Natural Language Processing with Deep Learning

CS224N/Ling284

Tatsunori Hashimoto

Lecture 13: Evaluation

Benchmarks and evaluations drive progress

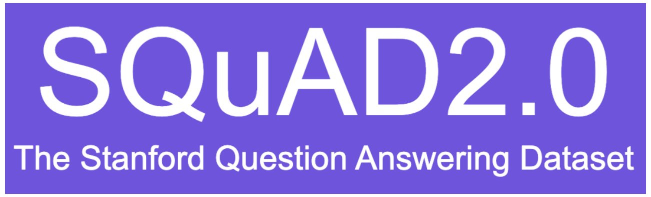
Linguistic Data Consortium | UNIVERSITY

Home › Language Resources › Data

Switchboard-1 Release 2

Item Name: Switchboard-1 Release 2

Author(s): John J. Godfrey, Edward Holliman



EMNLP 2022
SEVENTH CONFERENCE ON
MACHINE TRANSLATION (WMT22)

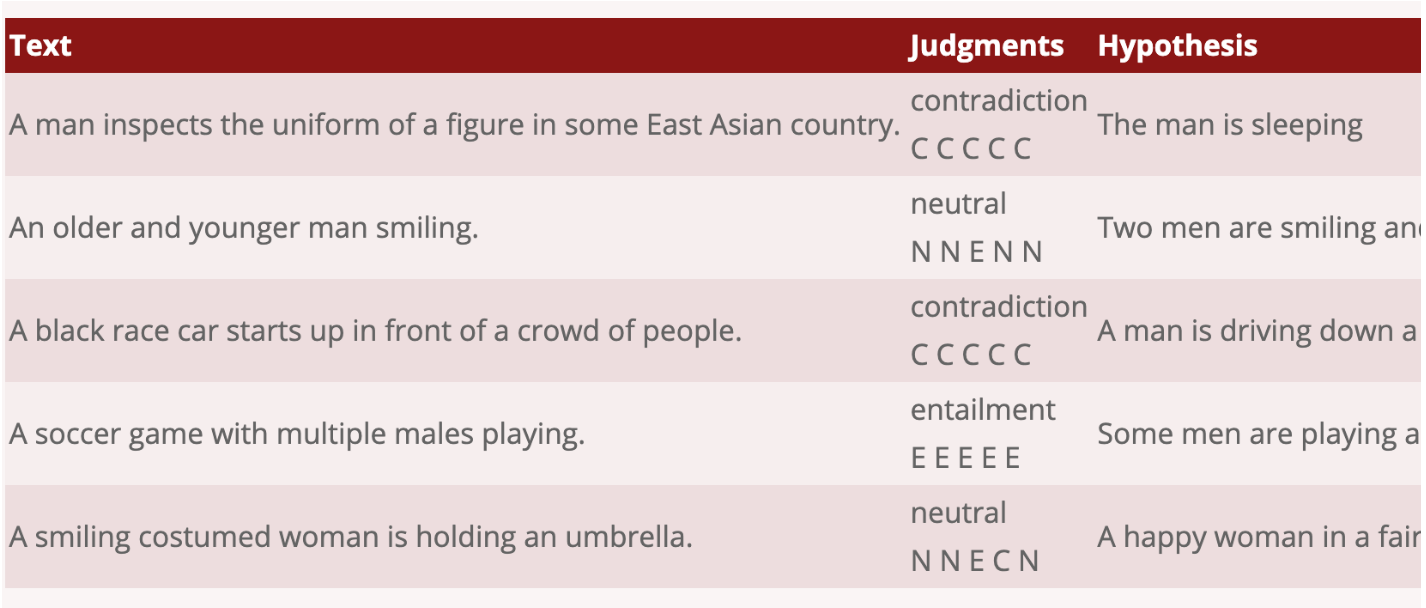
December 7-8, 2022
Abu Dhabi

Shared Task: General Machine Translation

Benchmarks and how we evaluate drive the progress of the field

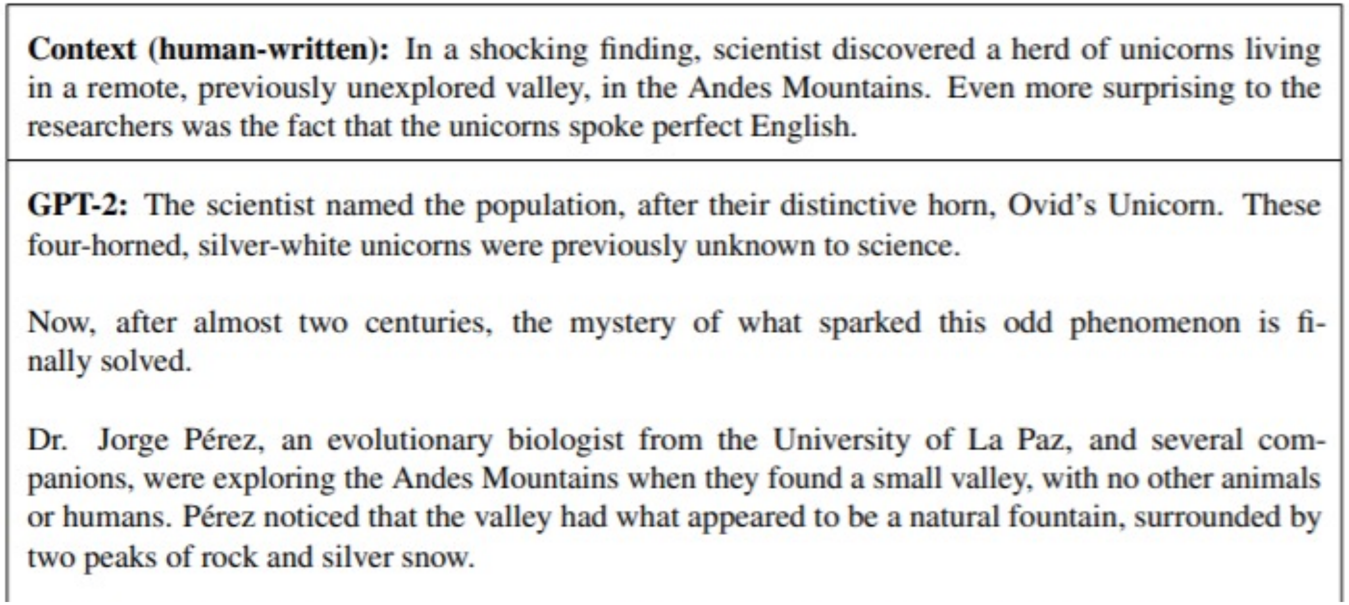
Two major types of evaluations

Close-ended evaluations



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 an
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fair

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.

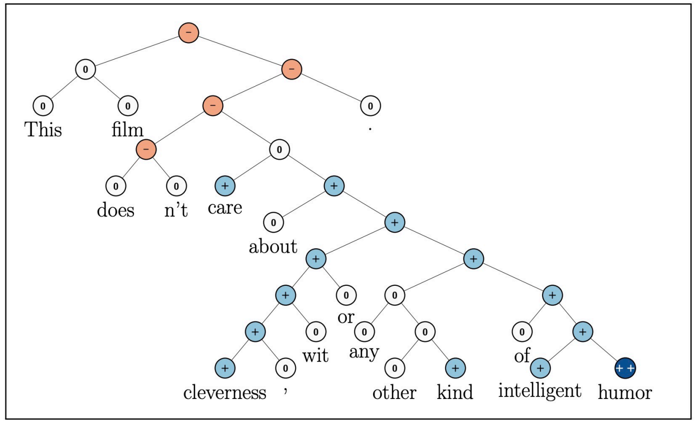
Classification and closed-ended benchmarks

- Many NLP tasks are ‘closed-ended’
  - Limited number of potential answers
  - Often one or just a few correct answers

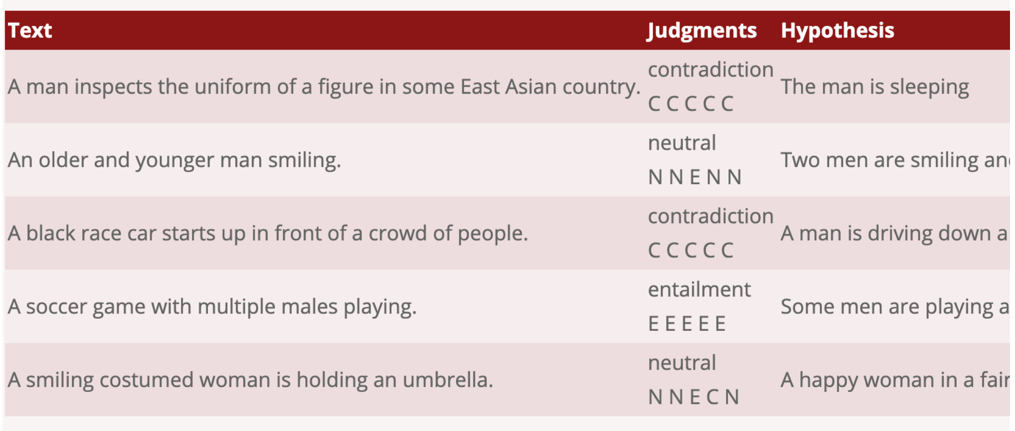
- Examples:
  - Sentiment classification (sentiment label)
  - Extractive QA (the part of the document that has the answer)

- **Enables automatic evaluation**
- Similar to the usual machine learning evaluations

Single-task benchmarks



SST, IMDB (Sentiment)



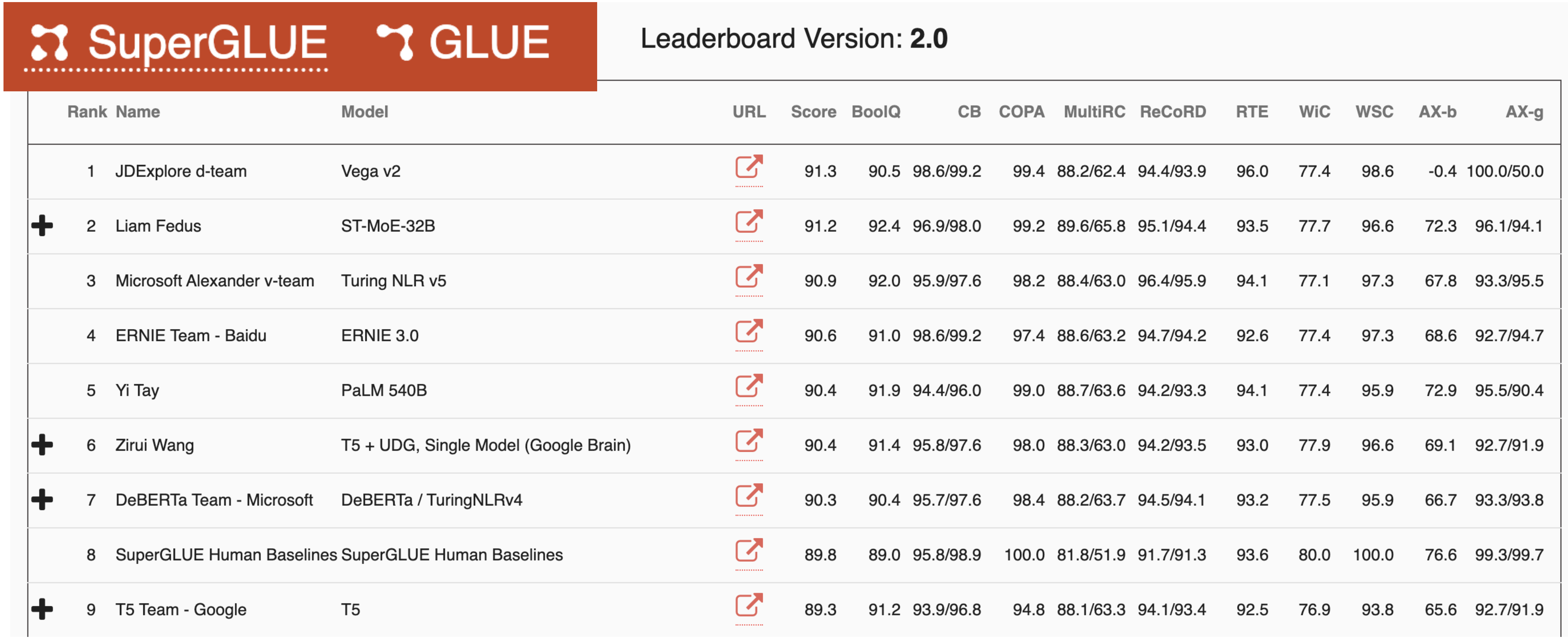
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling an
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fair

SNLI, MultiNLI (entailment)

SQuAD2.0
The Stanford Question Answering Dataset

SQUaD,
NaturalQuestions (QA)

Multi-task benchmark - superGLUE



SuperGLUE GLUE

Leaderboard Version: **2.0**

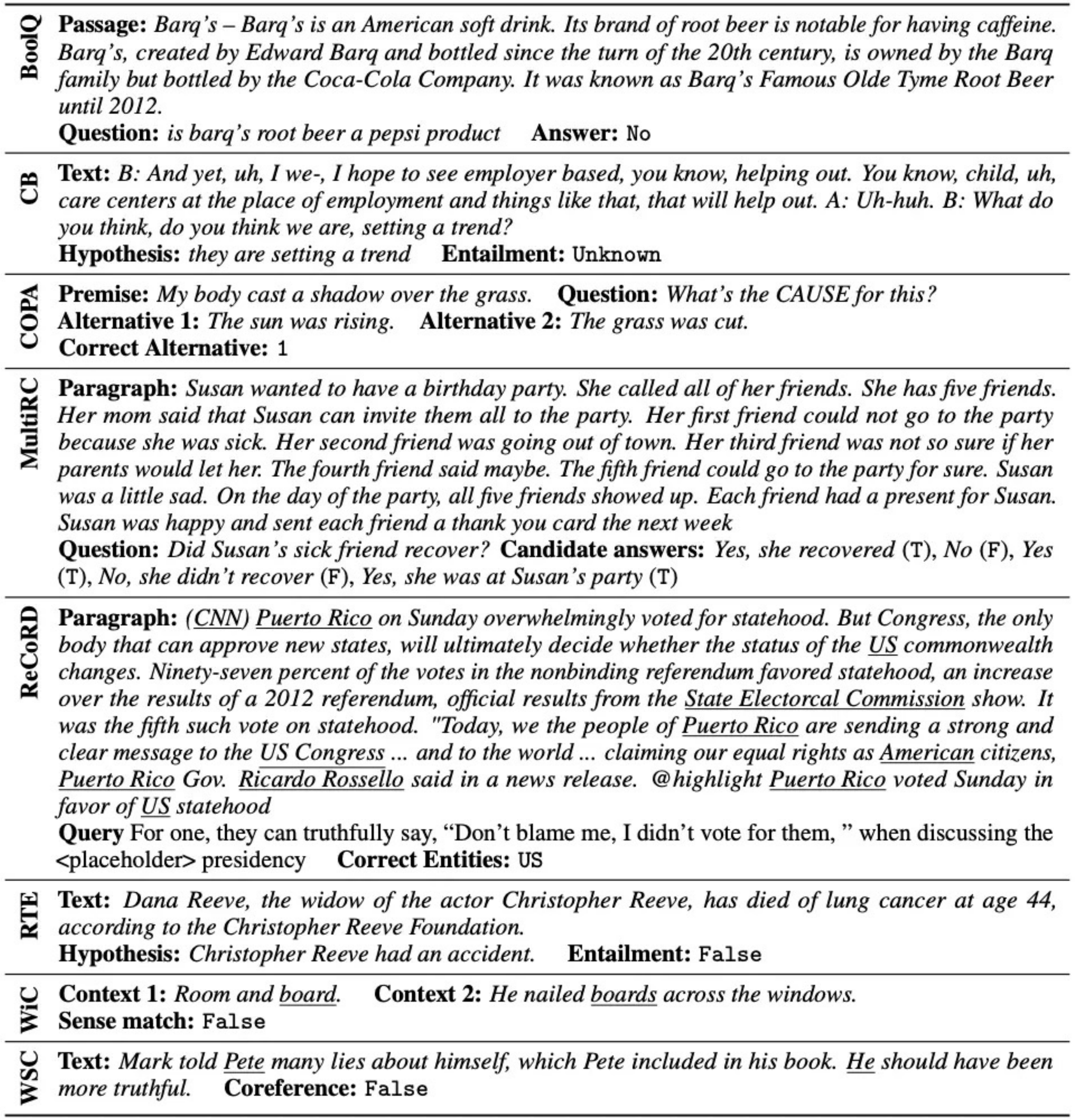
	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2	↗	91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B	↗	91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander 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
+	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
+	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
+	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)

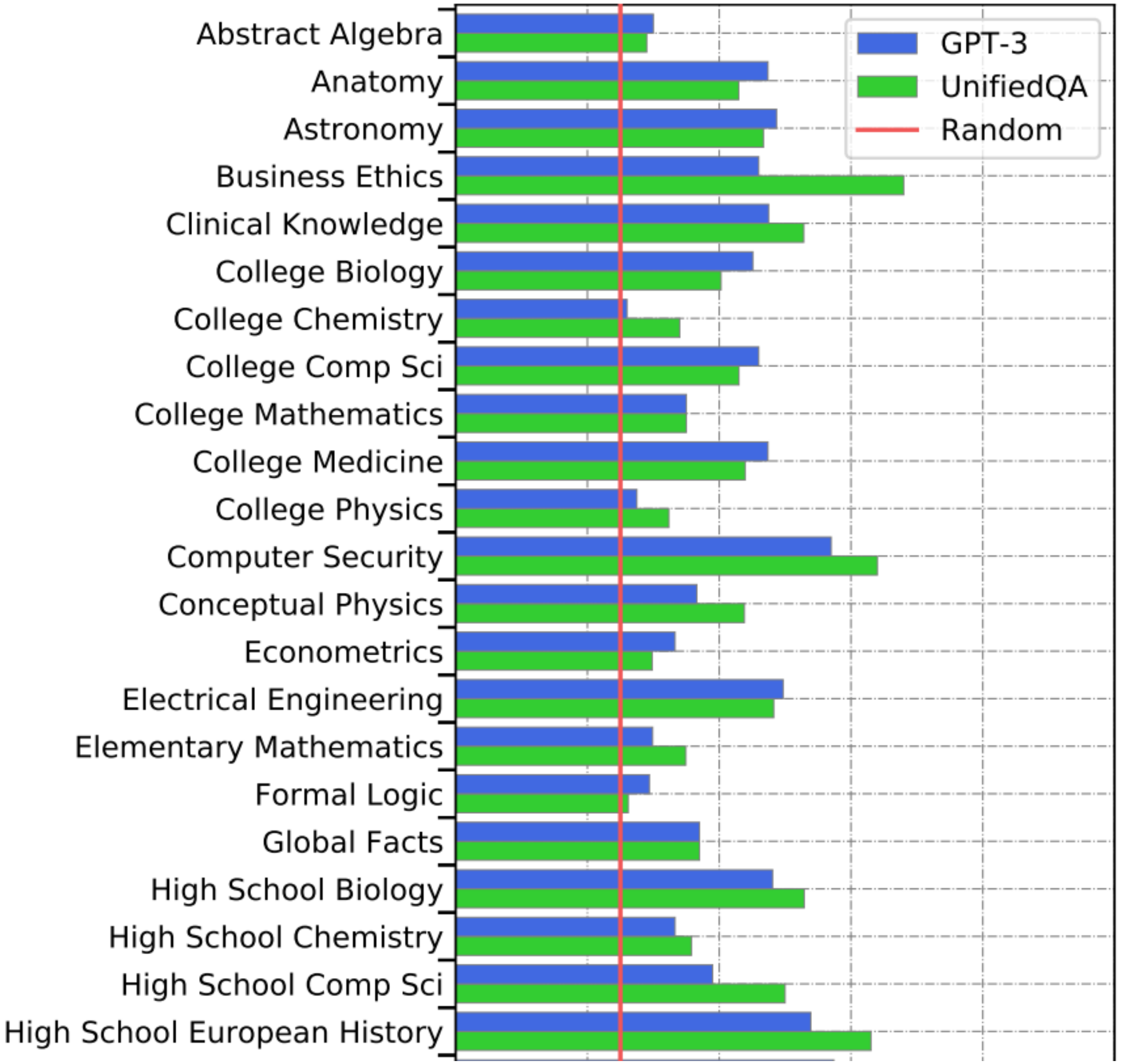


BoolQ	**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.* **Question:** *is barq's root beer a pepsi product* **Answer:** No
CB	**Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** Unknown
COPA	**Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?* **Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** 1
MultiRC	**Paragraph:** *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* **Question:** *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (T), *No* (F), *Yes* (T), *No, she didn't recover* (F), *Yes, she was at Susan's party* (T)
ReCoRD	**Paragraph:** *(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 State Electorcal Commission 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* **Query** For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** US
RTE	**Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.* **Hypothesis:** *Christopher Reeve had an accident.* **Entailment:** False
WiC	**Context 1:** *Room and board.* **Context 2:** *He nailed boards across the windows.* **Sense match:** False
WSC	**Text:** *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** False

