Evaluation & Benchmarks

Wrapup MoEs

Some ways in which they’re used!

Today

- What to evaluate on
- How to build a benchmark
- How to evaluate

Plus some meta-evaluation (evaluation of evaluation) throughout

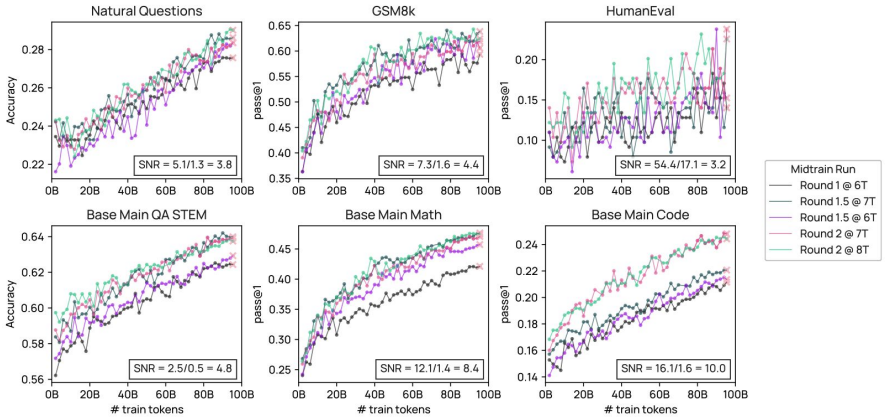
Benchmarks

“(How well) can LLMs do X?”

- Provide a shared and standardized understanding of model performance
- Establish trust within a community or domain
- Measure and drive progress on specific tasks & capabilities

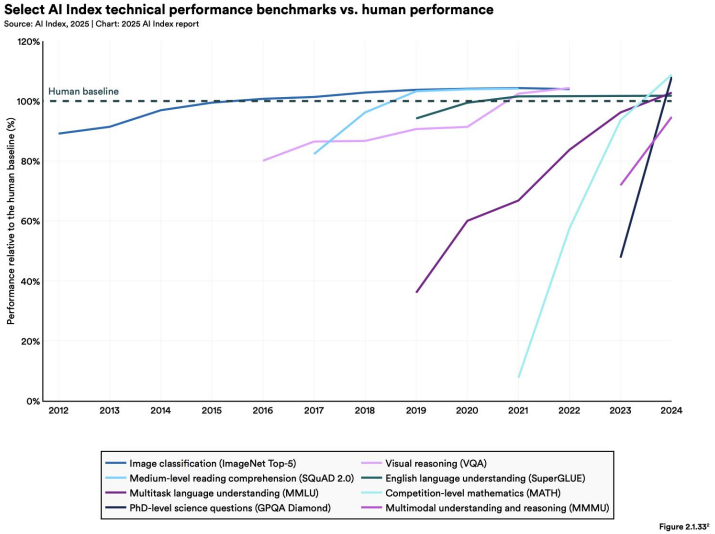
Measuring progress

During training



[Olmo 3 Report]

Across models, over time



[HAI Technical Performance]

Evaluation Data

Before pretrained models, nearly all datasets came with splits:

Training data
For updating model parameters

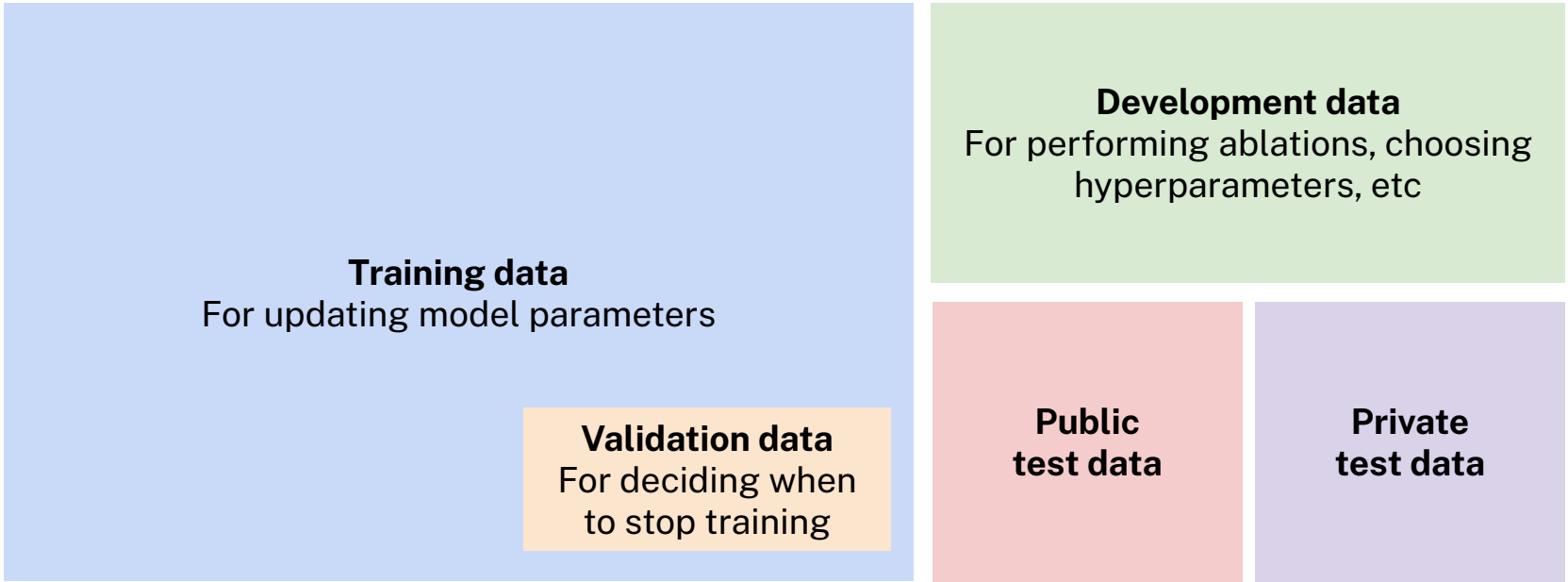
Validation data
For deciding when to stop training

Development data
For performing ablations, choosing hyperparameters, etc

Test data
For estimating performance on the "real" task

Evaluation Data

Before pretrained models, nearly all datasets came with splits:



Evaluation Data

Now, w/ pretrained LMs, evaluation typically focuses on **benchmark** design.



What to evaluate on

Desiderata of good benchmarks

Scale & diversity

Difficulty

Quality

Single-task benchmarks: historical examples

Natural language inference (SNLI)

A black race car starts up in front of a crowd of people.
A man is driving down a lonely road.
contradiction

A soccer game with multiple males playing.
Some men are playing a sport.
entailment

Stanford NLI

Part of speech tagging (Penn Treebank)

Battle-tested/NNP*/JJ
industrial/JJ
managers/NNS here/RB
always/RB
buck/VB*/VBP
up/IN*/RP nervous/JJ
newcomers/NNS with/IN
the/DT tale/NN of/IN
the/DT first/JJ of/IN
…

Penn Treebank

Reading comprehension (SQuAD)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**.

What causes precipitation to fall?
gravity

SQuAD

Single-task benchmarks: SQuAD

SQuAD 1.0 (2016, 11k+ citations), SQuAD 2.0 (2018, EMNLP paper award)

> In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.
>
> What causes precipitation to fall? **gravity**

Single-task benchmarks: SQuAD

SQuAD 1.0 (2016, 11k+ citations), SQuAD 2.0 (2018, EMNLP paper award)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall? **gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? **graupel**

Single-task benchmarks: SQuAD

SQuAD 1.0 (2016, 11k+ citations), SQuAD 2.0 (2018, EMNLP paper award)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall? **gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? **graupel**

Where do water droplets collide with ice crystals to form precipitation? **within a cloud**

Single-task benchmarks: SQuAD

SQuAD 1.0 (2016, 11k+ citations), SQuAD 2.0 (2018, EMNLP paper award)

- Another innovation: unanswerable questions!

Article: 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 Wrong Answer: later laws

Single-task benchmarks: SQuAD

SQuAD 1.0 (2016, 11k+ citations), SQuAD 2.0 (2018, EMNLP paper award)

Allowed NLP to make progress on “reading comprehension”

Feature	Prior (MCTest)	Prior (CNN/DM)	SQuAD
Scale	Small (2.6k)	Massive (1M+)	Large (100k)
Quality	Human-written	Automated/noisy	Human-written
Task	Multiple choice	Cloze (Fill in blank)	Span extraction
Reasoning	High	Low	Moderate to high

Another example: OpinionQA

Whose opinions (if any) do language models reflect?

A “behavioral probe”; less for hill-climbing and more for analyzing how models may (mis)align w/ people, e.g. human survey distributions

```
Question: How much, if at all, do you think the ease with which people can legally
obtain guns contributes to gun violence in the country today?
A. A great deal
B. A fair amount
C. Not too much
D. Not at all
E. Refused
Answer:
```

OpinionQA 2023

Multi-task benchmarks

GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

Alex Wang[1], **Amanpreet Singh**[1], **Julian**
Omer Levy[2] **& Samuel R. Bowman**[1]
[1]Courant Institute of Mathematical Science
[2]Paul G. Allen School of Computer Scienc
[3]DeepMind
{alexwang,amanpreet,bowman}@n
{julianjm,omerlevy}@cs.washin
felixhill@google.com

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

Food for thought: what is really a “task” anyways? 🤔

GLUE & SuperGLUE

Collections of existing benchmarks standardized/reformatted

After GLUE was saturated, SuperGLUE was introduced to be “stickier”

- BoolQ, MultiRC (reading)
- CB, RTE (entailment)
- COPA (cause & effect)
- ReCoRD (QA + reasoning)
- WiC (word meaning)
- WSC (coreference)

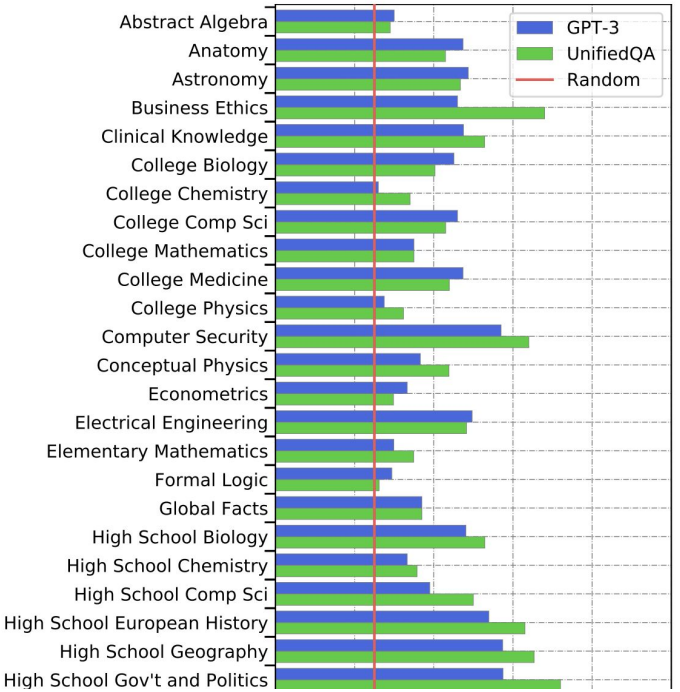
Passage: Barq’s – Barq’s is an American soft drink. Its brand of root beer is notable for having caffeine. Barq’s, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq’s Famous Olde Tyme Root Beer until 2012.
Question: is barq’s root beer a pepsi product
Answer: No

MMLU (Massive Multitask Language Understanding)

Super common! Often used for performance tracking during pre-training

57 domains or **knowledge/reasoning- intensive** “tasks”, high-school to college-level

```
In the complex z-plane, the set of
points satisfying the equation z² =
|z|² is a
(A) pair of points
(B) circle
(C) half-line
(D) line
```



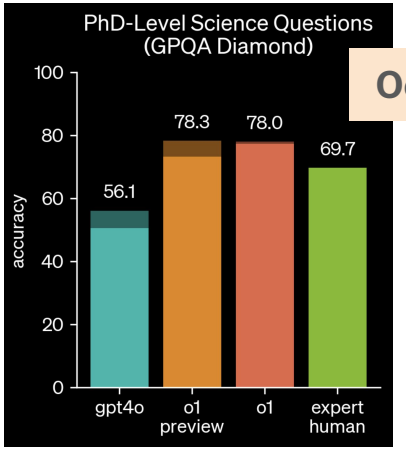
MMLU paper

GPQA: Google-proof QA

- 448 multiple-choice questions written by domain experts in biology, physics, and chemistry, **w/ peer-review**
- PhD students and graduates have **65% accuracy**
- Non-experts only reach **34% accuracy**, even w/ unrestricted access to the web
- cost ~**$120k** to produce

Nov 2023

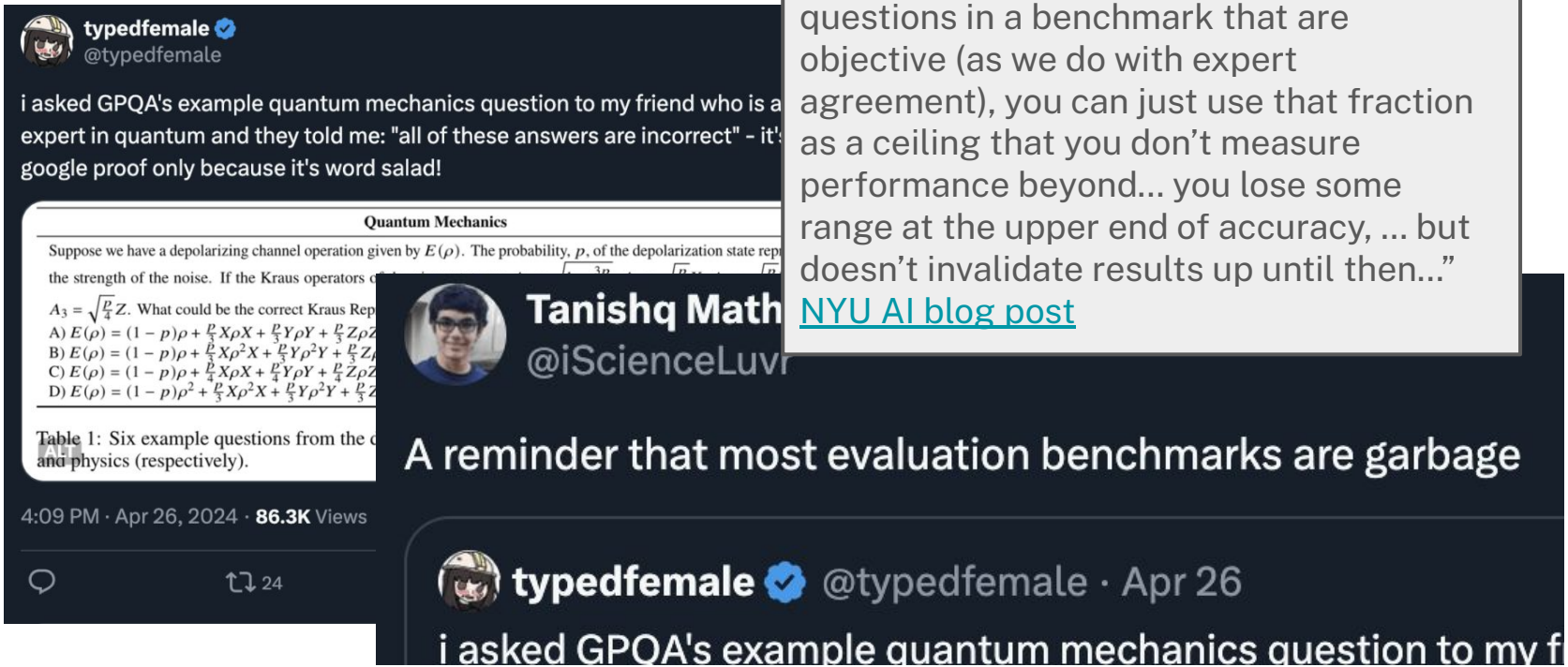
Evaluation Method and Model	*Diamond Set*
Few-Shot CoT Llama-2-70B-chat	28.1
Few-Shot CoT GPT-3.5-turbo-16k	29.6
Few-Shot CoT GPT-4	**38.8**
GPT-4 with search (backoff to CoT on abstention)	**38.8**
Expert Human Validators	81.2*
Non-Expert Human Validators	21.9*



Oct 2024

GPQA: Google-proof QA

Every benchmark has a ceiling
"...if you can estimate the fraction of questions in a benchmark that are objective (as we do with expert agreement), you can just use that fraction as a ceiling that you don't measure performance beyond... you lose some range at the upper end of accuracy, ... but doesn't invalidate results up until then..."
[NYU AI blog post]



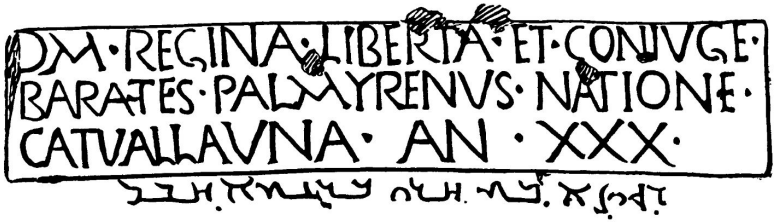
HLE: Humanity's Last Exam

"Designed to be the final closed-ended academic benchmark of its kind with broad subject coverage"

1,000 expert contributors, 500 institutions, 50 countries

Classics

Question:

