Evaluation, Benchmarks, & Experimental Design

CS6120: Natural Language Processing
Northeastern University

David Smith
with slides from Jaehun Jung, Tatsunori Hashimoto, and Chris Manning

Benchmarks and evaluations drive progress

Linguistic Data Consortium | UNIVERSITY
LDC
Linguistic Data Consortium
ABOUT
MEMBERS
COMMUNICATIONS
LANGUAGE RESOURCES
Data
Obtaining Data
Catalog
Home › Language Resources › Data

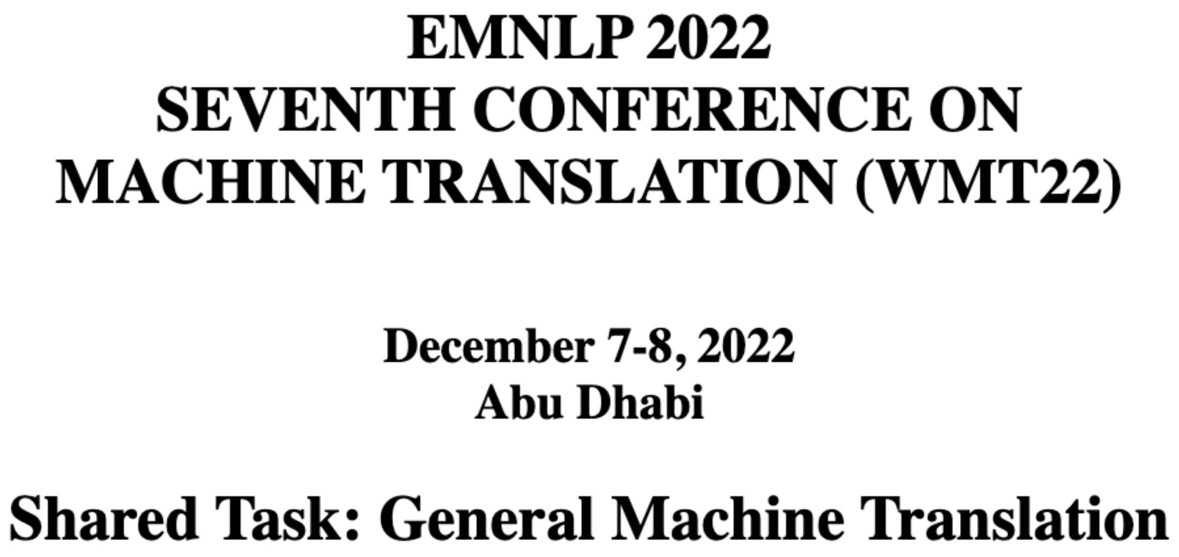
Switchboard-1 Release 2

Item Name: Switchboard-1 Release 2

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

SQuAD2.0

The Stanford Question Answering Dataset

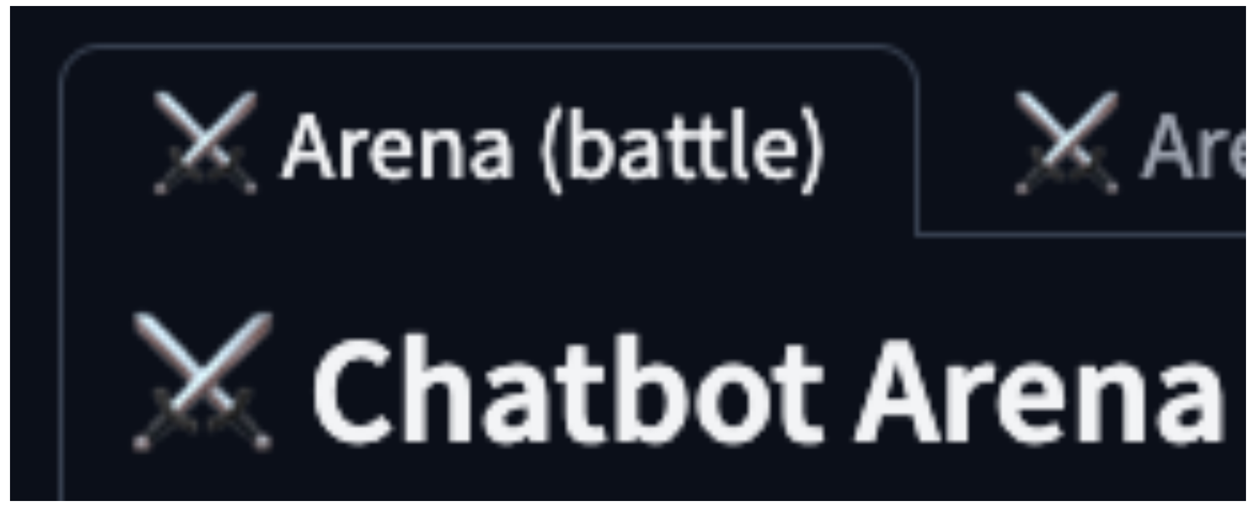


EMNLP 2022
SEVENTH CONFERENCE ON
MACHINE TRANSLATION (WMT22)

