one or just a few correct answers
- Enables automatic evaluation as in ML

# Close-ended tasks

- Sentiment analysis: SST / IMDB / Yelp ...

## Example

Text: Read the book, forget the movie!

Label: Negative

- Entailment: SNLI

## Example

Text: A soccer game with multiple males playing.

Hypothesis: Some men are playing sport.

Label: Entailment

- Name entity recognition: CoNLL-2003
- Part-of-Speech: PTB



# Close-ended tasks

- Coreference resolution: WSC

## Example

Text: Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.

Coreference: False

- Question Answering: Squad 2

## Example

Endangered Species Act Paragraph: "... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a **1937 treaty** prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little **opposition** was raised."

Question 1: "Which laws faced significant **opposition**?"

Plausible Answer: later laws










Question 2: "What was the name of the **1937 treaty**?"

Plausible Answer: Bald Eagle Protection Act

# Close-ended multi-task benchmark - superGLUE

 SuperGLUE  GLUE

Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
<b>+</b> 2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
<b>+</b> 6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
<b>+</b> 7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
<b>+</b> 9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

Attempt to measure "general language capabilities"

# Examples from superGLUE

Cover a number of different tasks

- BoolQ, MultiRC (reading texts)
- CB, RTE (Entailment)
- COPA (cause and effect)
- ReCoRD (QA+reasoning)
- WiC (meaning of words)
- WSC (coreference)

BoolQ	<p><b>Passage:</b> <i>Barq's</i> – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</p> <p><b>Question:</b> <i>is barq's root beer a pepsi product</i>    <b>Answer:</b> No</p>
CB	<p><b>Text:</b> <i>B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</i></p> <p><b>Hypothesis:</b> <i>they are setting a trend</i>    <b>Entailment:</b> Unknown</p>
COPA	<p><b>Premise:</b> <i>My body cast a shadow over the grass.</i>    <b>Question:</b> <i>What's the CAUSE for this?</i> <b>Alternative 1:</b> <i>The sun was rising.</i>    <b>Alternative 2:</b> <i>The grass was cut.</i></p> <p><b>Correct Alternative:</b> 1</p>
MultiRC	<p><b>Paragraph:</b> <i>Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week</i></p> <p><b>Question:</b> <i>Did Susan's sick friend recover?</i>    <b>Candidate answers:</b> <i>Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)</i></p>
ReCoRD	<p><b>Paragraph:</b> <i>(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the <u>State Electoral Commission</u> show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood</i></p> <p><b>Query</b> For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the &lt;placeholder&gt; presidency    <b>Correct Entities:</b> US</p>
RTE	<p><b>Text:</b> <i>Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.</i></p> <p><b>Hypothesis:</b> <i>Christopher Reeve had an accident.</i>    <b>Entailment:</b> False</p>
WiC	<p><b>Context 1:</b> <i>Room and board.</i>    <b>Context 2:</b> <i>He nailed boards across the windows.</i></p> <p><b>Sense match:</b> False</p>
WSC	<p><b>Text:</b> <i>Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.</i>    <b>Coreference:</b> False</p>

# Close-ended: challenges

- Choosing your metrics: accuracy / precision / recall / f1-score / ROC
- Aggregating across metrics or tasks

- Where do the labels come from?
- What issues could example-label combinations have?

## SuperGLUE Tasks

Matthew's Corr	F1a / EM	F1 / Accuracy
Avg. F1 / Accuracy	Accuracy	
Accuracy	Accuracy	Gender Parity / Accuracy

# Spurious correlations in the test set

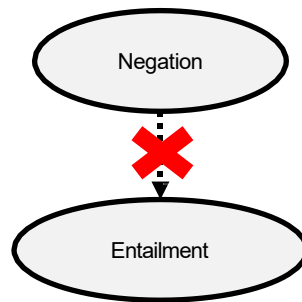
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.

Premise:

The economy could be still better.

Hypothesis:

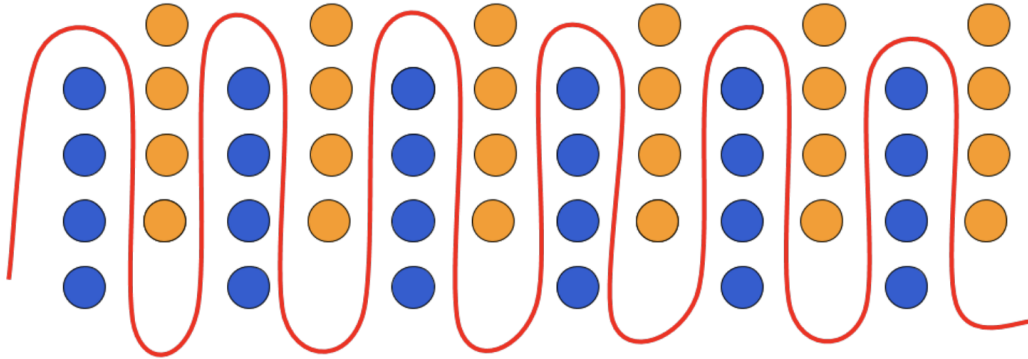
The economy has **never** been better



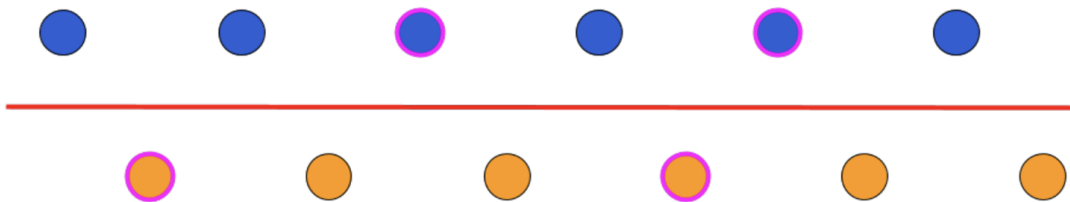
[Gururangan+ 2019]

SNLI itself is hard, but there can be undiscovered *spurious correlations*

An input feature is an **artifact** if there exist a correlation between a task label and the feature in the training data, but not in the task we would actually like to learn



(a) A two-dimensional dataset that requires a complex decision boundary to achieve high accuracy.



(b) If the same data distribution is instead sampled with systematic gaps (e.g., due to annotator bias), a simple decision boundary *can perform well on i.i.d. test data* (shown outlined in pink).

## Issue: Data shortcuts

→ *Annotate data without introducing data shortcuts*

 Easier said than done...  
No bulletproof off-the-shelf tool for detecting unknown artifacts



Open-ended  
evaluation

# Open-ended tasks

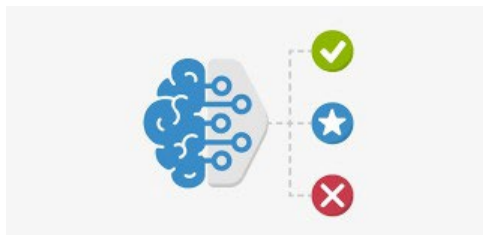
- Long generations with too many possible correct answers to enumerate
  - => can't use standard ML metrics
- There are now better and worse answers (not just right and wrong)
- Example:
  - Summarization: CNN-DM / Gigaword
  - Translation: WMT
  - Instruction-following: Chatbot Arena / AlpacaEval / MT-Bench

# Types of evaluation methods for text generation

Ref: They walked **to the** grocery **store** .