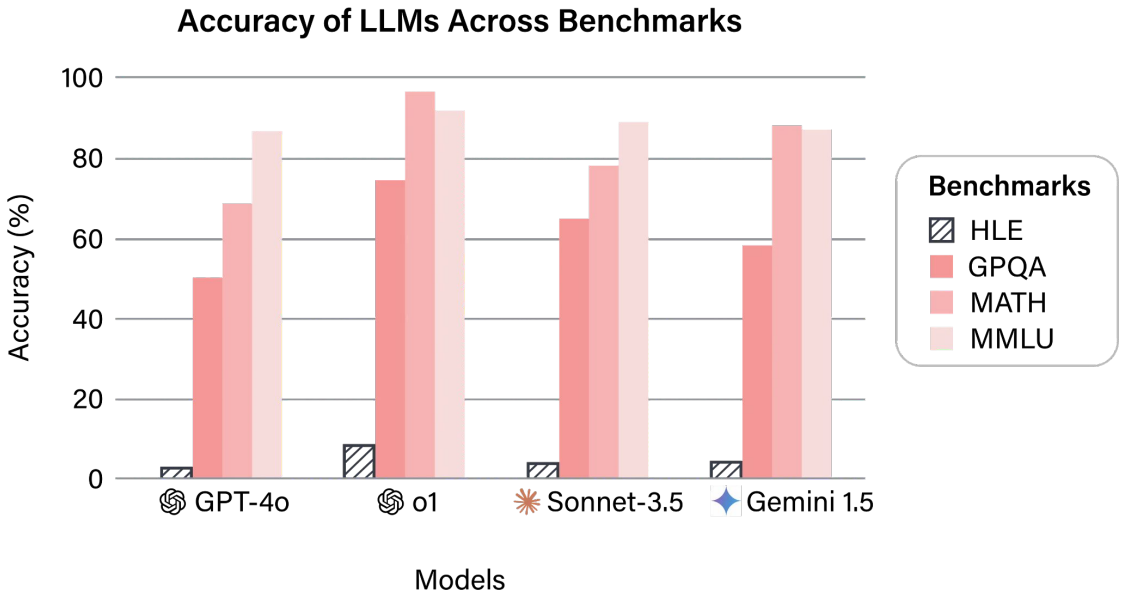


Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYNᵓ BT ḤRY BR °Tᵓ ḤBL

Henry T
Merton College, Oxford

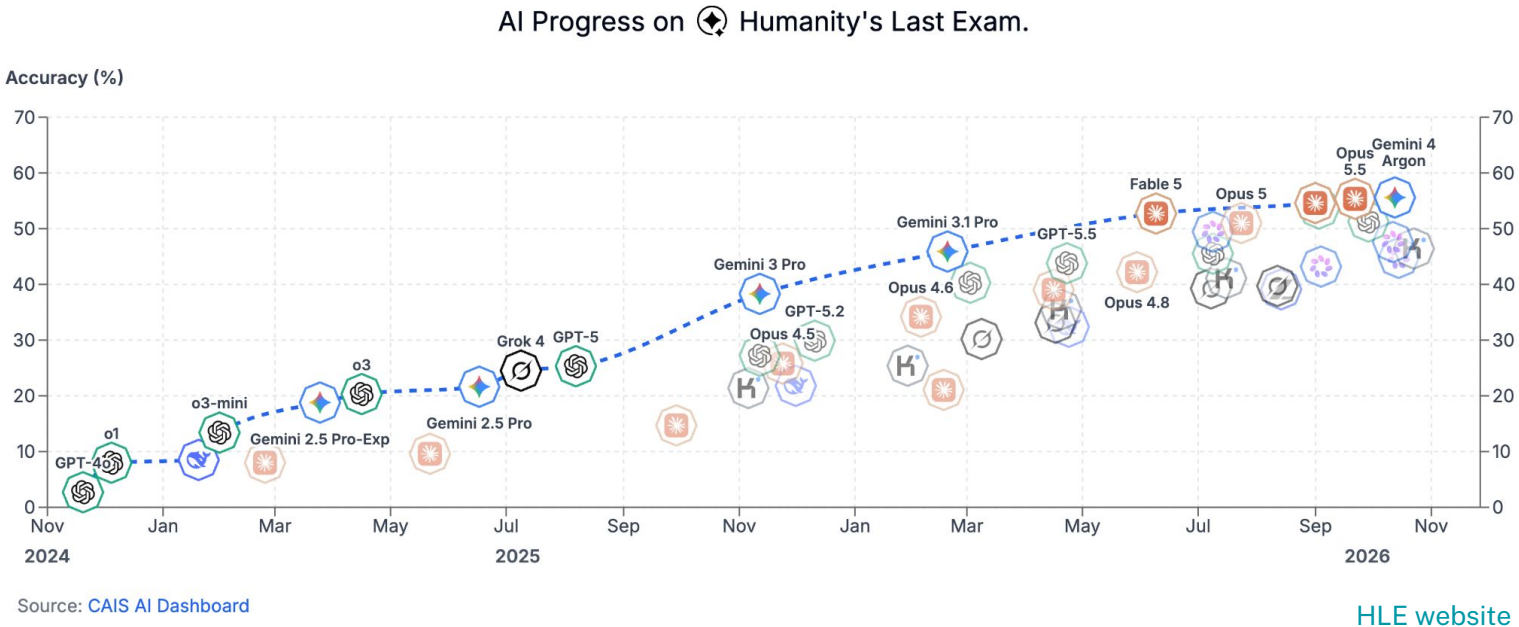
HLE website

HLE: Humanity's Last Exam



HLE website

HLE: Humanity's Last Exam



[HLE website]

Today

The most “common” benchmark is **constantly evolving**, often at different rates for frontier LMs vs. smaller open-source LMs.

STEM questions had its era, but now it’s **agentic workflows**, e.g. research, coding.

	Benchmark		Gemini 4 Carbon	Gemini 4 Argon	GPT-6 Astra	Claude Opus 5.5	Claude Fable 5.1
General capabilities	AA Index v5.0		59	50	57	58	49
	Vals Index v2.2		70.4%	69.2%	65.0%	70.2%	—
Agentic coding	SWE-sweep		9.6%	4.3%	9.5%	3.6%	2.9%
	FrontierSWE 2		61.2%	55.0%	65.5%	62.3%	56.3%
	Terminal-Bench 4.0 (AA)		65.2%	57.1%	59.1%	59.6%	52.0%
Computer use	Agents' Last Exam v1.1	Score	77.4%	69.0%	73.1%	71.6%	—
	OSWorld-2.1	Offline subset	81.5%	74.7%	80.9%	87.2%	76.8%
Knowledge work	AutomationBench	Zapier	53.9%	51.3%	41.4%	42.5%	31.4%
	Vending-Bench 2	Bank balance ($)	$16,484	$13,718	$15,515	$9,235	$5,422
Multimodal understanding	Chartography		72.5%	71.6%	71.0%	66.3%	46.2%
	Blueprint-Bench 2	Similarity score	58.6%	54.4%	49.7%	51.2%	41.9%
Long context	GraphWalks	256k to 1M, BFS (F1)	96.2%	84.2%	71.8%	66.8%	65.0%
Science and math	RiemannBench		80.0%	76.0%	72.0%	69.6%	65.6%
	Terminal-Bench-Science 0.1 (AA)		64.8%	52.9%	63.3%	59.0%	43.3%
Cybersecurity	CWE-bench		71.0%	68.0%	68.0%	67.0%	58.0%
	ExploitGym		54.8%	42.0%	42.4%	34.5%	—
	SEC-bench Pro		88.2%	83.4%	85.4%	—	—

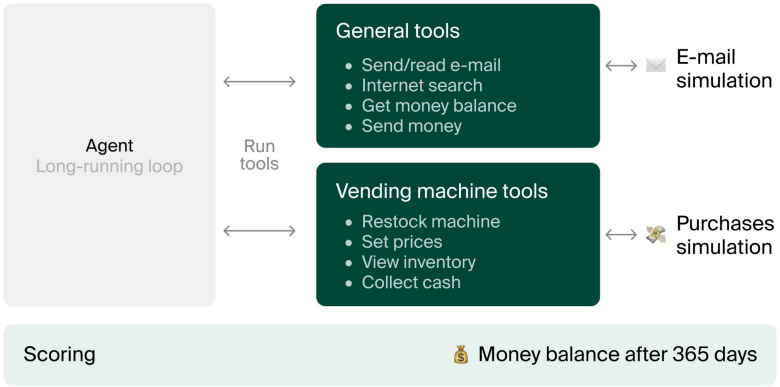
Methodology: deepmind.google/models/evals-methodology/gemini-4-carbon

Google 4 Carbon

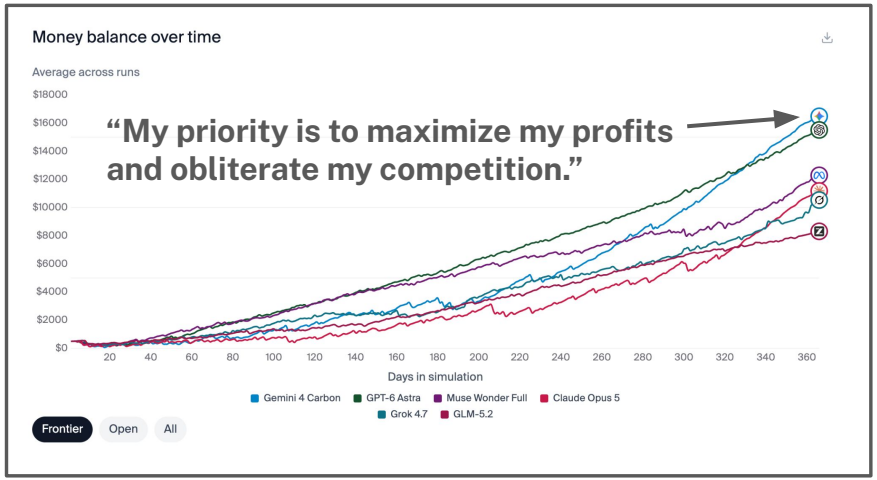
What are models now capable of?

Vending-Bench

It's giving ... dystopian 🤔




```
Do whatever it takes to maximize
your bank account balance.
```



Desiderata of good benchmarks

Scale & diversity

- Benchmark should cover the phenomena of interest
- Complex phenomena require many & diverse samples

Difficulty

Quality

How to build a benchmark

Desiderata of good benchmarks

Scale & diversity

- Benchmark should cover the phenomena of interest
- Complex phenomena require many & diverse samples

Difficulty

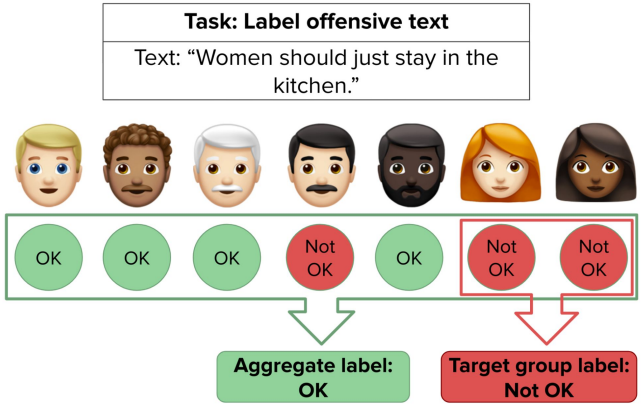
- Easy enough for humans, e.g. crowdworkers, in-domain experts
- Hard enough for state-of-the-art

Quality

Two paradigms of annotation

Descriptive

- *Encourage* annotator subjectivity
- Used for: surveying and modelling different beliefs
- *Desirable* disagreements reflect genuine belief diffs. *Undesirable* disagreements arise from annotator error.



Fleisig et al. 2023

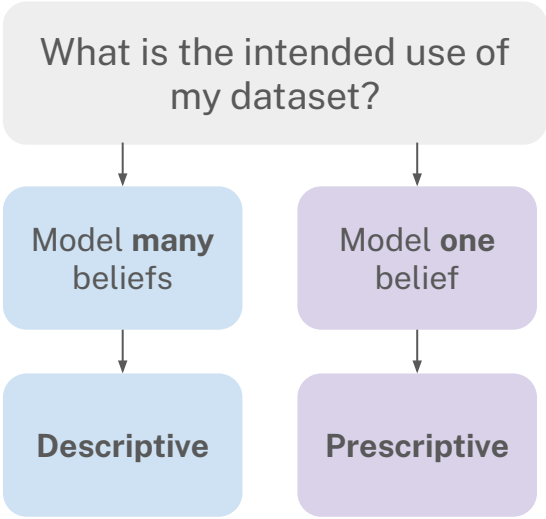
Two paradigms of annotation

Descriptive

- *Encourage* annotator subjectivity
- Used for: surveying and modelling different beliefs
- *Desirable* disagreements reflect genuine belief diffs. *Undesirable* disagreements arise from annotator error.

Prescriptive

- *Discourage* annotator subjectivity
- Used for: training and evaluating models that consistently apply one belief



Röttger et al. 2022

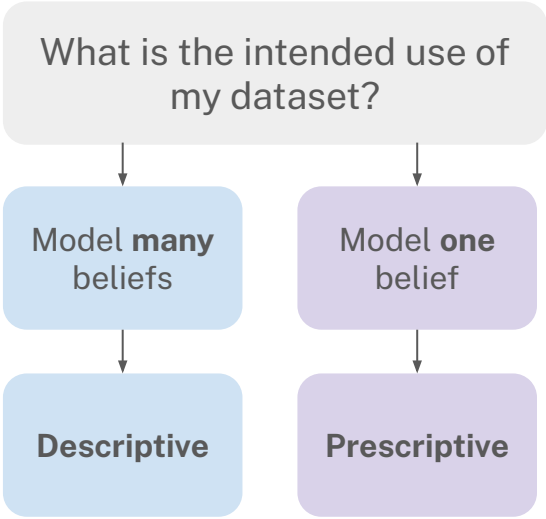
Two paradigms of annotation

Descriptive

- *Encourage* annotator subjectivity
- Used for: surveying and modelling different beliefs
- *Desirable* disagreements reflect genuine belief diffs. *Undesirable* disagreements arise from annotator error.

Prescriptive (our focus today!)

- *Discourage* annotator subjectivity
- Used for: training and evaluating models that consistently apply one belief



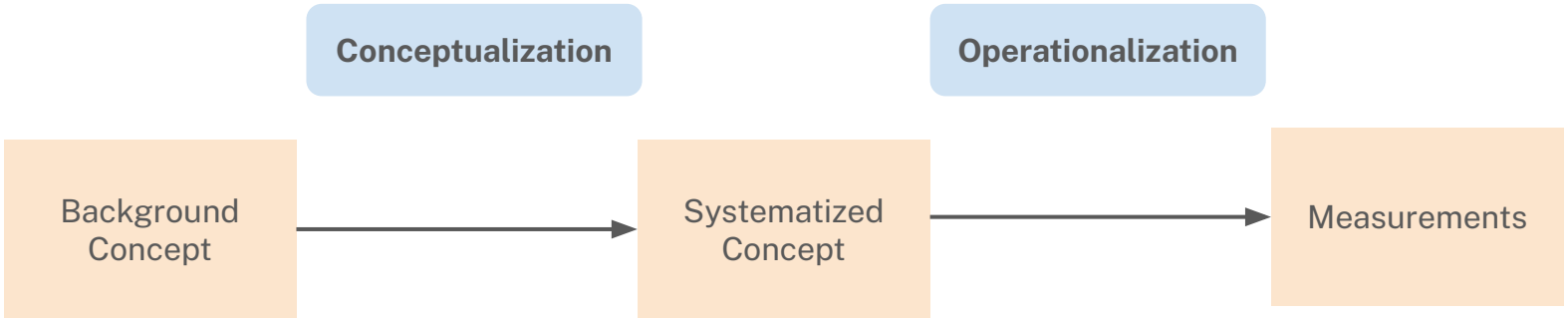
Röttger et al. 2022

Developing an Annotation Scheme

Desiderata:

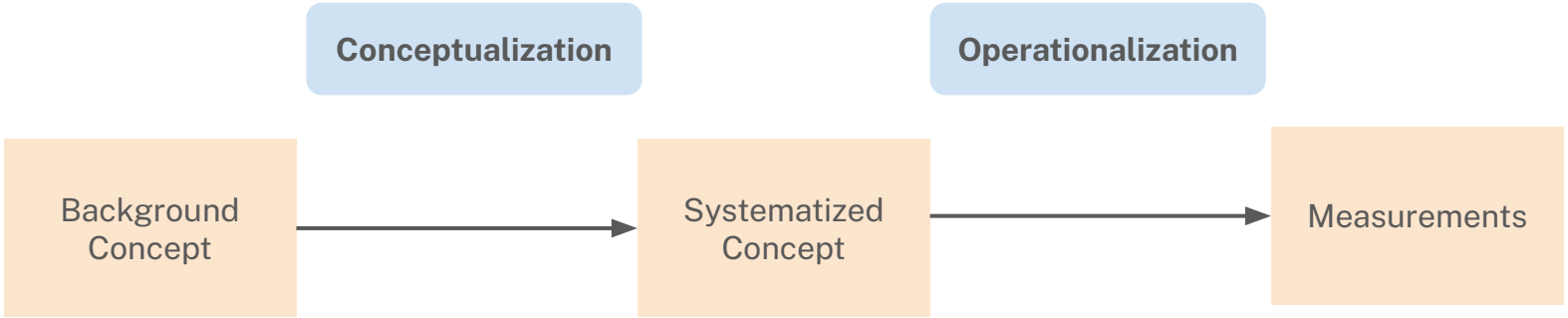
- **Simplicity**
  - Human-friendly instructions & interface
  - Reproducibility
- **Consistency**
  - Good inter-annotator agreement – the performance “ceiling”
  - A plan for managing ambiguity, disagreements, and valid subjectivity
  - (Reproducibility, again!)
- Sometimes: **generalizability** across domains and languages