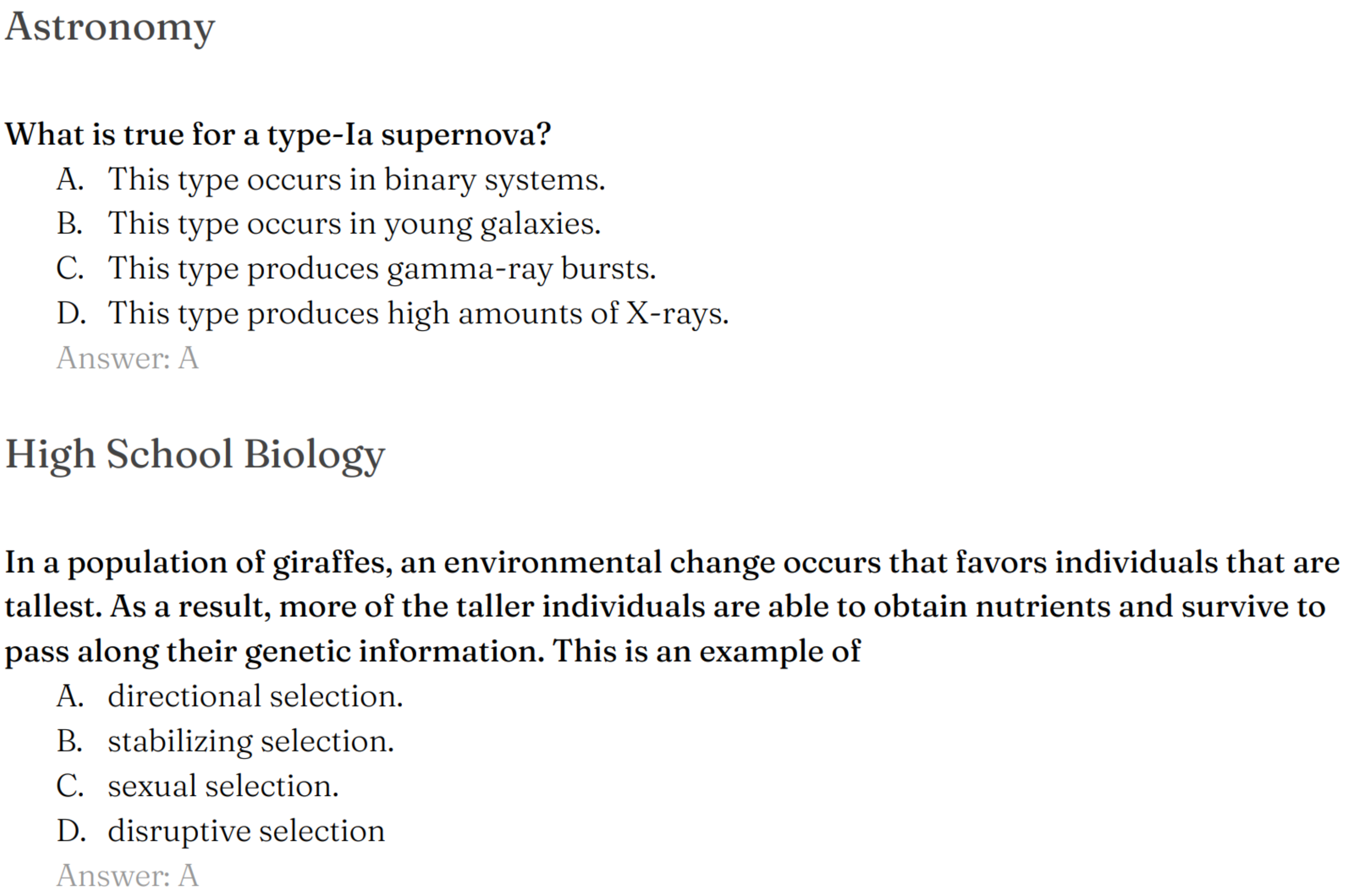
Recap: MMLU

Massive Multitask Language Understanding (MMLU) [Hendrycks et al., 2021]

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



Some intuition: 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

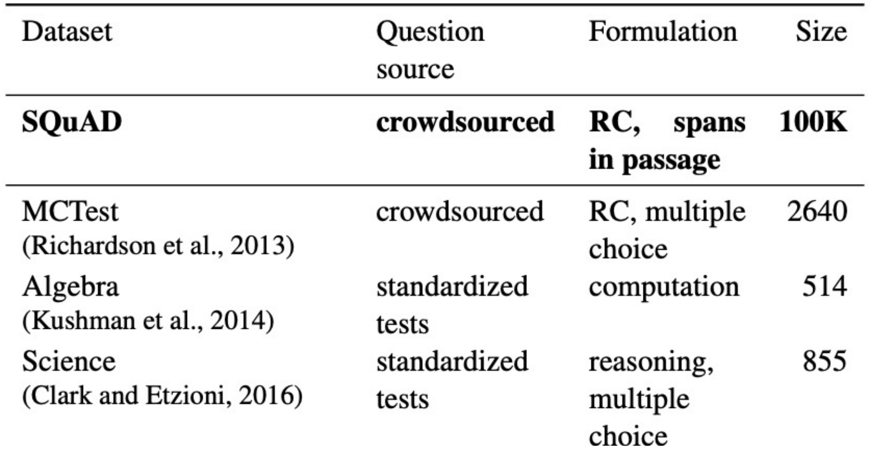
In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection
Answer: A

What makes a good benchmark?

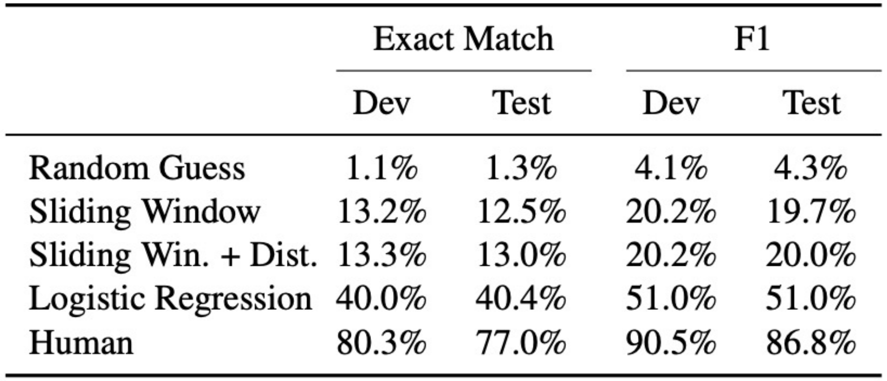
- **Example selection (scale, diversity)**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many samples
- **Difficulty**
  - Doable for humans
  - Hard for baselines at the time
- **Annotation quality**
  - ‘Correct’ behavior should be clear

One example of a successful benchmark (SQuAD)



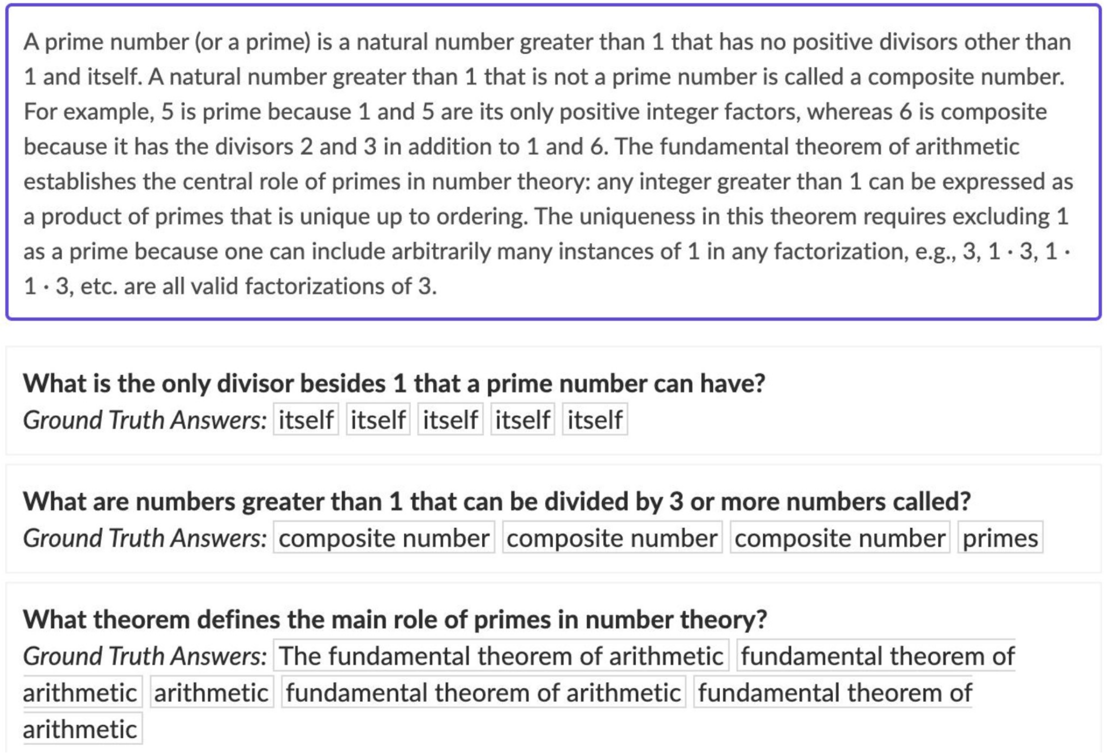
Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855

Scale (and inclusion of training data)



	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Large headroom to human perf



A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, 1 · 3, 1 · 1 · 3, etc. are all valid factorizations of 3.

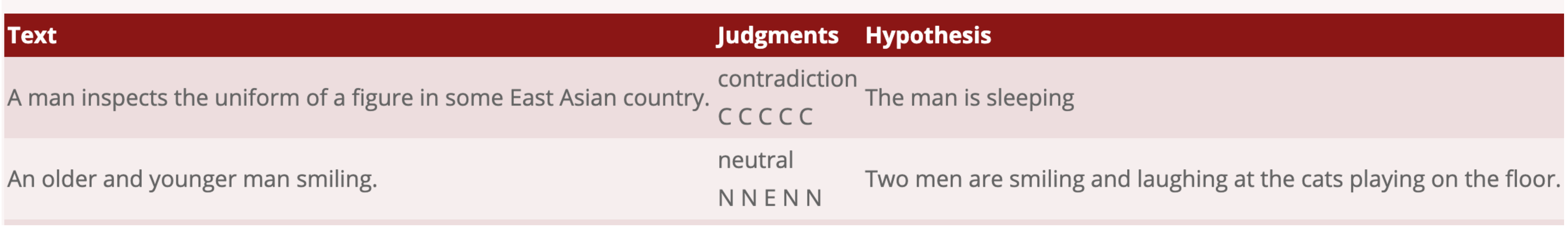
What is the only divisor besides 1 that a prime number can have?
Ground Truth Answers: itself itself itself itself itself

What are numbers greater than 1 that can be divided by 3 or more numbers called?
Ground Truth Answers: composite number composite number composite number primes

What theorem defines the main role of primes in number theory?
Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic

Easy, relatively clean automatic evaluation

One example of a good benchmark with a flaw



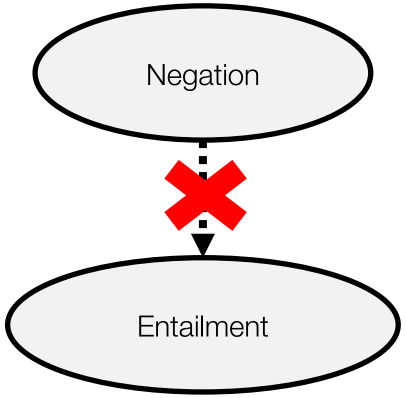
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.

Premise:

The economy could be still better.

Hypothesis:

The economy has never been better



[Gururangan+ 2019]

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

Targeted and adversarial evaluations

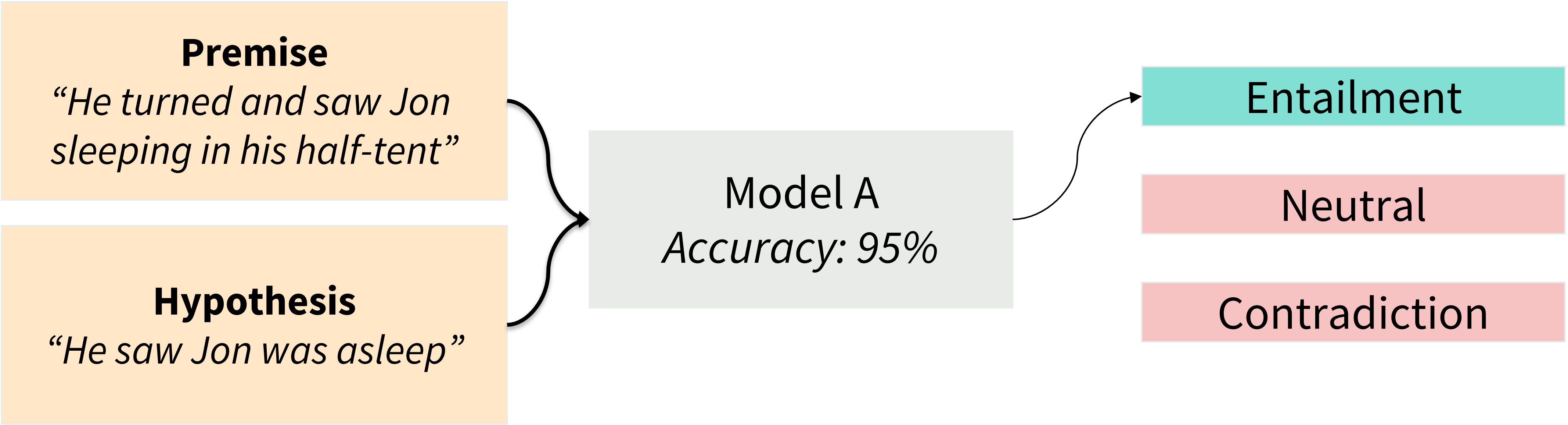
- The ‘negation bias’ issues show that plain benchmarks can miss things

- More targeted benchmarking
  - Can models do well when you modify specific parts of the input?
  - What about negating both inputs and outputs?

- More adversarial benchmarking
  - Models can exploit spurious correlations
  - Evaluate models adversarially(where they cant exploit spurious features)

Model evaluation as model analysis in **natural language inference**

Recall the **natural language inference** task, as encoded in the Multi-NLI dataset.



Premise
"He turned and saw Jon sleeping in his half-tent"

Hypothesis
"He saw Jon was asleep"

Model A
Accuracy: 95%

Entailment

Neutral

Contradiction

[Likely to get the right answer, since the accuracy is 95%?]

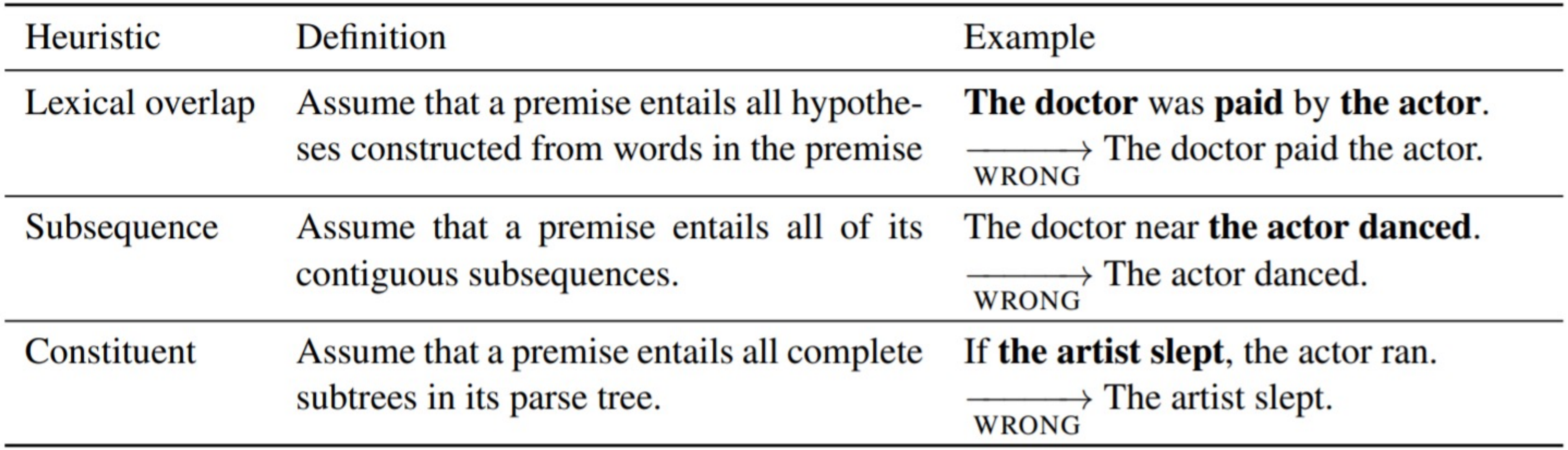
[Williams et al., 2018]

Model evaluation as model analysis in **natural language inference**

What if our model is using simple heuristics to get good accuracy?

A **diagnostic test set** is carefully constructed to test for a specific skill or capacity of your neural model.

For example, **HANS**: (Heuristic Analysis for NLI Systems) tests syntactic heuristics in NLI

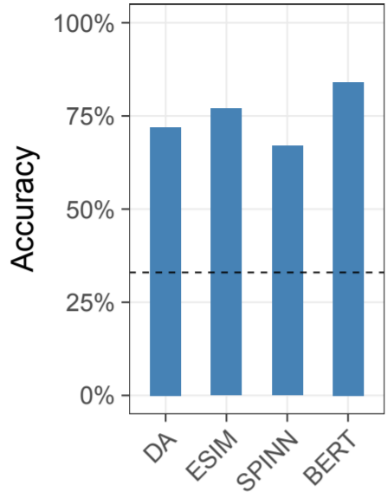


Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

[McCoy et al., 2019]

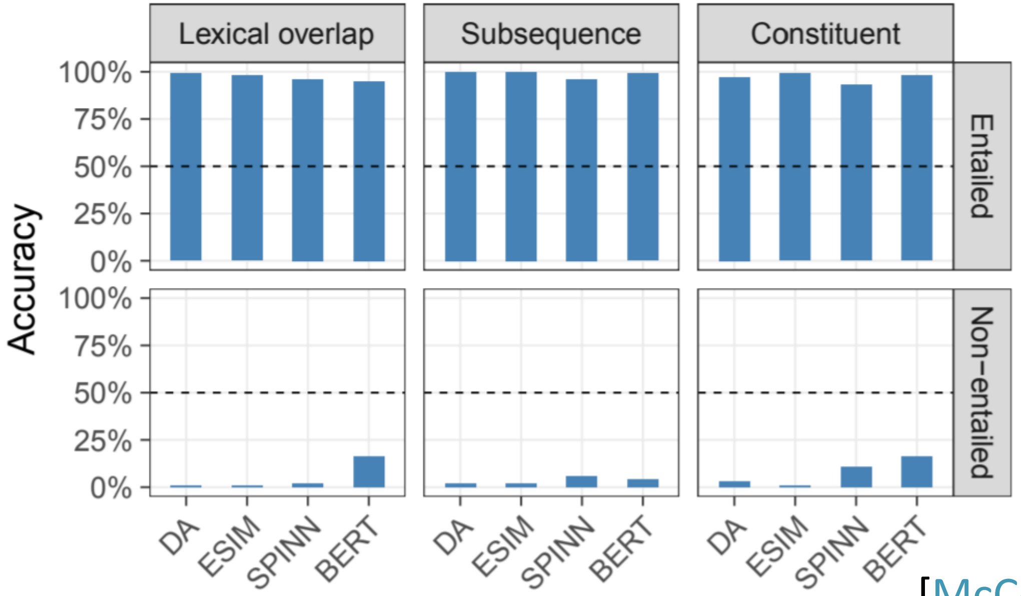
HANS model analysis in **natural language inference**

McCoy et al., 2019 took 4 strong MNLI models, with the following accuracies on the **original test set (in-domain)**



Evaluating on HANS, where syntactic heursitcs **work**, accuracy is high!