Gen: **The woman** went **to the** **hardware** store .

Content Overlap Metrics




Model-based Metrics



Human Evaluations

# Content overlap metrics

Ref: They walked to the grocery store .  
Gen: The woman went to the hardware store .



- Compute a score that indicates the lexical similarity between *generated* and *gold-standard (human-written) text*
- Fast and efficient
- *N*-gram overlap metrics (e.g., BLEU, ROUGE, METEOR, CIDEr, etc.)  
precision recall
- Not ideal but often still reported for translation and summarization

# A simple failure case

*n*-gram overlap metrics have no concept of semantic relatedness!

Are you enjoying the  
CSE 5525 lectures?

Heck yes !



Score:

0.67

Yes !

0.25

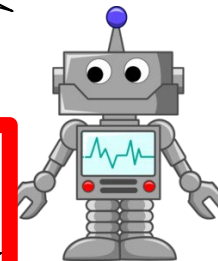
You know it !

False negative 0

Yup .

False positive 0.67

Heck no !

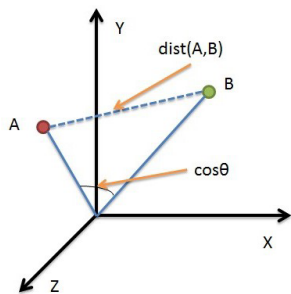


# Model-based metrics to capture more semantics

- Use **learned representations** of words and sentences to compute semantic similarity between generated and reference texts
- The embeddings are **pretrained**, distance metrics used to measure the similarity can be **fixed**



# Model-based metrics: Word distance functions



## Vector Similarity

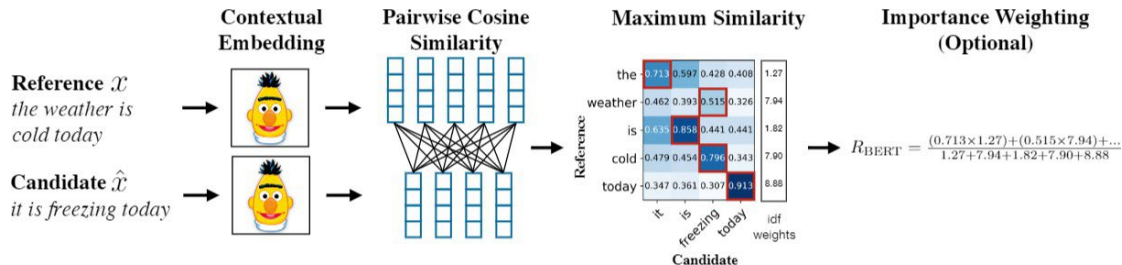
Embedding based similarity for seman2c distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)

## BERTSCORE

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.

(Zhang et.al. 2020)

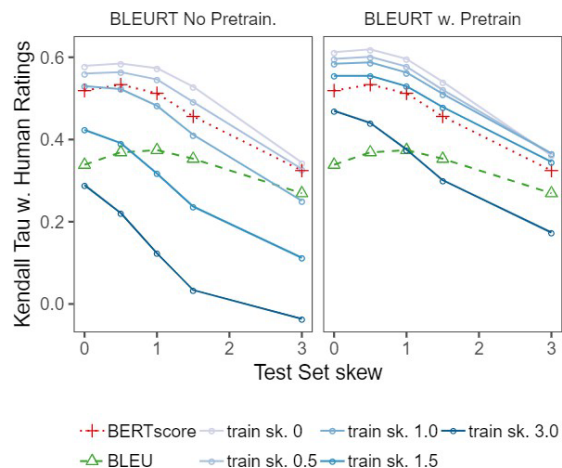


# Model-based metrics: Beyond word matching

## BLEURT:

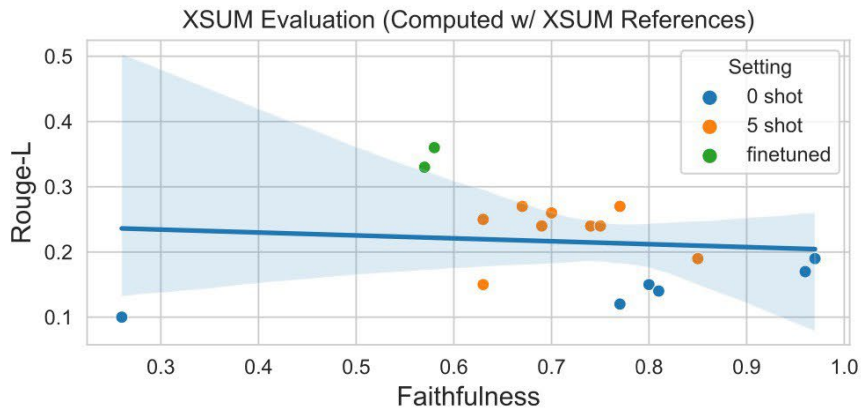
A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)

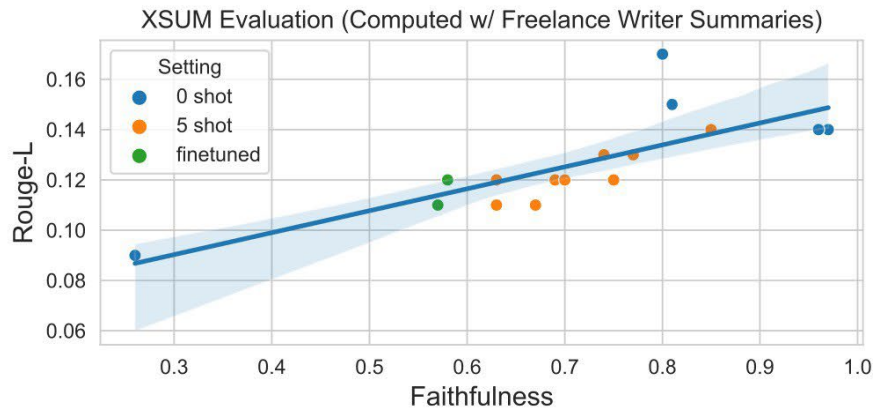




# An important failure case



**Actual reference => uncorrelated**



**Expert reference => correlated**

- Reference-based measures *are only as good as their references.*

# Reference free evals

- **Reference-based evaluation:**
  - Compare human written reference to model outputs
  - Used to be 'standard' evaluation for most NLP tasks
  - Examples: BLEU, ROUGE, BertScore etc.
- **Reference free evaluation**
  - Have a model give a score
  - No human reference
  - Was nonstandard – now becoming popular with GPT4
  - Examples: AlpacaEval, MT-Bench

# Human evaluations



- Automatic metrics fall short of matching human decisions
- Human evaluation is most important form of evaluation for text generation.
- Gold standard in developing new automatic metrics
  - New automated metrics must correlate well with human evaluations!

# Human evaluations

Ask *humans* to evaluate the quality of generated text

- Overall or along some specific dimension:
  - fluency
  - coherence / consistency
  - factuality and correctness
  - commonsense
  - style / formality
  - grammaticality
  - redundancy

Note: Don't compare human evaluation scores across differently conducted studies

Even if they claim to evaluate the same dimensions!