December 7-8, 2022
Abu Dhabi

Shared Task: General Machine Translation

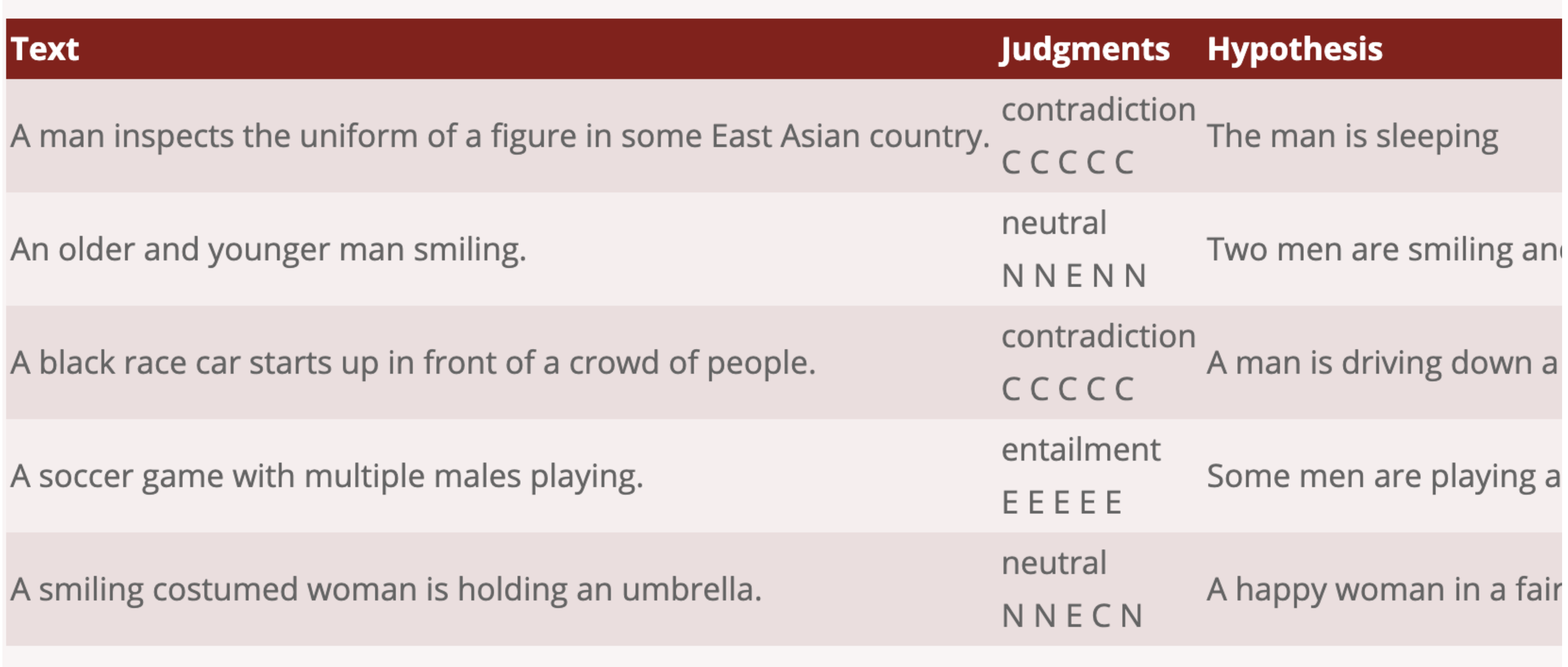
GLUE



Benchmarks and how we evaluate drive the progress of the field

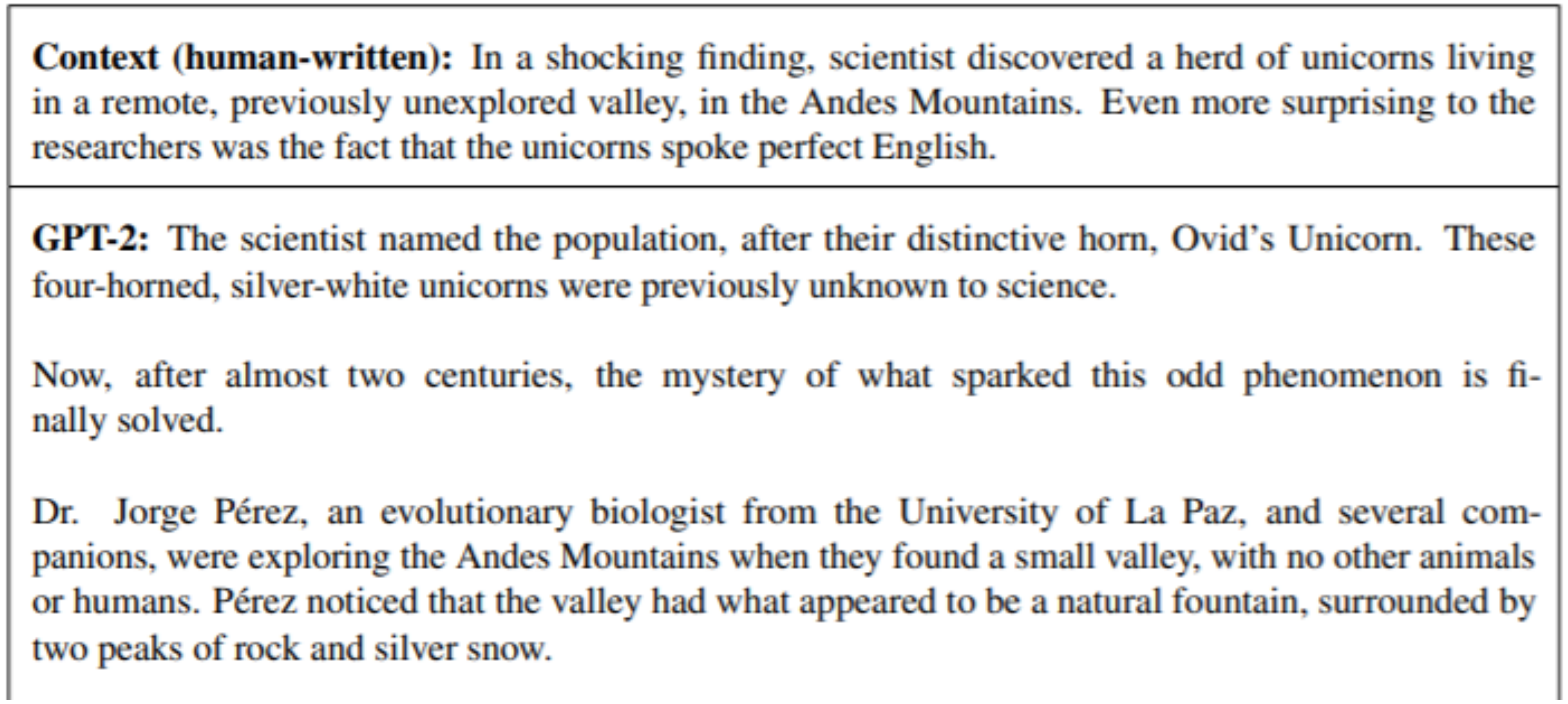
Two major types of evaluations

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 fai

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

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)