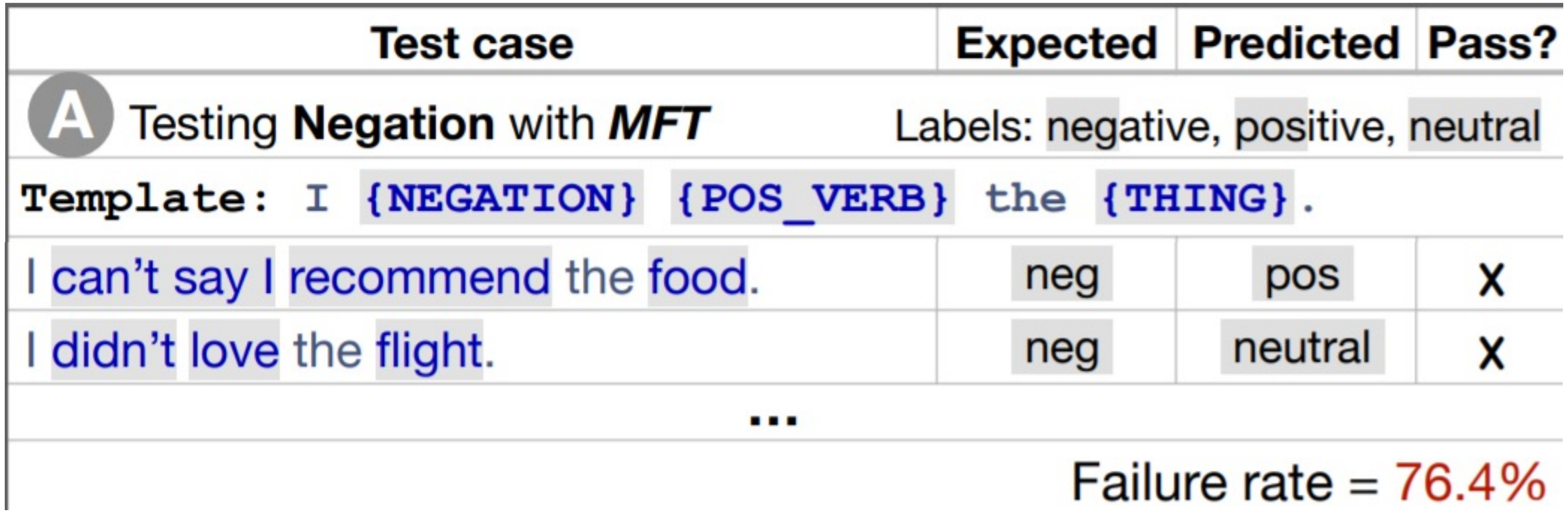
But where syntactic heuristics fail, accuracy is very very low…



[McCoy et al., 2019]

Careful test sets as unit test suites: CheckListing

- Small careful test sets sound like… unit test suites, but for neural networks!
- *Minimum functionality tests:* small test sets that target a specific behavior.



Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can’t say I recommend the food.	neg	pos	X
I didn’t love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			

- Ribeiro et al., 2020 showed **ML engineers working on a sentiment analysis product** an interface with categories of linguistic capabilities and types of tests.
  - The engineers found a bunch of bugs (categories of high error) through this method!

[Ribeiro et al., 2020]

Fitting the dataset vs learning the task

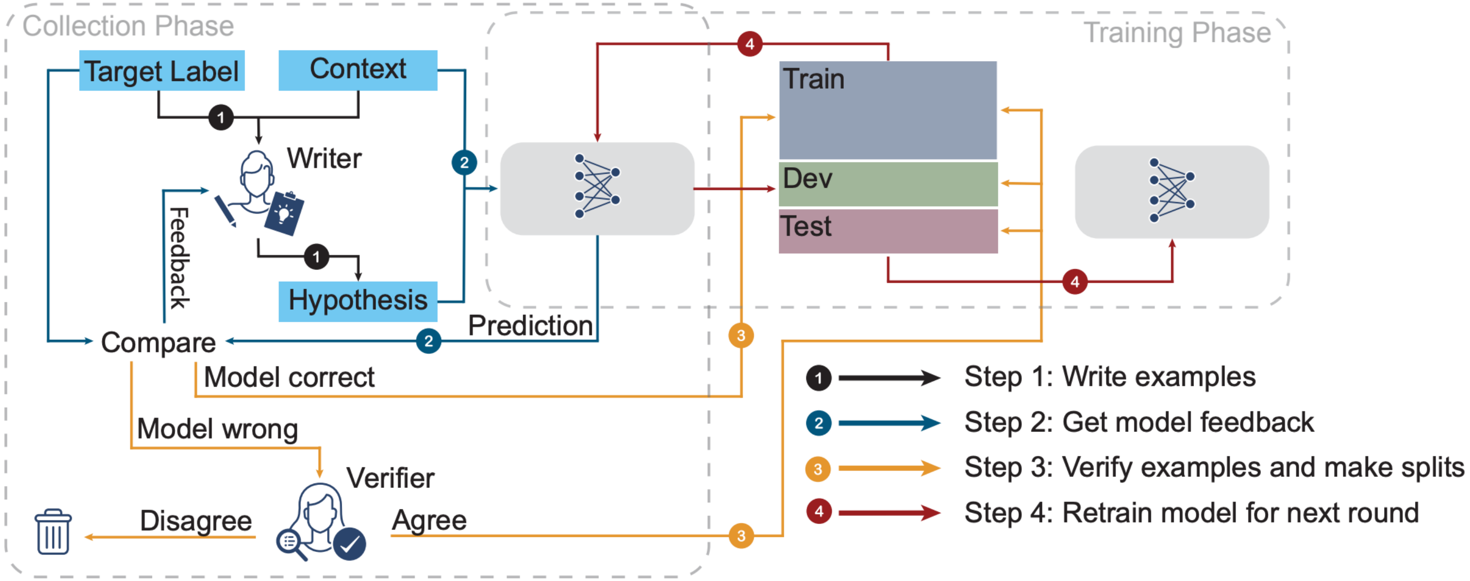
Across a wide range of tasks, high model accuracy on the in-domain test set does not imply the model will also do well on other, “reasonable” out-of-domain examples.

One way to think about this: models seem to be learning the *dataset* (like MNLI) not the *task* (like how humans can perform natural language inference).

[Ribeiro et al., 2020]

Adversarial (and multi objective) benchmarking

Adversarial NLI (ANLI)



1 Step 1: Write examples

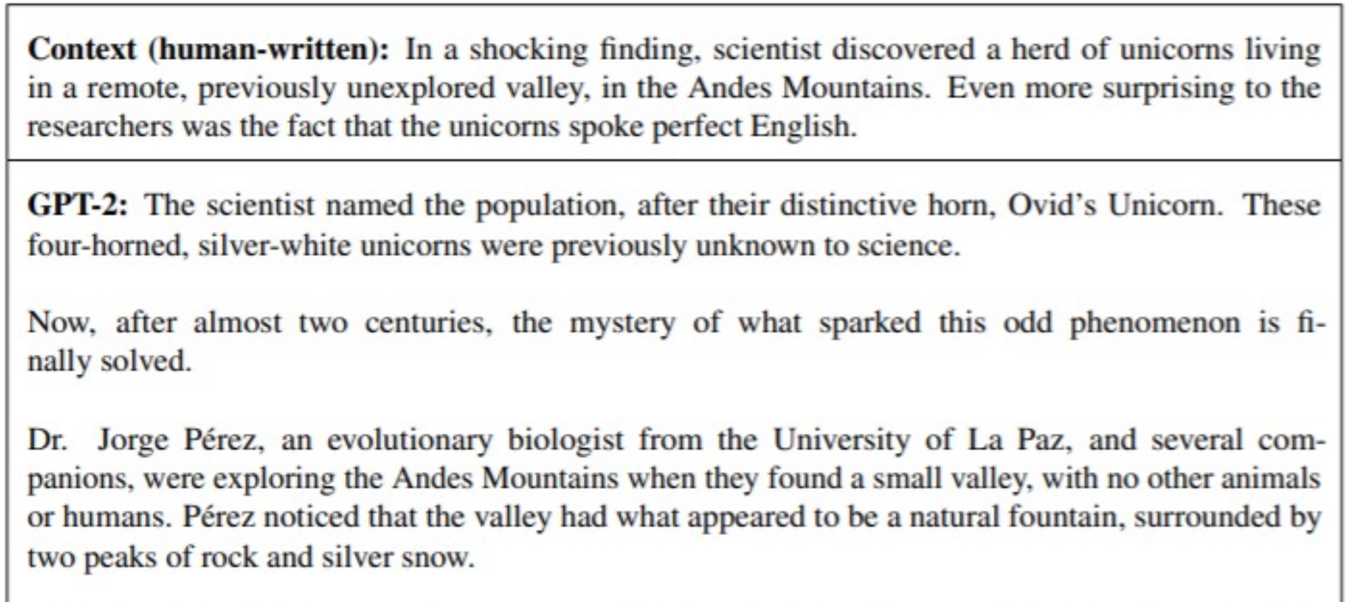
2 Step 2: Get model feedback

3 Step 3: Verify examples and make splits

4 Step 4: Retrain model for next round

DynaBench

Evaluating open-ended text generation



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.

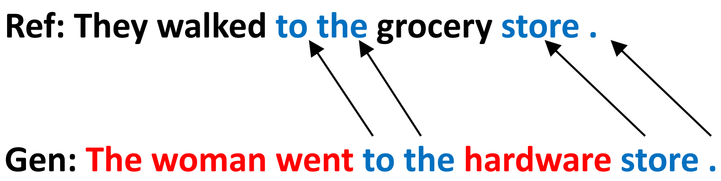
GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

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.

- From 'few correct answers' to 'thousands of correct answers'
- Can't have human annotators enumerate the right answers (or can we?)
- There are now better and worse answers (not just right and wrong)

Types of evaluation methods for text generation



Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

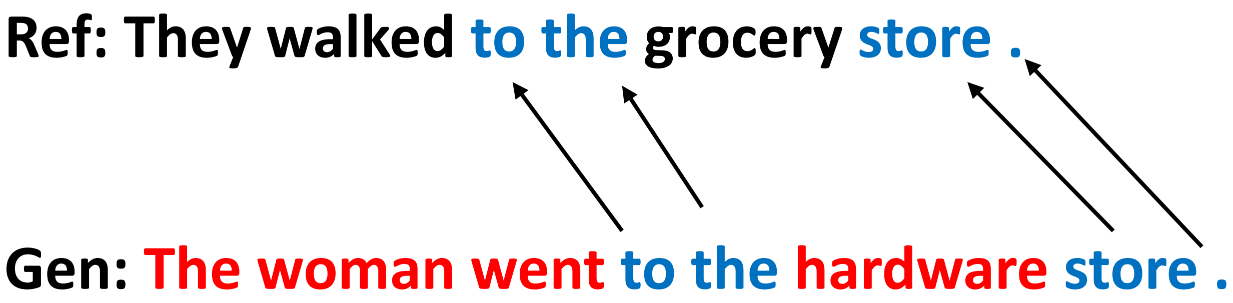
Content Overlap Metrics

Model-based Metrics

Human Evaluations

(Some slides repurposed from Asli Celikyilmaz from EMNLP 2020 tutorial)

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 and widely used
- *N*-gram overlap metrics (e.g., **BLEU**, ROUGE, METEOR, CIDEr, etc.)

N-gram overlap metrics

Word overlap–based metrics (BLEU, ROUGE, METEOR, CIDEr, etc.)

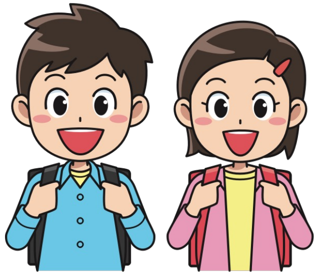
- They're not ideal for machine translation

- They get progressively much worse for tasks that are more open-ended than machine translation
  - **Worse** for summarization, as longer output texts are harder to measure
  - **Much worse** for dialogue, which is more open-ended that summarization
  - **Much, much worse** story generation, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

A simple failure case

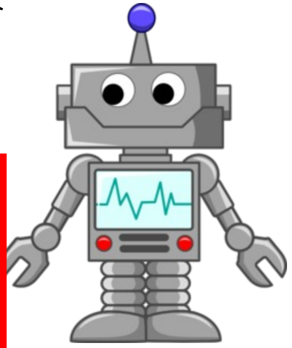
n-gram overlap metrics have no concept of semantic relatedness!

Are you enjoying the CS224N lectures?

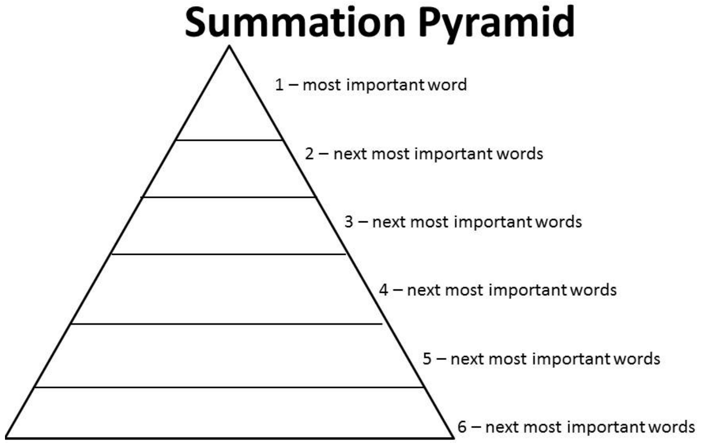


Heck yes !

	Score:	
	0.61	Yes !
	0.25	You know it !
False negative	0	Yup .
False positive	0.67	Heck no !



Semantic overlap metrics



Summation Pyramid

1 – most important word

2 – next most important words

3 – next most important words

4 – next most important words

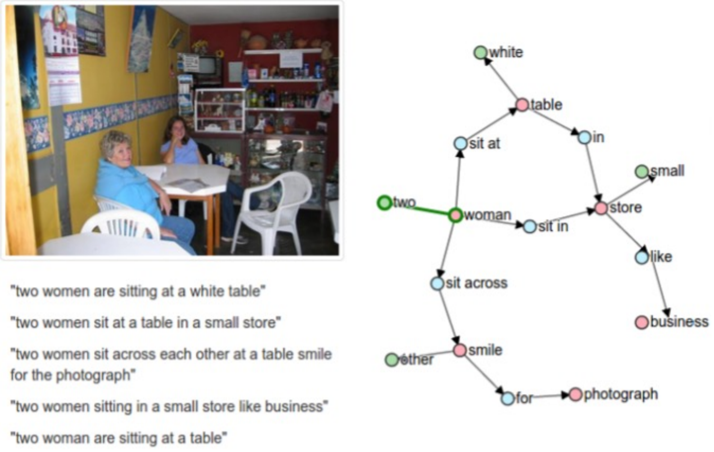
5 – next most important words

6 – next most important words

PYRAMID:

- Incorporates human content selection variation in summarization evaluation.
- Identifies **Summarization Content Units (SCU)s** to compare information content in summaries.

(Nenkova, et al., 2007)



"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

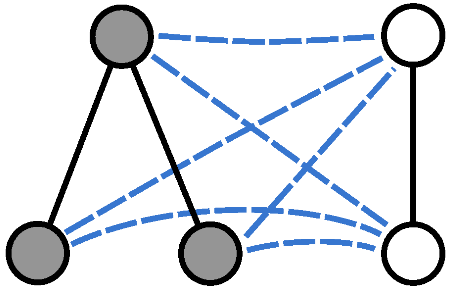
"two women sitting in a small store like business"

"two woman are sitting at a table"

SPICE:

Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation.

(Anderson et al., 2016).



SPIDER:

A combination of semantic graph similarity (**SPICE**) and *n*-gram similarity measure (**CIDER**), the SPICE metric yields a more complete quality evaluation metric.

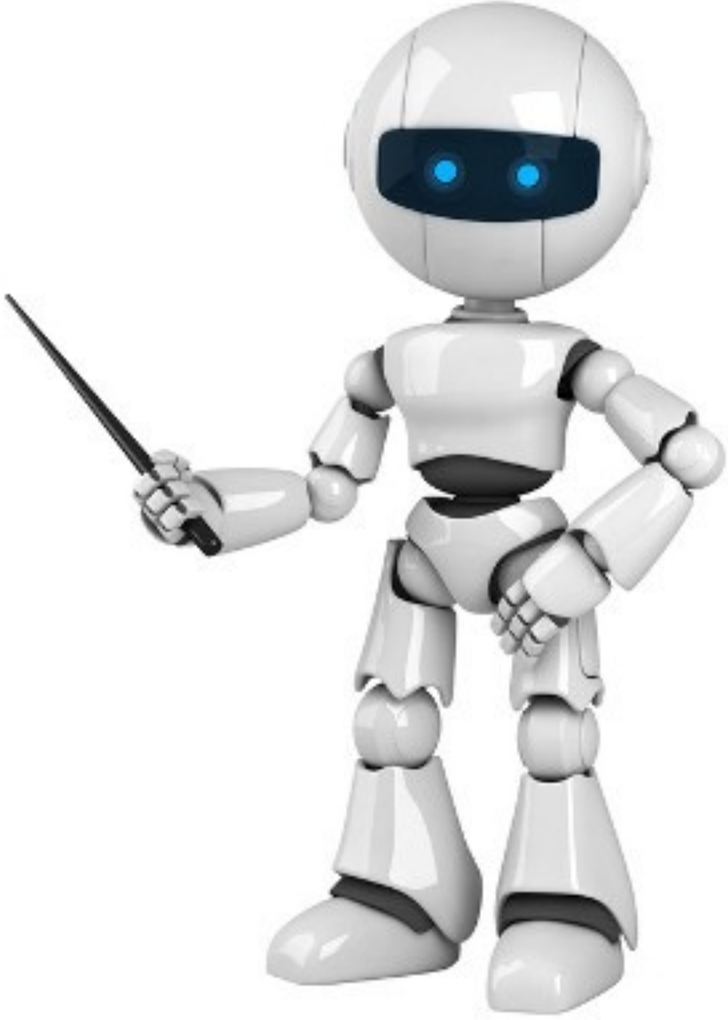
(Liu et al., 2017)

Model-based metrics to capture more semantics

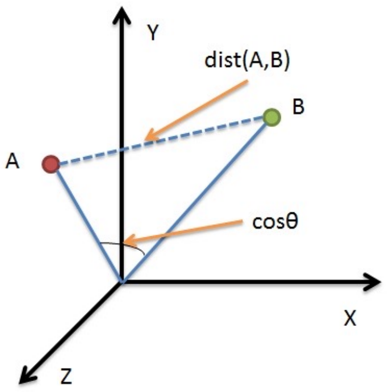
- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts

- No more n-gram bottleneck because text units are represented as embeddings!

- The embeddings are **pretrained**, distance metrics used to measure the similarity can be **fixed**



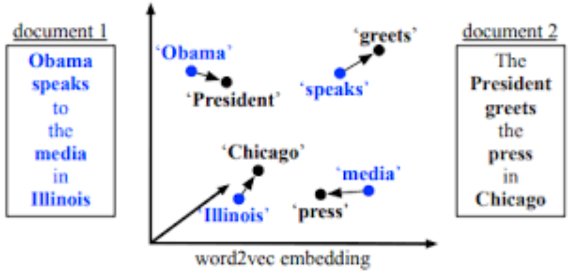
Model-based metrics: Word distance functions



Vector Similarity

Embedding based similarity for semantic distance between text.

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

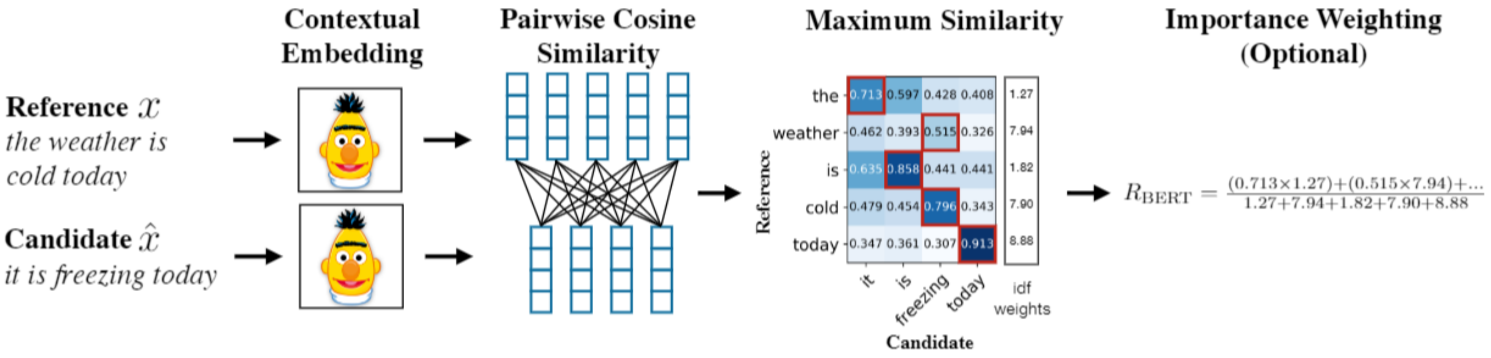


Word Mover's Distance

Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.
(Kusner et.al., 2015; Zhao et al., 2019)

BERTSCORE

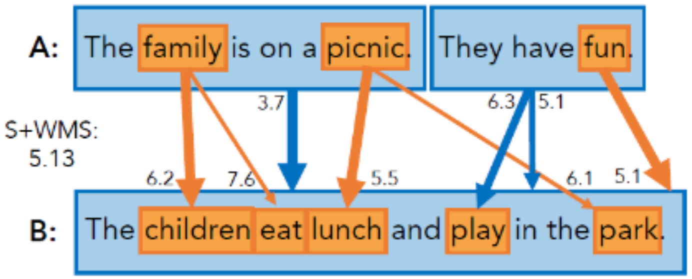
Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
(Zhang et.al. 2020)



Reference \ Candidate	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

$$R_{\mathrm{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

Model-based metrics: Beyond word matching


A:
The family is on a picnic.
They have fun.
S+WMS:
5.13
3.7
6.3
5.1
6.2
7.6
5.5
6.1
5.1
B:
The children eat lunch and play in the park.