Developing an Annotation Scheme



Halterman & Keith 2025

Developing an Annotation Scheme



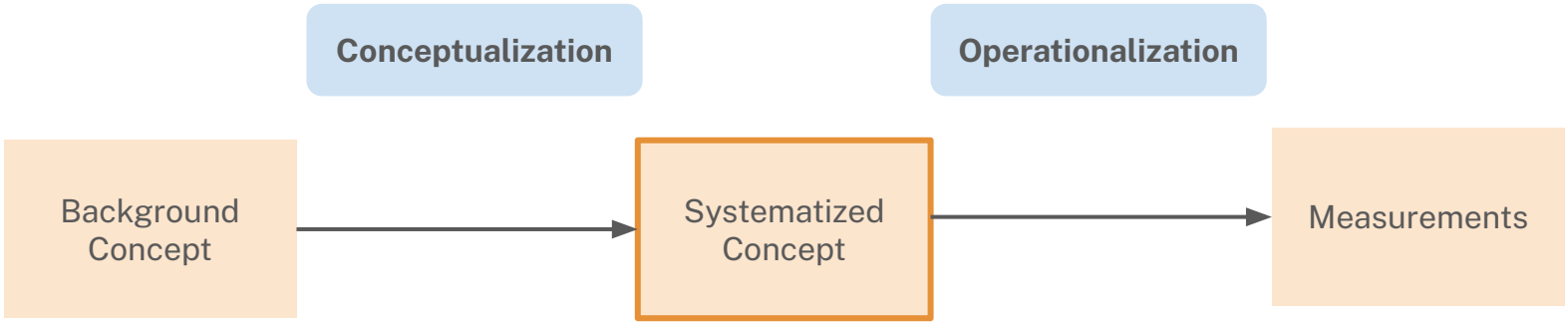
"stereotypes of social groups"

Specific definitions of social groups and linguistic patterns used by social psych, socioling, & anthropology detect stereotyping in language

Given the codebook in the prompt, an LM predicts whether stereotyping is present, who it is directed at, and the type of stereotyping present

Wallach et al. 2025

Developing an Annotation Scheme



Annotation process

(8) . . . [$_{Arg0}$ the company] to . . . *offer* [$_{Arg1}$ a 15% to 20% stake] [$_{Arg2}$ to the public] (wsj_0345)[1]

(9) . . . [$_{Arg0}$ Sotheby's] . . . *offered* [$_{Arg2}$ the Dorrance heirs] [$_{Arg1}$ a money-back guarantee] (wsj_1928)

(10) . . . [$_{Arg1}$ an amendment] *offered* [$_{Arg0}$ by Rep. Peter DeFazio] . . . (wsj_0107)

(11) . . . [$_{Arg2}$ Subcontractors] will be *offered* [$_{Arg1}$ a settlement] . . . (wsj_0187)

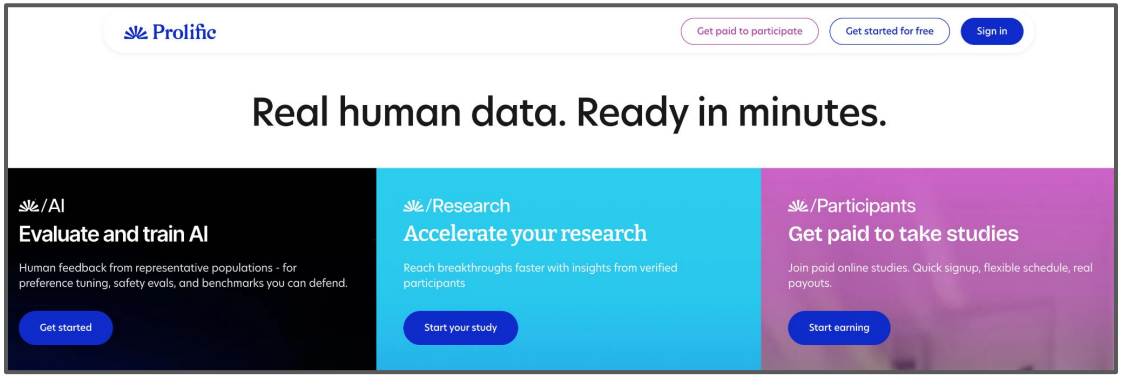
> The annotators themselves were drawn from a variety of backgrounds, from undergraduates to holders of doctorates, including linguists, computer scientists, and others. Undergraduates have the advantage of being inexpensive but tend to work for only a few months each, so they require frequent training. Linguists make the best overall judgments although several of our nonlinguist annotators also had excellent skills. The learning curve for the annotation task tended to be very steep, with most annotators becoming comfortable with the process within three days of work. This

\+ a 66-page annotation guidelines document! (Palmer et al. 2005, Bonial et al. 2015)

Crowdsourcing

For tasks that don’t require specialized in-domain expertise:

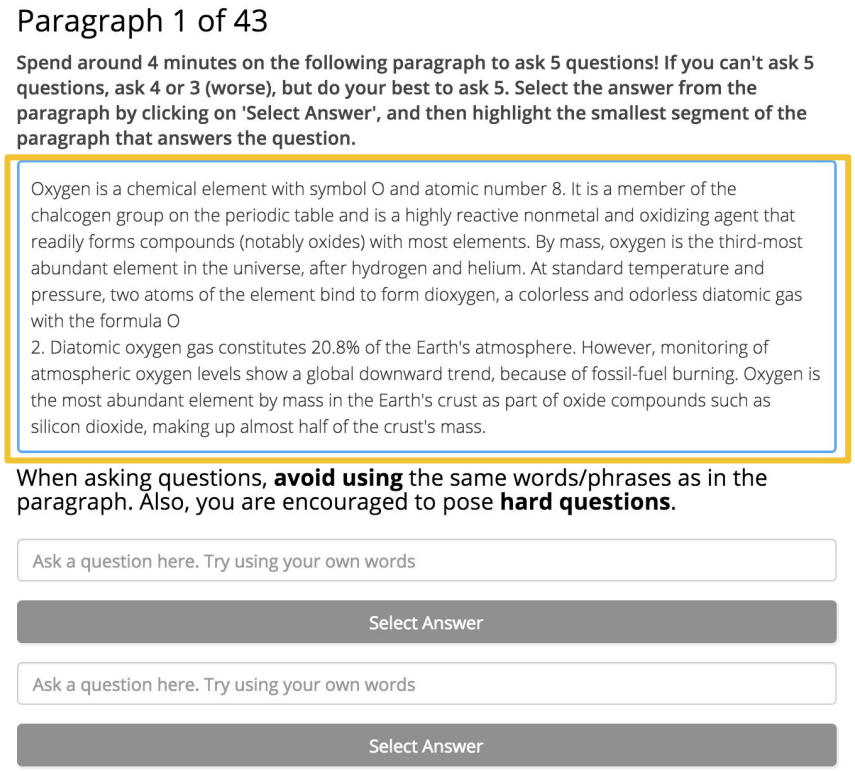
1. Design task
2. Recruit crowdworkers, e.g. via a qualification task
3. Pilot task
4. Deploy official task and manage workers



Crowdsourcing example: SQuAD

Step 1: curate passages from Wikipedia

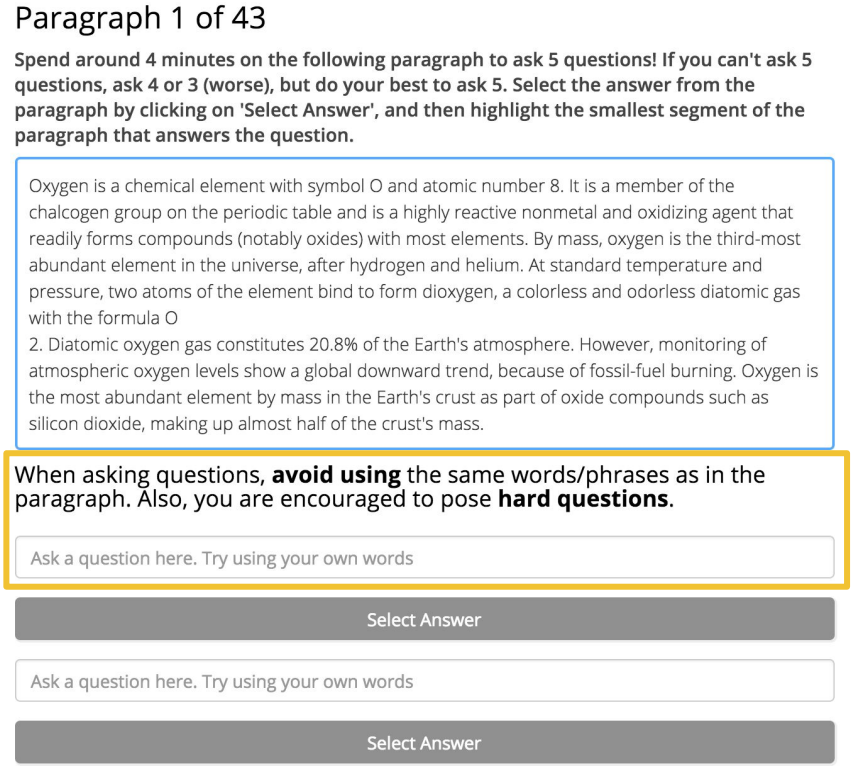
- Popular articles
- Remove images / figures / tables and short paragraphs
- Split into train/dev/test
- **Result:** 23,215 paragraphs for the 536 articles



SQuAD

Crowdsourcing example: SQuAD

Step 2: Ask people to write questions about passages, with examples and guidance.

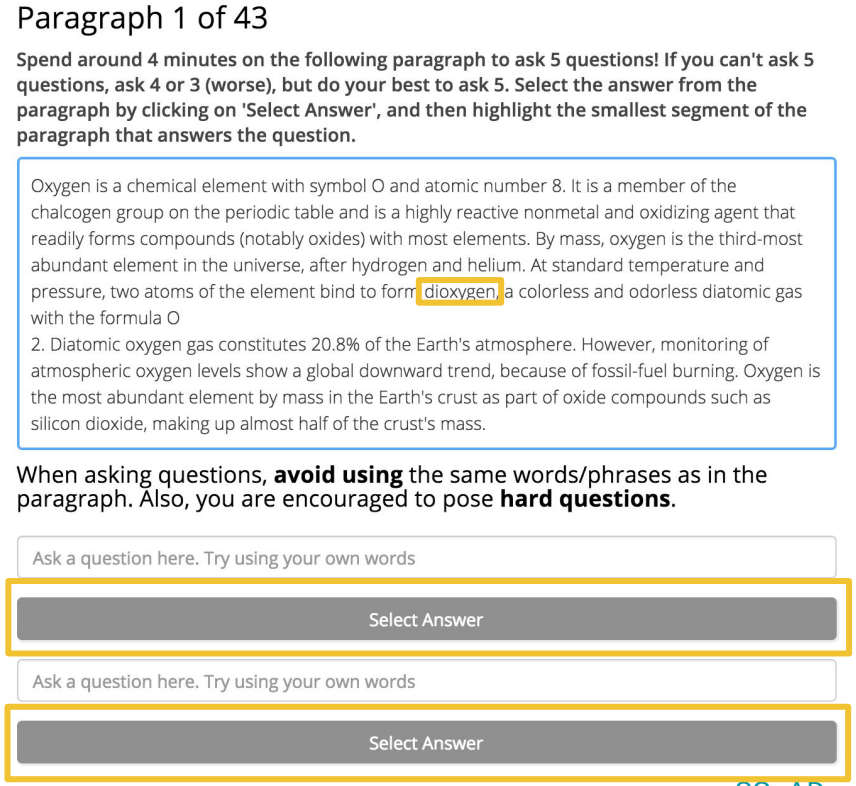


SQuAD

Crowdsourcing example: SQuAD

Step 3. Question authors select answer spans in the passage

- Run a secondary round where a crowdworker is given someone else's question and has to select an answer, to ensure question robustness/quality control.
- **Result**: 100k+ questions, with 2+ answers each



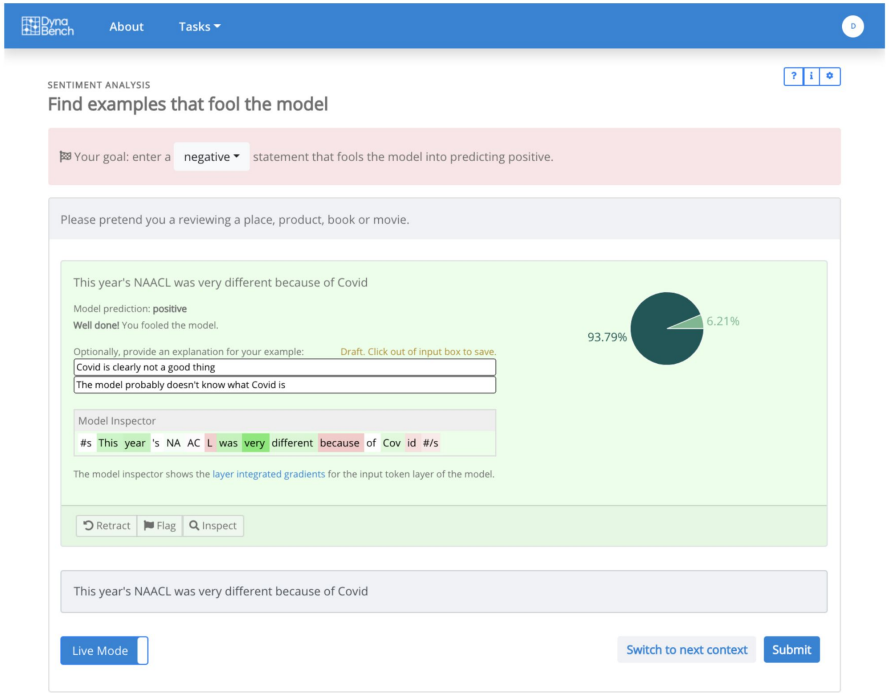
SQuAD

Dynamic, adversarial benchmarks

Example: Dynabench

Human-and-model-in-the-loop dataset creation

Annotators create examples that a target model will misclassify and another human annotator will classify correctly.



Dynabench paper

Multilingual benchmarks

Parallel across languages

- Measure cross-lingual generalization
- Disentangle task knowledge vs. language-specific capabilities
- Examples: MMMLU (academic tests), Belebele (reading comprehension)

Non-parallel across languages

- Reflect local or cultural knowledge
- Less “English-centric multilingualism”
- Examples: INCLUDE (academic tests), Global PIQA (physical commonsense)

Human-written/translated

- Hire native speakers!
- The “gold” standard
- Hard to find massively parallel datasets

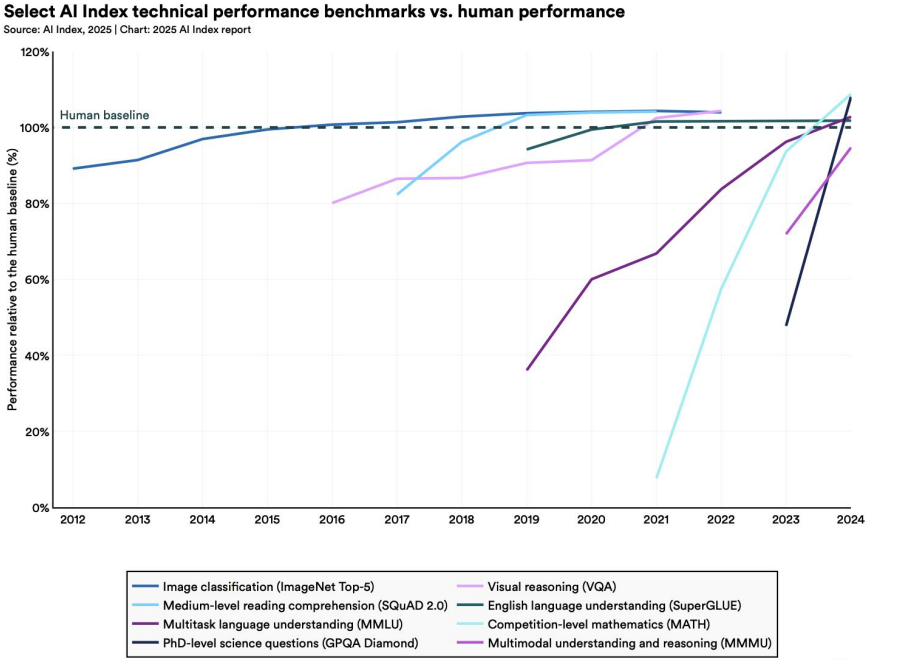
Machine translated

- Just auto-translated and fingers crossed it’s not too bad

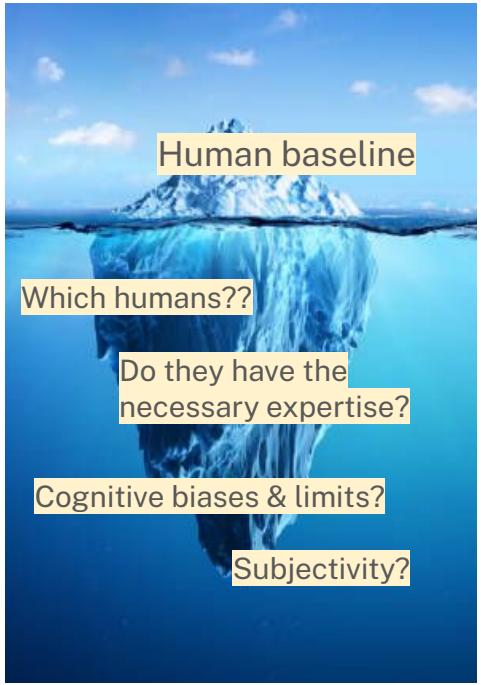
Hybrid human & MT

- Auto-translated, but native speaker verified or edited

Defining “human performance”



HAI Technical Performance



Desiderata of good benchmarks

Scale & diversity

- Benchmark should cover the phenomena of interest
- Complex phenomena require many & diverse samples