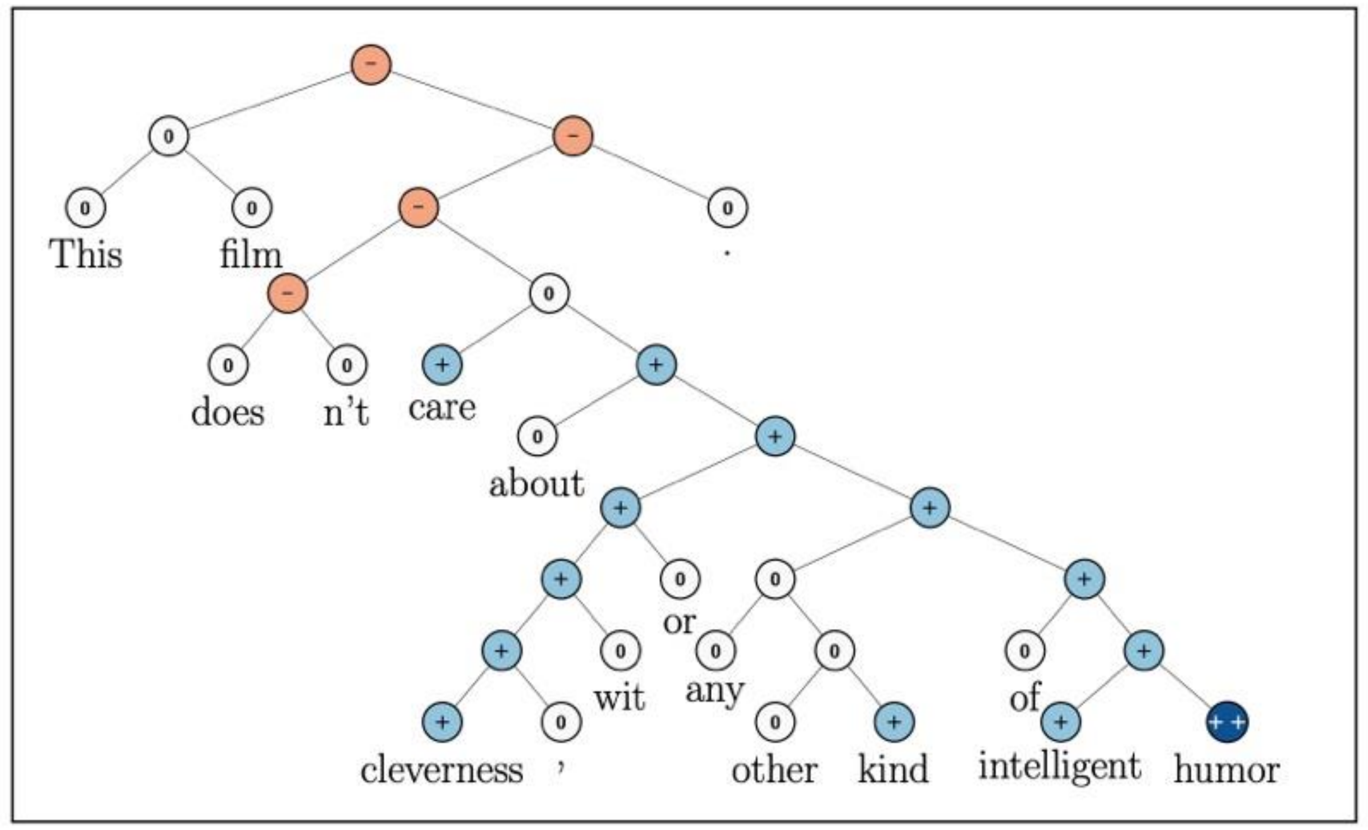
Classification and closed-ended benchmarks

- 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**

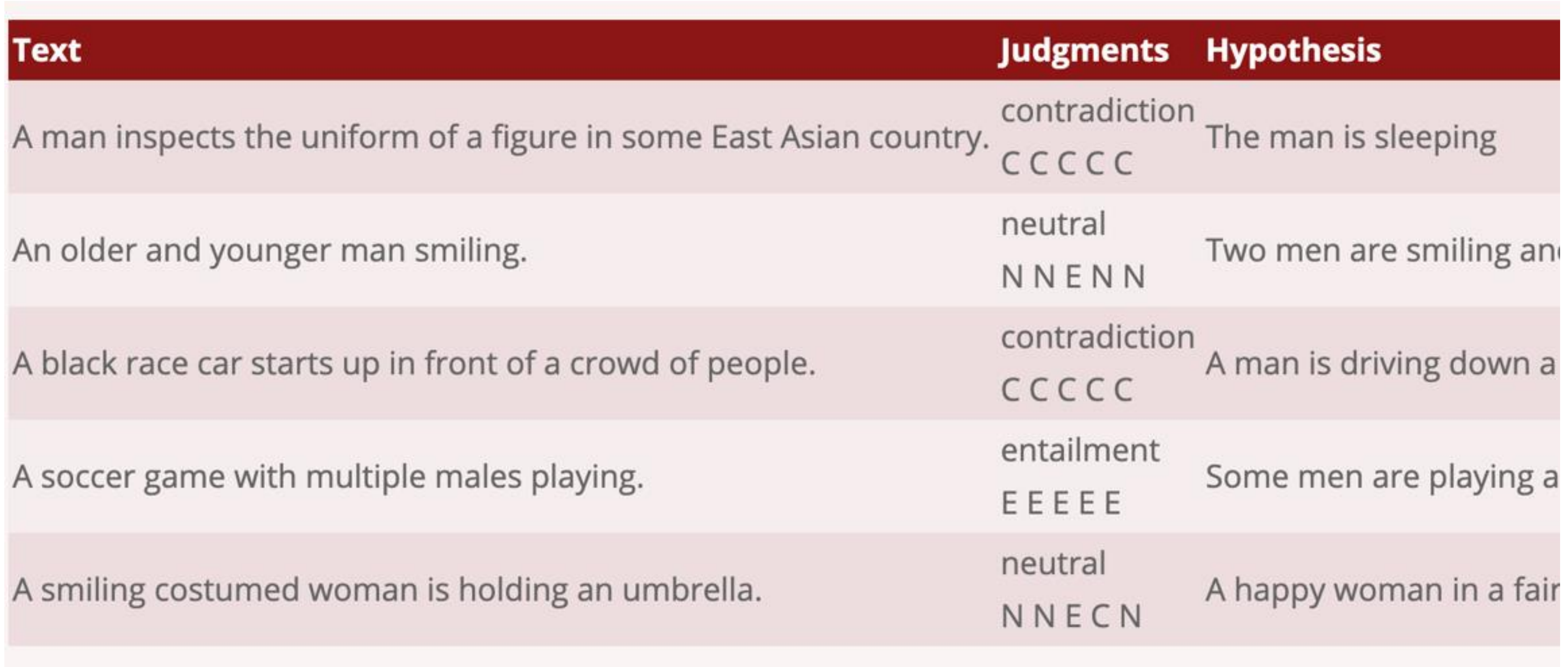
Classification and closed-ended benchmarks

- 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)