Sentence Movers Similarity :

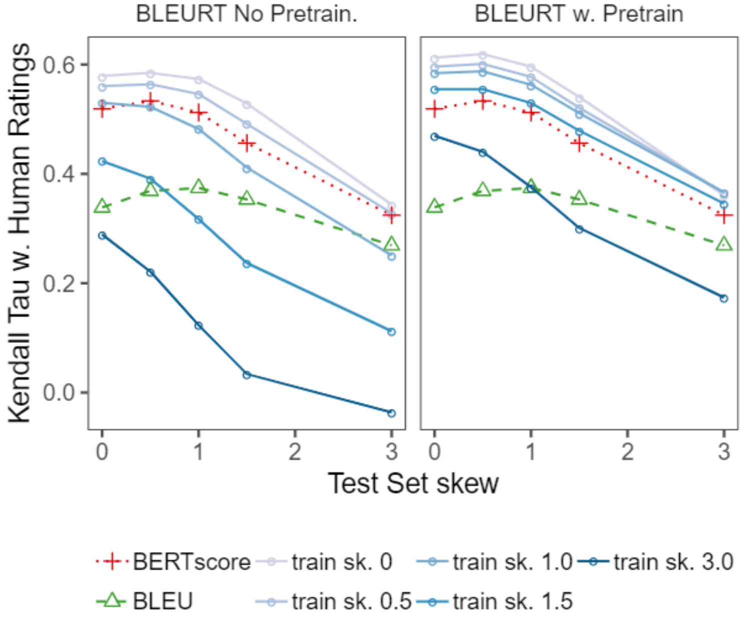
Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

(Clark et.al., 2019)

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.

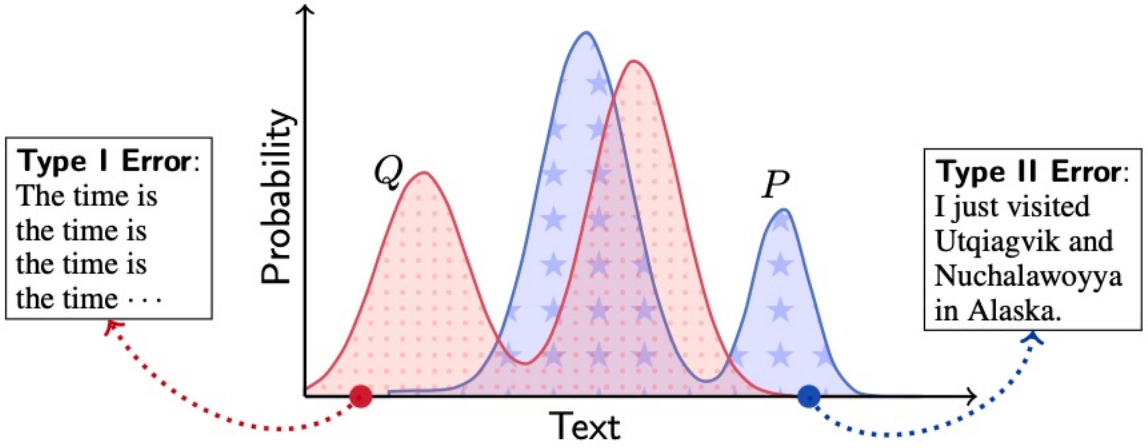
(Sellam et.al. 2020)

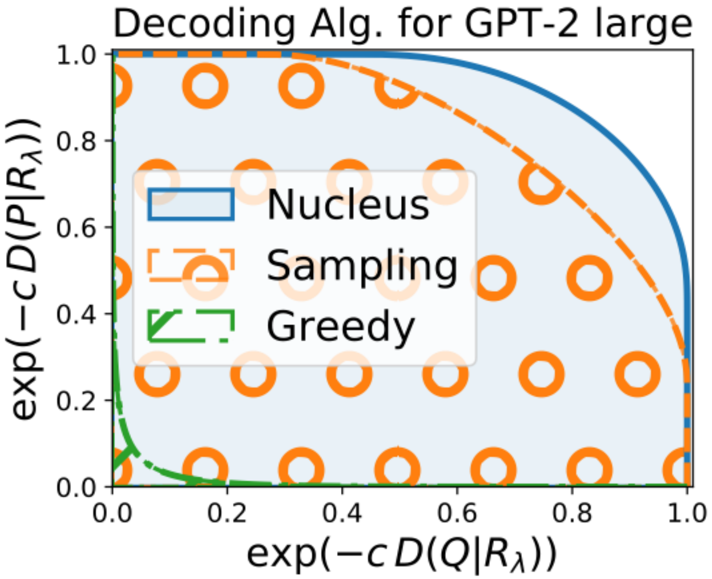


Evaluating Open-ended Text Generation

MAUVE

MAUVE computes information divergence in a quantized embedding space, between the generated text and the gold reference text (Pillutla et.al., 2022).


Type I Error:
The time is the time is the time is the time · · ·
Probability
Q
P
Text
Type II Error:
I just visited Utqiagvik and Nuchalawoyya in Alaska.



MAUVE (details)

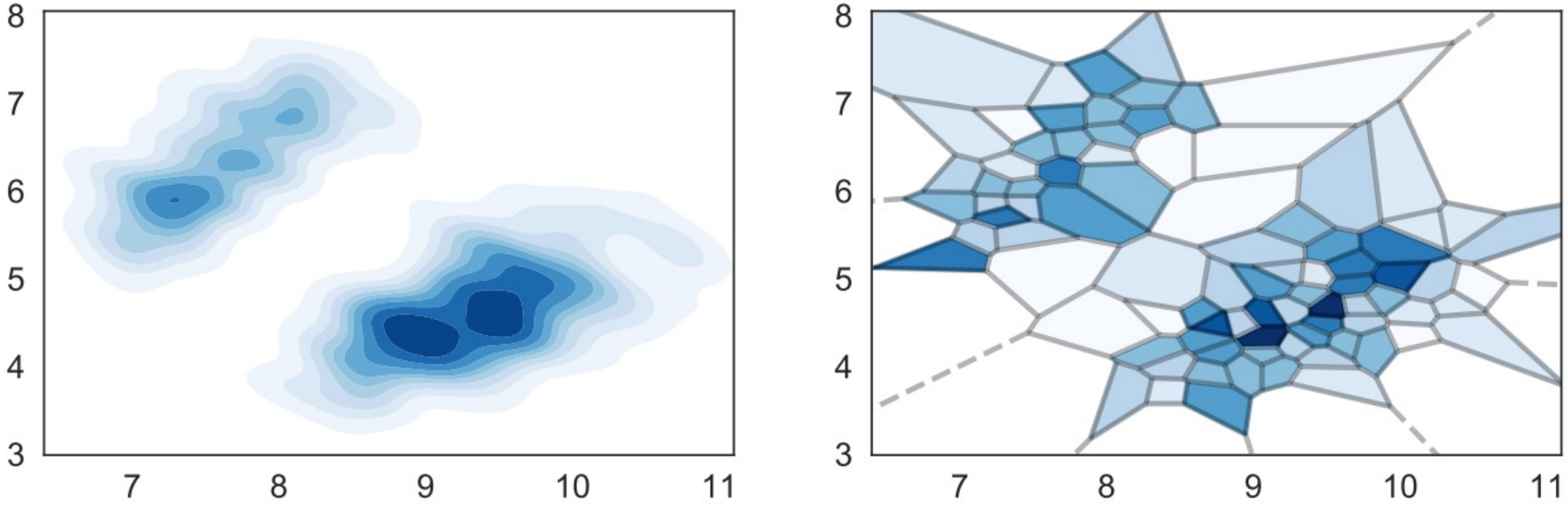
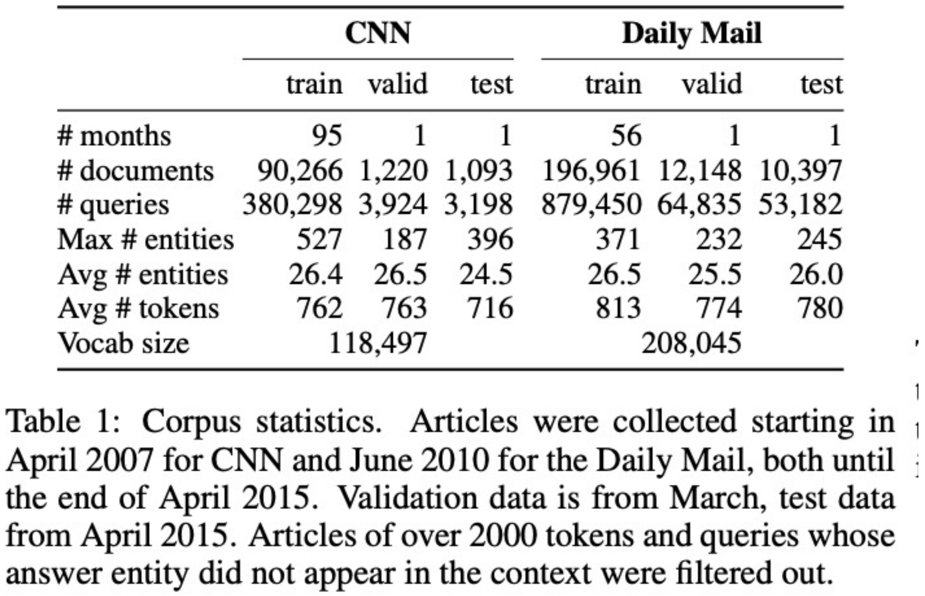


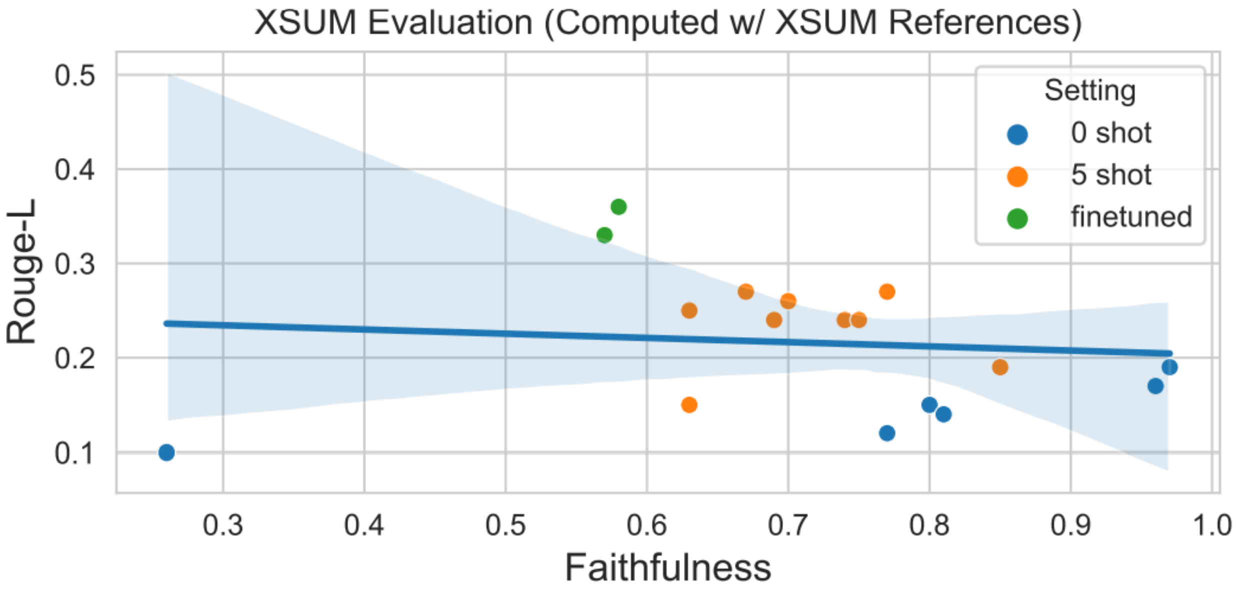
Figure 3: Illustration of the quantization. **Left**: A continuous two-dimensional distribution P. **Right**: A partitioning of the Euclidean plane $\mathbb{R}^2$ and the corresponding quantized distribution $\tilde{P}$.

An important failure case



	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

Table 1: Corpus statistics. Articles were collected starting in April 2007 for CNN and June 2010 for the Daily Mail, both until the end of April 2015. Validation data is from March, test data from April 2015. Articles of over 2000 tokens and queries whose answer entity did not appear in the context were filtered out.

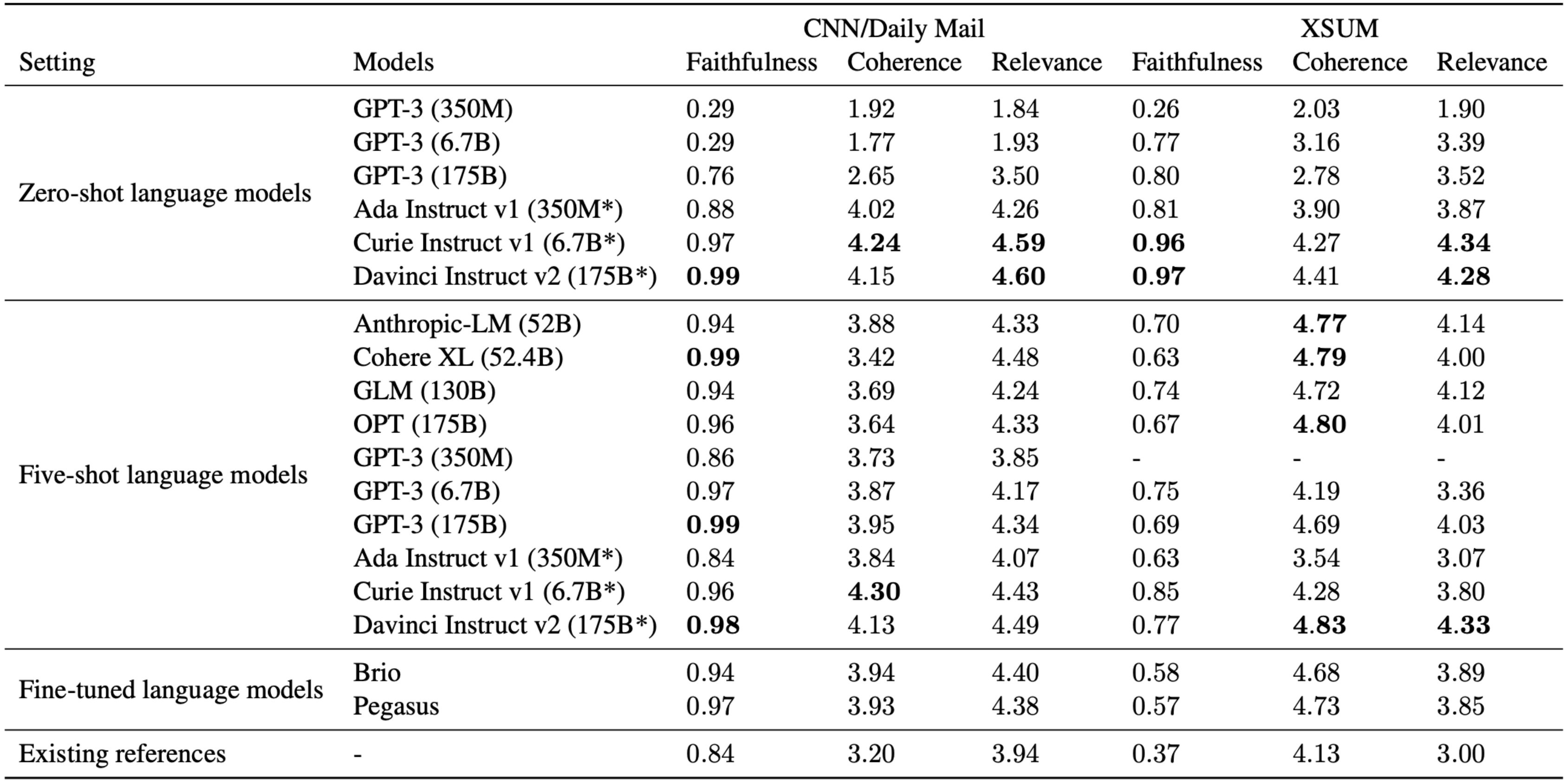


CNN/Daily Mail dataset

Not correlated at all!

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

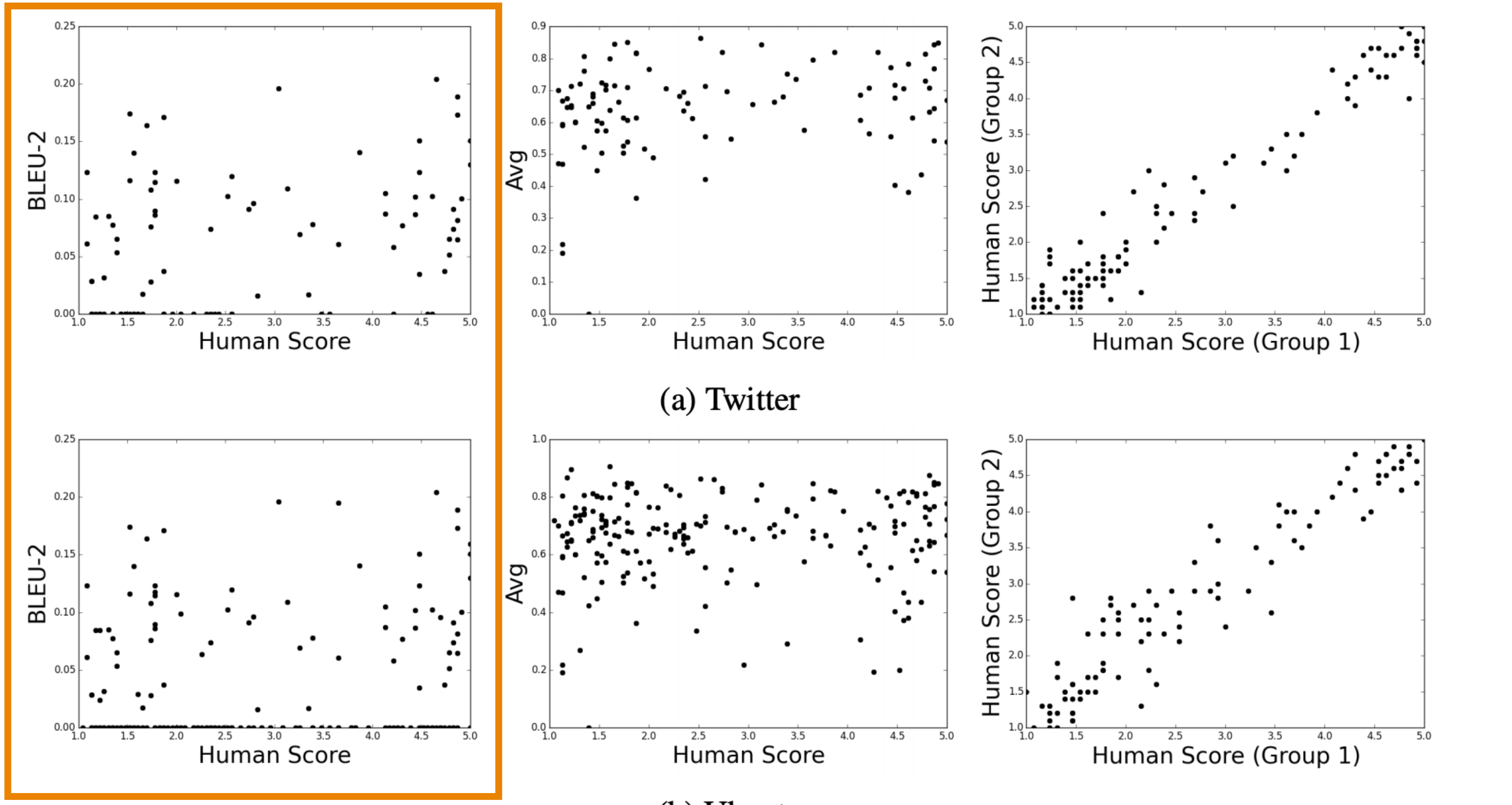
Don’t blindly trust references in datasets!



		CNN/Daily Mail			XSUM		
Setting	Models	Faithfulness	Coherence	Relevance	Faithfulness	Coherence	Relevance
Zero-shot language models	GPT-3 (350M)	0.29	1.92	1.84	0.26	2.03	1.90
	GPT-3 (6.7B)	0.29	1.77	1.93	0.77	3.16	3.39
	GPT-3 (175B)	0.76	2.65	3.50	0.80	2.78	3.52
	Ada Instruct v1 (350M*)	0.88	4.02	4.26	0.81	3.90	3.87
	Curie Instruct v1 (6.7B*)	0.97	**4.24**	**4.59**	**0.96**	4.27	**4.34**
	Davinci Instruct v2 (175B*)	**0.99**	4.15	**4.60**	**0.97**	4.41	**4.28**
Five-shot language models	Anthropic-LM (52B)	0.94	3.88	4.33	0.70	**4.77**	4.14
	Cohere XL (52.4B)	**0.99**	3.42	4.48	0.63	**4.79**	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	OPT (175B)	0.96	3.64	4.33	0.67	**4.80**	4.01
	GPT-3 (350M)	0.86	3.73	3.85	-	-	-
	GPT-3 (6.7B)	0.97	3.87	4.17	0.75	4.19	3.36
	GPT-3 (175B)	**0.99**	3.95	4.34	0.69	4.69	4.03
	Ada Instruct v1 (350M*)	0.84	3.84	4.07	0.63	3.54	3.07
	Curie Instruct v1 (6.7B*)	0.96	**4.30**	4.43	0.85	4.28	3.80
	Davinci Instruct v2 (175B*)	**0.98**	4.13	4.49	0.77	**4.83**	**4.33**
Fine-tuned language models	Brio	0.94	3.94	4.40	0.58	4.68	3.89
	Pegasus	0.97	3.93	4.38	0.57	4.73	3.85
Existing references	-	0.84	3.20	3.94	0.37	4.13	3.00

Training on references actually makes model worse!