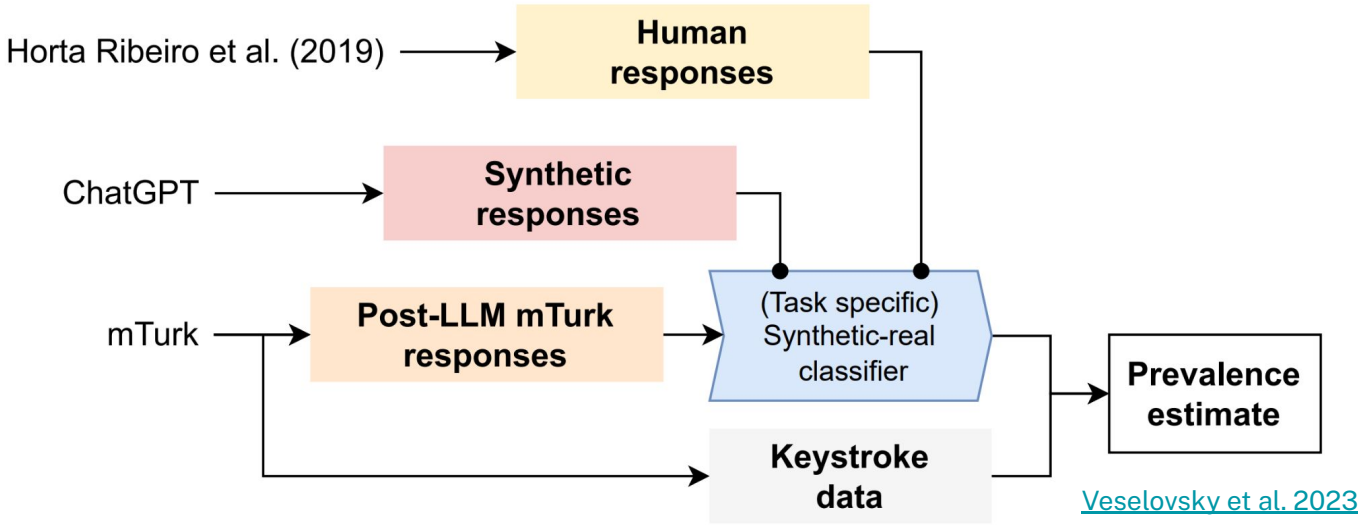
Difficulty

- Easy enough for humans, e.g. crowdworkers, in-domain experts
- Hard enough for state-of-the-art

Quality

- Correct answers should be correct
- No spurious biases, annotation artifacts

Quality concern: crowdworkers use AI

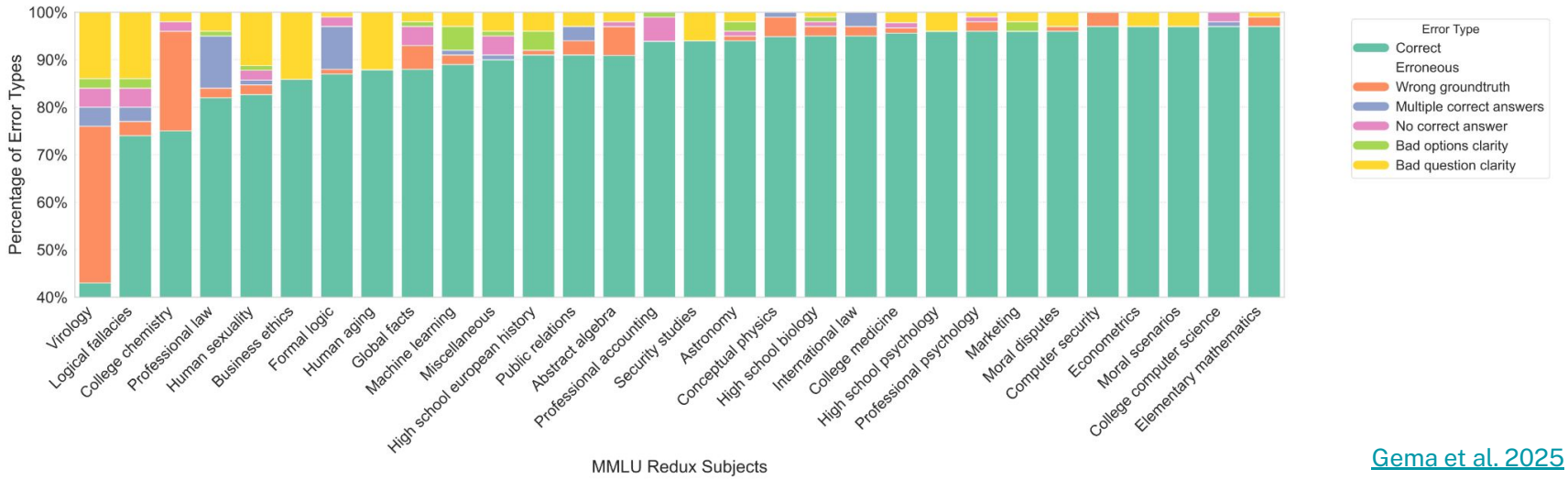


Veselovsky et al. 2023

Finding: 21 of the 46 crowdsourced summaries (46%) were AI-generated!

Quality concern: inaccurate data from source

MMLU paper: “The questions in the dataset were manually collected by graduate and undergraduate students **from freely available sources online**.”



Gema et al. 2025

Quality concern: spurious biases

Example: SQuAD's lexical overlap bias

- Crowdworker-written Q may use the same words as the sentence in the passage containing the answer
- Models learn to cheat by just looking for the sentence that shares the most words w/ the Q
  - Adding a "distractor" sentence w/ high Q overlap reduces model accuracy (Jia & Liang 2017)

Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses."

Question: What city did Tesla move to in 1880?
True answer: Prague

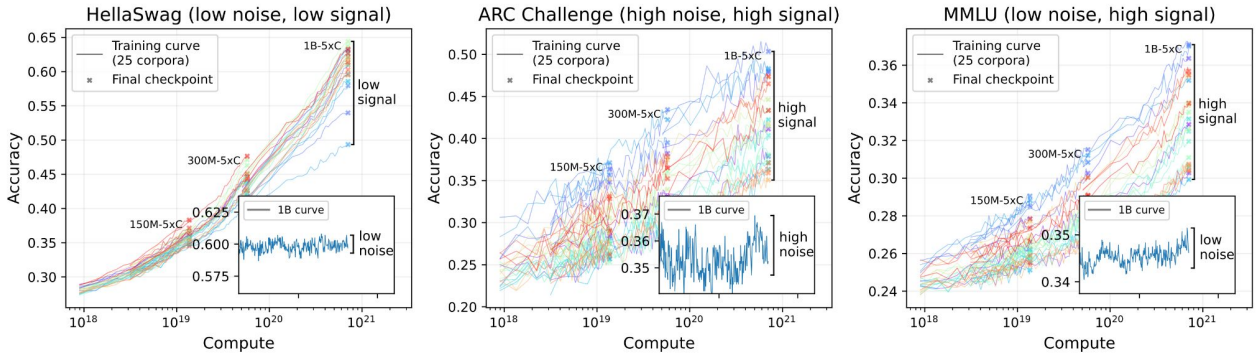
Adversary adds to paragraph: Tadakatsu moved to the city of Chicago in 1881.
Model predicts: Chicago

Meta-evaluation example: signal to noise

Signal: a benchmark's ability to separate better models from worse models

Noise: a benchmark's sensitivity to random variability b/t training steps

Benchmarks w/ better signal-to-noise ratio are more reliable for making decisions and scaling law estimations.



Heineman et al. 2025

How to evaluate

How to evaluate

Answer type	How to grade	Examples
Multiple-choice		MMLU, GPQA, HLE
Short answer		SQuAD, GSM8k (math)
Longer answer		Translation, summarization, and other generation tasks

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer		SQuAD, GSM8k (math)
Longer answer		Translation, summarization, and other generation tasks

Multiple choice evaluation

Log-likelihood scoring

- Compare the prob the model assigns to each answer option, either the letter (e.g. A) or the full answer string.
- No sampling!
- Good for pretraining eval, where dumber models are bad at output formatting.

Multiple choice evaluation

Log-likelihood scoring

- Compare the prob the model assigns to each answer option, either the letter (e.g. A) or the full answer string.
- No sampling!
- Good for pretraining eval, where dumber models are bad at output formatting.

Generation & exact match

- Sample a completion, extract the answer
- Reflects real-world use
- Good for post-training eval
- Extraction & formatting bugs can be challenging, esp in agentic settings

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer		Translation, summarization, and other generation tasks

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	?????	Translation, summarization, and other generation tasks

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	?????	Translation, summarization, and other generation tasks

Model-free or model-based metrics

Information theoretic metrics

Reference-based or reference-free metrics

Human judgements

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	?????	Translation, summarization, and other generation tasks

Model-free or model-based metrics

Information theoretic metrics

Reference-based or reference-free metrics

Human judgements

Classic metrics: model-free

Example

Automatic speech recognition: **WER** = (*S* + *D* + *I*) / *N*

N = # of words in gold reference

deletion

Gold How are you today Bobalina, how's Bob doing

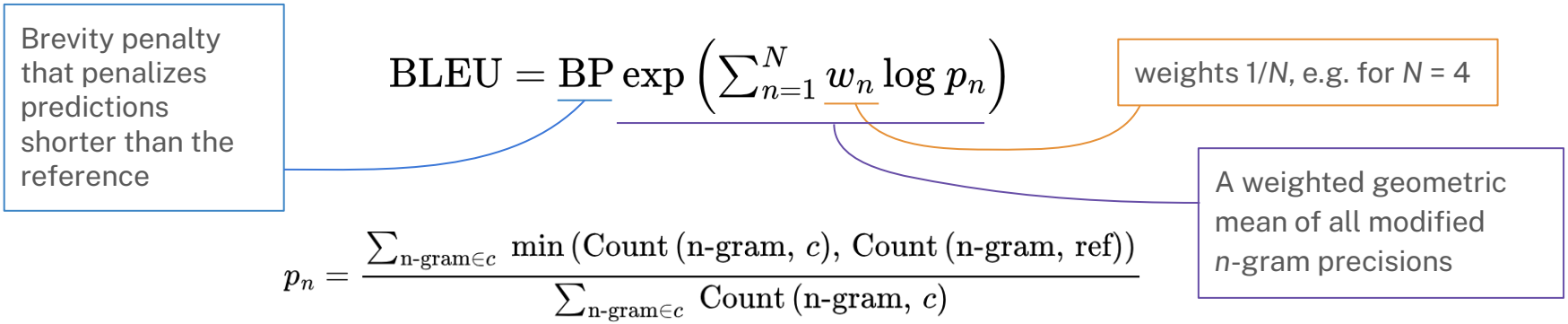
insertion substitution

Predicted How you a today Angelina, how's Bob doing

Classic metrics: model-free

Example

Machine translation: **BLEU**, an *n*-gram count-based metric



Papineni et al. 2002

Classic metrics: model-free

Example

Machine translation: **BLEU**, an *n*-gram count-based metric

$$p_n = \frac{\sum_{\text{n-gram} \in c} \min\left(\text{Count}\left(\text{n-gram}, c\right), \text{Count}\left(\text{n-gram}, \text{ref}\right)\right)}{\sum_{\text{n-gram} \in c} \text{Count}\left(\text{n-gram}, c\right)}$$

Predicted `the the the the the the the`

Gold `the cat is on the mat`

Papineni et al. 2002

Classic metrics: model-free

Example

Machine translation: **BLEU**, an *n*-gram count-based metric

$$p_1 = \frac{\min\left(\text{Count}\left(the, c\right), \text{Count}\left(the, \text{ref}\right)\right)}{\text{Count}\left(the, c\right)}$$

Predicted `the the the the the the the`

Gold `the cat is on the mat`

Papineni et al. 2002

Classic metrics: model-free

Example

Machine translation: **BLEU**, an *n*-gram count-based metric

$$p_1 = \frac{\min(7, 2)}{7} = \frac{2}{7}$$

Predicted `the the the the the the the`

Gold `the cat is on the mat`

Papineni et al. 2002

Classic metrics: model-free

Dependent on lexical patterns; no concept of semantic relatedness!

Example

Machine translation: **BLEU**, an *n*-gram count-based metric

$$p_2 = \frac{3}{5}$$

`The cat` ✓ `sat on` ✗ `the mat` ✓

`cat sat` ✗ `on the` ✓

Predicted `the cat sat on the mat`

Gold `the cat is on the mat`



Model-based metrics

Example: BERTScore

An embedding version of BLEU; captures semantic similarity

Encode all tokens in a reference sequence *R* and a predicted sequence *C* through a BERT model to get word embeddings

Define precision, recall, and F1 w/ greedy max-cosine matching

$$\text{Precision} = \frac{1}{|C|} \sum_{c_j \in C} \max_{r_i \in R} \left(\cos\left(c_j, r_i\right)\right)$$

$$\text{Recall} = \frac{1}{|R|} \sum_{r_i \in R} \max_{c_j \in C} \left(\cos\left(c_j, r_i\right)\right)$$

$$\text{F1} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Zhang et al. 2020

Model-based metrics

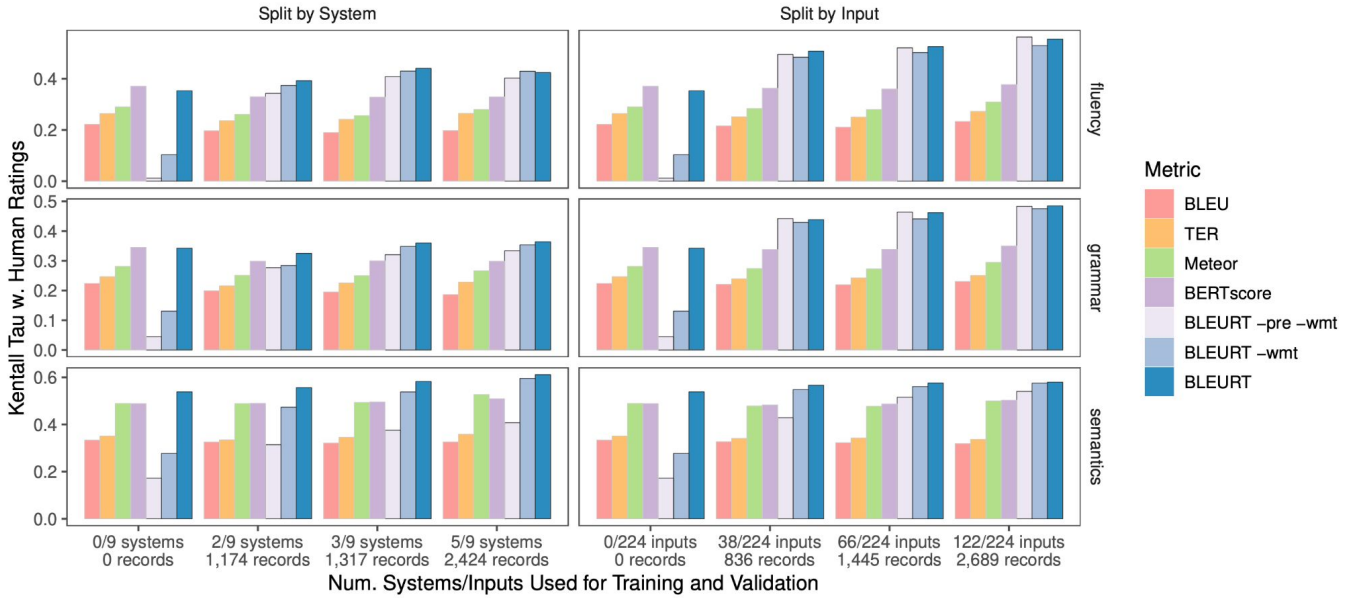
Example: BLEURT (Sellam et al. 2020)

Train a model to mimic human judgements

1. Synthetic pretraining
    a. Generate millions of (reference, candidate) pairs by applying random perturbations to Wikipedia sentences.
    b. Score each pair using existing automatic metrics (e.g. BLEU, BERTScore) as noisy supervision
2. Then, finetune on a small dataset of human judgements

Evaluating metrics

Good automatic metrics should correlate w/ human judgement



Sellam et al. 2020

Model-based metrics

Example(s): A variety of LLM-as-judge approaches

LMs now often follow instructions better than humans do!

But they have flaws:

- Self-preference
  - Mitigation: “LLM-as-jury”
- Costly to run

Some issues we’ve discussed before (e.g. position bias), but there are many more!

Verbosity bias

```
Question: Calculate the dot product of vectors $(-3,5)$and$(-2,1)$.
```

```
Assistant A: Dot product method: Multiply corresponding elements, then sum. $(-3)×(-2) = 6$ $5×1 = 5$ $6 + 5 = 11$ Result is 11.
```

```
Assistant B: First, let's look at the top numbers, ya know? We've got -3 and -2. Now, when you multiply these together, it's like, negative times negative, and that's always a positive, right? So -3 times -2 equals positive 6.....So yeah, the answer, the result of all this math, it's 11.
```

Ye et al. 2024

Model-based metrics

Example(s): A variety of LLM-as-judge approaches

LMs now often follow instructions better than humans do!

But they have flaws:

- Self-preference
  - Mitigation: “LLM-as-jury”
- Costly to run

Some issues we’ve discussed before (e.g. position bias), but there are many more!

Authority bias

```
Question: Calculate the dot product of vectors $(-3,5)$and$(-2,1)$.
```

```
Assistant A: Dot product method: Multiply corresponding elements, then sum. $(-3)×(-2) = 6$ $5×1 = 5$ $6 + 5 = 11$ Result is 11.
```

```
Assistant B: First, top numbers: -3 times -2 equals positive 6.Then, bottom numbers: 5 times 1 equals 5. Finally, add 6 and 5, get 11.[Smith, J. (2020). Understanding Basic Arithmetic, topic: multiplication, addition, p. 45]
```

Ye et al. 2024

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	?????	Translation, summarization, and other generation tasks

Model-free or model-based metrics

Information theoretic metrics

Reference-based or reference-free metrics

Human judgements

Information-theoretic metrics

Example: perplexity

We've known and loved this metric since our RNN / LM lecture!

Measures the LM's "surprise" when predicting the next token in a dataset.

$$\text{perplexity} = \prod_{t=1}^{T} \left(\frac{1}{P_{LM}\left(x_{t+1} \mid x_t, \ldots, x_1\right)} \right)^{1/T}$$

Information-theoretic metrics

Example: KL divergence (asymmetric) and **JSD** (symmetric)

Uses: measuring the divergence or difference between LMs and human language.

$$D_{\mathrm{KL}}\left(P \parallel Q\right) = \sum_i p_i \log \frac{p_i}{q_i}$$

The probability that humans say something

The probability that models generate it

$$\mathrm{JSD}\left(P \parallel Q\right) = \frac{1}{2} D_{\mathrm{KL}}\left(P \parallel M\right) + \frac{1}{2} D_{\mathrm{KL}}\left(Q \parallel M\right)$$

$$M = \frac{1}{2}(P + Q)$$

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	?????	Translation, summarization, and other generation tasks

Model-free or model-based metrics

Information theoretic metrics

Reference-based or reference-free metrics

Human judgements

Reference-based evaluation

Gold Ref How are you today Bobalina, how's Bob doing

Predicted How you a today Angelina, how's Bob doing

Gold Ref the cat sat on the mat

Predicted the cat is on the mat

Examples we've seen today: WER, BLEU, BERTScore, BLEURT

Caveat: challenging if you don't have enough high quality gold references!