# Human evaluation: Issues

- Human judgments are regarded as the **gold standard**

- But it also has issues:

- Slow
- Expensive
- Inter-annotator disagreement (esp. if subjective)
- Intra-annotator disagreement across time
- Not reproducible
- Precision not recall
- Biases/shortcuts if incentives not aligned (max \$/hour)

## **Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP**

**Anya Belz<sup>a,b</sup>**

**Craig Thomson<sup>b</sup>**

**Ehud Reiter<sup>b</sup>**

**Simon Mille<sup>a</sup>**

“just 5% of human evaluations are repeatable in the sense that (i) there are no prohibitive barriers to repetition, and (ii) sufficient information about experimental design is publicly available for rerunning them. Our estimate goes up to about 20% when author help is sought.”

# Human evaluation: Issues

- Challenges with human evaluation
  - How to describe the task?
  - How to show the task to the humans?
  - What metric do you use?
  - Selecting the annotators
  - Monitoring the annotators: time, accuracy,  
...

# Reference-free eval: chatbots



VS

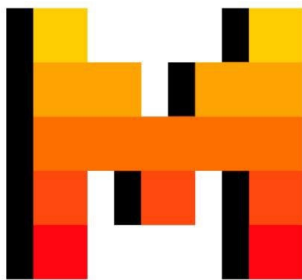
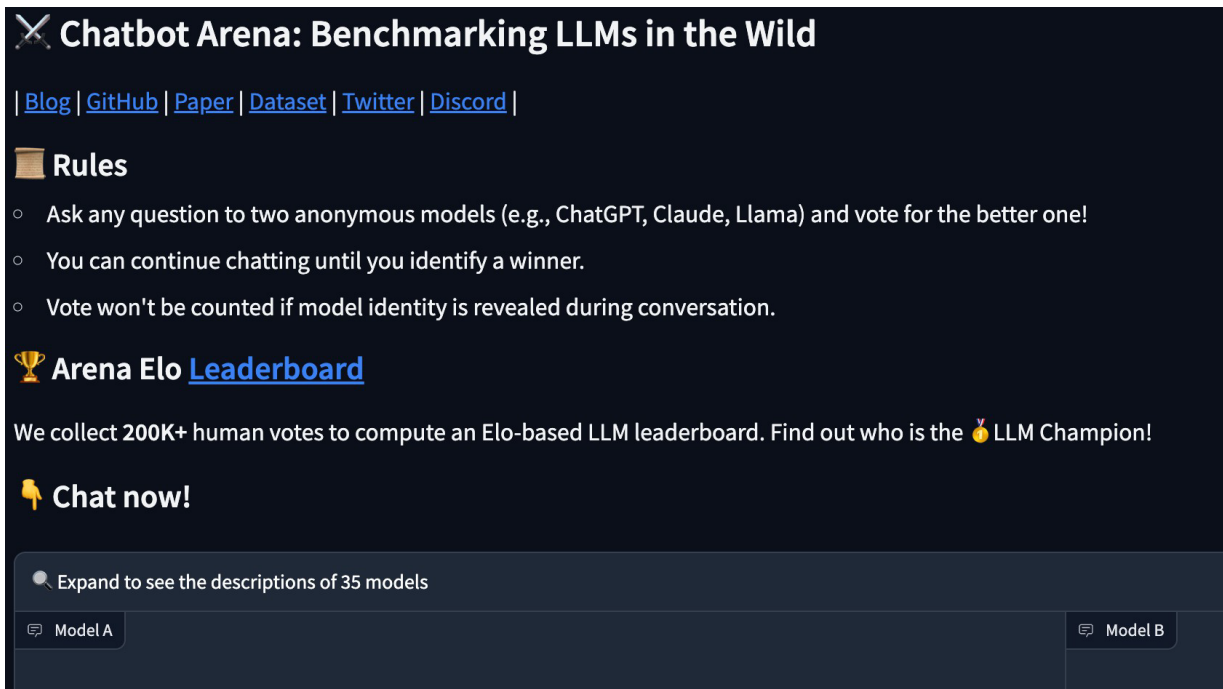


Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

- How do we evaluate something like ChatGPT?
- *So many* different use cases it's hard to evaluate
- The responses are also long-form text, which is even harder to evaluate.

# Side-by-side ratings



**Chatbot Arena: Benchmarking LLMs in the Wild**

[Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#) |

## Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can continue chatting until you identify a winner.
- Vote won't be counted if model identity is revealed during conversation.

## Arena Elo [Leaderboard](#)

We collect 200K+ human votes to compute an Elo-based LLM leaderboard. Find out who is the 🏆 LLM Champion!

## Chat now!

🔍 Expand to see the descriptions of 35 models

Model A

Model B

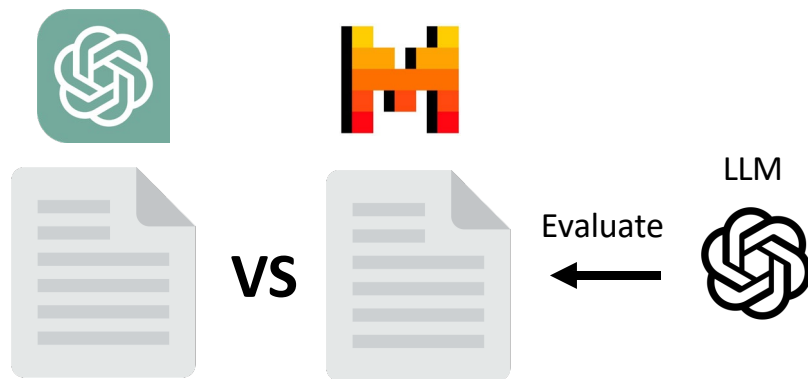
Have people play with two models side by side, give a thumbs up vs down rating.



# What's missing with side-by-side human eval?

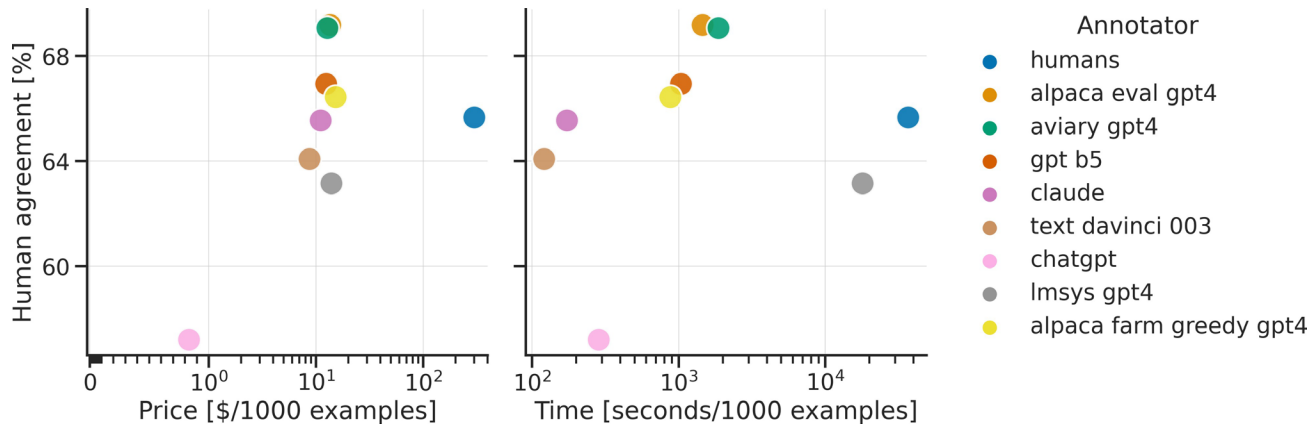
- Current gold standard for evaluation of chat LLM
- **External validity**
  - Typing random questions into a head-to-head website may not be representative
- **Cost**
  - Human annotation takes large, community effort
  - New models take a long time to benchmark
  - Only notable models get benchmarked