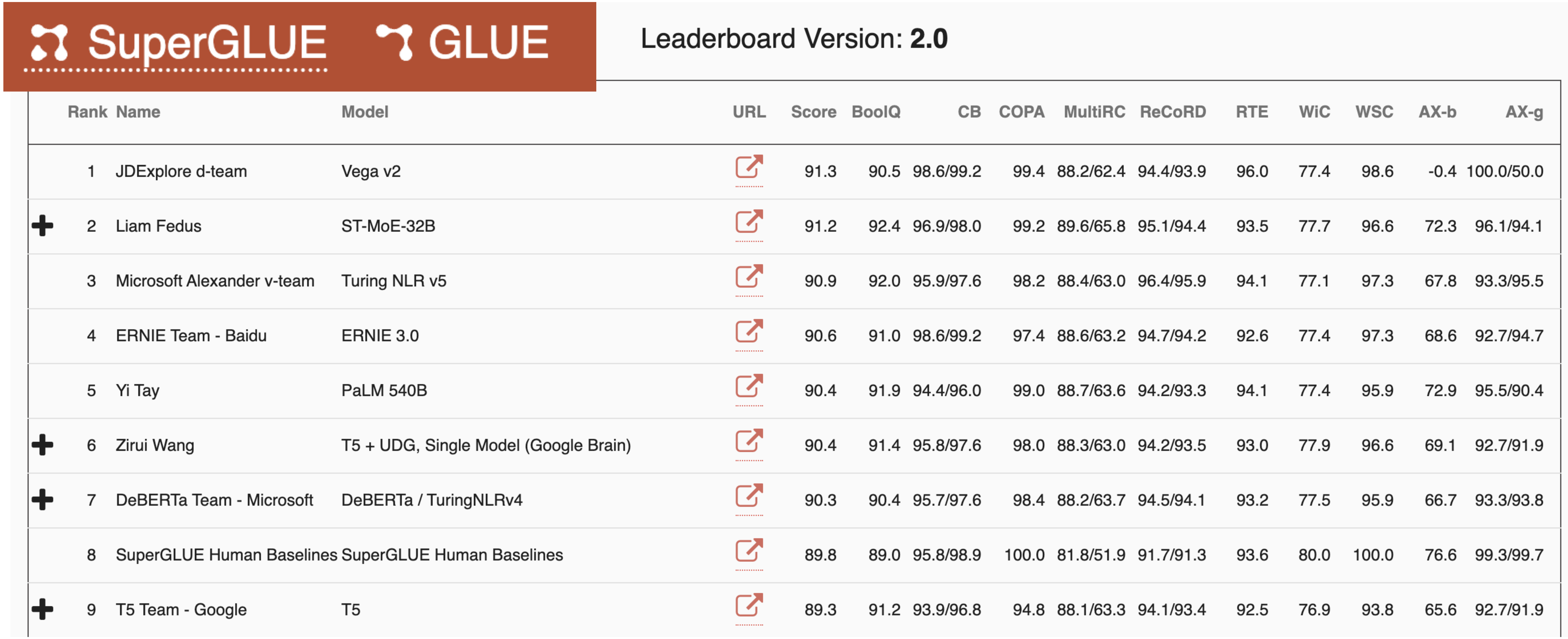
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling an
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is 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 benchmarks



SuperGLUE GLUE

Leaderboard Version: **2.0**

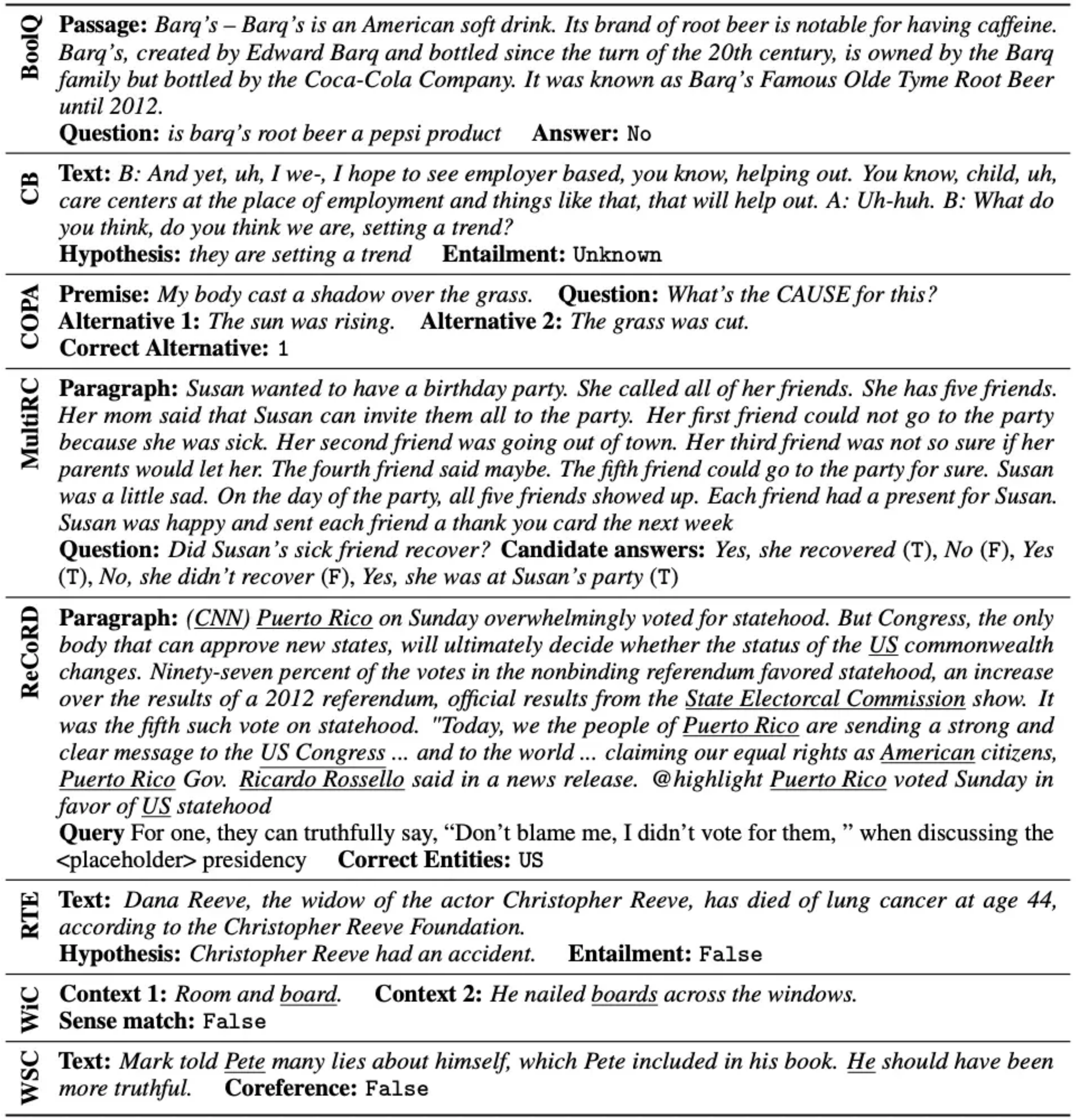
	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2	↗	91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B	↗	91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander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