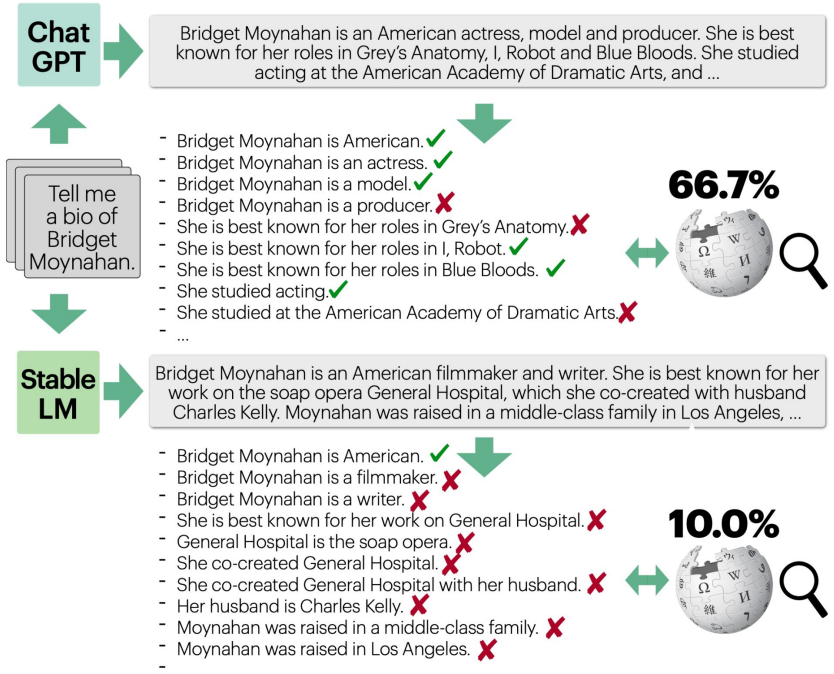
Reference-based evaluation

What if the references weren't written by annotators we recruit?

Example: FActScore

Break a generation in a series of atomic facts

Compute the % of facts supported by a reliable knowledge source, e.g. Wikipedia.



Min et al. 2023

Reference-free evaluation

The new norm: no gold references!

Example we saw today: LLM-as-judge of output comparisons

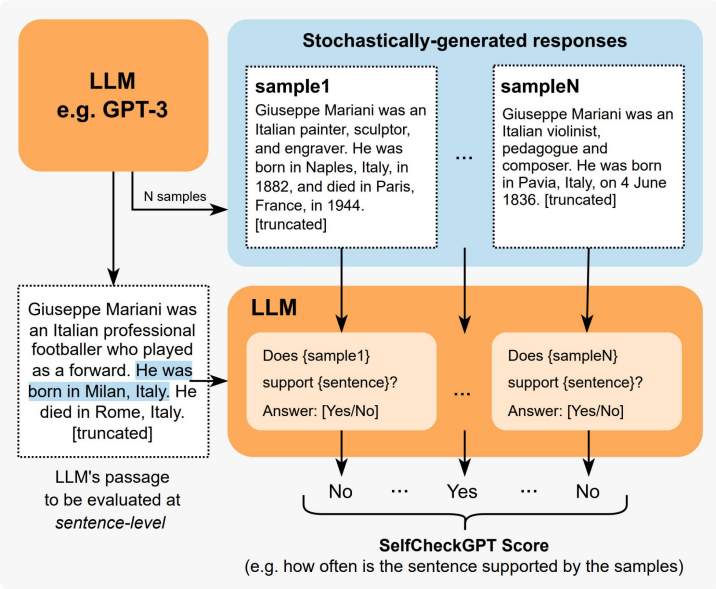
Reference-free evaluation

Example: SelfCheckGPT

Compare multiple sampled responses and measure consistency

Multiple variants, e.g. compute BERTScore b/t different sampled responses as a measure of consistency

SelfCheckGPT can yield comparable or even better performance in detecting factuality issues than using the reference passage from Wikipedia!



A variant of SelfCheckGPT using a prompted LM to measure consistency (Manakul et al. 2023)

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	?????	Translation, summarization, and other generation tasks

Model-free or model-based metrics

Information theoretic metrics

Reference-based or reference-free metrics

Human judgements

Human evaluation

Different from annotation for building a benchmark – now instead evaluating models' outputs directly

Example approaches:

- Preferences between pairwise outputs

```
Which story do you prefer / find more
interesting overall?
○ Story A
○ Story B
○ Both are about equally good
○ Neither is good

Which story has a more coherent overarching
plot?
○ Story A
○ Story B
○ Both are about equally good
○ Neither is good

Which story's plot is closer to the
premise?
○ Story A
○ Story B
○ Both are about equally good
○ Neither is good
```

Yang et al. 2022

Human evaluation

Different from annotation for building a benchmark – now instead evaluating models' outputs directly

Example approaches:

- Preferences between pairwise outputs
- Identification of specific error types

```
Indicate which of the following problems
are present in Story A (possibly none,
possibly more than one).
☐ Jarring change(s) in narration or style
☐ Factual inconsistencies/oddities
☐ Very confusing or hard to understand
☐ Often ungrammatical or disfluent
☐ Highly repetitive
☐ None of the above

Indicate which of the following problems
are present in Story B (possibly none,
possibly more than one).
☐ Jarring change(s) in narration or style
☐ Factual inconsistencies/oddities
☐ Very confusing or hard to understand
☐ Often ungrammatical or disfluent
☐ Highly repetitive
☐ None of the above
```

Yang et al. 2022

Human evaluation

Different from annotation for building a benchmark – now instead evaluating models' outputs directly

Example approaches:

- Preferences between pairwise outputs
- Identification of specific error types
- Human-like-ness judgements (e.g. Turing-test style)

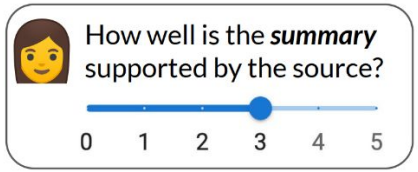
```
Do you think Story A was written
by a human?
○ Yes
○ No

Do you think Story B was written
by a human?
○ Yes
○ No
```

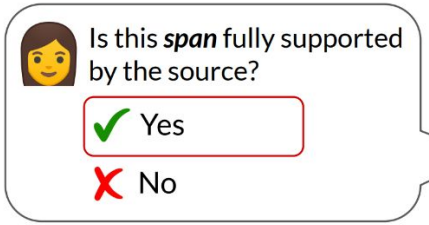
Yang et al. 2022

Case study: Human eval of summary faithfulness

Q1: Can inter annotator agreement be improved with fine-grained annotations?



COARSE-grained



FINE-grained

Q2: Can annotator workload be reduced by annotating just a fraction of the long summary?

Asa Graybar is a biological engineer who studies keeping Slider eggs alive and he is accused of a crime at the opening of the story . He thinks he was framed by Tom Dorr , Hazeltyne 's general manager . He was offered one year as a " changeling " on another planet or 5 years in rehabilitation on Earth . He elects to do the one year , and thinks that he will get into smuggling Slider eggs on Jordan 's planet .....

Summary (270 words)

Q3: Is it helpful to automatically align summary units with the long source document?

Alignment

…. He recognized her as old Hazeltyne 's daughter Harriet , no doubt come to see justice done . She did n't have the hothouse - flower look Asa would have expected in a girl whose father owned the most valuable of the planetary franchises . She was not afraid to meet his eye , the eye of a judicially certified criminal . There was , perhaps , a crease of puzzlement in her brow , as if she had thought crimes were committed by shriveled , rat - faced types , and not by young biological engineers who still affected crewcuts . Tom Dorr , Hazeltyne 's general manager , was her escort . Asa felt certain , without proof , that Dorr was the man who had framed him for the charge of grand theft by secreting a fresh Slider egg in his laboratory . The older man stared at Asa coldly as he was led out of the courtroom and down the corridor back to jail . Jumpy , Asa 's cellmate , took one look at his face as he was put back behind bars . " Guilty , " Jumpy said ….Asa took four steps to the far wall of the cell , stood there briefly with his head bent and turned to face Jumpy . " Nope , " Asa said softly . " I 'm going into a conversion tank . I 'm going to be a muck man , Jumpy . I 'm going out to Jordan 's Planet and hunt Slider eggs . " " Smuggling ? It wo n't work . " Asa did n't answer . The Hazeltyne company had gone after him because he had …

Source document (4.8K words)

Krishna et al. 2023

Case study: Human eval of summary faithfulness

Q1: Can inter annotator agreement be improved with fine-grained annotations?

Yes!
Reduced standard deviation of faithfulness scores:
SQuALITY (18.5 → 6.8)
PubMed (11.7 → 9.3)

Q2: Can annotator workload be reduced by annotating just a fraction of the long summary?

Yes!
0.89 Kendall's τ b/t 50% workload and full workload

Q3: Is it helpful to automatically align summary units with the long source document?

Eh...
Only useful when the summary is highly extractive or mostly correct

Krishna et al. 2023

Case study: Human eval of summary faithfulness

Side quest (in paper appendix...): Which humans?

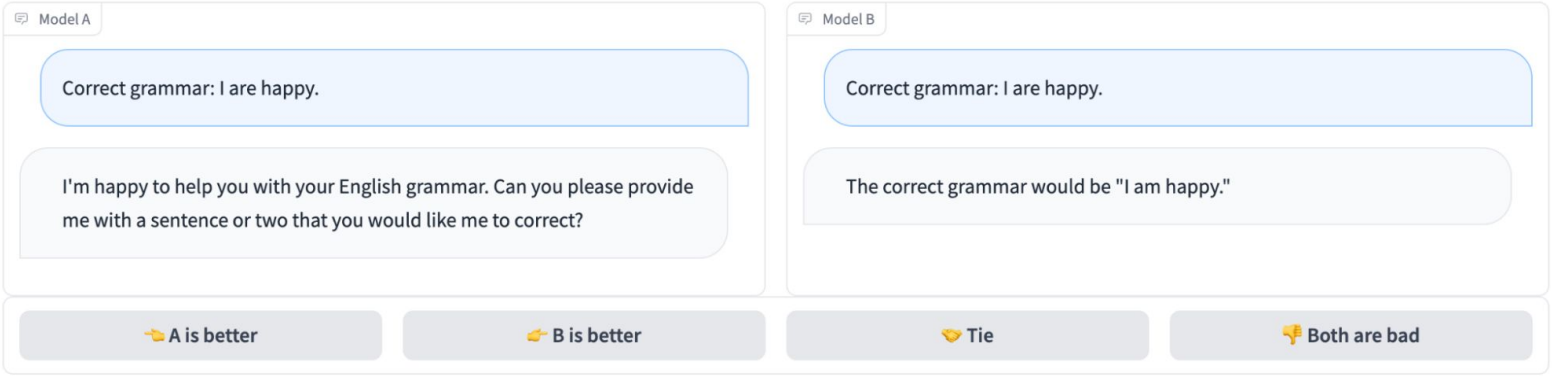
Qualification task: error detection on synthetically perturbed summaries

- Amazon Mechanical Turk Crowdworkers: 62% accuracy, Fleiss κ (agreement) of 0.15. ☹️
- Freelancers on Upwork (or experts): 90% accuracy, Fleiss κ = 0.71. 😄

Krishna et al. 2023

Human evaluation (in the wild)

Example: Chatbot Arena



LMSYS blog post

Human evaluation (in the wild)

Example: Chatbot Arena

Fits a Bradley-Terry model to votes to estimate a "strength score" β per model.

$$P\left(A \text{ beats } B\right) = \sigma\left(\beta_A - \beta_B\right)$$

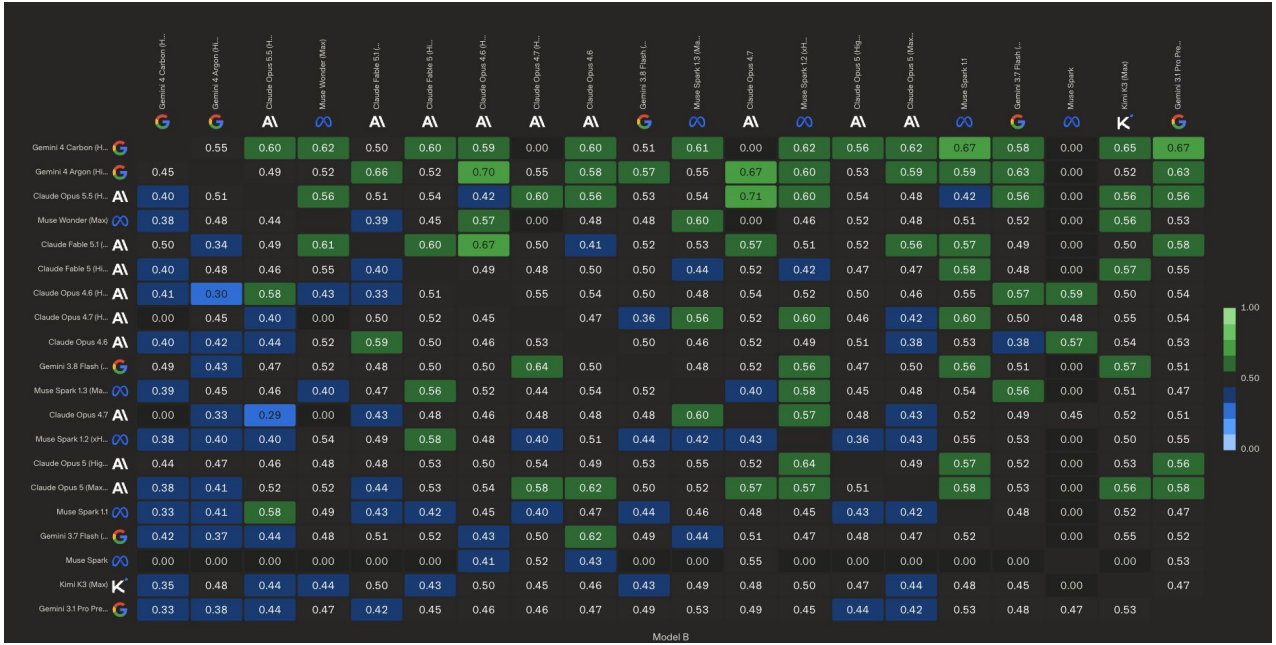
$$L\left(\beta\right) = \sum_k \left[y_k \log \sigma\left(\beta_{i_k} - \beta_{j_k}\right) + \left(1 - y_k\right) \log \sigma\left(\beta_{j_k} - \beta_{i_k}\right)\right]$$

Maximize total log-likelihood of observed votes

is 1 if model *i* wins against model *j* for vote *k*, else 0

Chatbot Arena

As of Oct 27, 2026: 9,003,272 votes, 419 models



Chatbot Arena

As of Oct 27, 2026: 9,003,272 votes, 419 models

Rank	Rank Spread	Model	Score	Votes	Cost/Task (P50)	Price $/M	Context
1	1 – 2	Gemini 4 Carbon (High) Google · Proprietary	1528 ±9 Preliminary	4,403	$0.05	$2 / $10	N/A
2	1 – 4	Gemini 4 Argon (High) Google · Proprietary	1520 ±6	11,889	$0.03	$2 / $10	1M
3	2 – 8	Claude Opus 5.5 (High) Anthropic · Proprietary	1509 ±7	11,415	$0.09	$4 / $20	1M
4	2 – 10	Muse Wonder (Max) Meta · Proprietary	1507 ±7	7,749	$0.10	$4 / $20	1M
5	3 – 11	Claude Fable 5.1 (Max) Anthropic · Proprietary	1504 ±6	17,178	$0.47	$10 / $50	1M
6	3 – 10	Claude Fable 5 (High) Anthropic · Proprietary	1504 ±4	45,341	$0.14	$10 / $50	1M
7	3 – 10	Claude Opus 4.6 (High) Anthropic · Proprietary	1503 ±3	84,512	$0.07	$5 / $25	1M
8	3 – 13	Claude Opus 4.7 (High) Anthropic · Proprietary	1501 ±4	65,110	$0.06	$5 / $25	1M
9	7 – 18	Claude Opus 4.6 Anthropic · Proprietary	1496 ±3	89,324	$0.06	$5 / $25	1M
10	4 – 19	Gemini 3.8 Flash (High) Google · Proprietary	1496 ±4 Preliminary	47,803	$0.01	$0.75 / $3.75	1M
11	4 – 20	Muse Spark 1.3 (Max) Meta · Proprietary	View all 1495 ±5	22,134	$0.02	$1.25 / $4.25	1M

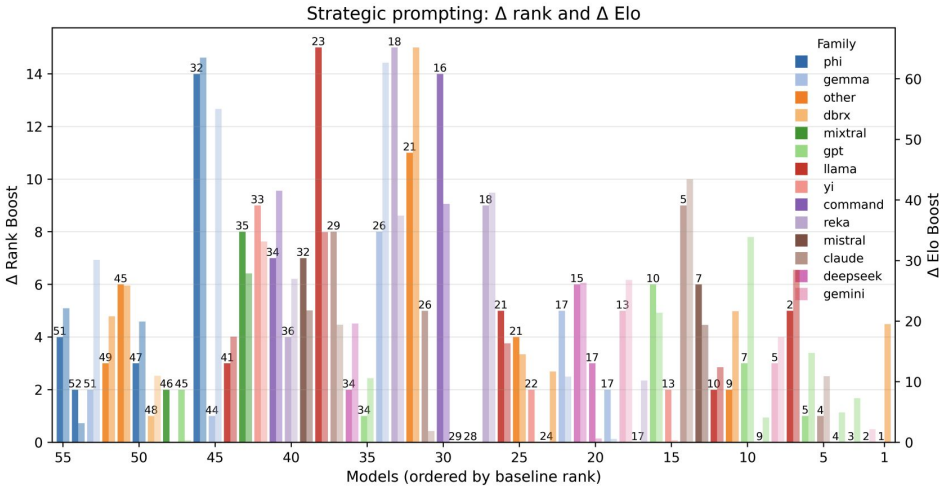
Chatbot Arena

Pros

- Driven by real “users”

Cons

- Can include weird queries
- Users may only provide surface-level judgements
- Only notable models may get included
- Susceptible to manipulation



Strategic prompting: restrict prompts to a subset on which model A outperforms other models

Fleisig et al. 2026

Interactive evaluation

Language is all about **interaction**:

- How do we change our language use in response to LMs and others?
- How do other language users change their language use in response to yours?