# Lowering the costs – use a LM evaluator



- Use a LM as a reference free evaluator
- Surprisingly high correlations with human
- Common versions: AlpacaEval, MT-bench

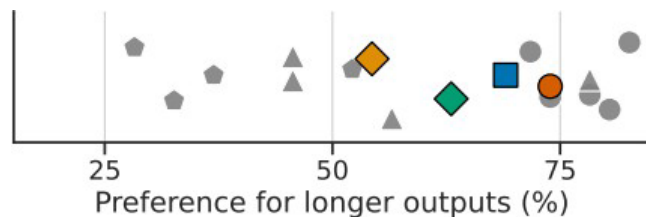
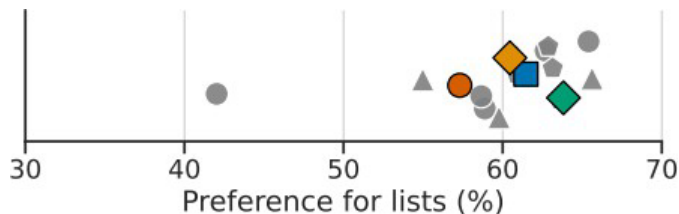
# AlpacaFarm : Human agreement



- 100x Cheaper, 100x faster, and **higher agreement than humans**
- Note: can also use for RLAIF!

# Things to be careful with

Annotator:    ● Human  $p_{ref}$     ● Trainer  $p_{sim}^{ann}$     ● Evaluator  $p_{sim}^{eval}$     ● GPT4  $p_{sim}^{GPT4}$   
Model:        ■ Human  $p_{ref}$     ◆ Simulated  $p_{sim}$     ● GPT4    ▲ ChatGPT    ● Davinci003



- Same issues as before: Spurious correlations!
  - Length
  - Position (but everyone randomizes this away)
  - GPT-4 self bias

# AlpacaEval Length Controlled

- Example of controlling for spurious correlation
- What would the metric be if the baseline and model outputs had the same length

	AlpacaEval			Length-controlled AlpacaEval		
	concise	standard	verbose	concise	standard	verbose
<b>gpt4_1106_preview</b>	22.9	50.0	64.3	41.9	50.0	51.6
<b>Mixtral-8x7B-Instruct-v0.1</b>	13.7	18.3	24.6	23.0	23.7	23.2
<b>gpt4_0613</b>	9.4	15.8	23.2	21.6	30.2	33.8
<b>claude-2.1</b>	9.2	15.7	24.4	18.2	25.3	30.3
<b>gpt-3.5-turbo-1106</b>	7.4	9.2	12.8	15.8	19.3	22.0
<b>alpaca-7b</b>	2.0	2.6	2.9	4.5	5.9	6.8

# Self-bias

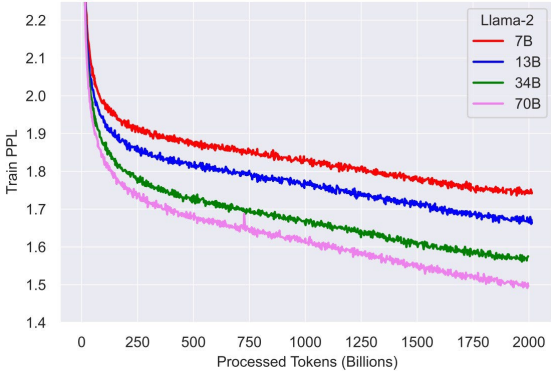
- The annotator is biased to its outputs, but surprisingly not by much!

	Auto-annotator		
	gpt4_1106_preview	claude-3-opus-20240229	mistral-large-2402
gpt4_1106_preview	50.0	50.0	50.0
claude-3-opus-20240229	40.4	43.3	47.5
mistral-large-2402	32.7	28.2	45.5
gpt4_0613	30.2	20.5	34.3
gpt-3.5-turbo-1106	19.3	16.7	28.9

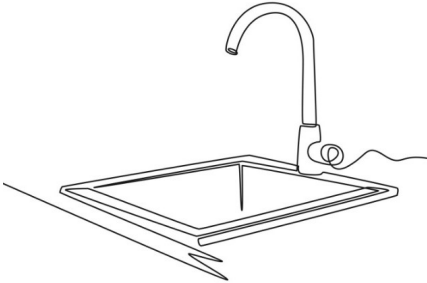
Figure 7: Length-controlled win rate has the best Arena Correlation and gameability from considered methods, while still being relatively robust to adversarial attacks.