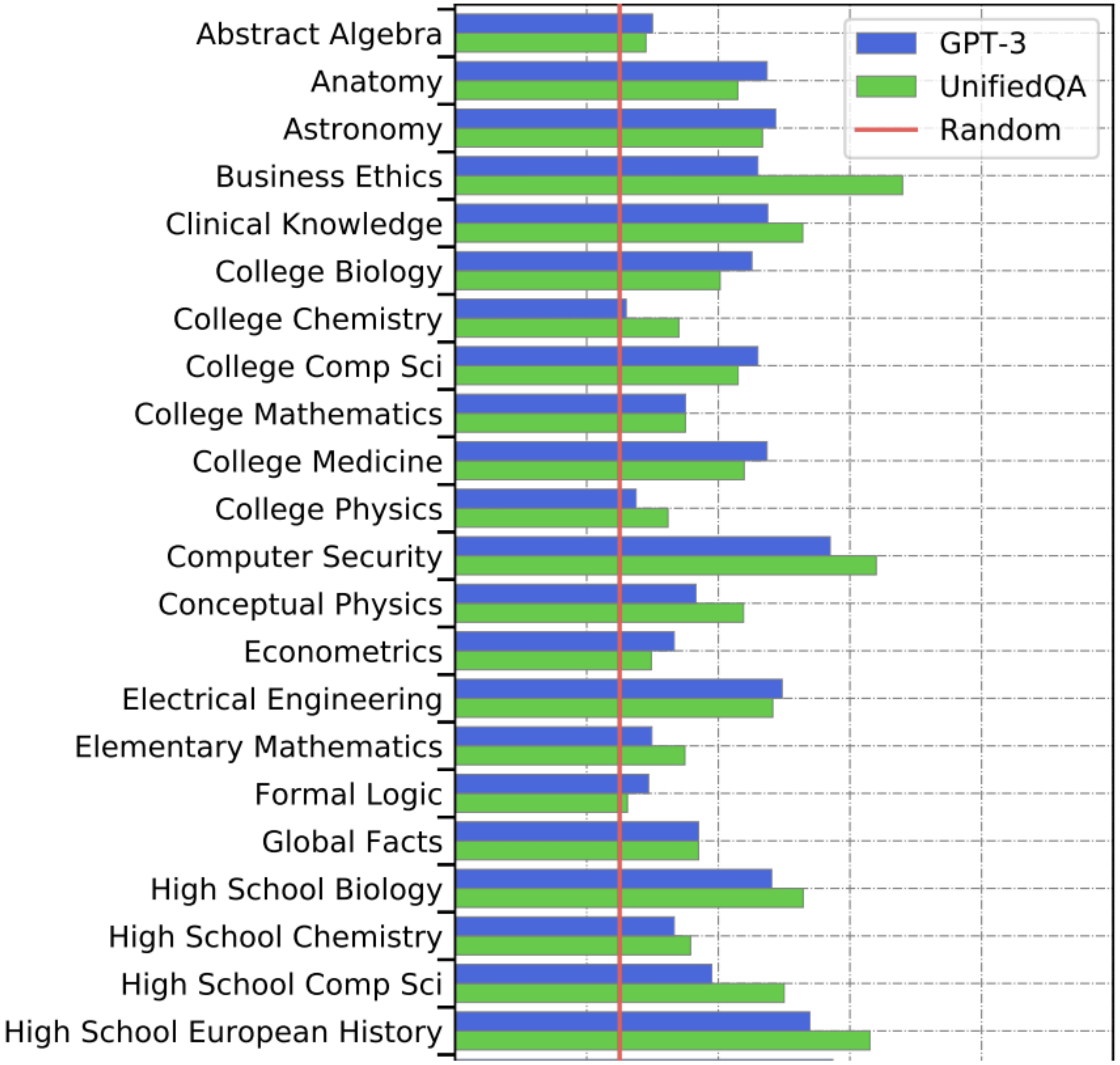


BoolQ	**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme 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`

Another multi-task benchmark: MMLU

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

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



What makes a good benchmark?