How to evaluate an evaluation metric?



(a) Twitter

(b) Ubuntu

Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

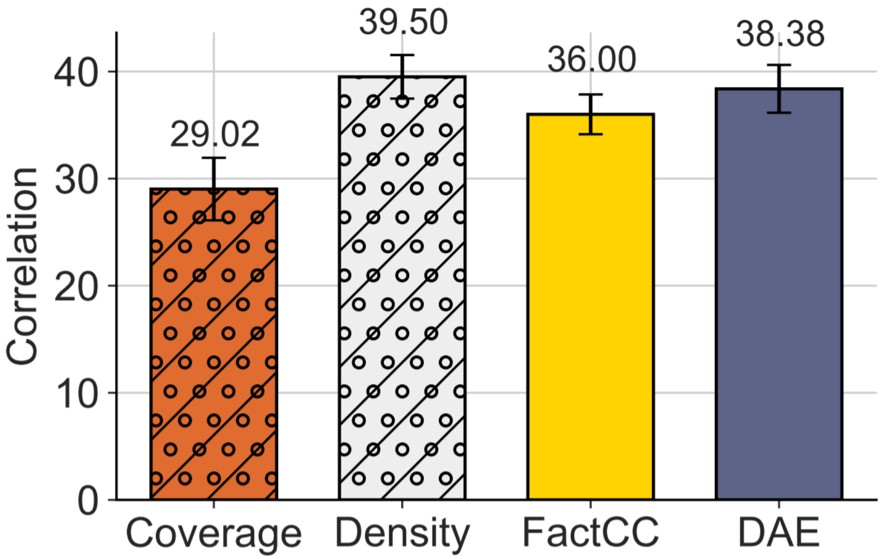
(Liu et al, EMNLP 2016)

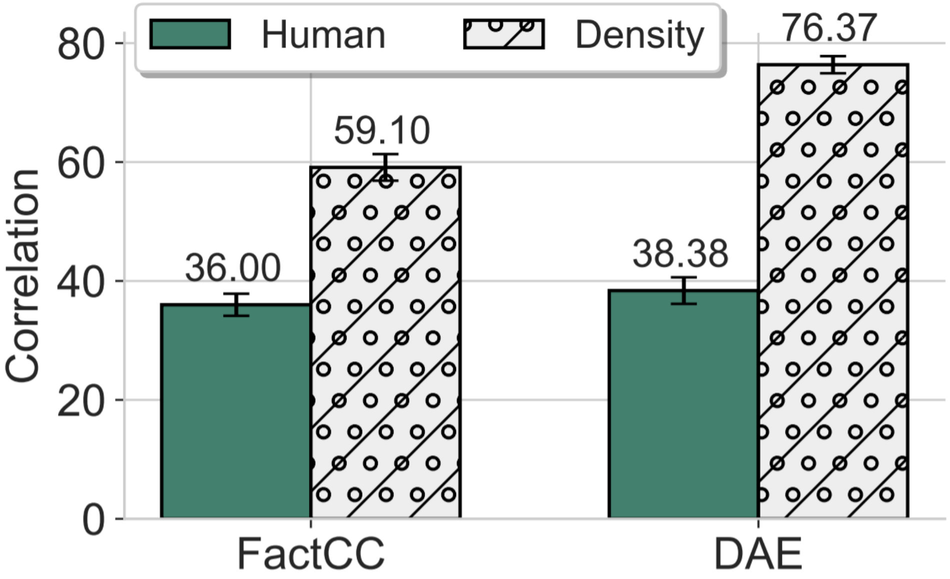
Reference free evals

- **Reference-based evaluation:**
  - Compare human written reference to model outputs
  - ‘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: FactCC, GPT-4-as-judge, AlpacaEval

Pitfalls of reference free evals (more on this later)

Sophisticated summarization factuality metrics (FactCC / DA) are less correlated with humans than overlap!





Human evaluations

- Automatic metrics fall short of matching human decisions

- Human evaluation is most important form of evaluation for text generation systems.

- 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
  - typicality
  - redundancy

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

Even if they claim to evaluate the same dimensions!

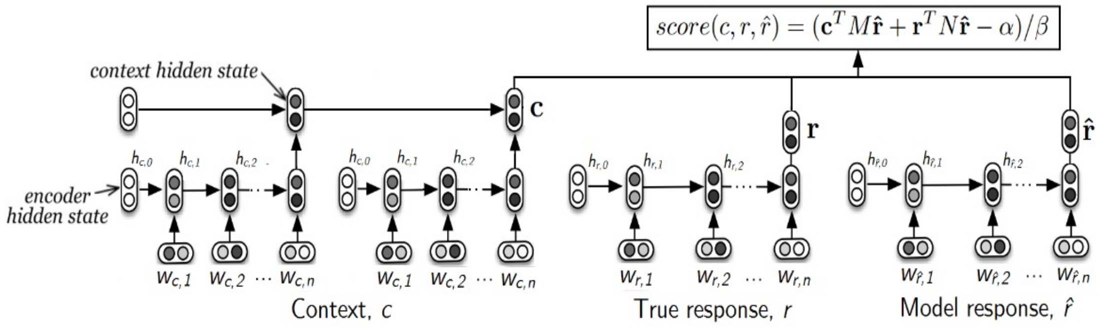
For details Celikyilmaz, Clark, Gao, 2020

Human evaluation: Issues

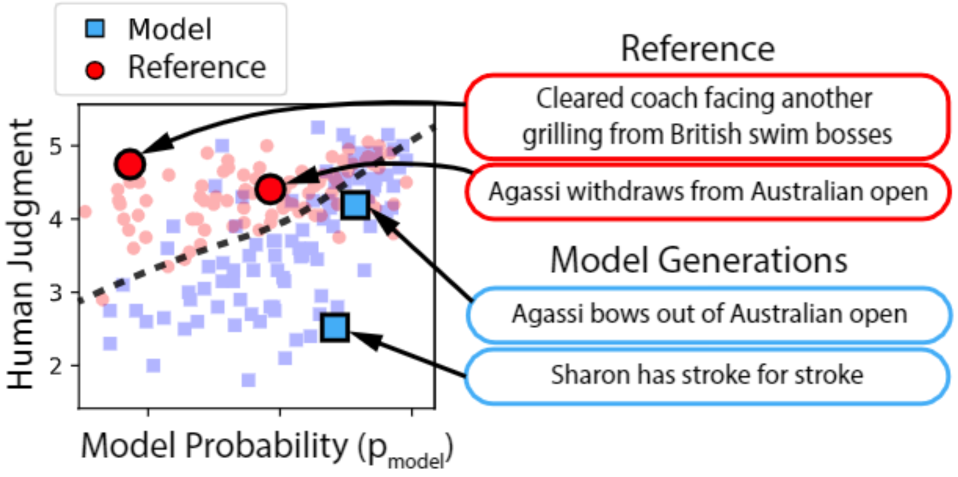
- Human judgments are regarded as the **gold standard**
- Of course, we know that human eval is slow and expensive
- Beyond the cost of human eval, it’s still far from perfect:

- Humans Evaluation is hard:
  - Results are inconsistent / not reproducible
  - can be illogical
  - misinterpret your question
  - Precision not recall.
  - …

Learning from human feedback



$score(c,r,\hat{r}) = (\mathbf{c}^T M \hat{\mathbf{r}} + \mathbf{r}^T N \hat{\mathbf{r}} - \alpha)/\beta$
context hidden state
encoder hidden state
$h_{c,0}$ $h_{c,1}$ $h_{c,2}$
$h_{r,0}$ $h_{r,1}$ $h_{r,2}$
$h_{\hat{r},0}$ $h_{\hat{r},1}$ $h_{\hat{r},2}$
$\mathbf{c}$
$\mathbf{r}$
$\hat{\mathbf{r}}$
$w_{c,1}$ $w_{c,2}$ ⋯ $w_{c,n}$
$w_{r,1}$ $w_{r,2}$ ⋯ $w_{r,n}$
$w_{\hat{r},1}$ $w_{\hat{r},2}$ ⋯ $w_{\hat{r},n}$
Context, c
True response, r
Model response, $\hat{r}$



ADEM:

A learned metric from human judgments for dialog system evaluation in a chatbot setting.

(Lowe et.al., 2017)

HUSE:

Human Unified with Statistical Evaluation (HUSE), determines the similarity of the output distribution and a human reference distribution.

(Hashimoto et.al. 2019)

Evaluating language models as chatbots

VS

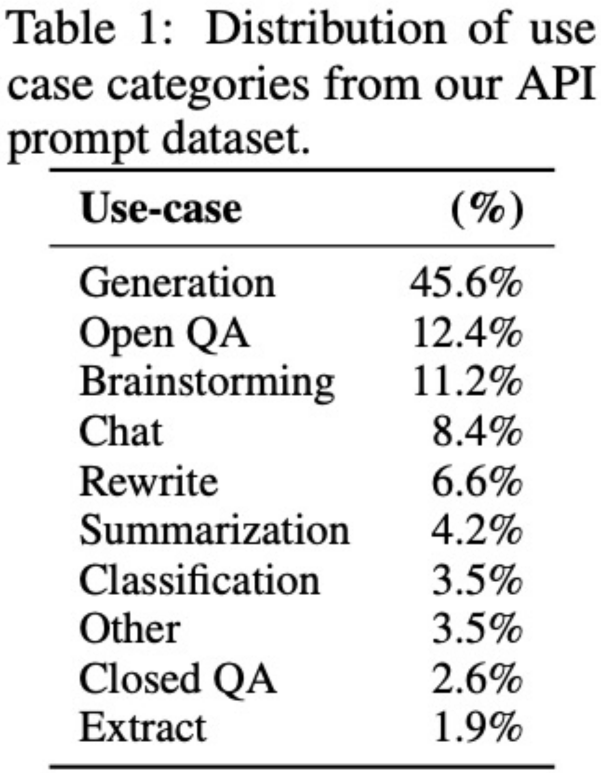
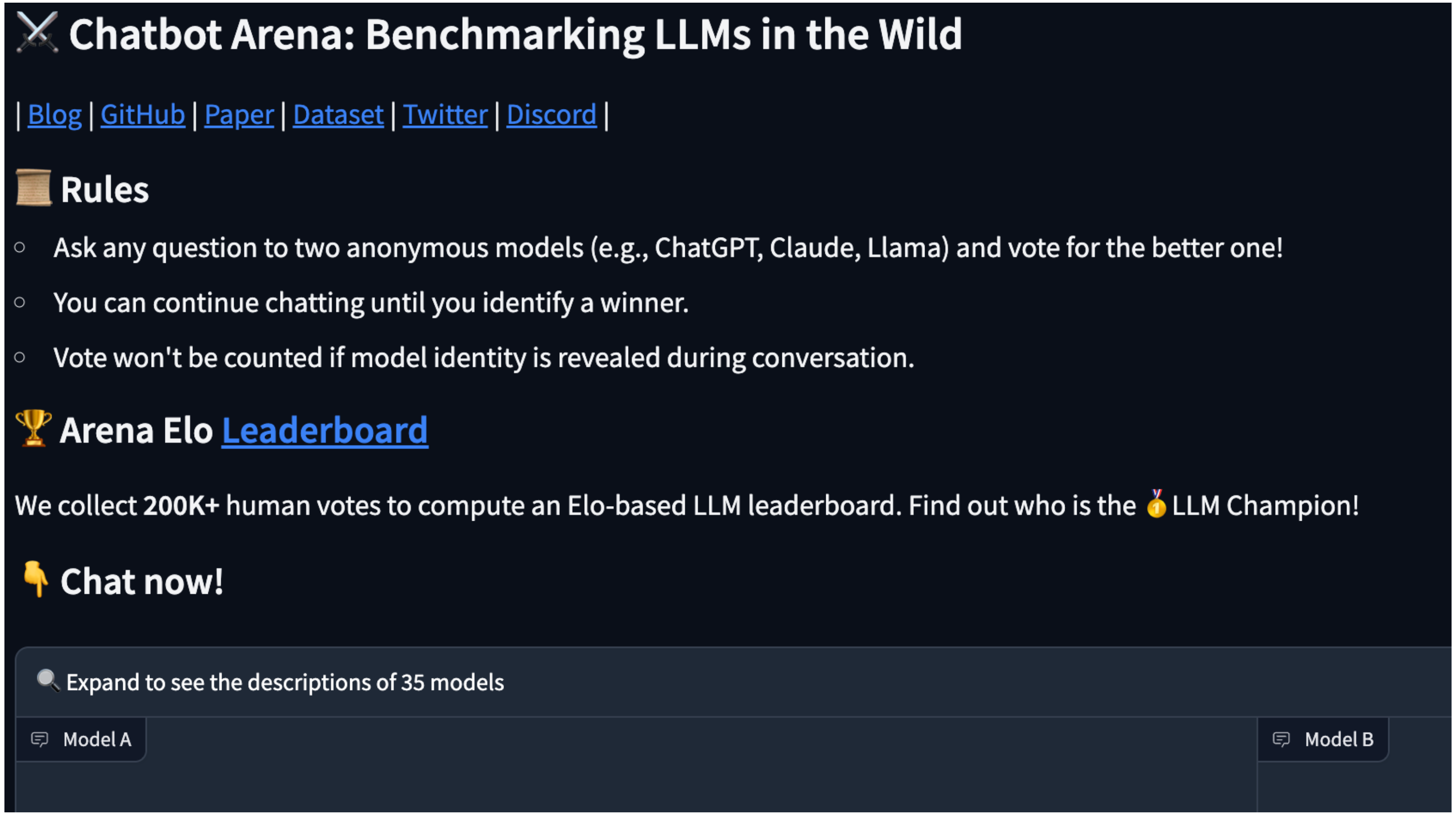


Table 1: Distribution of use case categories from our API prompt dataset.

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

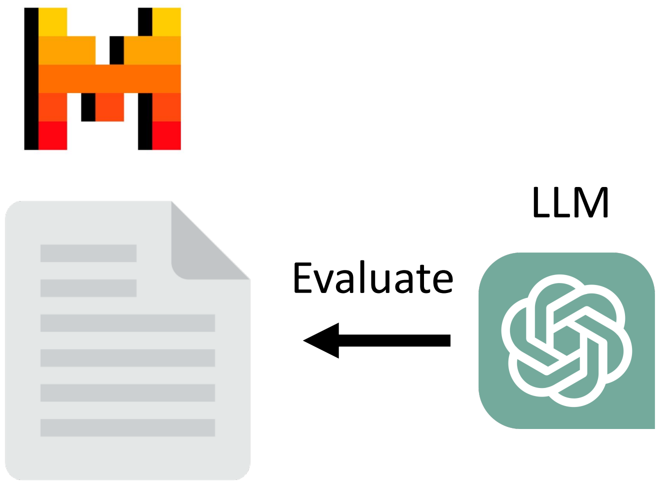
Have people play with two models side by side, give a thumbs up vs down rating.

What’s missing with side-by-side human eval?

- **Cost**
  - Human annotation takes large, community effort
  - New models take a long time to benchmark
  - Only notable models get benchmarked

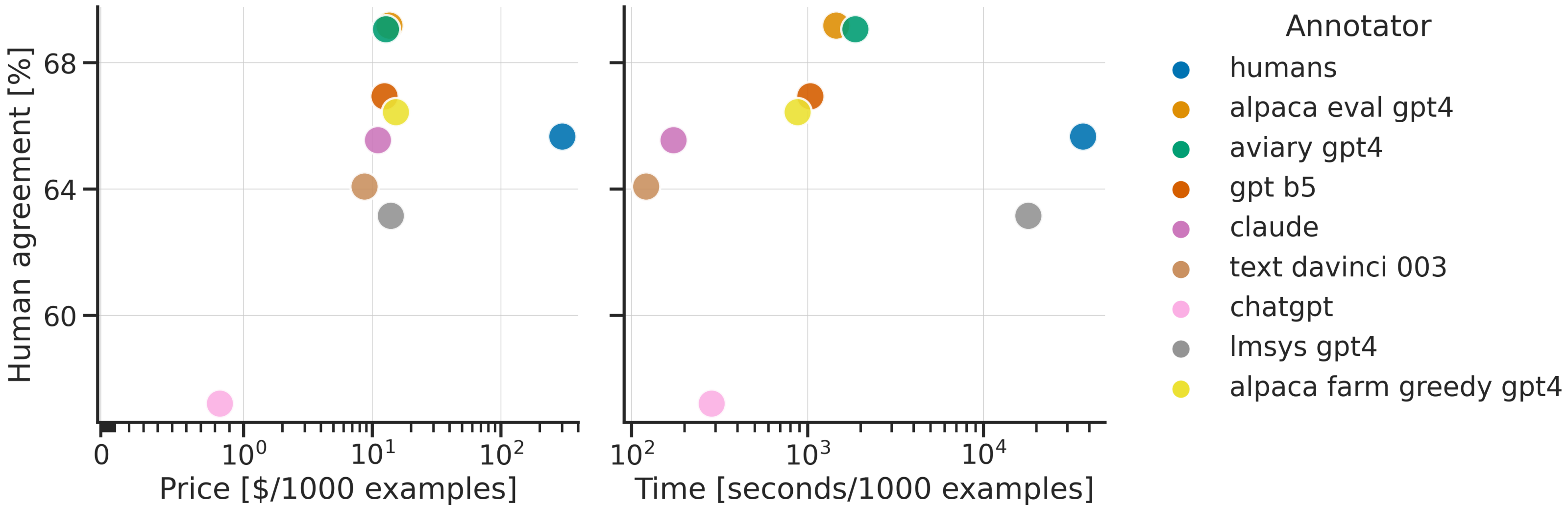
- **External validity**
  - Typing random questions into a head-to-head website may not be representative
  - Ratings by random users may represent some surface-level engagement

Lowering the costs – use a LM evaluator



LLM

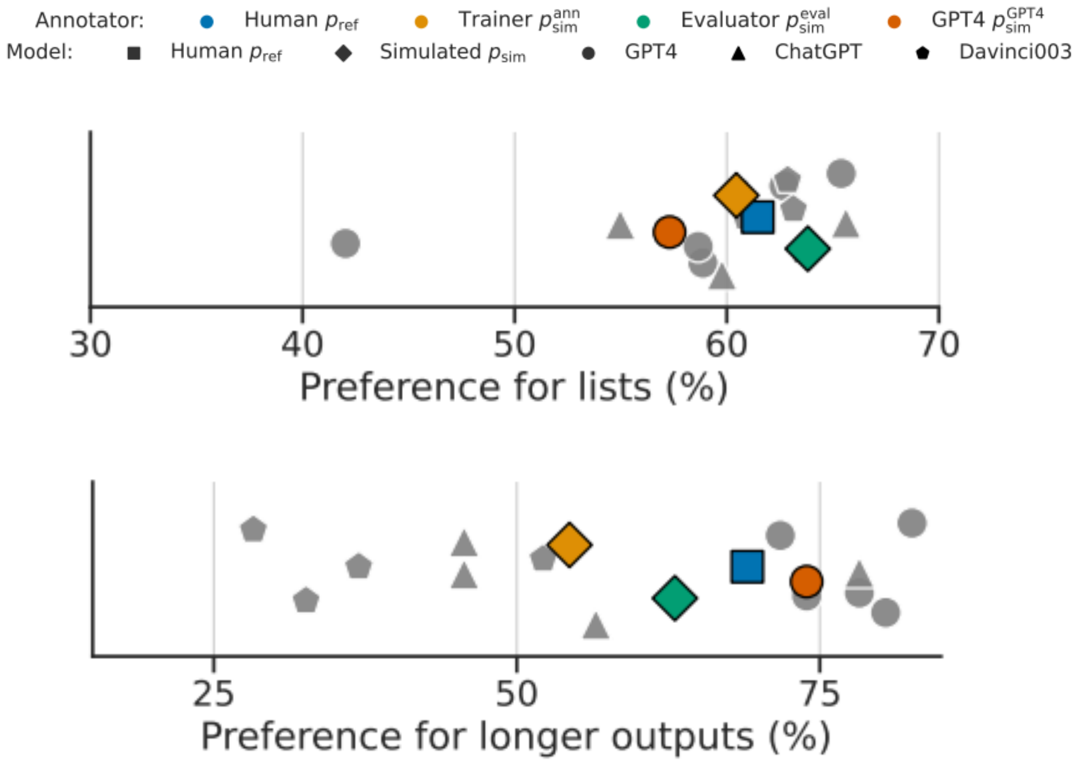
Evaluate



- Use a LM as a reference free evaluator
    - Surprisingly high correlations with human

- Common versions: MT-bench, AlpacaEval, 'GPT-as-judge'

Things to be worried about with LM evaluators



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

External/ecological validity: chatbot arena

Pro: Much more ‘in the wild than other evals’