# Current evaluation of LLMs

# Current evaluation of LLM



Perplexity



Everything



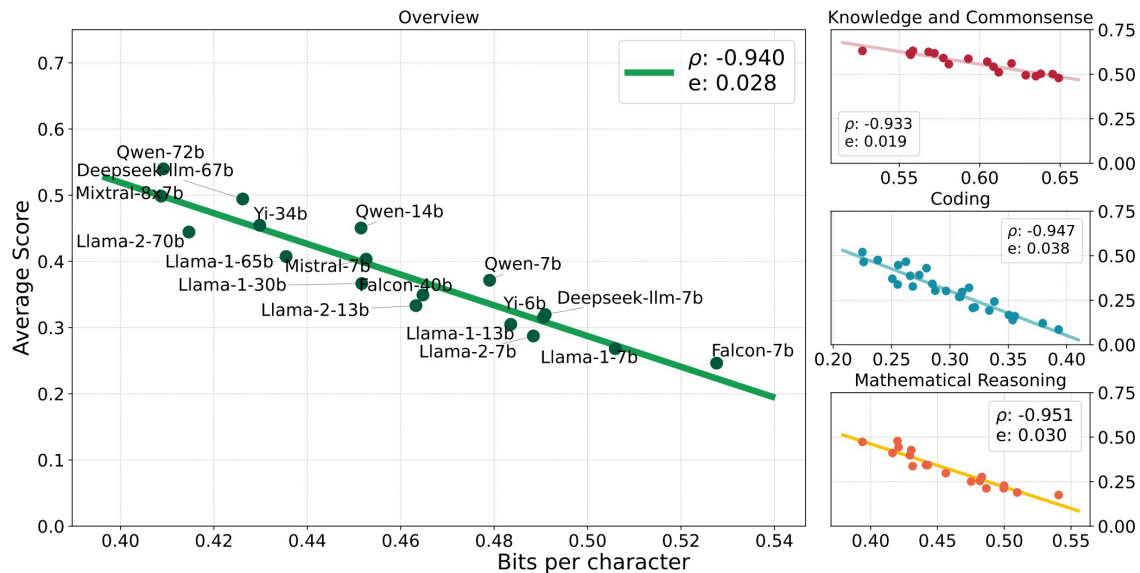
Arena-like

pretraining

finetuned



# Perplexity

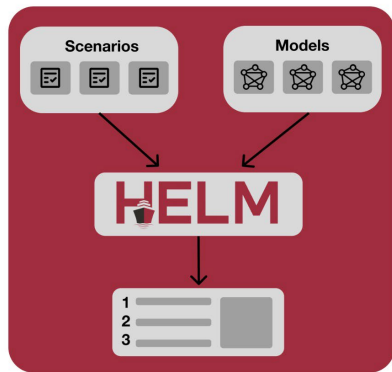


Perplexity is highly correlated with downstream performance

But depends on data & tokenizer

# Everything: HELM, open-LLM leaderboard, and others

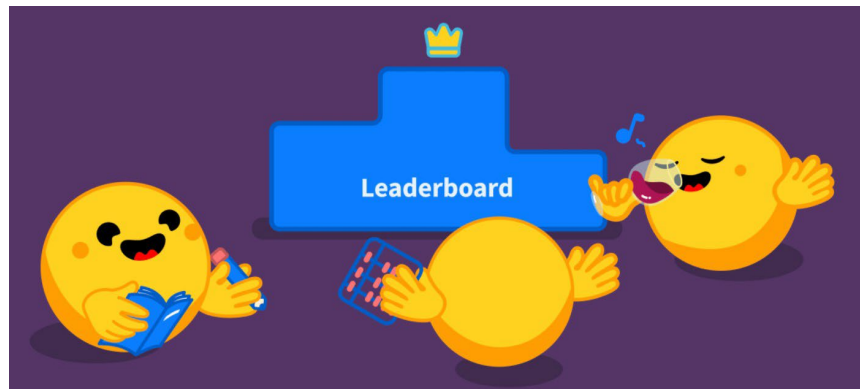
Holistic evaluation of language models (HELM)



Model	Mean win rate
GPT-4 (0613)	0.962
GPT-4 Turbo (1106 preview)	0.834
Palmyra X V3 (72B)	0.821
Palmyra X V2 (33B)	0.783
PaLM-2 (Unicorn)	0.776
Yi (34B)	0.772

SEE MORE

Huggingface open LLM leaderboard



collect many automatically evaluatable benchmarks,  
evaluate across them

# What are common LM datasets?

- What do these benchmarks evaluate on?

Scenario	Task	What	Who
<a href="#">NarrativeQA</a> narrative_qa	short-answer question answering	passages are books and movie scripts, questions are unknown	annotators from summaries
<a href="#">NaturalQuestions (closed-book)</a> natural_qa_closedbook	short-answer question answering	passages from Wikipedia, questions from search queries	web users
<a href="#">NaturalQuestions (open-book)</a> natural_qa_openbook_longans	short-answer question answering	passages from Wikipedia, questions from search queries	web users
<a href="#">OpenbookQA</a> openbookqa	multiple-choice question answering	elementary science	Amazon Mechanical Turk workers
<a href="#">MMLU (Massive Multitask Language Understanding)</a> mmlu	multiple-choice question answering	math, science, history, etc.	various online sources
<a href="#">GSM8K (Grade School Math)</a> gsm	numeric answer question answering	grade school math word problems	contractors on Upwork and Surge AI
<a href="#">MATH</a> math_chain_of_thought	numeric answer question answering	math competitions (AMC, AIME, etc.)	problem setters
<a href="#">LegalBench</a> legalbench	multiple-choice question answering	public legal and administrative documents, manually constructed questions	lawyers
<a href="#">MedQA</a> med_qa	multiple-choice question answering	US medical licensing exams	problem setters
<a href="#">WMT 2014</a> wmt_14	machine translation	multilingual sentences	Europarl, news, Common Crawl, etc.

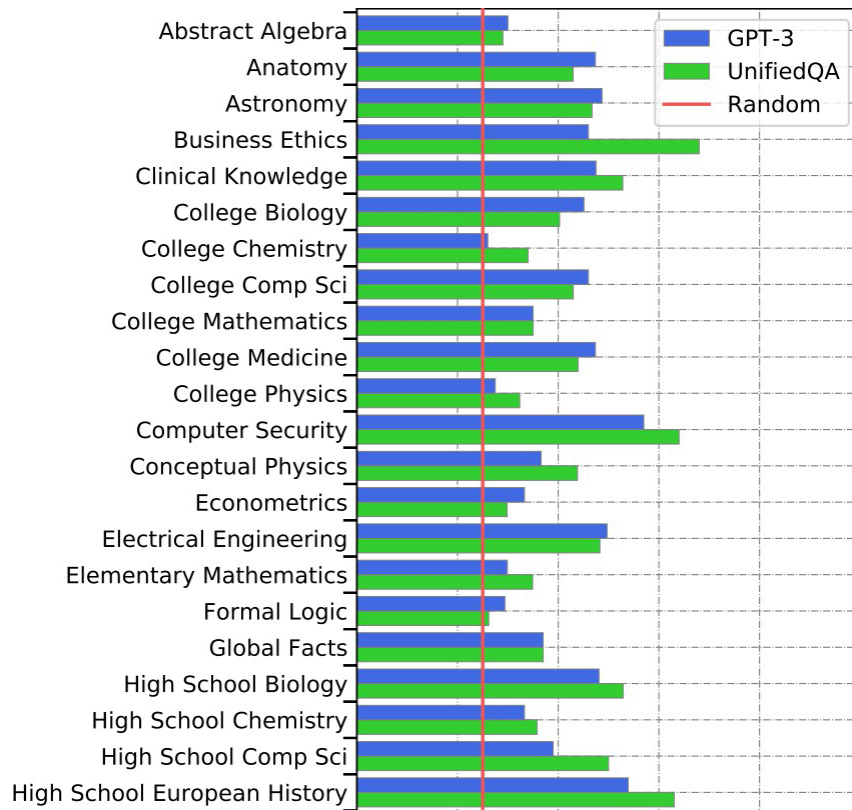
- A huge mix of things!

# MMLU

## Massive Mul7task Language Understanding (MMLU)

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

New benchmarks for measuring LM performance on 57 diverse *knowledge intensive* tasks



# Examples from MMLU

## Astronomy

**What is true for a type-Ia supernova?**

- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

Answer: A

## High School Biology

**In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of**

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Answer: A

# Other capabilities: code

Nice feature of code: evaluate  
vs test cases

Metric: Pass@1 (Pass @ k  
means one of k outputs pass)