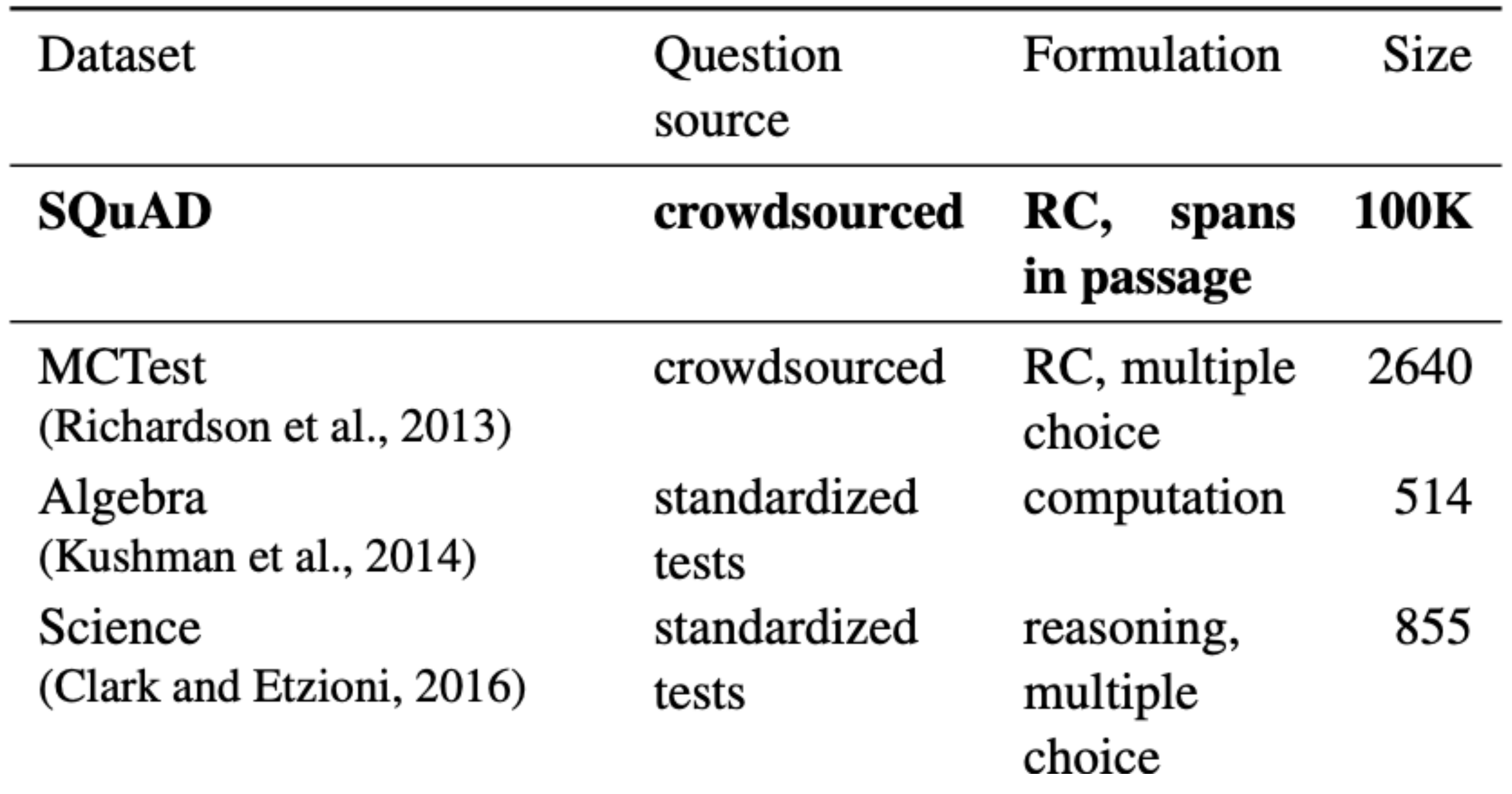
What makes a good benchmark?

- **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 the benchmark was created)

What makes a good benchmark?

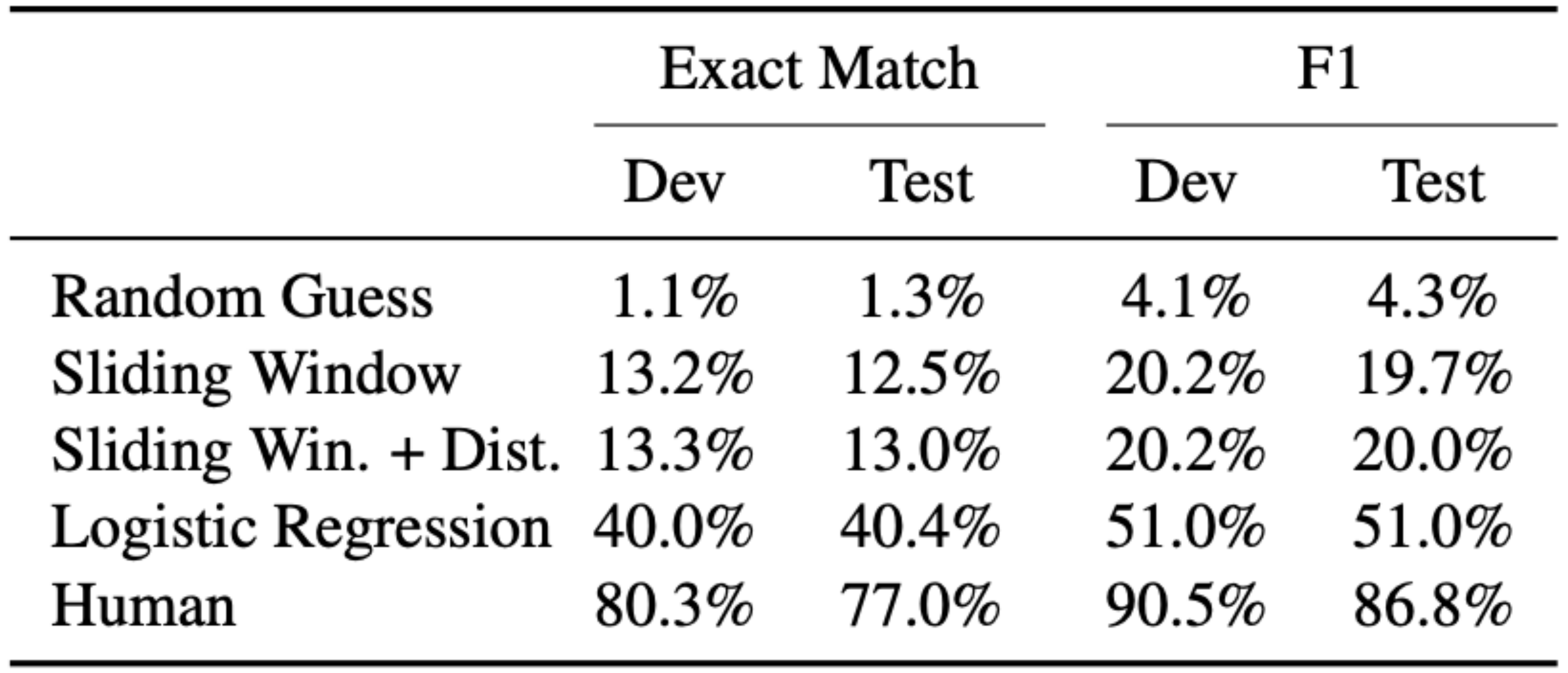
- **Example selection (scale, diversity)**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many samples
- **Difficulty**
  - Doable for humans
  - Hard for baselines (at the time the benchmark was created)
- **Annotation quality and consistency**
  - ‘Correct’ behavior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