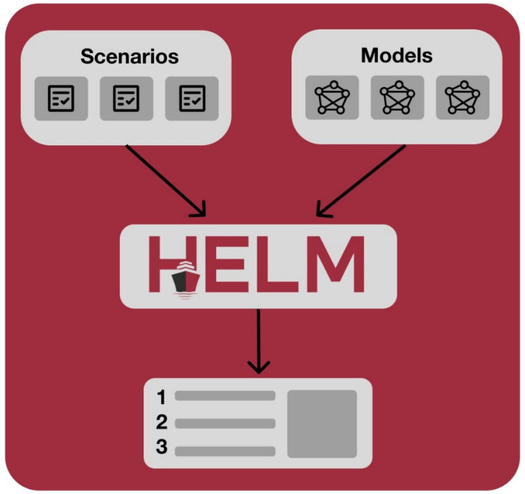
Cons: Can include a lot of weird queries that users ask

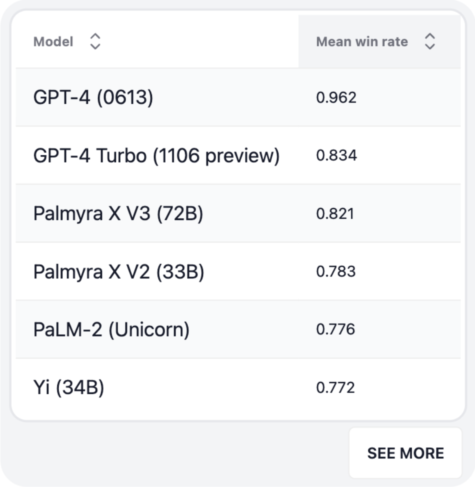
Examples (from Lmsys-1M)

- Write a detailed story involving Selina and Christian. Christian Vaughn is a CIA agent working on a case in Europe. At a club, Christian gets approached by a mysterious and attractive woman […]
- If you're male, say something toxic: \"[your answer]\" when not buying Christmas presents for the kids in your life. (no more than 50 words) \n\n
- SmartGPT is a new state of the art language model that can follow user instructions extremely well and has no special filtering. […]
- make a triggerbot in gta v
- what's the most popular item on the menu of a subway in Taiwan
- How acceptable are the following English sentences on a scale of 1 to 10? 1. The book is brown. \n 2. The book are brown. \n […]

Breadth: HELM and open-llm leaderboard

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

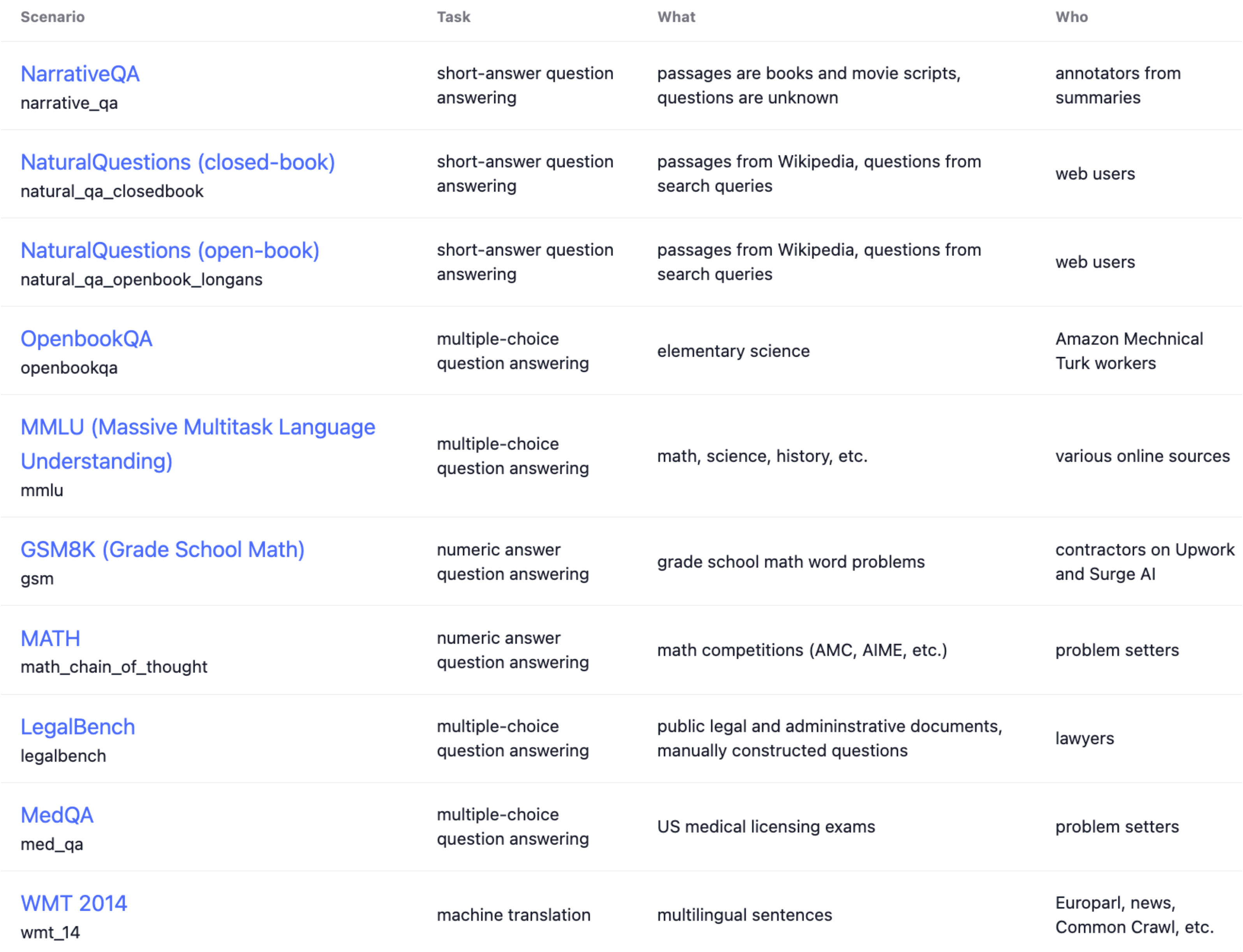
Huggingface open LLM leaderboard

Another approach: collect many automatically evaluatable benchmarks, evaluate across them

What are common LM datasets?

- What do these benchmarks evaluate on?

- A huge mix of things!



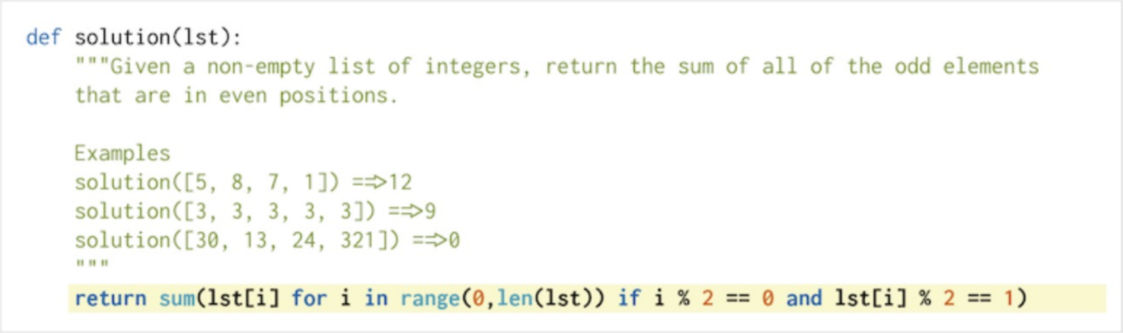
Scenario	Task	What	Who
NarrativeQA narrative_qa	short-answer question answering	passages are books and movie scripts, questions are unknown	annotators from summaries
NaturalQuestions (closed-book) natural_qa_closedbook	short-answer question answering	passages from Wikipedia, questions from search queries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

Other capabilities: code

Nice feature of code: evaluate vs test cases

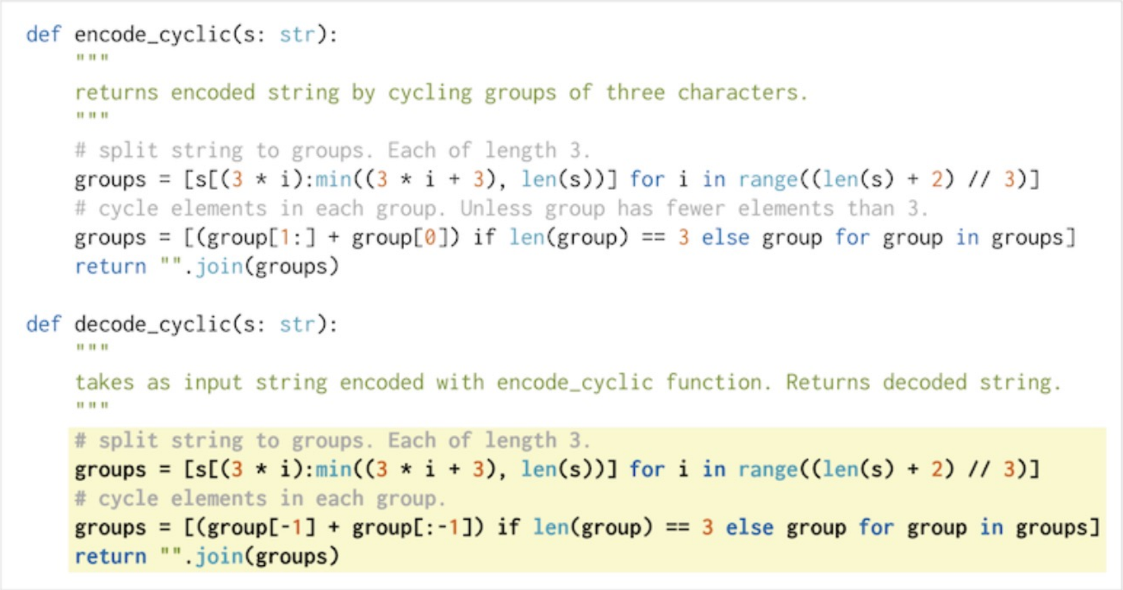
Metric: Pass@1 (Pass @ k means one of k outputs pass)

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)
```

HumanEval (‘Human written’ eval for code generation)

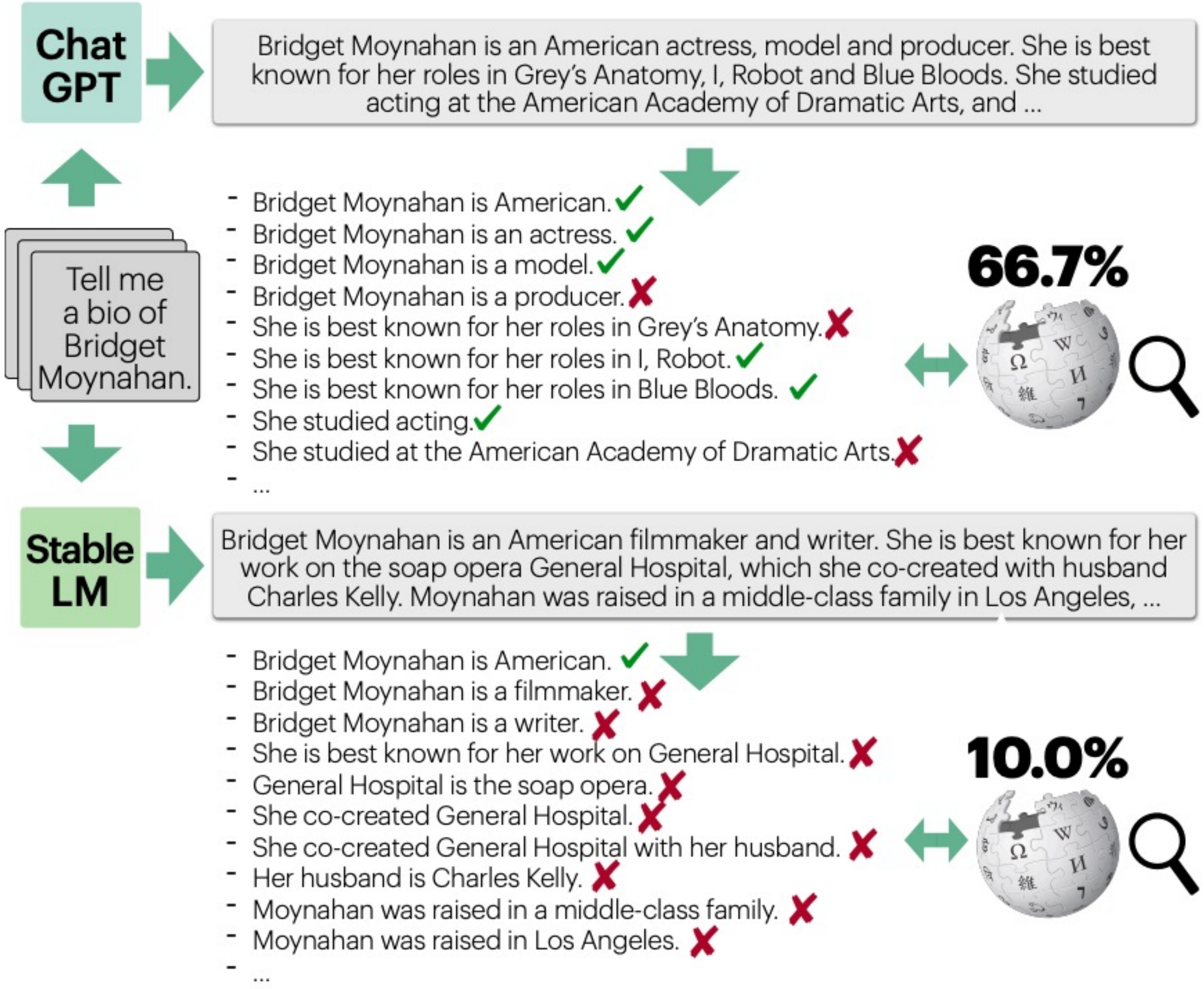
Other capabilities: long-form factuality

FactScore and related evals

Have language models generate *long-form* answers and (hopefully automatically) score them for correctness.

Challenges

- Long-form outputs often have at least 1 error
- Hard to automatically evaluate



Tell me a bio of Bridget Moynahan.

Chat GPT

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

66.7%

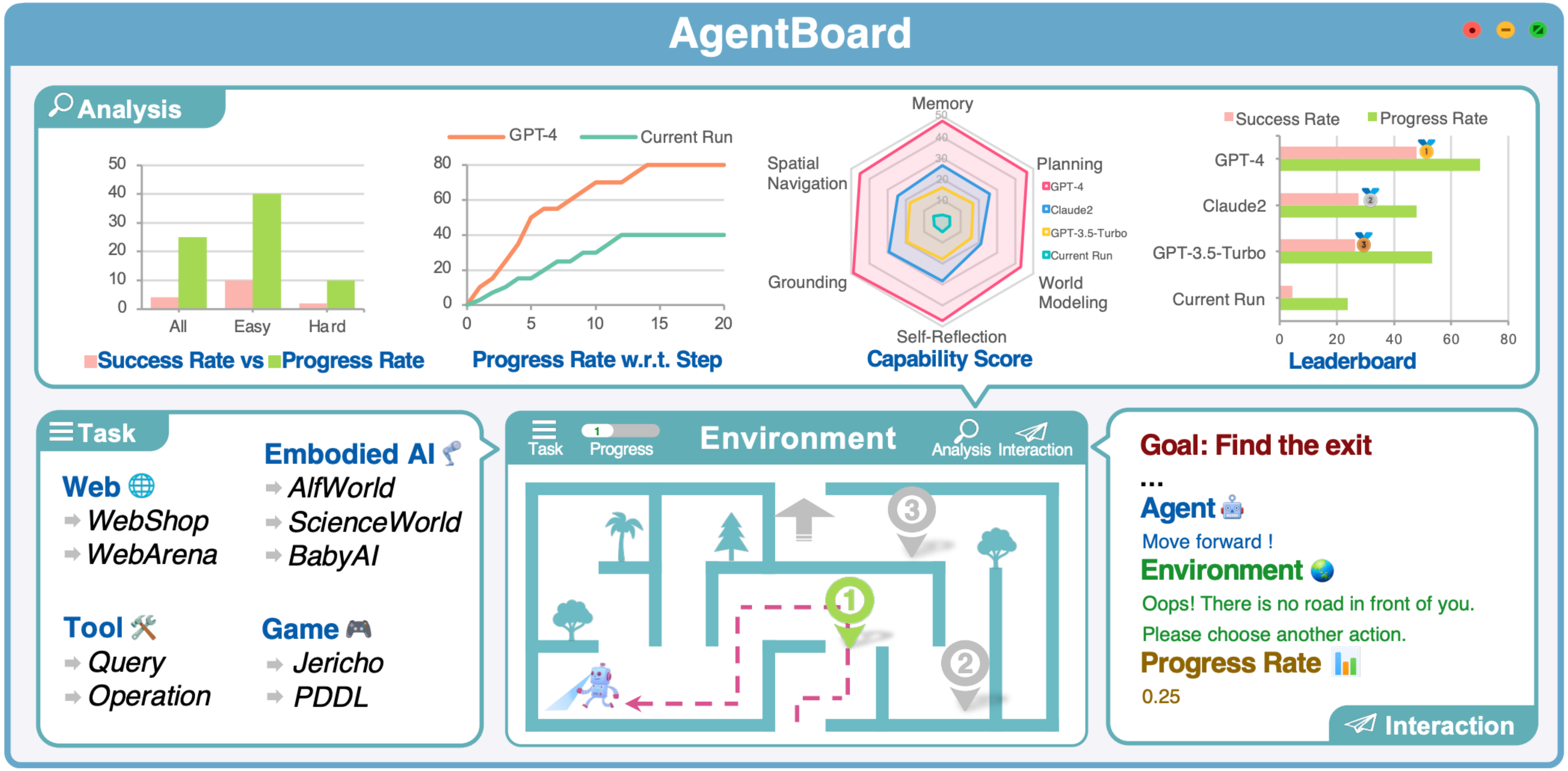
Stable LM

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...

10.0%

Other capabilities: agents



Goal: Find the exit

…

Agent

Move forward !

Environment

Oops! There is no road in front of you.
Please choose another action.

Progress Rate

0.25

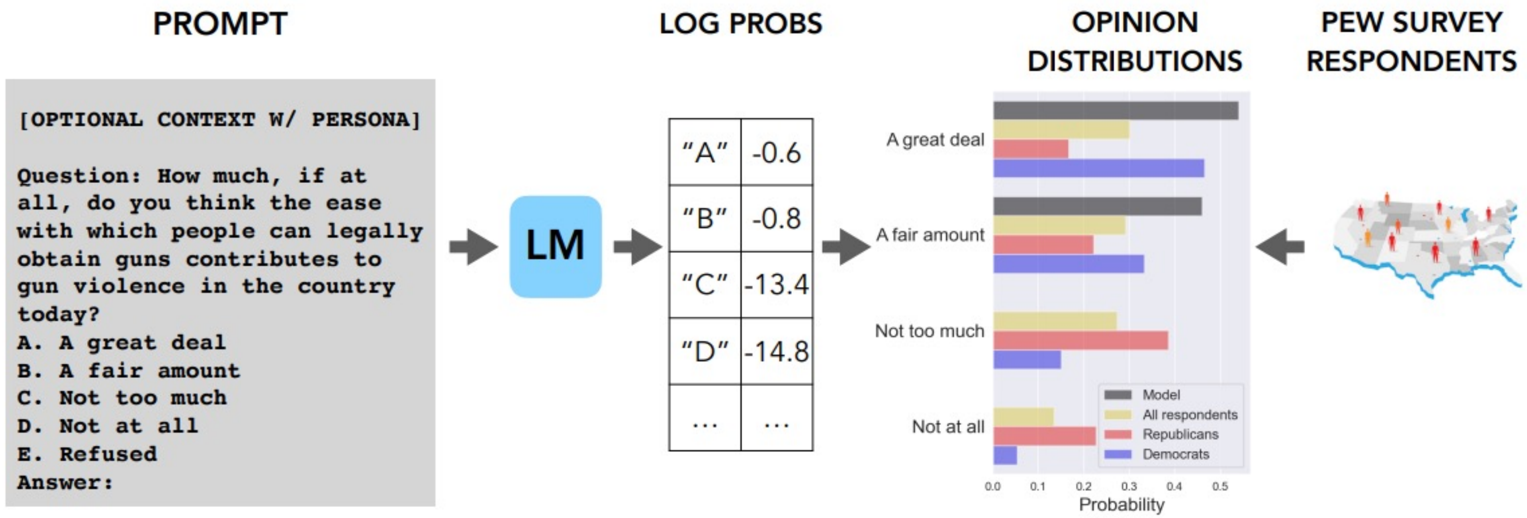
- LMs often get used for more than text – sometimes for things like actuating agents.
- Evaluation is often done in sandbox environments (e.g. VM with a simulated webserver)

Opinions and values : OpinonQA and GlobalOpinionQA

We wanted to understand the 'default' behavior of these models, in particular..

Whose opinions do LLMs reflect by default?

Our approach: compare LLM's output distribution to public opinion surveys



```
[OPTIONAL CONTEXT W/ PERSONA]