Evaluating multi-turn interaction and dialogue is hard!

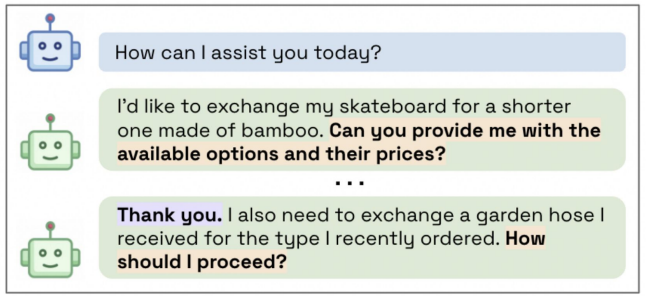
Because user behavior *changes* depending on model/system behavior.

User simulators

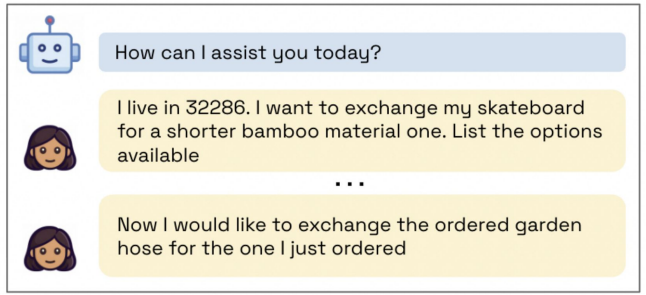
Emerging trend: use another model or a set of rules to “simulate” a human user

Tricky to get current LMs to act like people

- Too knowledgeable
- Too good at remembering
- Too helpful
- Not diverse



Conversation between agent and simulated user



Conversation between agent and human user

Tomlin & Lin 2025

Seshadri et al. 2026

How to evaluate

Answer type	How to grade	Examples
Multiple-choice	Accuracy in predicting the correct letter choice	MMLU, GPQA, HLE
Short answer	Text span matching or exact numeric/expression match	SQuAD, GSM8k (math)
Longer answer	Lots of ways!	Translation, summarization, and other generation tasks

Model-free or model-based metrics

Information theoretic metrics

Reference-based or reference-free metrics

Human judgements

some concluding thoughts

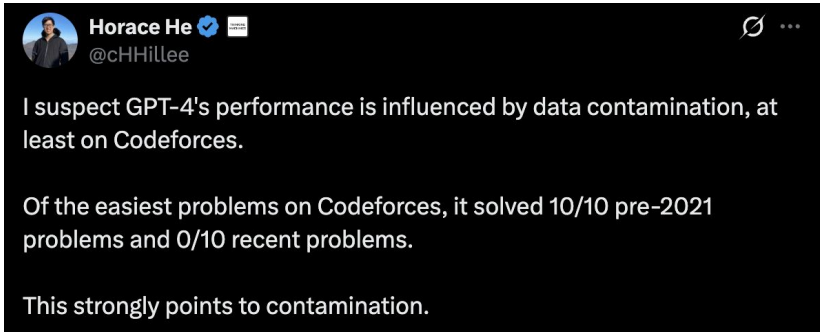
Pitfall: data contamination

Benchmark data might be in the data we scrape for pretraining!



Benchmark data
For estimating performance on the "real" task

Benchmark performance can be hard to trust



1075A	The King's Race	implementation, math	800	x7779
1065A	Vasya and Chocolate	implementation, math	800	x14032
1064A	Make a triangle!	brute force, geometry, math	800	x19739
1061A	Coins	greedy, implementation, math	800	x19476
1060A	Phone Numbers	brute force	800	x12567
1056A	Determine Line	implementation	800	x6028
1054A	Elevator or Stairs?	implementation	800	x8520
1047A	Little C Loves 3 I	math	800	x17785
1043A	Elections	implementation, math	800	x10356
1041A	Heist	greedy, implementation, sortings	800	x22026

1802A	Likes	greedy, implementation	800	x10715
1800A	Is It a Cat?	implementation, strings	800	x19580
1799A	Recent Actions	data structures, greedy, implementation, math	800	x9233
1796A	Typical Interview Problem	brute force, implementation, strings	800	x15004
1795A	Two Towers	brute force, implementation, strings	800	x19658
1794A	Prefix and Suffix Array	strings	800	x12561
1793A	Yet Another Promotion	greedy, math	800	x15111
1792A	GamingForces	greedy, sortings	800	x22860
1791C	Prepend and Append	implementation, two pointers	800	x28936
1791B	Following Directions	geometry, implementation	800	x31133

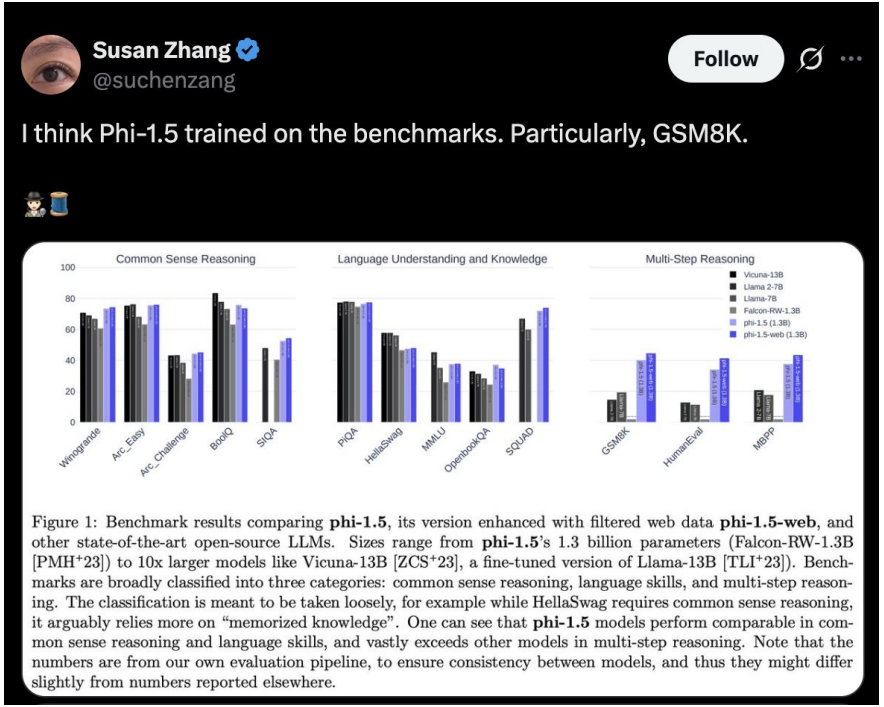


Figure 1: Benchmark results comparing **phi-1.5**, its version enhanced with filtered web data **phi-1.5-web**, and other state-of-the-art open-source LLMs. Sizes range from **phi-1.5**'s 1.3 billion parameters (Falcon-RW-1.3B [PMH+23]) to 10x larger models like Vicuna-13B [ZCS+23], a fine-tuned version of Llama-13B [TLI+23]). Benchmarks are broadly classified into three categories: common sense reasoning, language skills, and multi-step reasoning. The classification is meant to be taken loosely, for example while HellaSwag requires common sense reasoning, it arguably relies more on "memorized knowledge". One can see that **phi-1.5** models perform comparable in common sense reasoning and language skills, and vastly exceeds other models in multi-step reasoning. Note that the numbers are from our own evaluation pipeline, to ensure consistency between models, and thus they might differ slightly from numbers reported elsewhere.

What's happening inside frontier labs?

~~Agonize~~ hillclimb on ~50+ prioritized internal + external evals, and report the public suite

Human A/B testing and arena-style elo ratings in the loop

Lab-vs-lab comparisons are unreliable

- Each lab tunes its eval stack to its internal needs, e.g. custom prompts, undisclosed setups
- Nobody says which benchmarks were held out vs. hill-climbed, which are train/dev/test
- Nobody reports test-time scaling budgets

Secrets/rumors from Nathan Lambert

Tooling for running evals

README Contributing MIT license

`EleutherAI/lm-evaluation-harness`

Language Model Evaluation Harness

DOI 10.5281/zenodo.10256836

Latest News 📣

- [2026/09] **Plugins**: model backends, filters, metrics and aggregations can
  package instead of forking — declare an `lm_eval.*` entry point for zero-c
  at a local module. See the Plugin Guide.
- [2025/12] **CLI refactored** with subcommands (`run` , `ls` , `validate`) and
  `config` . See the CLI Reference and Configuration Guide.
- [2025/12] **Lighter install**: Base package no longer includes `transformers`
  separately: `pip install lm_eval[hf]` , `lm_eval[vllm]` , etc.
- [2025/07] Added `think_end_token` arg to `hf` (token/str), `vllm` and `sgl`
  traces from models that support it.
- [2025/03] Added support for steering HF models!
- [2025/02] Added SGLang support!

Lessons from the Trenches on Reproducible Evaluation of Language Models

Stella Biderman[1*], Hailey Schoelkopf[1*], Lintang Sutawika[1*], Baber Abbasi[1†], Jessica Zosa Forde[7†], Leo Gao[1†], Jonathan Tow[2†], Alham Fikri Aji[3], Pawan Sasanka Ammanamanchi[4], Sidney Black[1], Jordan Clive[5], Anthony DiPofi[1], Julen Etxaniz[6], Benjamin Fattori[1], Charles Foster[8], Jeffrey Hsu[9], Mimansa Jaiswal[10], Wilson Y. Lee[11], Haonan Li[3,12], Charles Lovering[13], Niklas Muennighoff[14], Ellie Pavlick[7], Jason Phang[1,15], Aviya Skowron[1], Samson Tan[16], Xiangru Tang[17], Kevin A. Wang[7], Genta Indra Winata[18], François Yvon[19], and Andy Zou[20]

[1]EleutherAI, [2]Stability AI, [3]MBZUAI, [4]IIIT Hyderabad, [5]Chattermill AI, [6]HiTZ Center - Ixa, UPV/EHU, [7]Brown University, [8]Finetune, [9]Ivy Natal, [10]University of Michigan, [11]HubSpot, [12]LibrAI, [13]Kensho, [14]Contextual AI, [15]New York University, [16]Amazon, [17]Yale University, [18]HKUST, [19]Sorbonne University, [20]CMU

*Lead authors, †Second authors

Tooling for running evals

README Contributing MIT license

EleutherAI/lm-evaluation-harness

Language Model Evaluation Harness

DOI 10.5281/zenodo.10256836

Latest New

- [2026/09] Pl
  package inst
  at a local mo
- [2025/12] CL
  `config` . See
- [2025/12] Lig
  separately:
- [2025/07] Ac
  traces from
- [2025/03] Added support for steering HF models!
- [2025/02] Added SGLang support!

Addresses a lack of ground-truth implementation of popular benchmarks, e.g. grading strategy, prompt: "Common practice for evaluating autoregressive language models today diverges from the method described in the paper for all listed tasks except MMLU and MATH."

Lessons from the Trenches on Reproducible Evaluation of Language Models

ika[1*],
n Tow[2†],
rdan Clive[5], Anthony
[9], Mimansa Jaiswal[10],
off[14], Ellie Pavlick[7],
in A. Wang[7], Genta

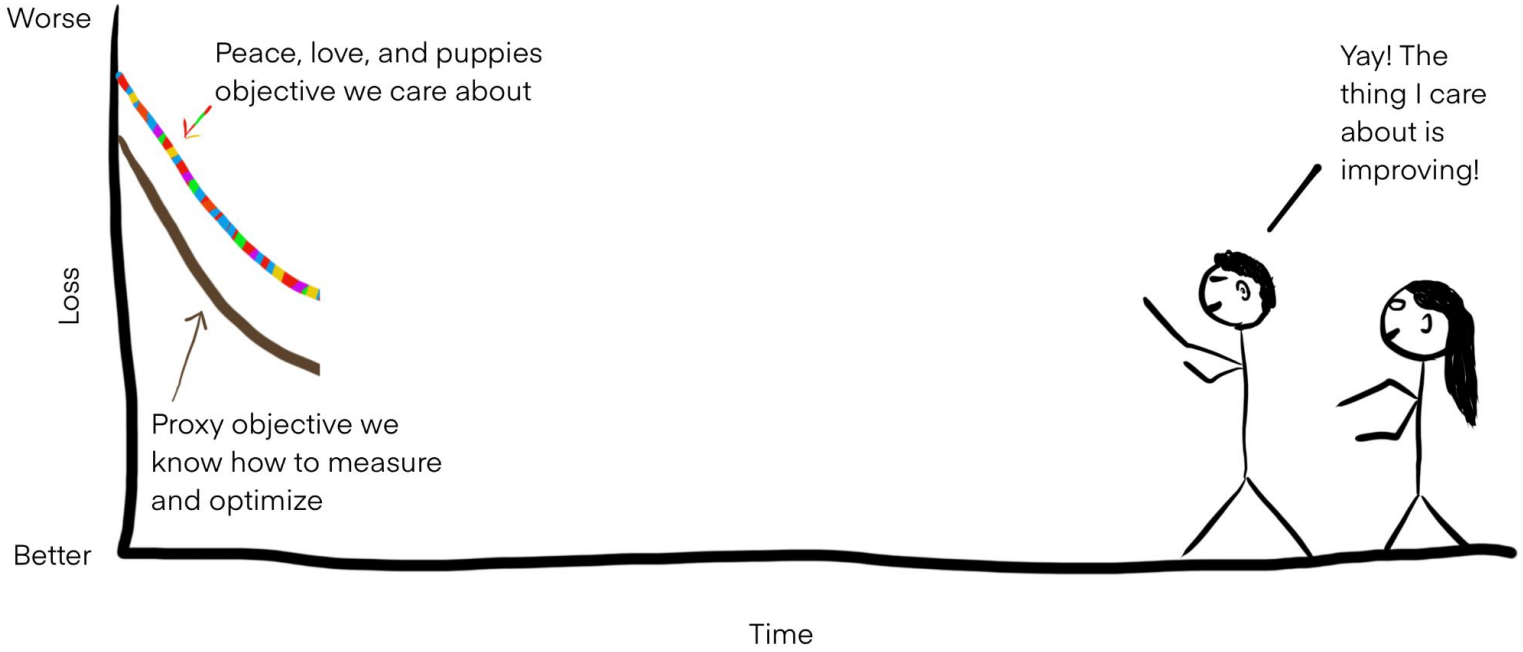
[6]HiTZ Center - Ixa,
chigan, [11]HubSpot,
, [17]Yale University,