A successful benchmark: SQuAD



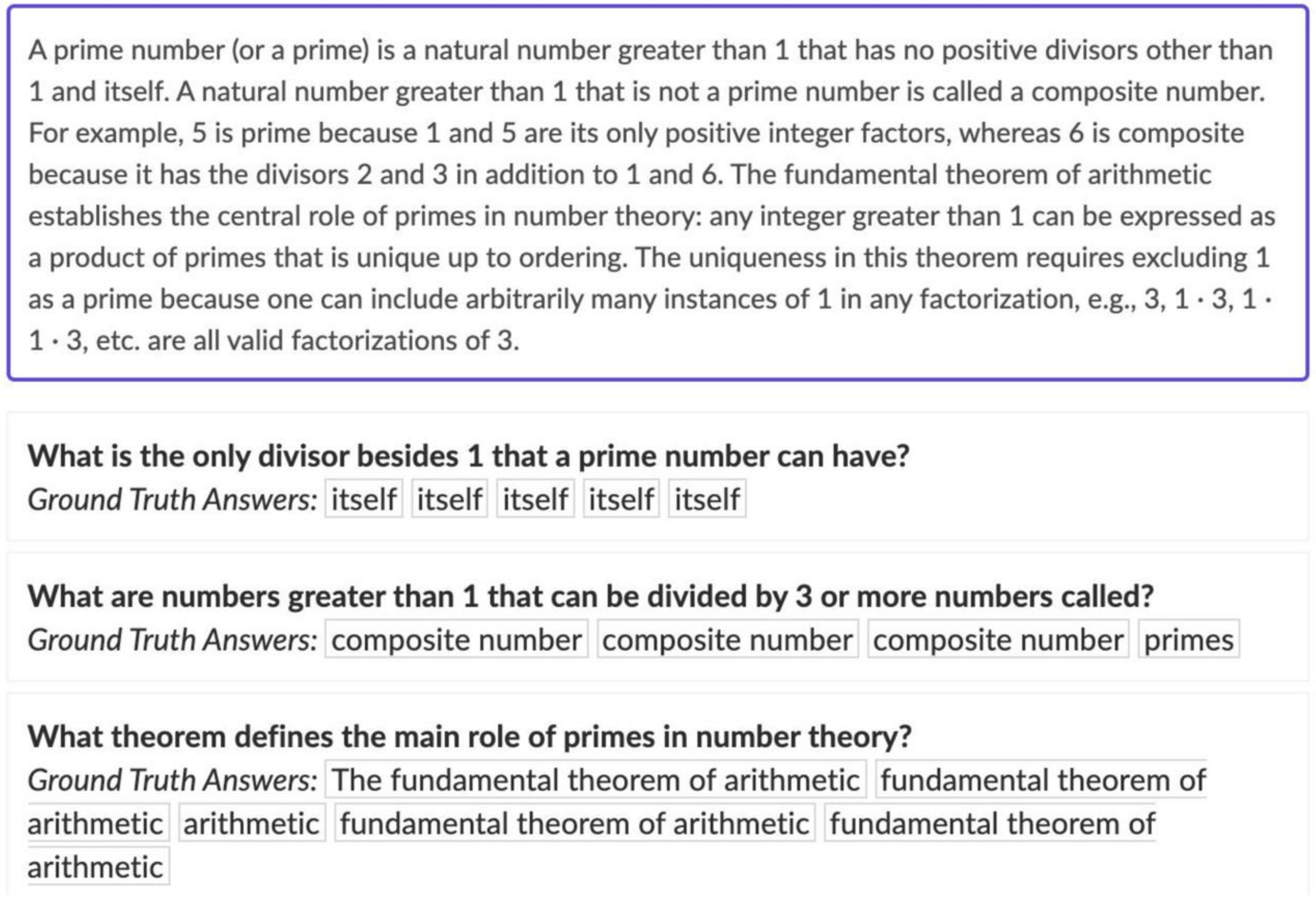
Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple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

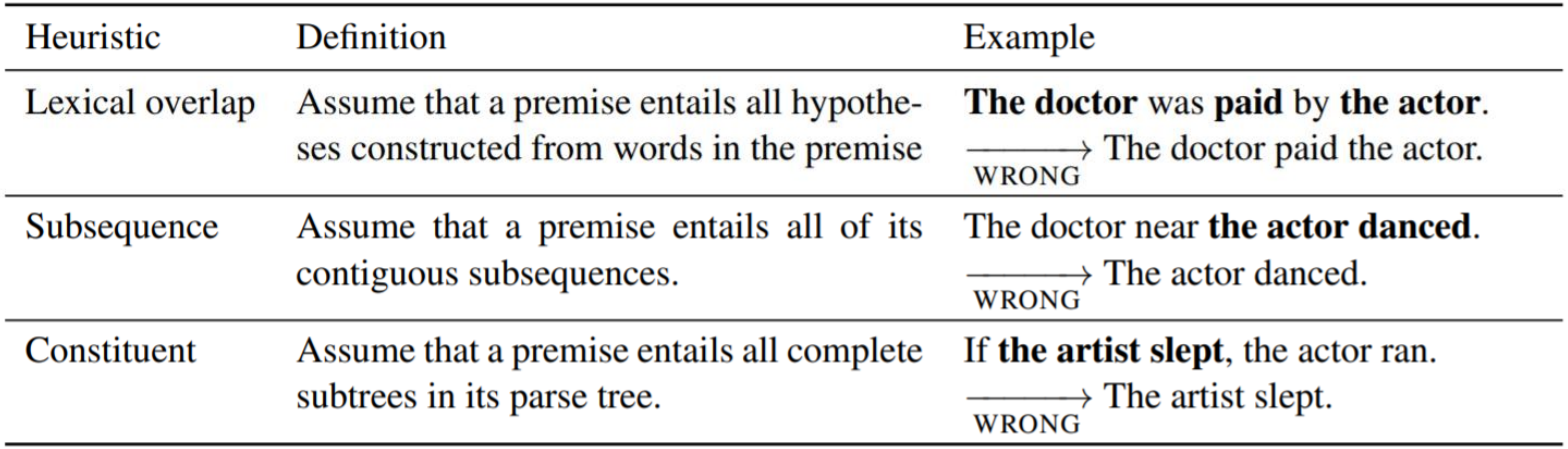
Easy, relatively clean automatic evaluation

Finding model shortcuts with diagnostic tests

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.