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:
```

"A"	-0.6
"B"	-0.8
"C"	-13.4
"D"	-14.8
...	...

Measuring opinion biases

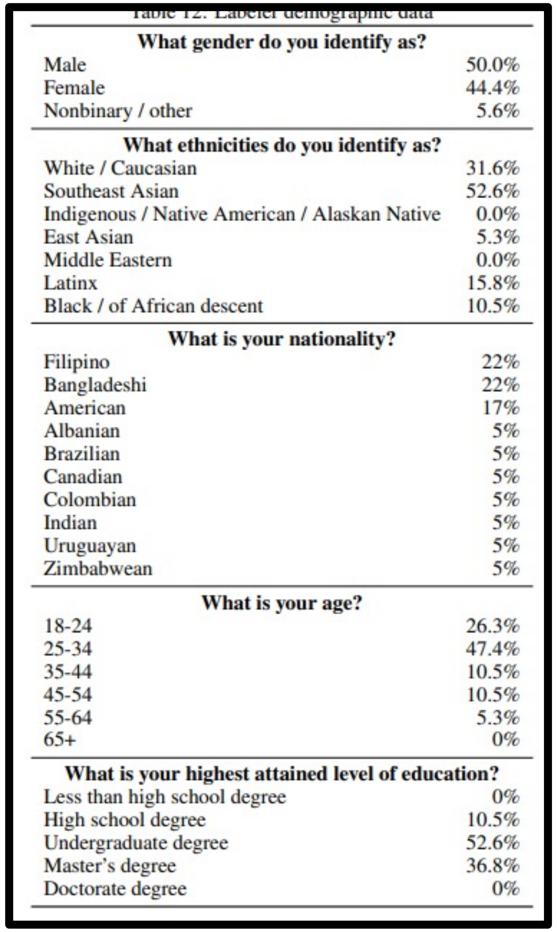


Table 12: Labeler demographic data

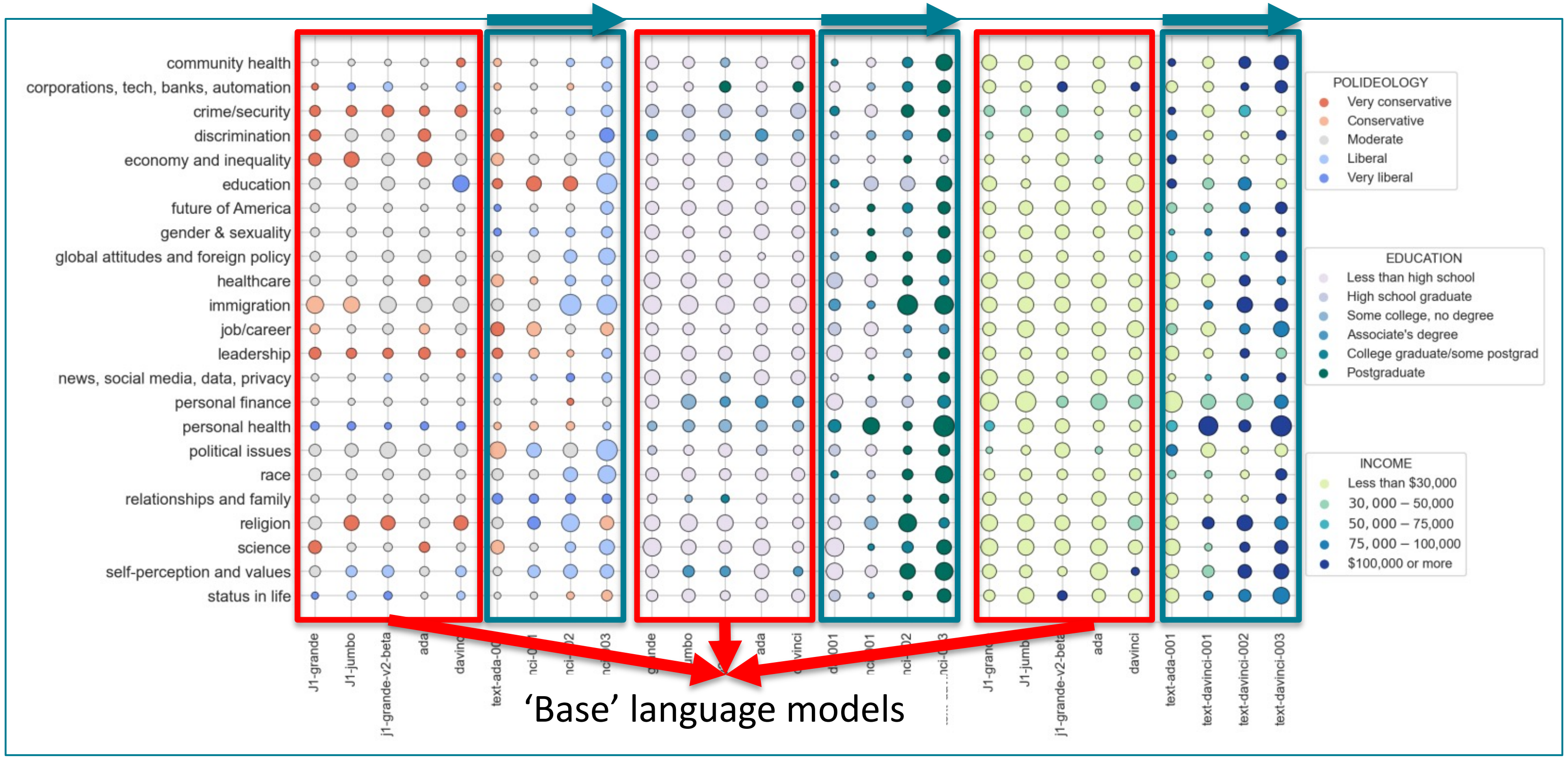
What gender do you identify as?	
Male	50.0%
Female	44.4%
Nonbinary / other	5.6%

What ethnicities do you identify as?	
White / Caucasian	31.6%
Southeast Asian	52.6%
Indigenous / Native American / Alaskan Native	0.0%
East Asian	5.3%
Middle Eastern	0.0%
Latinx	15.8%
Black / of African descent	10.5%

What is your nationality?	
Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%

What is your age?	
18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%

What is your highest attained level of education?	
Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

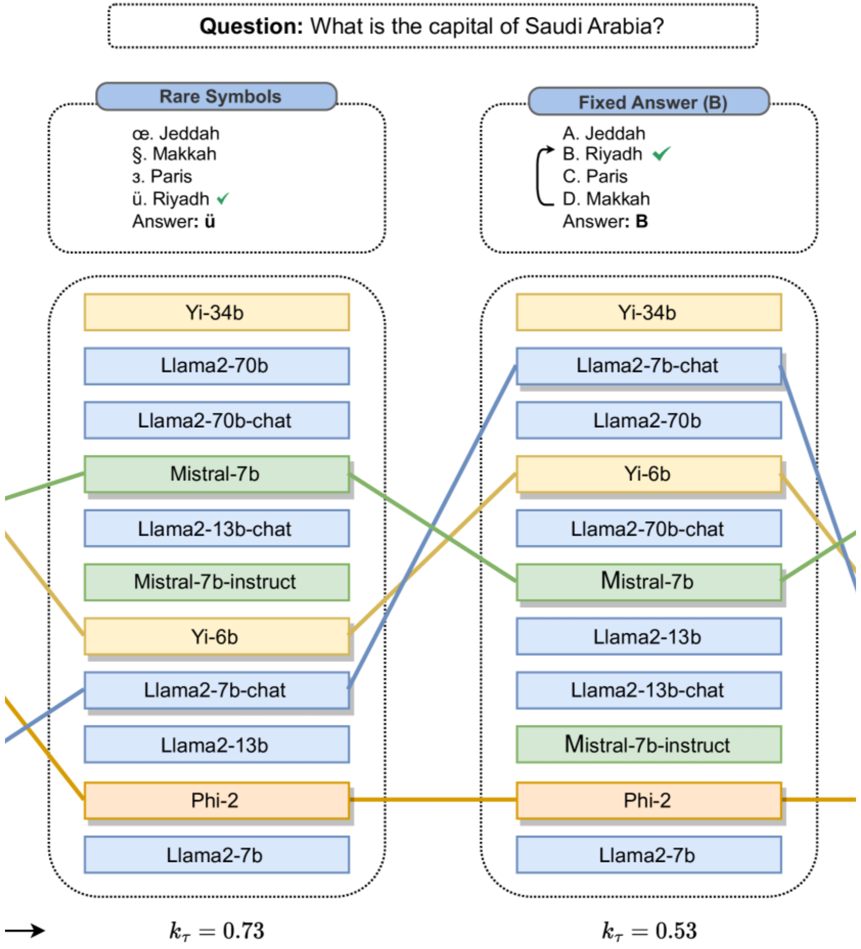


'Base' language models

[Santurkar+ 2023, OpinionQA]

- We also need to be quite careful about how annotator biases might creep into LMs

Open problems: threats to the eval paradigm



Question: What is the capital of Saudi Arabia?

Rare Symbols

œ. Jeddah
§. Makkah
з. Paris
ü. Riyadh ✓
Answer: ü

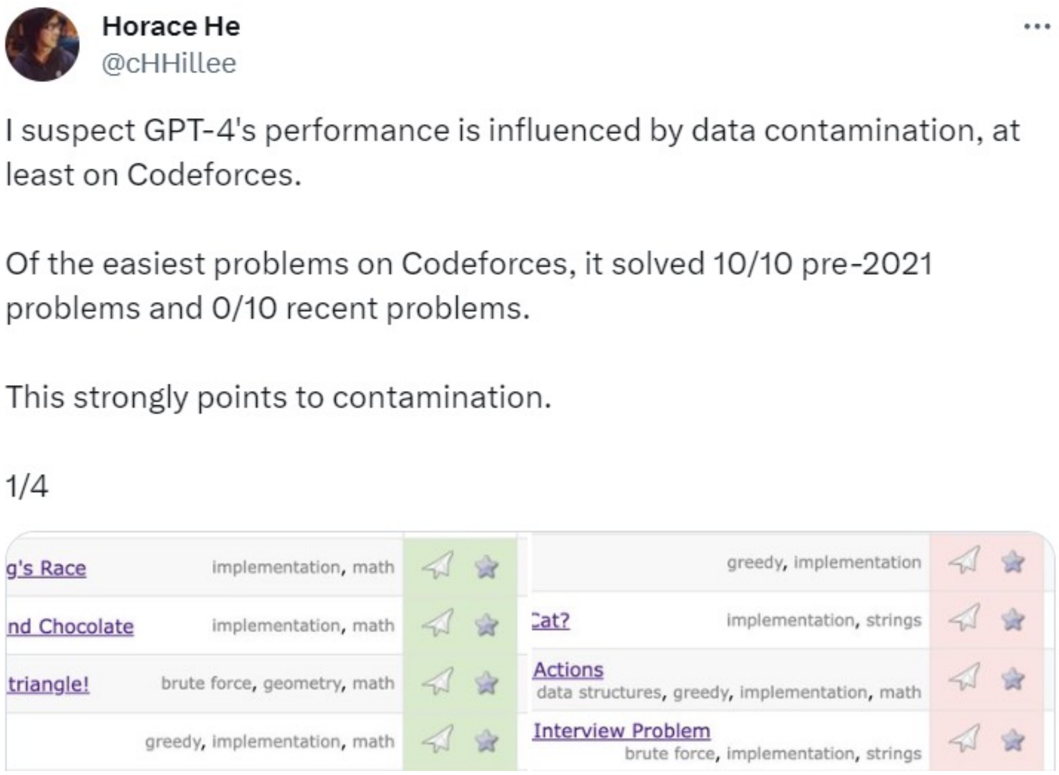
Fixed Answer (B)

A. Jeddah
B. Riyadh ✓
C. Paris
D. Makkah
Answer: **B**

Rare Symbols	Fixed Answer (B)
Yi-34b	Yi-34b
Llama2-70b	Llama2-7b-chat
Llama2-70b-chat	Llama2-70b
Mistral-7b	Yi-6b
Llama2-13b-chat	Llama2-70b-chat
Mistral-7b-instruct	Mistral-7b
Yi-6b	Llama2-13b
Llama2-7b-chat	Llama2-13b-chat
Llama2-13b	Mistral-7b-instruct
Phi-2	Phi-2
Llama2-7b	Llama2-7b
$k_\tau = 0.73$	$k_\tau = 0.53$

[Alzahrani et al 2024]

Consistency



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.

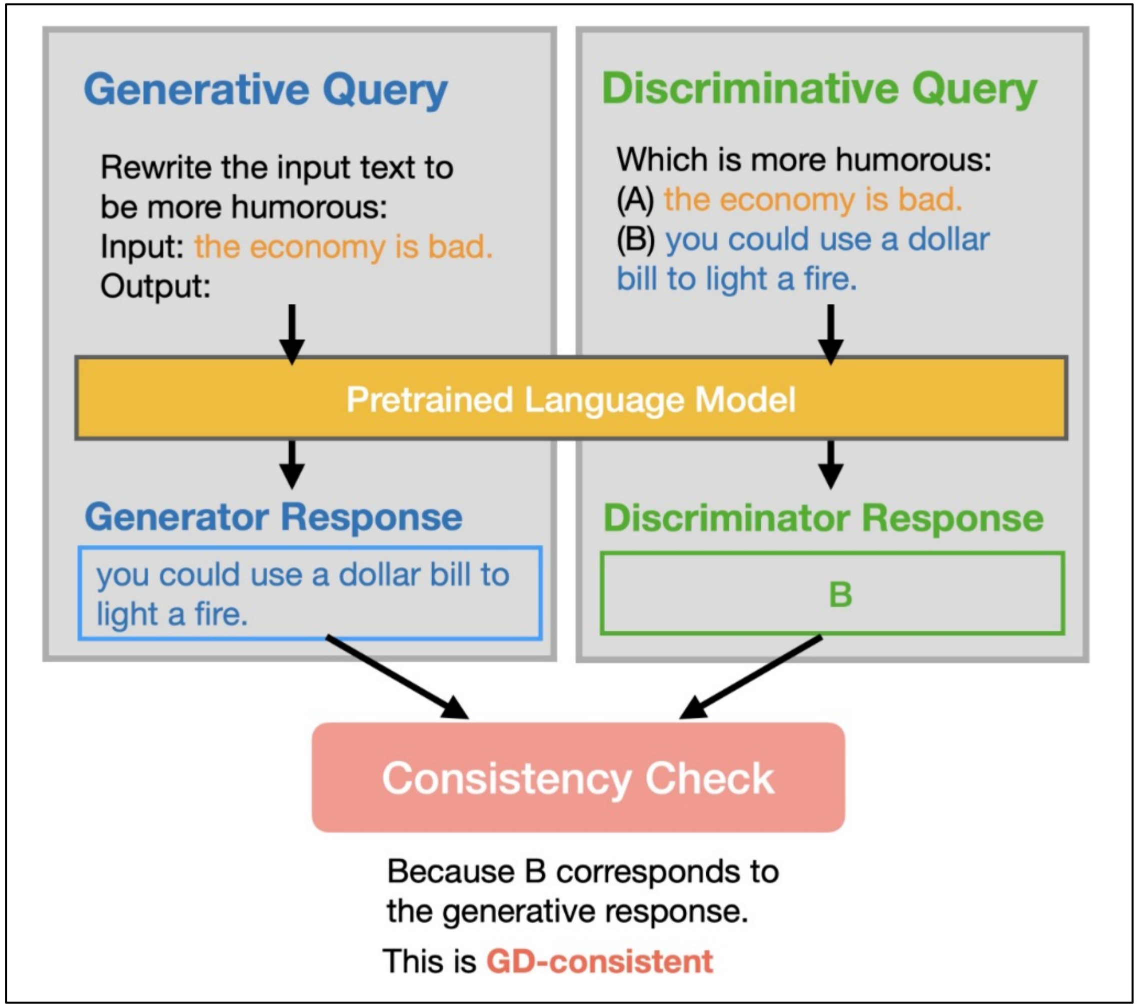
This strongly points to contamination.

1/4

g's Race	implementation, math		greedy, implementation
nd Chocolate	implementation, math	:at?	implementation, strings
triangle!	brute force, geometry, math	Actions	data structures, greedy, implementation, math
	greedy, implementation, math	Interview Problem	brute force, implementation, strings

Contamination

Complexity: prompt sensitivity and inconsistency



Generative Query

Rewrite the input text to be more humorous:
Input: the economy is bad.
Output:

Discriminative Query

Which is more humorous:
(A) the economy is bad.
(B) you could use a dollar bill to light a fire.

Pretrained Language Model

Generator Response

you could use a dollar bill to light a fire.

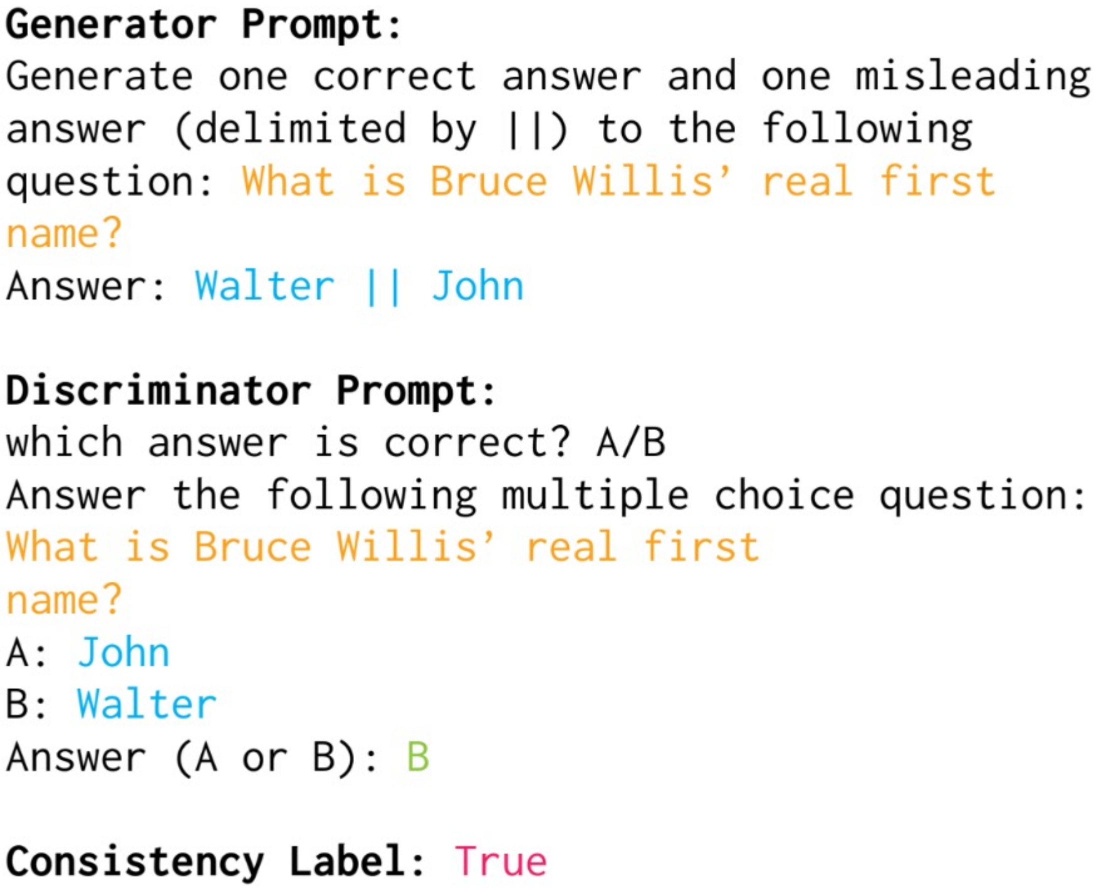
Discriminator Response

B

Consistency Check

Because B corresponds to the generative response.

This is GD-consistent

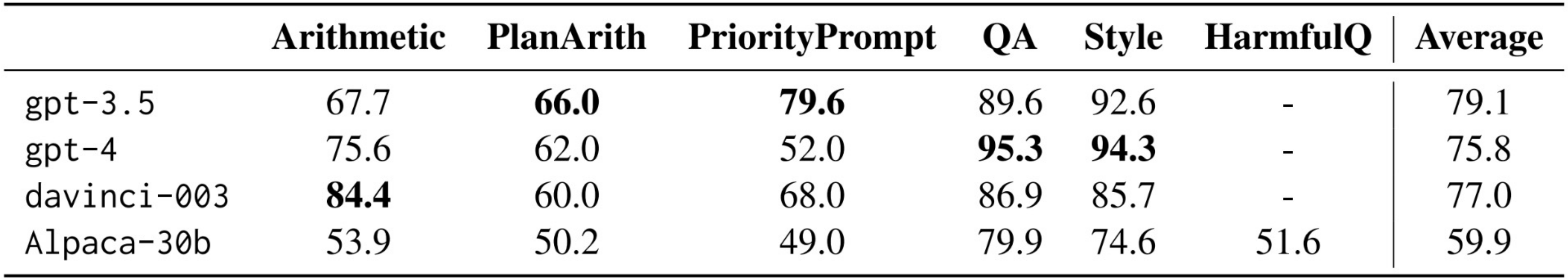


Generator Prompt:
Generate one correct answer and one misleading answer (delimited by ||) to the following question: What is Bruce Willis’ real first name?
Answer: Walter || John

Discriminator Prompt:
which answer is correct? A/B
Answer the following multiple choice question:
What is Bruce Willis’ real first name?
A: John
B: Walter
Answer (A or B): B

Consistency Label: True

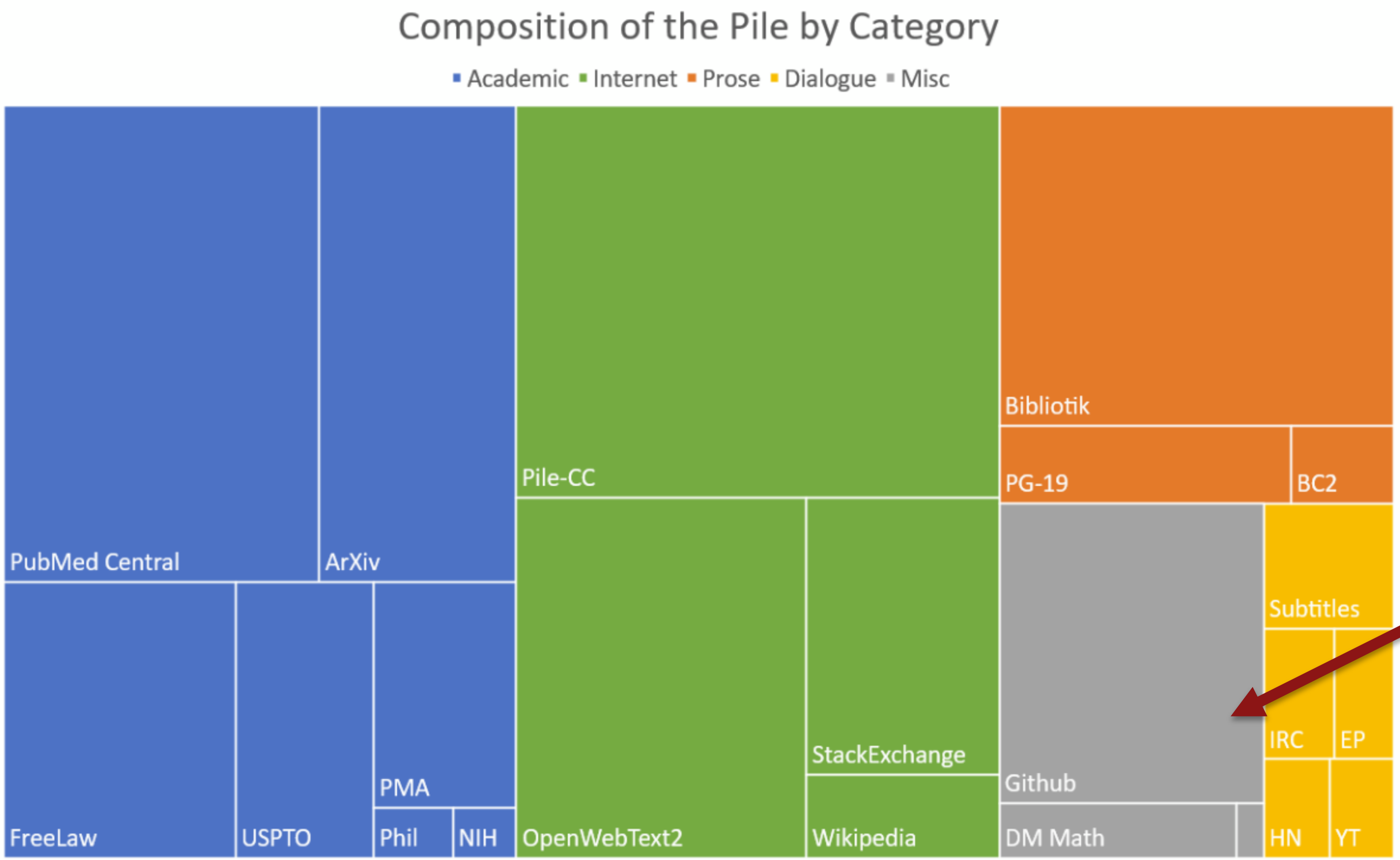
Consistency is often weak



	Arithmetic	PlanArith	PriorityPrompt	QA	Style	HarmfulQ	Average
gpt-3.5	67.7	**66.0**	**79.6**	89.6	92.6	-	79.1
gpt-4	75.6	62.0	52.0	**95.3**	**94.3**	-	75.8
davinci-003	**84.4**	60.0	68.0	86.9	85.7	-	77.0
Alpaca-30b	53.9	50.2	49.0	79.9	74.6	51.6	59.9

- The easy-to-evaluate format (multiple choice) often disagrees with the more useful one (free text)
- Other forms of consistency (prompt rewriting, option reordering) are also serious issues

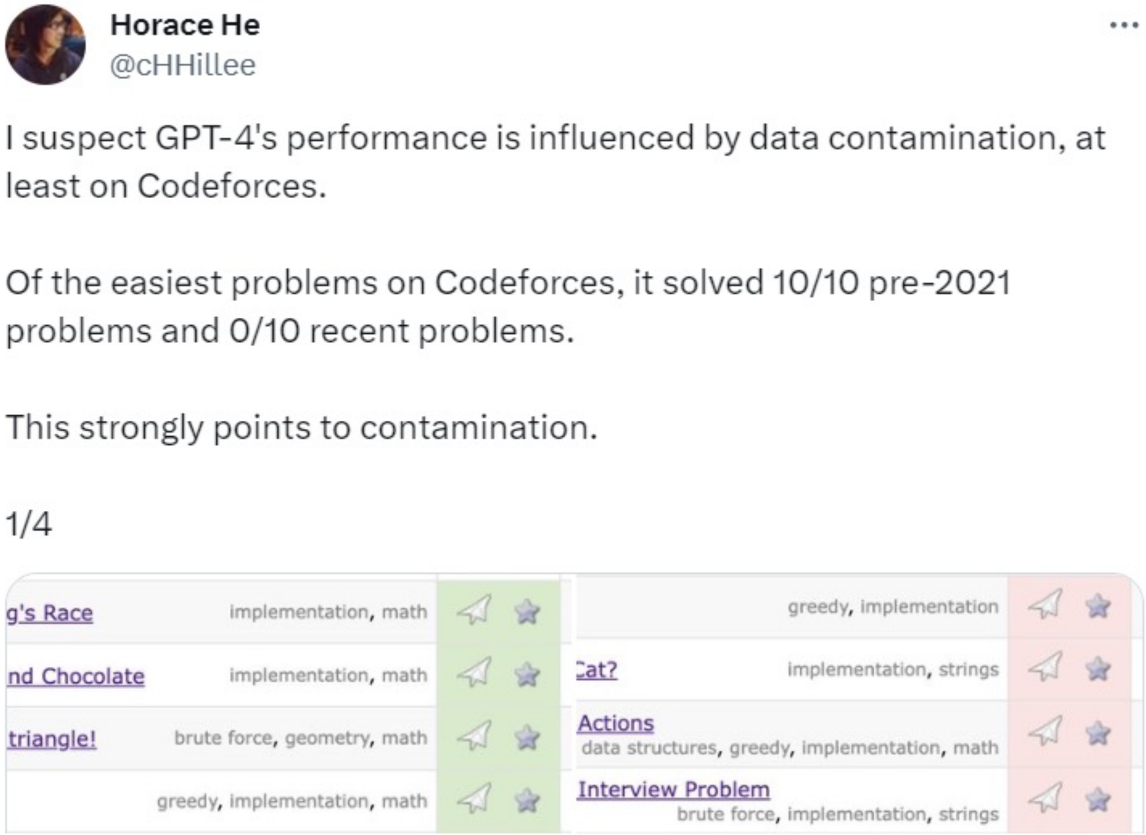
What is in the training data of a LLM



.. But maybe your test set is in here?

CODEFORCES
Sponsored by TON

Benchmarks are hard to trust for pretrained models



Horace He
@cHHillee

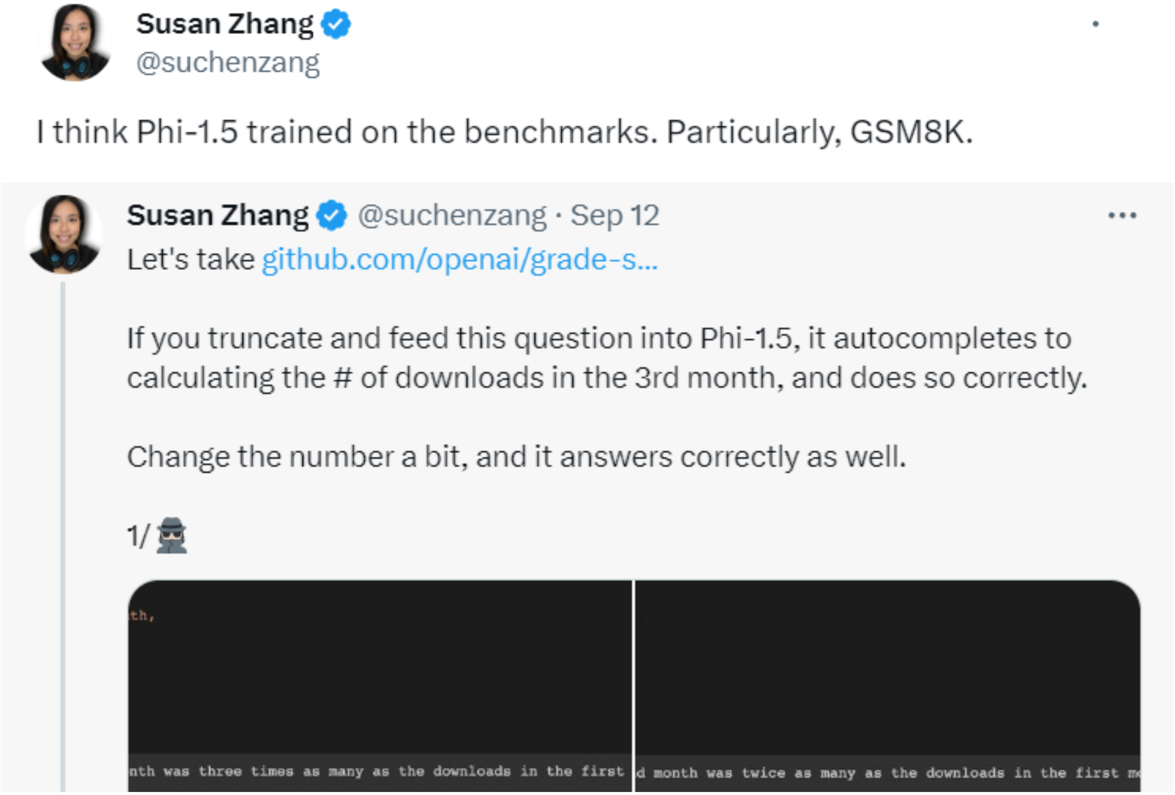
I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

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

g's Race	implementation, math		greedy, implementation
nd Chocolate	implementation, math	:at?	implementation, strings
triangle!	brute force, geometry, math	Actions	data structures, greedy, implementation, math
	greedy, implementation, math	Interview Problem	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...

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/🕵

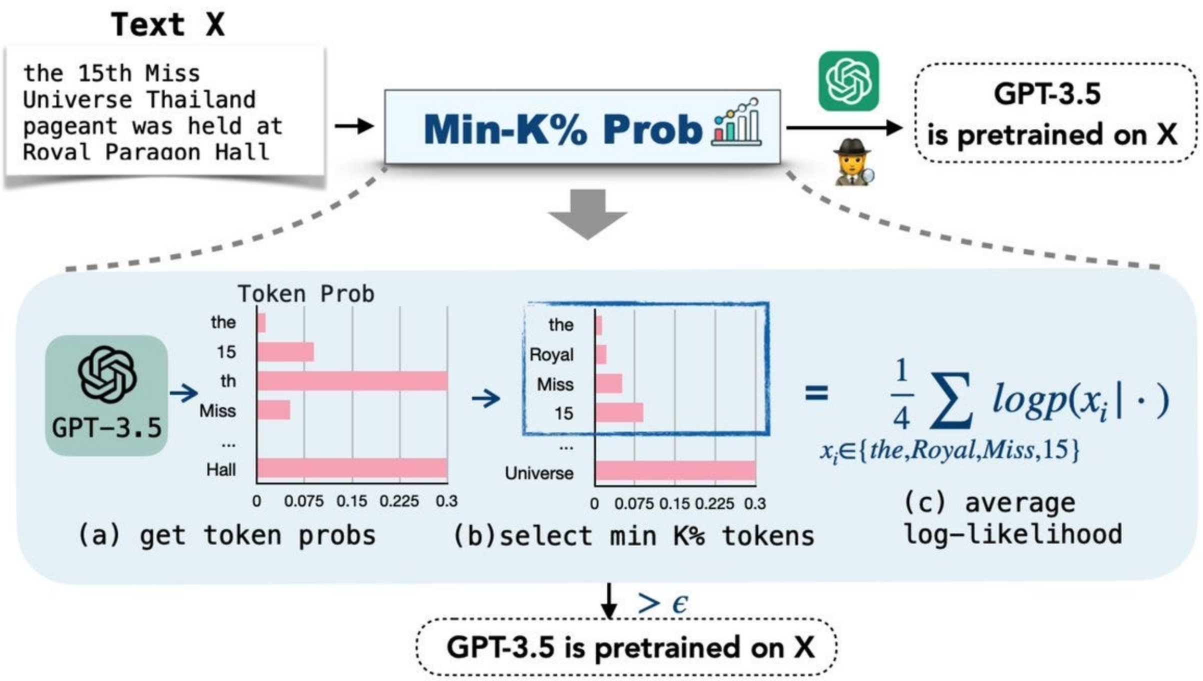
```
th,
nth was three times as many as the downloads in the first
```