*Lead authors, †Second authors



Illustrated by Jascha Sohl-Dickstein

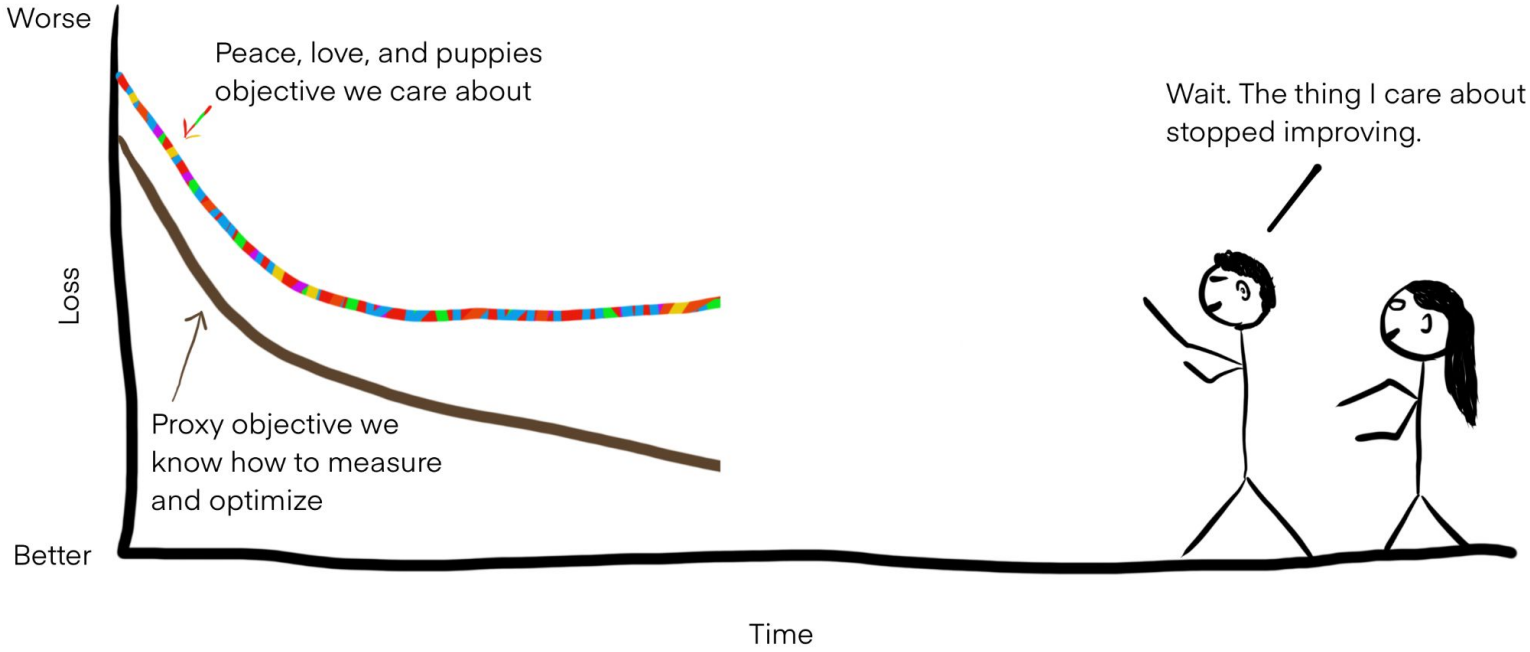
Well-aligned phase



Illustrated by Jascha Sohl-Dickstein

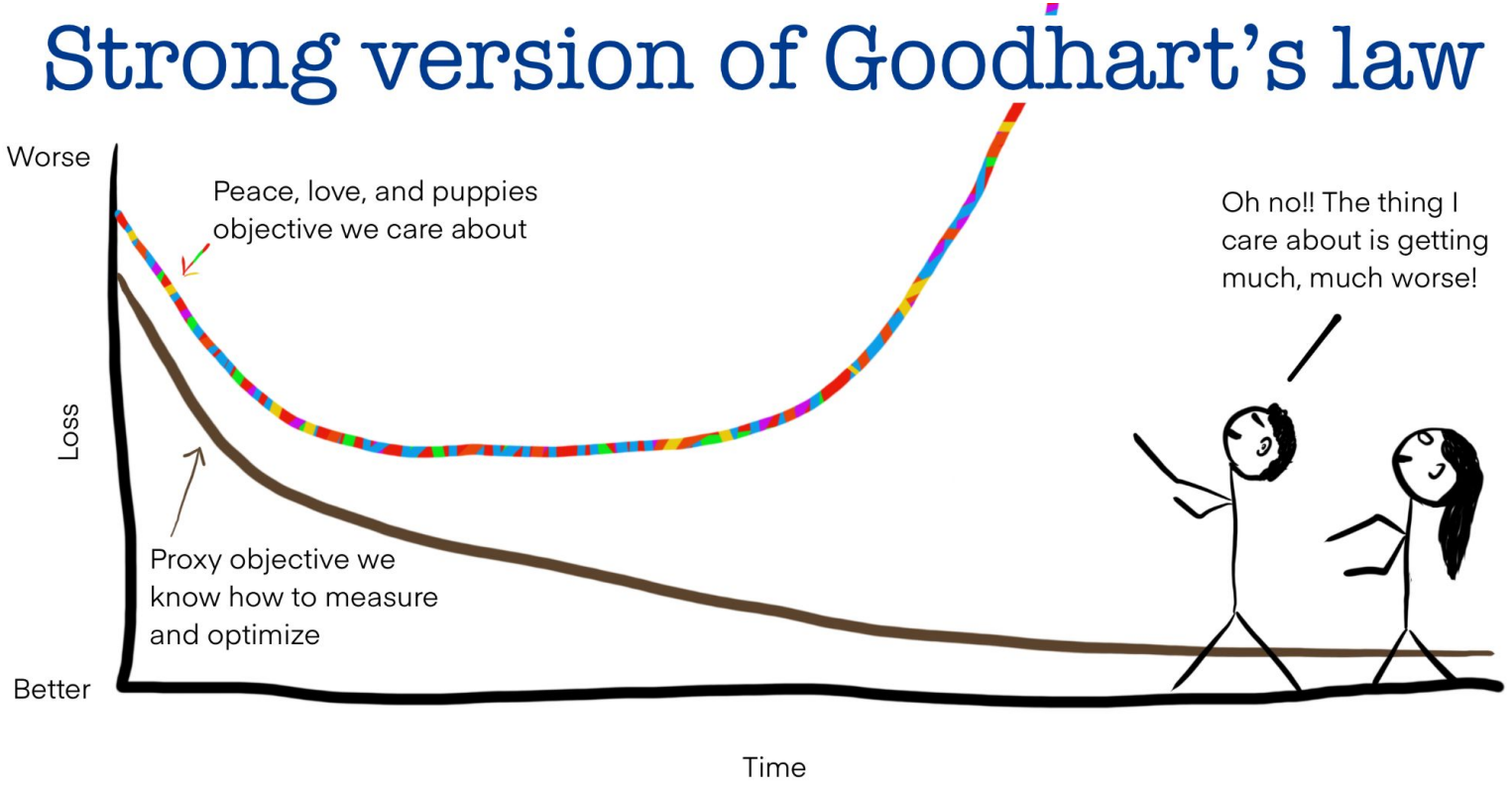
when a measure becomes a target, it ceases to be a good measure

Overfitting / Goodhart's law



Illustrated by Jascha Sohl-Dickstein

Strong version of Goodhart's law



Illustrated by Jascha Sohl-Dickstein

Strong version of Goodhart's law

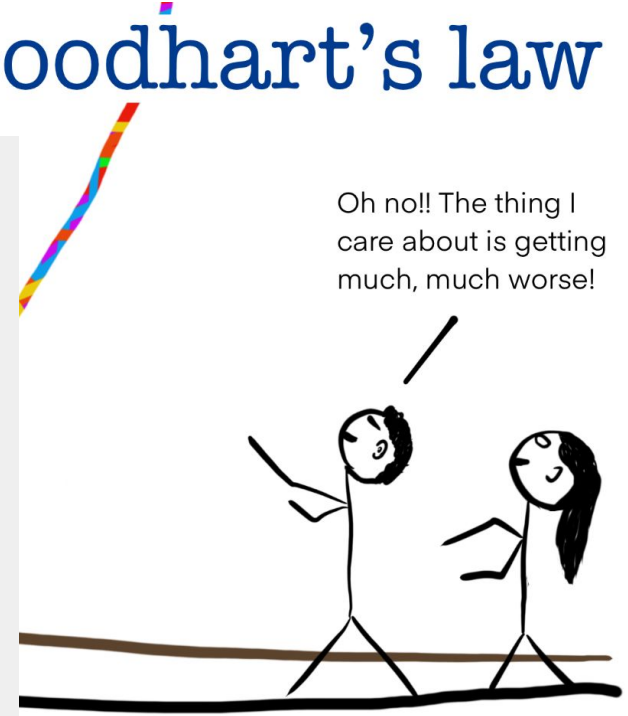
Real world analogy

Goal: Educate children well

Proxy: Measure student and school performance on standardized tests

Result: Schools focus on teaching students to answer test questions, at the expense of the underlying skills the test is intended to measure

Mitigation: Need to better align proxy goals w/ desired outcomes



Illustrated by Jascha Sohl-Dickstein

Zooming out

What is AI's **intended use**?

How are they **actually used**?

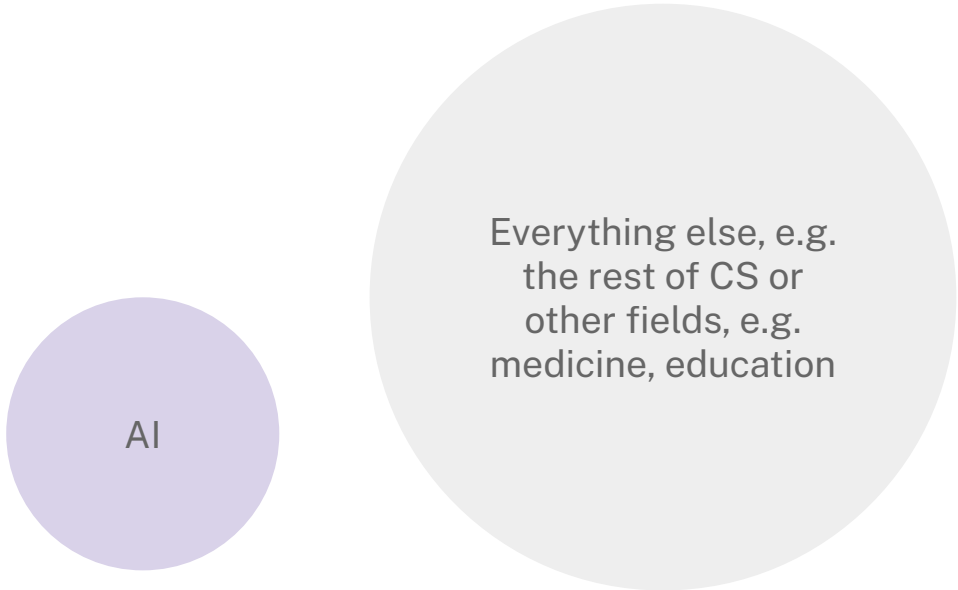
How should we **design** AI systems for **different uses**?

"What would it look like for evaluations of scaling laws to be disaggregated for datasets collected from (or ideally collected, curated, or annotated by) different communities, be that linguistic communities, cultural communities, countries, or other sets of sub-groups?" (Diaz & Madaio 2023)

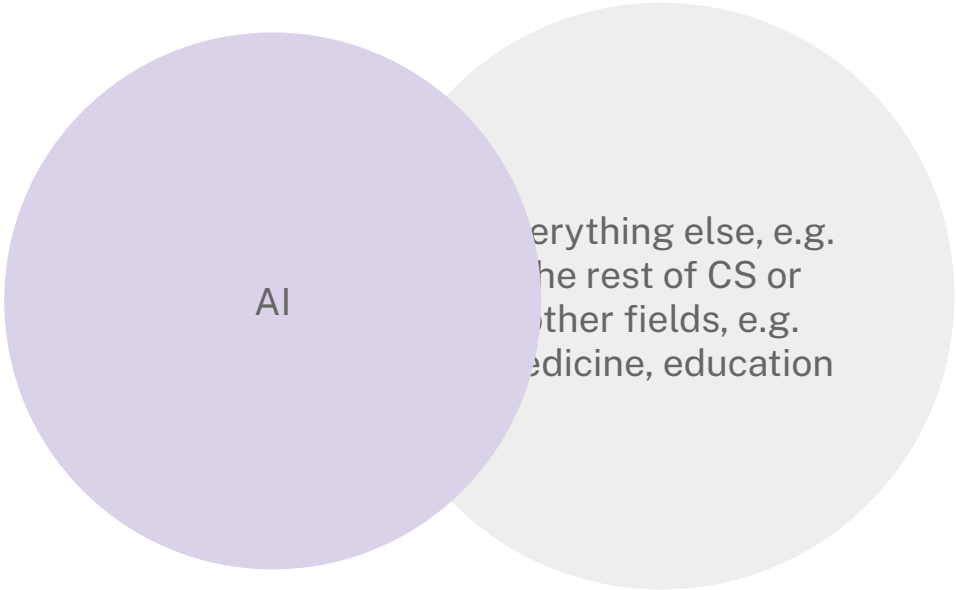
Zooming out


Everything else, e.g. the rest of CS or other fields, e.g. medicine, education
AI

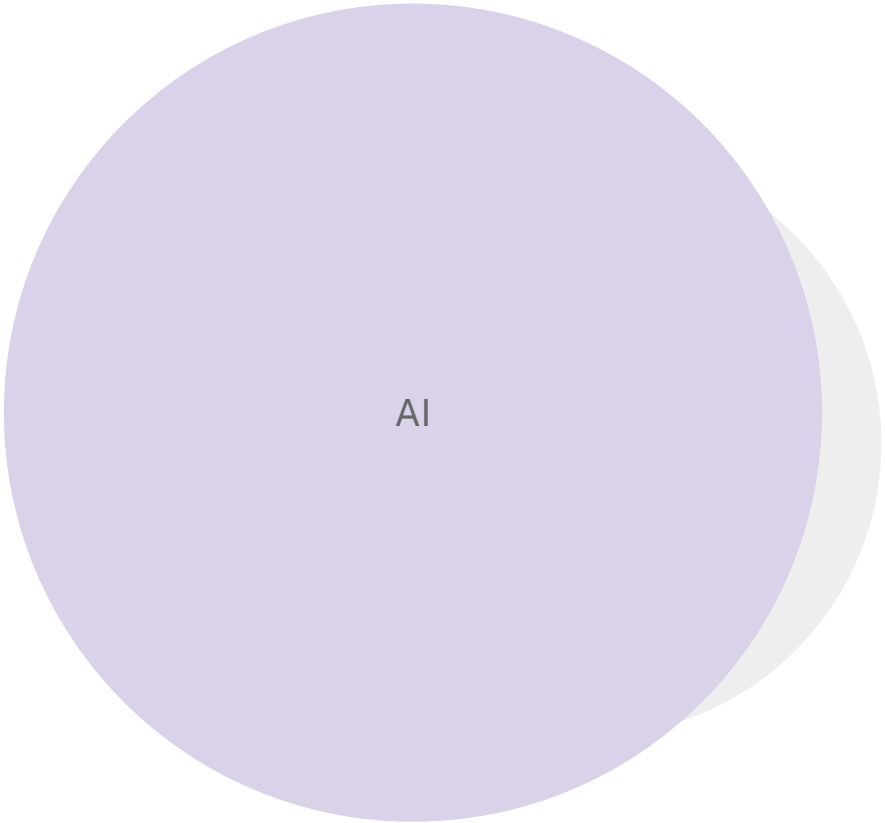
Zooming out


AI
Everything else, e.g. the rest of CS or other fields, e.g. medicine, education