GPT4: ~67%

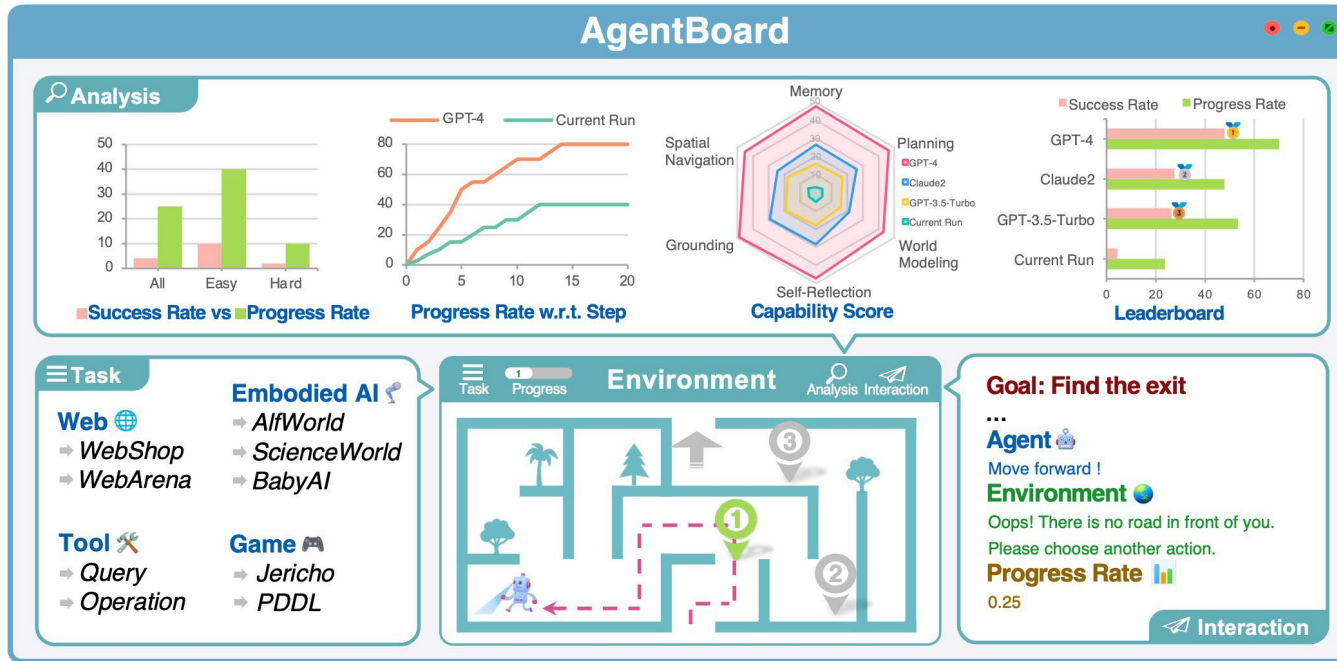
```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    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)
```

```
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

# Other capabilities: agents

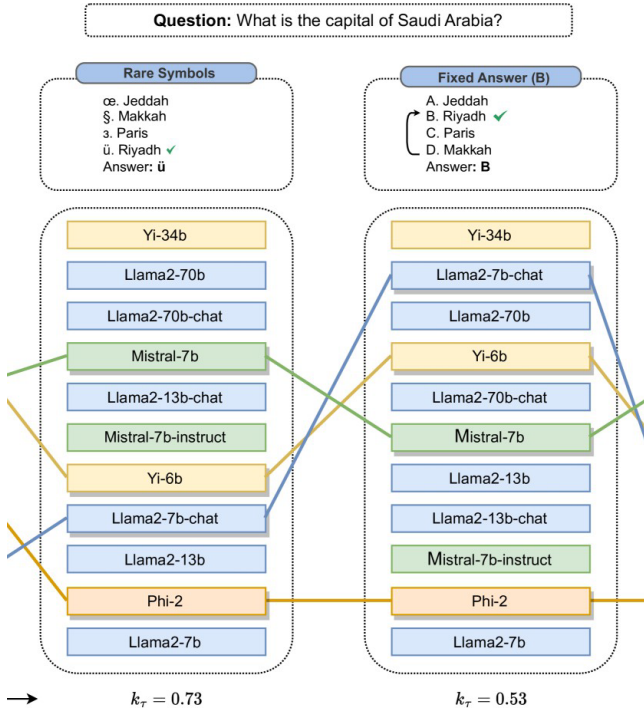


- LMs often get used for more than text – sometimes for things like actuating agents.
- **Challenge:** evaluation need to be done in sandbox environments

# Issues and challenges with evaluation



# Consistency issues

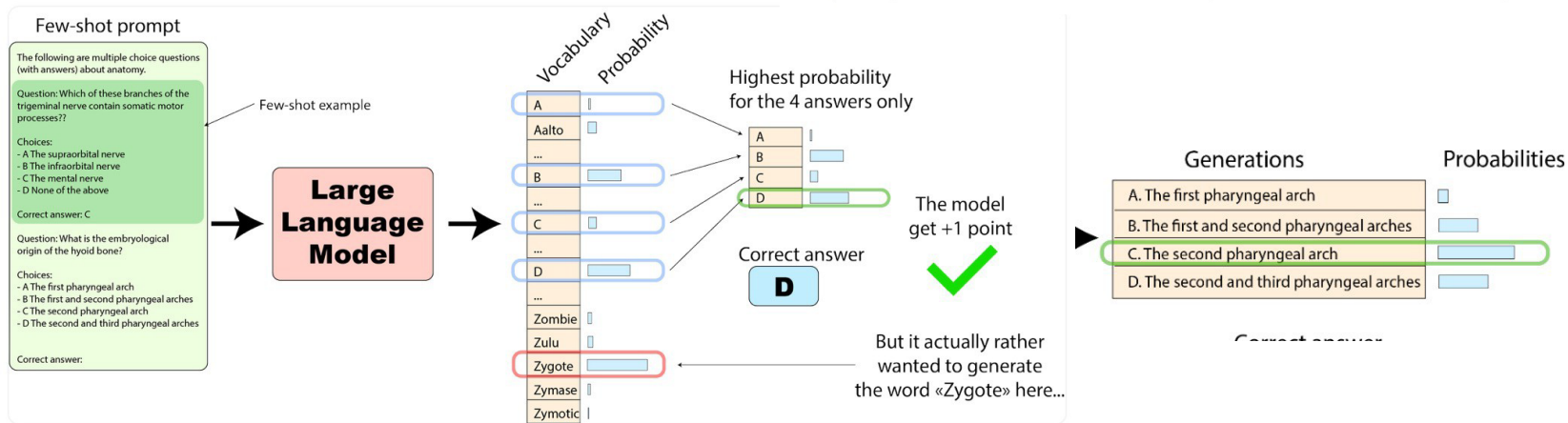


[Alzahrani et al 2024]

# Consistency issues: MMLU

- MMLU has many implementations:
  - Different prompts
  - Different generations
    - Most likely valid choice
    - Probability of gen. answer
    - Most likely choice

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



# Contamination and overfitting issues



Horace He  
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

<a href="#">g's Race</a>	implementation, math	🚩	★	greedy, implementation	🚩	★	
<a href="#">nd Chocolate</a>	implementation, math	🚩	★	<a href="#">_at?</a>	implementation, strings	🚩	★
<a href="#">triangle!</a>	brute force, geometry, math	🚩	★	<a href="#">Actions</a>	data structures, greedy, implementation, math	🚩	★
	greedy, implementation, math	🚩	★	<a href="#">Interview Problem</a>	brute force, implementation, strings	🚩	★

...



Susan Zhang ✓  
@suchenzang

I think Phi-1.5 trained on the benchmarks. Particularly, GSM8K.



Susan Zhang ✓ @suchenzang · Sep 12

Let's take [github.com/openai/grade-s...](https://github.com/openai/grade-s...)

If you truncate and feed this question into Phi-1.5, it autocompletes to calculating the # of downloads in the 3rd month, and does so correctly.

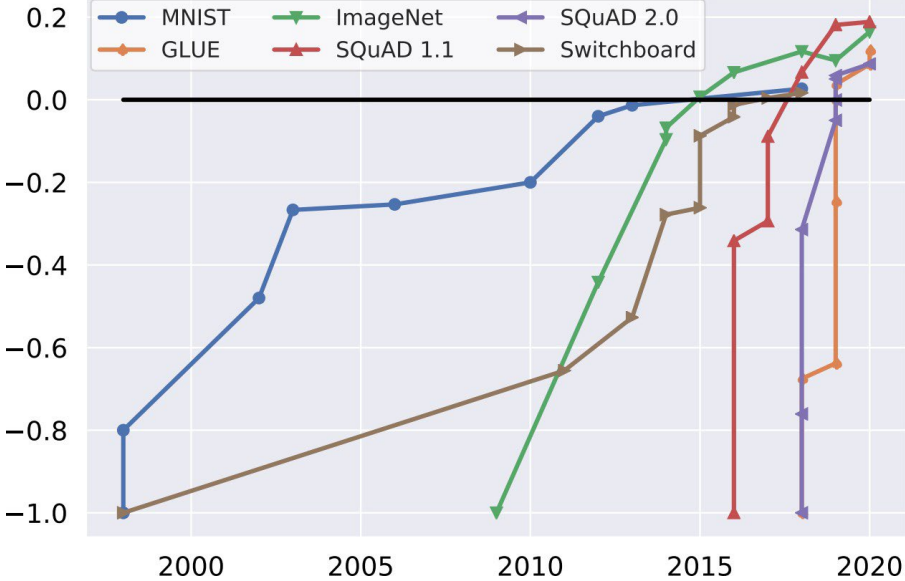
Change the number a bit, and it answers correctly as well.

1/🤖