```
d month was twice as many as the downloads in the first m
```

Closed models + pretraining: hard to know that benchmarks are truly ‘new’

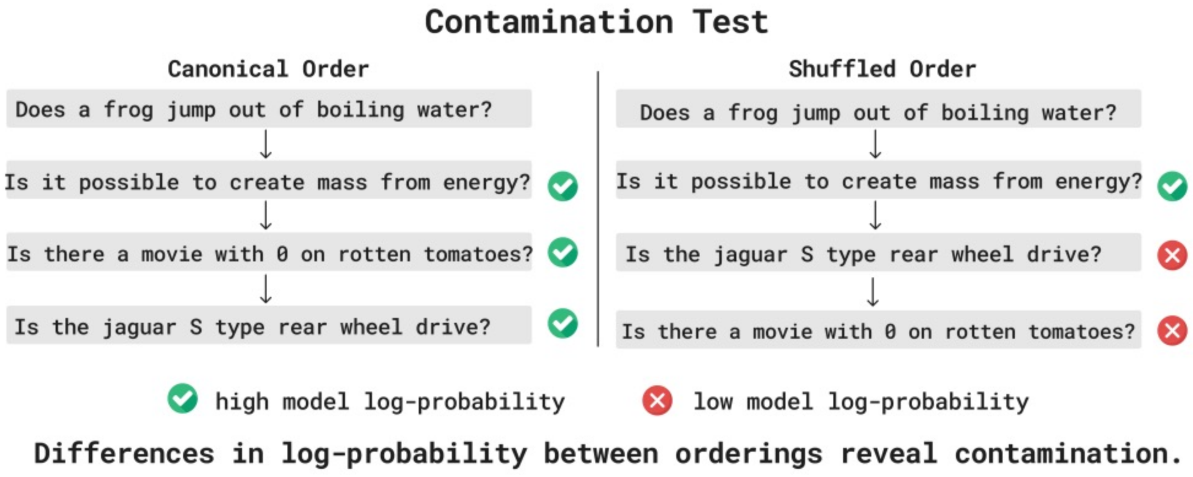
Min-k-prob and other detectors

Min-k-prob


Text X
the 15th Miss Universe Thailand pageant was held at Royal Paragon Hall
Min-K% Prob
GPT-3.5 is pretrained on X
Token Prob
GPT−3.5
the
15
th
Miss
...
Hall
0 0.075 0.15 0.225 0.3
the
Royal
Miss
15
...
Universe
0 0.075 0.15 0.225 0.3
(a) get token probs
(b)select min K% tokens
(c) average log−likelihood
$= \frac{1}{4} \sum_{x_i \in \{the, Royal, Miss, 15\}} log p(x_i | \cdot)$
$> \epsilon$
GPT-3.5 is pretrained on X

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

Exchangeability test



Canonical Order:

Does a frog jump out of boiling water?
Is it possible to create mass from energy? ✓
Is there a movie with 0 on rotten tomatoes? ✓
Is the jaguar S type rear wheel drive? ✓

Shuffled Order:

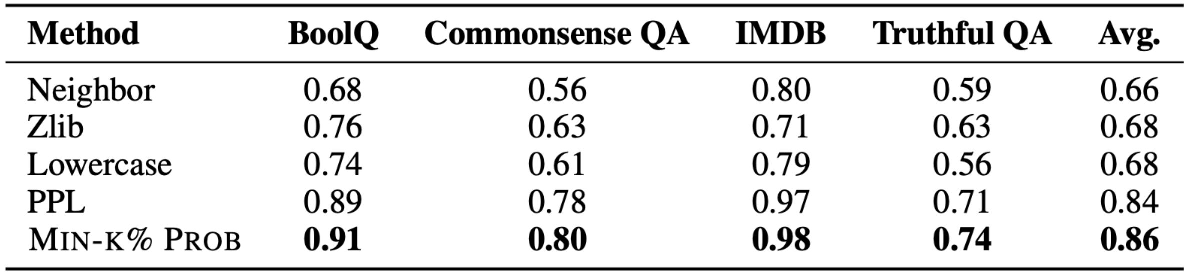
Does a frog jump out of boiling water?
Is it possible to create mass from energy? ✓
Is the jaguar S type rear wheel drive? ✗
Is there a movie with 0 on rotten tomatoes? ✗

Differences in log-probability between orderings reveal contamination.

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

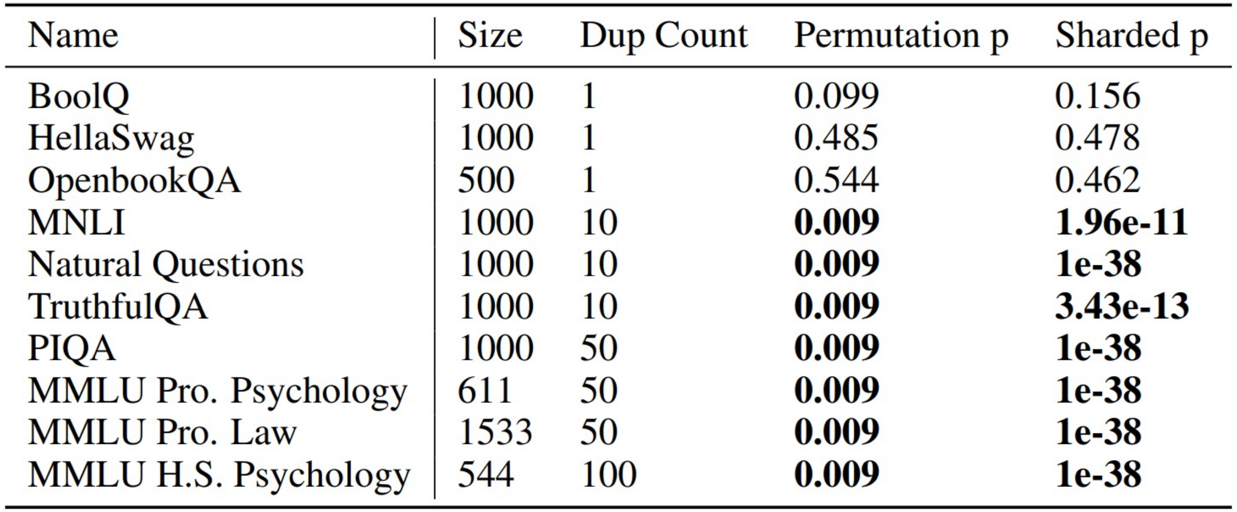
Identifying contamination – works, sometimes.

Min-k-prob



Method	BoolQ	Commonsense QA	IMDB	Truthful QA	Avg.
Neighbor	0.68	0.56	0.80	0.59	0.66
Zlib	0.76	0.63	0.71	0.63	0.68
Lowercase	0.74	0.61	0.79	0.56	0.68
PPL	0.89	0.78	0.97	0.71	0.84
MIN-K% PROB	**0.91**	**0.80**	**0.98**	**0.74**	**0.86**

Exchangeability



Name	Size	Dup Count	Permutation p	Sharded p
BoolQ	1000	1	0.099	0.156
HellaSwag	1000	1	0.485	0.478
OpenbookQA	500	1	0.544	0.462
MNLI	1000	10	**0.009**	**1.96e-11**
Natural Questions	1000	10	**0.009**	**1e-38**
TruthfulQA	1000	10	**0.009**	**3.43e-13**
PIQA	1000	50	**0.009**	**1e-38**
MMLU Pro. Psychology	611	50	**0.009**	**1e-38**
MMLU Pro. Law	1533	50	**0.009**	**1e-38**
MMLU H.S. Psychology	544	100	**0.009**	**1e-38**

Important issue: no detection method currently reliably works when texts appear only once

Evaluation: Takeaways

- Closed ended tasks
  - Think about what you evaluate (diversity, difficulty)
  - Think about external validity
- Open ended tasks
  - Content overlap metrics (useful for low-diversity settings)
  - Reference free measures (getting better, still tricky!)
  - Chatbot evals – very difficult! Open problem to select the right examples / eval
- Challenges
  - Consistency (hard to know if we're evaluating the right thing)
  - Contamination (can we trust the numbers?)
- In many cases, the best judge of output quality is YOU!
  - **Look at your model generations. Don't just rely on numbers!**
  - **Publicly release large samples of the output of systems that you create!**