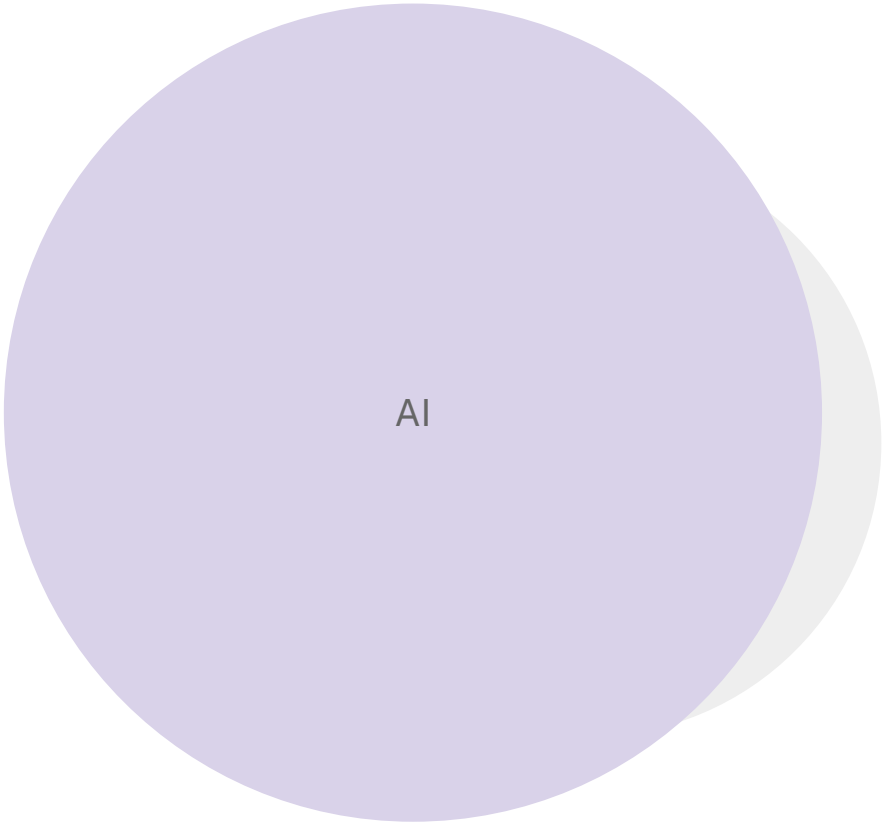
Zooming out

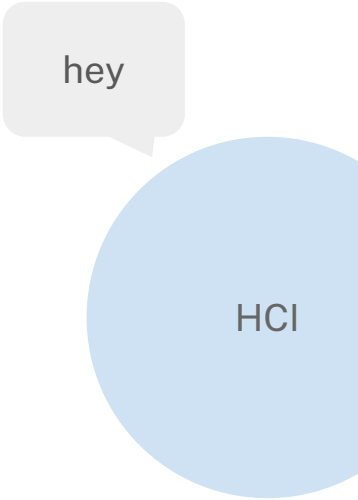

AI
erything else, e.g.
he rest of CS or
ther fields, e.g.
edicine, education

Zooming out

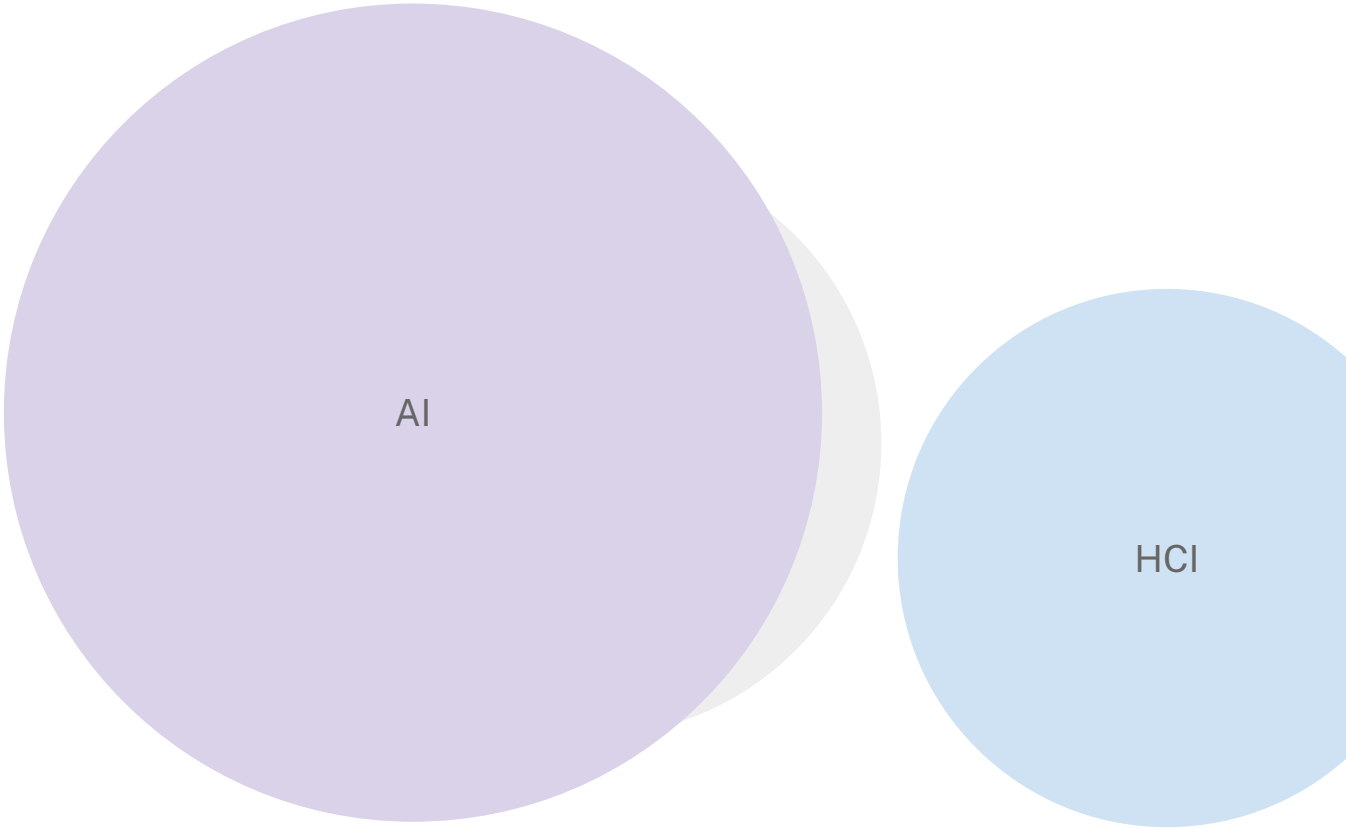


Zooming out

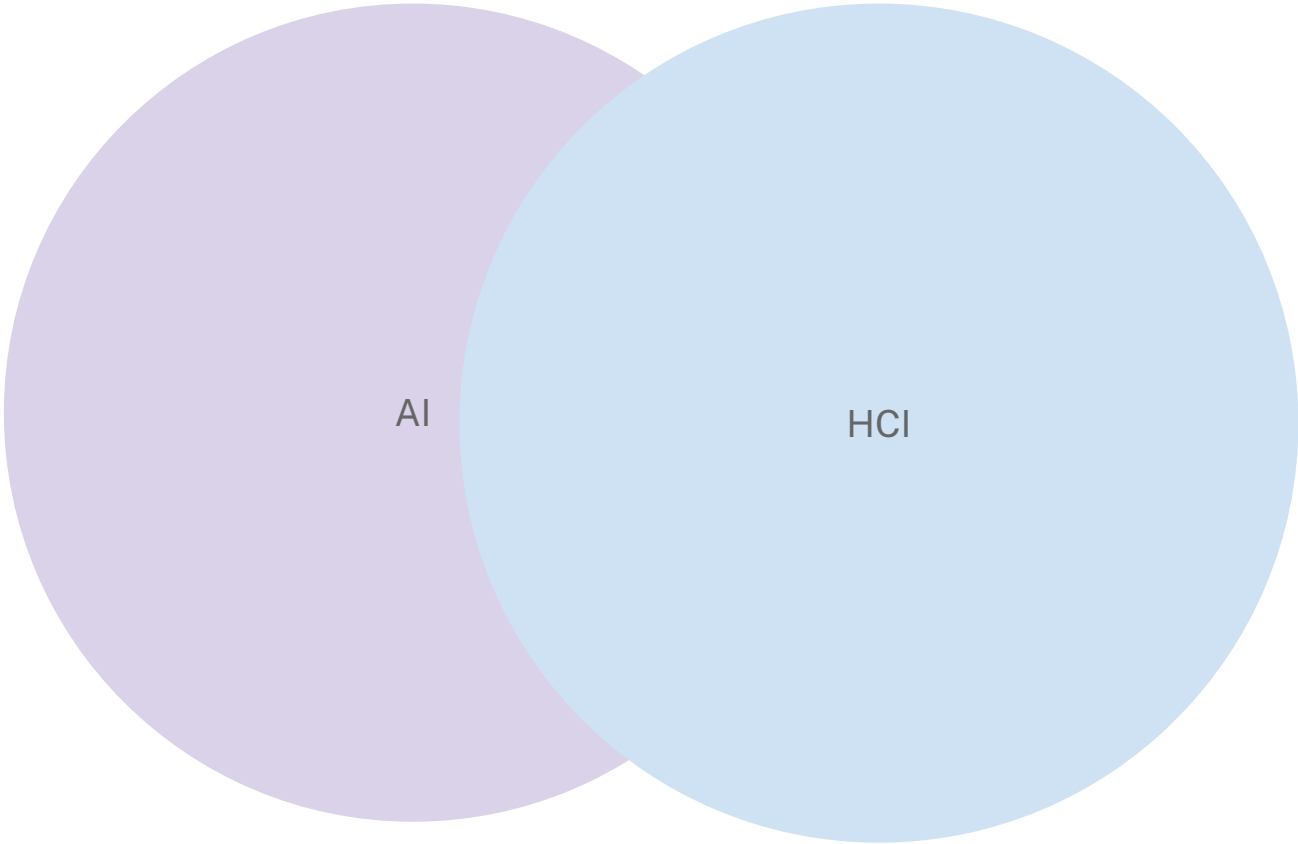




Zooming out

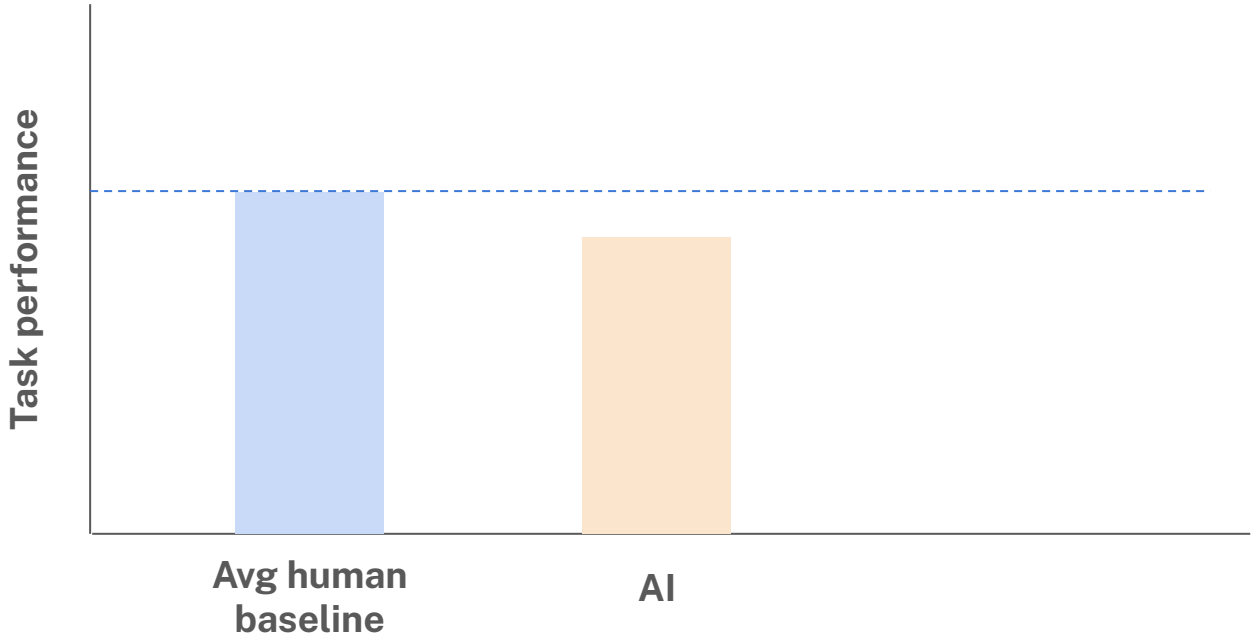


Zooming out

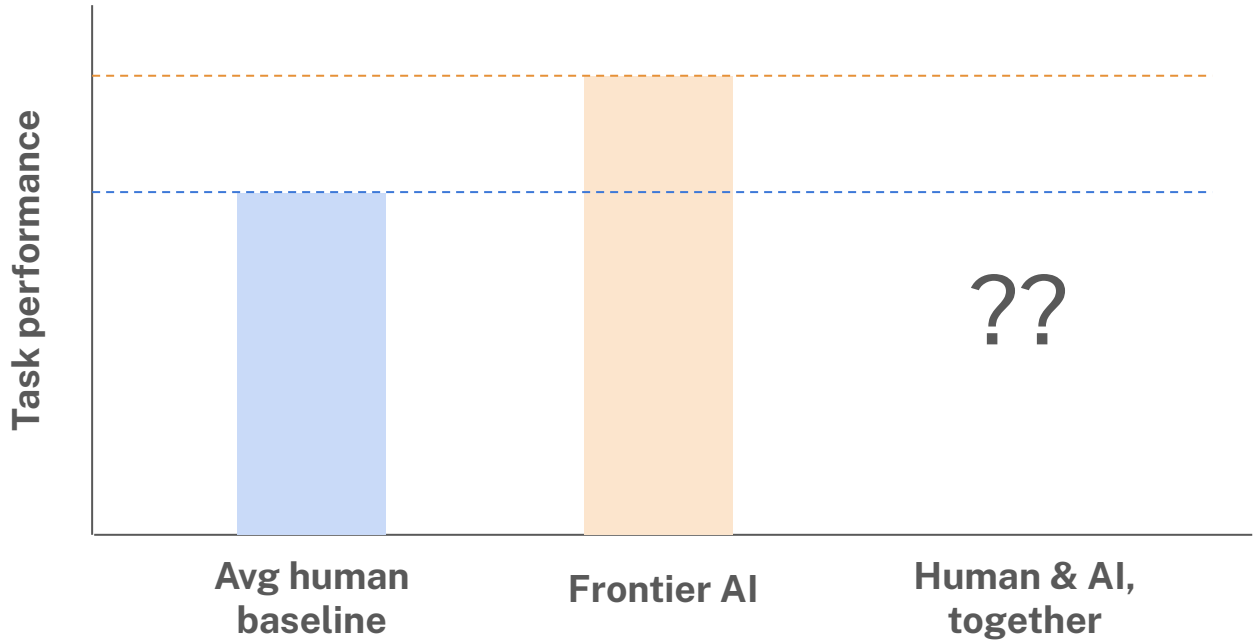


Open question: as AI becomes more capable and integrated into our daily lives, will HCI cannibalize AI?

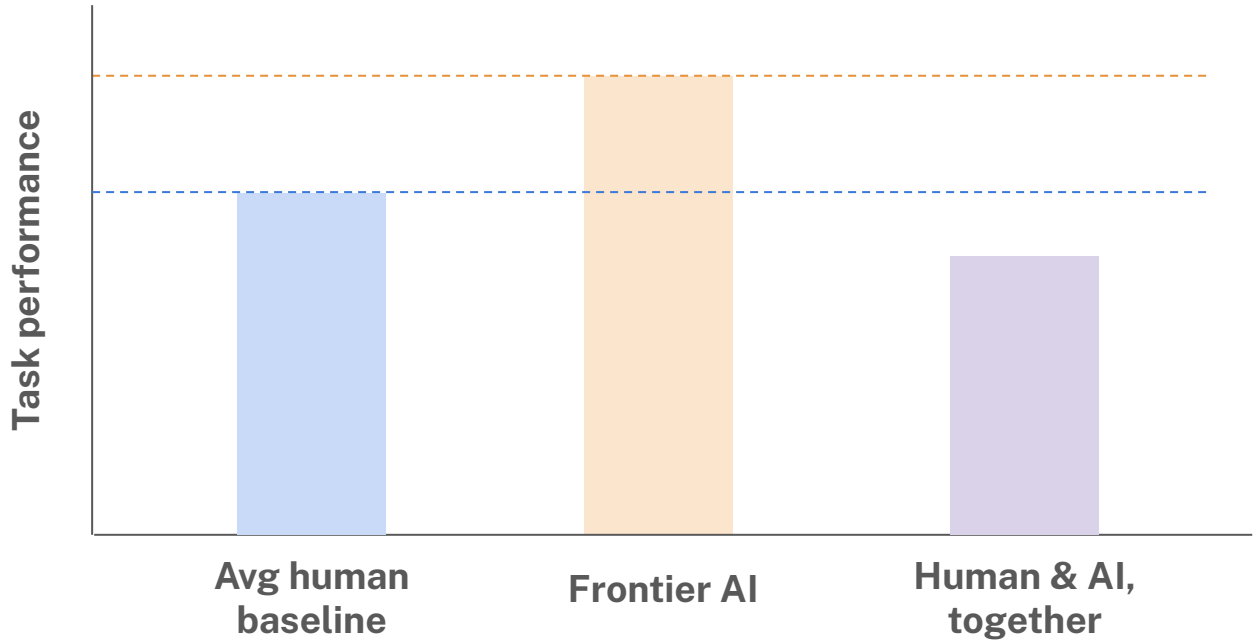
Which future?



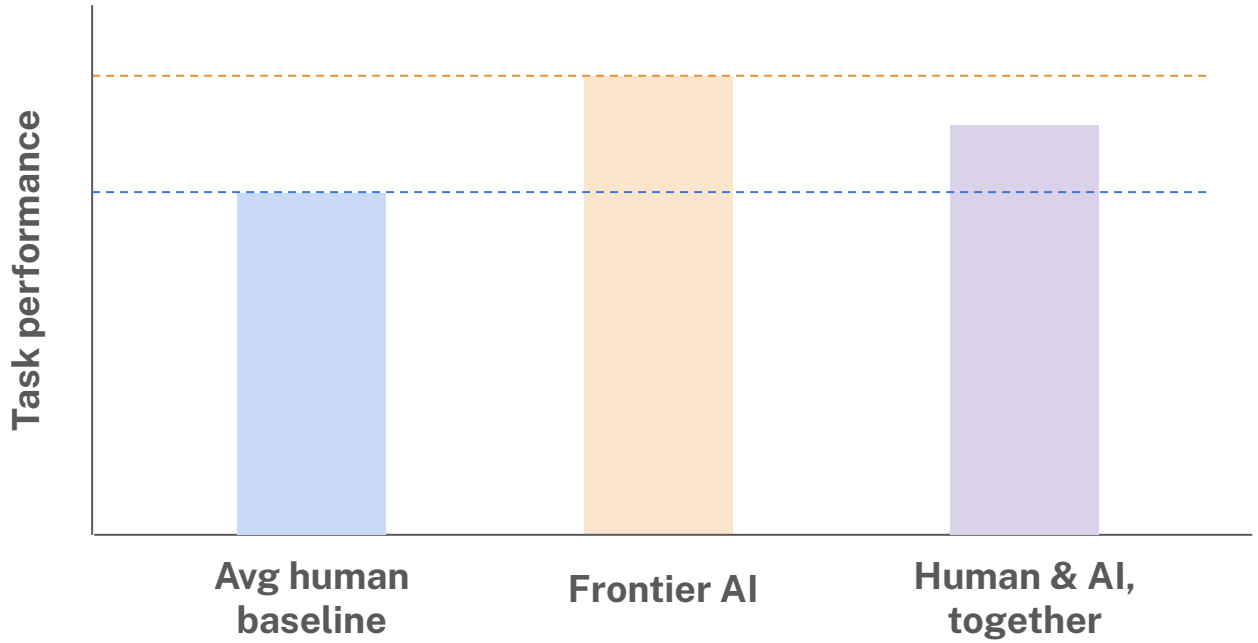
Which future?



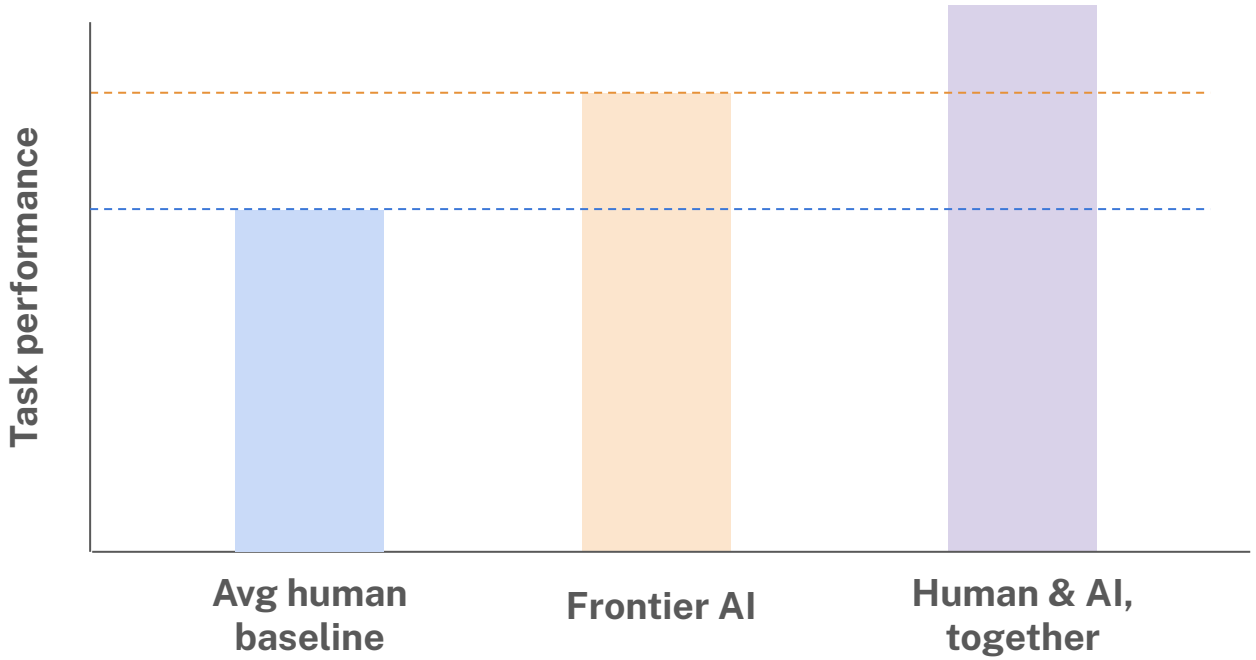
Which future?



Which future?



Which future?



Today

- What to evaluate on
- How to build a benchmark
- How to evaluate

Plus some meta-evaluation (evaluation of evaluation) throughout

Logistics

Apologies for being slow in commenting on and recording grades for presentations!

Peer feedback due tonight

More guest lectures coming up!

We have only a little over a month left together 😢