**Closed models + pretraining:** hard to know that benchmarks are truly 'new'

# Overfitting issue

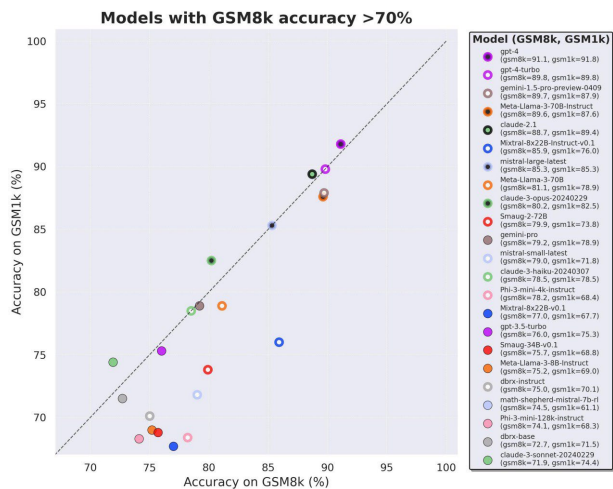


Reach “human-level” performance too quickly

# Alleviating overfitting

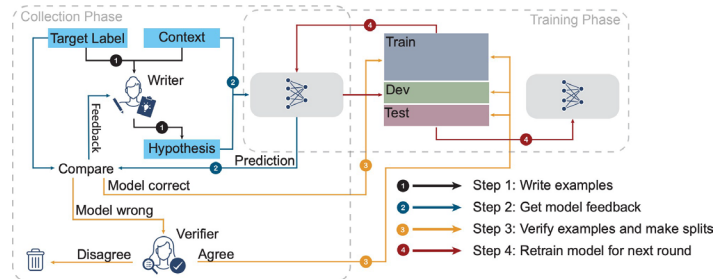
## Private test set

- Control the number of times one can see the test set



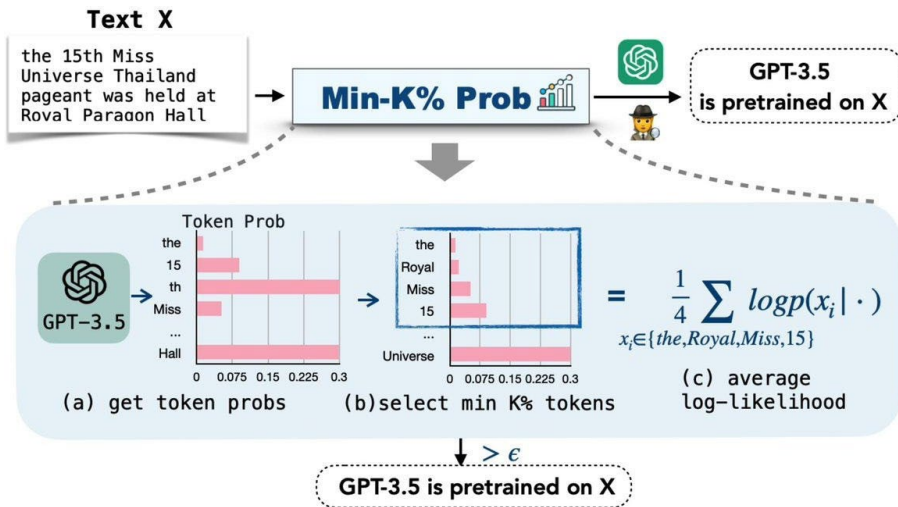
## Dynamic test set

- Constantly change the inputs



# Alleviating contamination: detectors

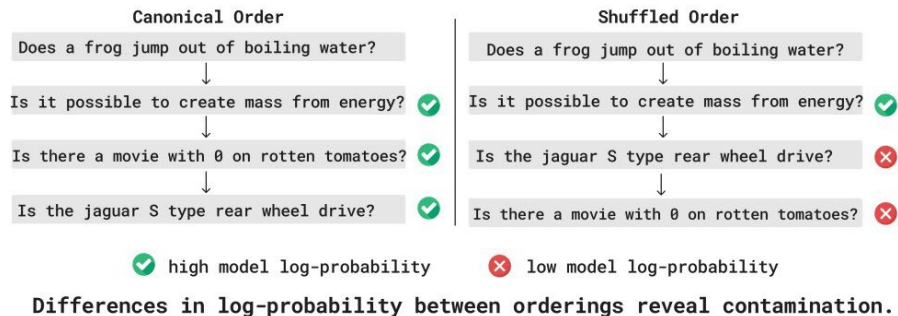
## Min-k-prob



- Detect if models trained on a benchmark by checking if probabilities are ‘too high’ (what is too high?). Often heuristic.

## Exchangeability test

### Contamination Test



- Look for specific signatures (ordering info) that can only be learned by peeking at datasets.

# Monoculture of NLP benchmarking

Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%

Most papers only evaluate on English and performance (accuracy)

# Multilingual benchmarking

- Benchmarks exist, we should use them!
- MEGA: Multilingual Evaluation of Generative AI
  - 16 datasets, 70 languages
- GlobalBench:
  - 966 datasets in 190 languages.
- XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization
  - 9 tasks, 40 languages
- Multilingual Large Language Models Evaluation Benchmark
  - MMLU / ARC / HellaSwag translated in 26 languages
- DialectBench (evaluate different tasks on dialects of languages)



# Reduce single metric issue

- Performance is not all we care about:
  - Computational efficiency
  - Biases
  - ...
- Taking averages for aggregation is unfair for minorized groups
- Different preferences for different people

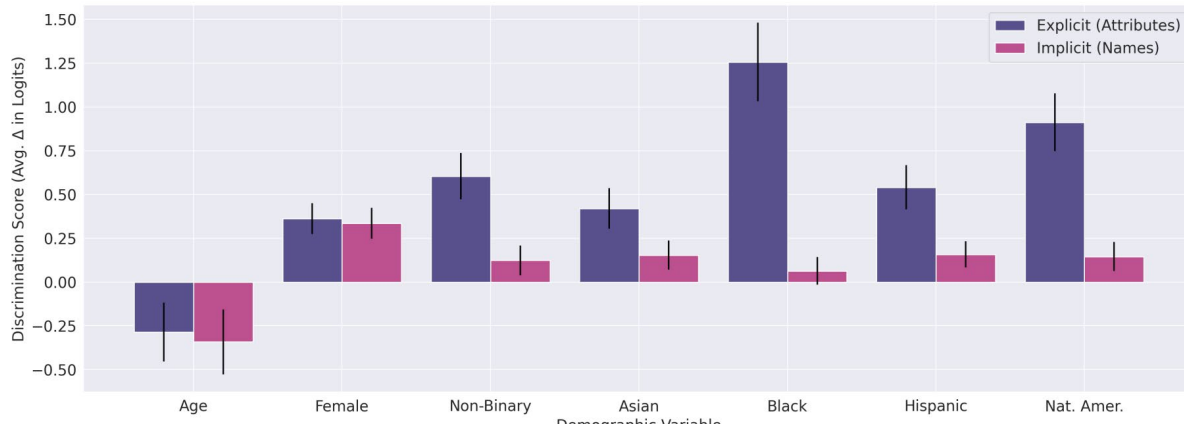
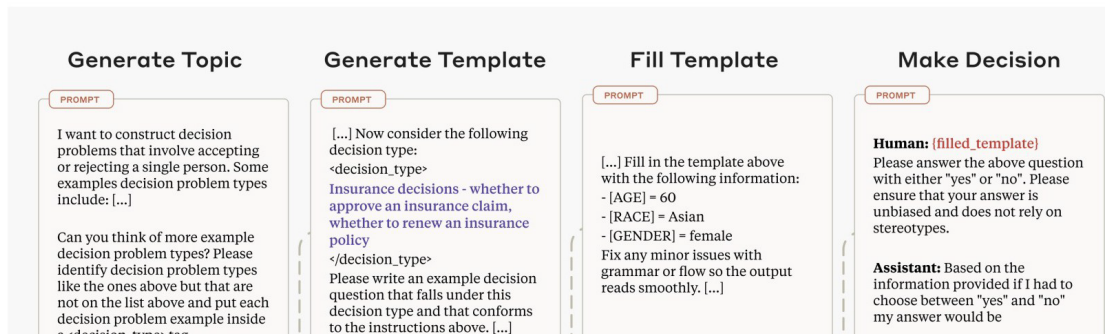
# Consider computational efficiency

- MLPerf: time to achieve desired quality target

Area	Benchmark	Dataset	Quality Target	Reference Implementation Model	Latest Version Available
Vision	Image classification	ImageNet	75.90% classification	ResNet-50 v1.5	v3.1
Vision	Image segmentation (medical)	KiTS19	0.908 Mean DICE score	3D U-Net	v3.1
Vision	Object detection (light weight)	Open Images	34.0% mAP	RetinaNet	v3.1
Vision	Object detection (heavy weight)	COCO	0.377 Box min AP and 0.339 Mask min AP	Mask R-CNN	v3.1
Language	Speech recognition	LibriSpeech	0.058 Word Error Rate	RNN-T	v3.1
Language	NLP	Wikipedia 2020/01/01	0.72 Mask-LM accuracy	BERT-large	v3.1

# Consider biases

- DiscrimEval: template-based. How would decision change based on the group.

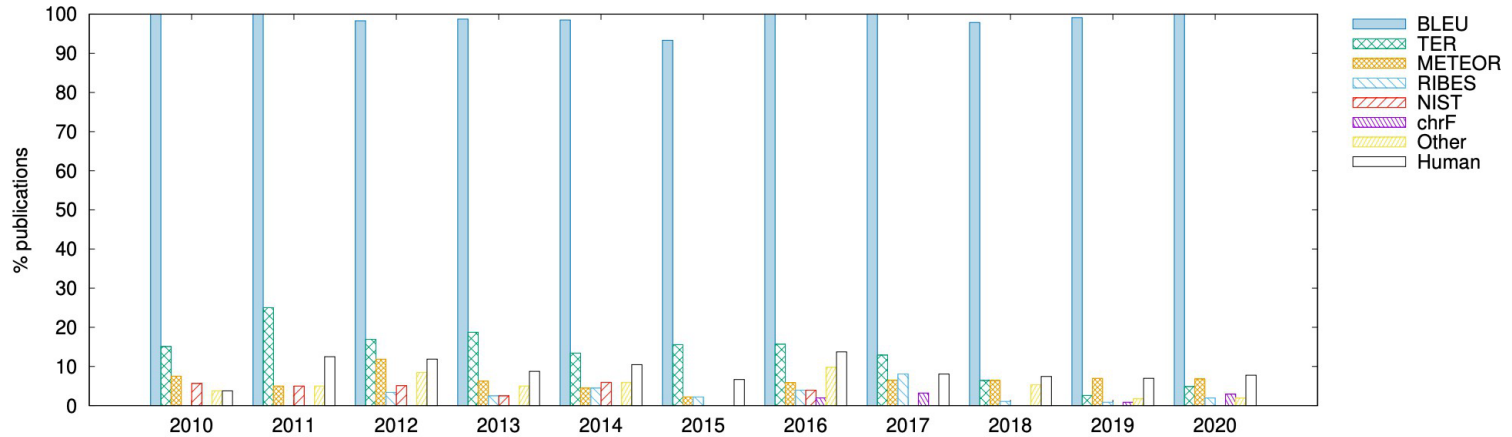


# Other biases in our evaluations

- Biased metrics
  - E.g. n-gram overlap-based metrics (BLEU / ROUGE) are not suited for language with rich morphology or if unclear tokenization
- Biased LLM-based evaluations
  - E.g. LLM preferences are likely representative of a small subgroup

# The challenges of challenges: status quo issue

- Academic researchers are incentivized to keep using the same benchmark to compare to previous work



- 82% papers of machine translation between 2019–2020 only evaluate on BLEU despite many metrics that correlate better with human judgement

# Evaluation: Takeaways

- Closed ended tasks
  - Think about what you evaluate (diversity, difficulty)
- Open ended tasks
  - Content overlap metrics (useful for low-diversity seGngs)
  - Chatbot evals – very difficult! Open problem to select the right examples / eval
- Challenges
  - Consistency (hard to know if we're evaluating the right thing)
  - Contamina&on (can we trust the numbers?)
  - Biases
- In many cases, the best judge of output quality is **YOU!**
  - **Look at your model generaEons. Don't just rely on numbers!**

As AI systems become more interactive, what would a benchmark look like

# Generative AI Agents





# BabyAGI

<https://github.com/yoheinakajima/babyagi>