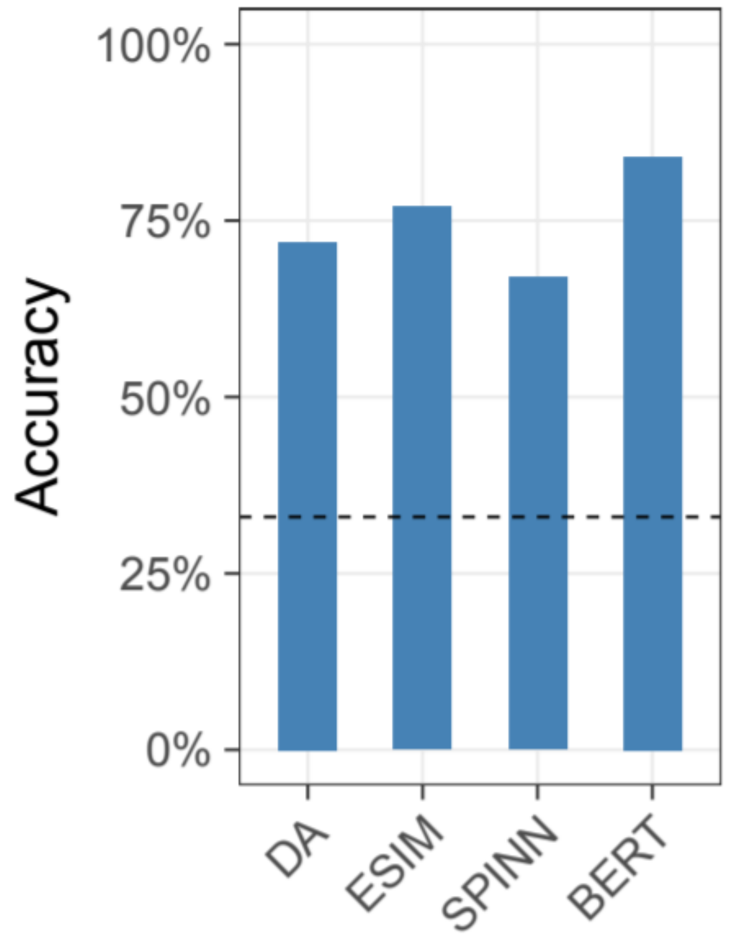
[McCoy et al., 2019]

Clever Hans

1.1 a 1.2 ä 1.3 au 1.4 äu 1.5 au 1.6 b 1.7 f
2.1 ch 2.2 d 2.3 e 2.4 ei 2.5 eu 2.6 f 2.7 g
1 f 3.2 i 3.3 ie 3.4 j 3.5 k 3.6 l 3.7 m
4.2 o 4.3 ö 4.4 p 4.5 q 4.6 r 4.7 s
5.4 ß 5.5 ß 5.6 ß 5.7 t
6.4 u 6.5 ü 6.6 v 6.7 w
2/3 + 3/4 =
26743 : 8 =
129 86 × 3 =

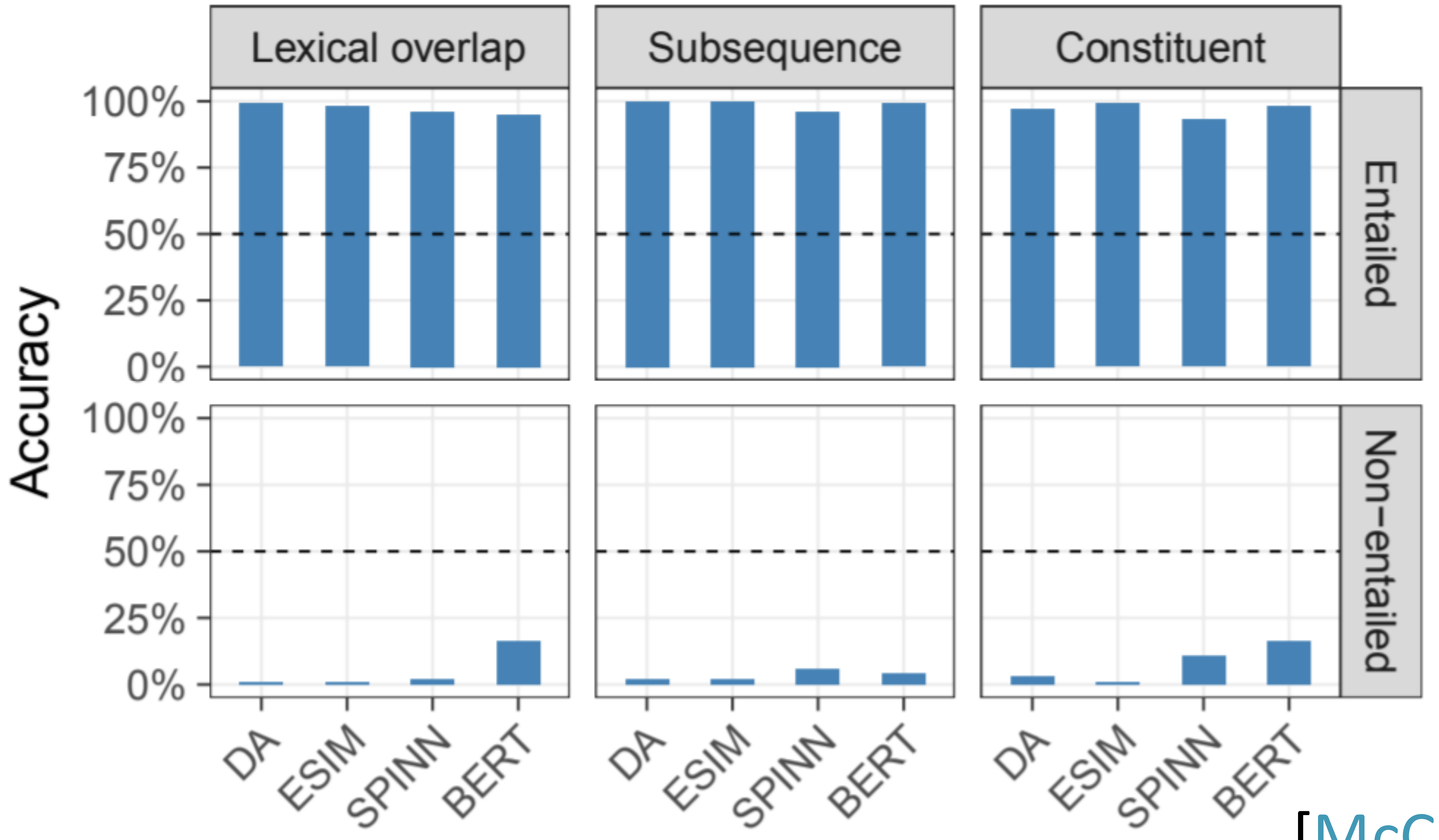
HANS model analysis in natural language inference

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!

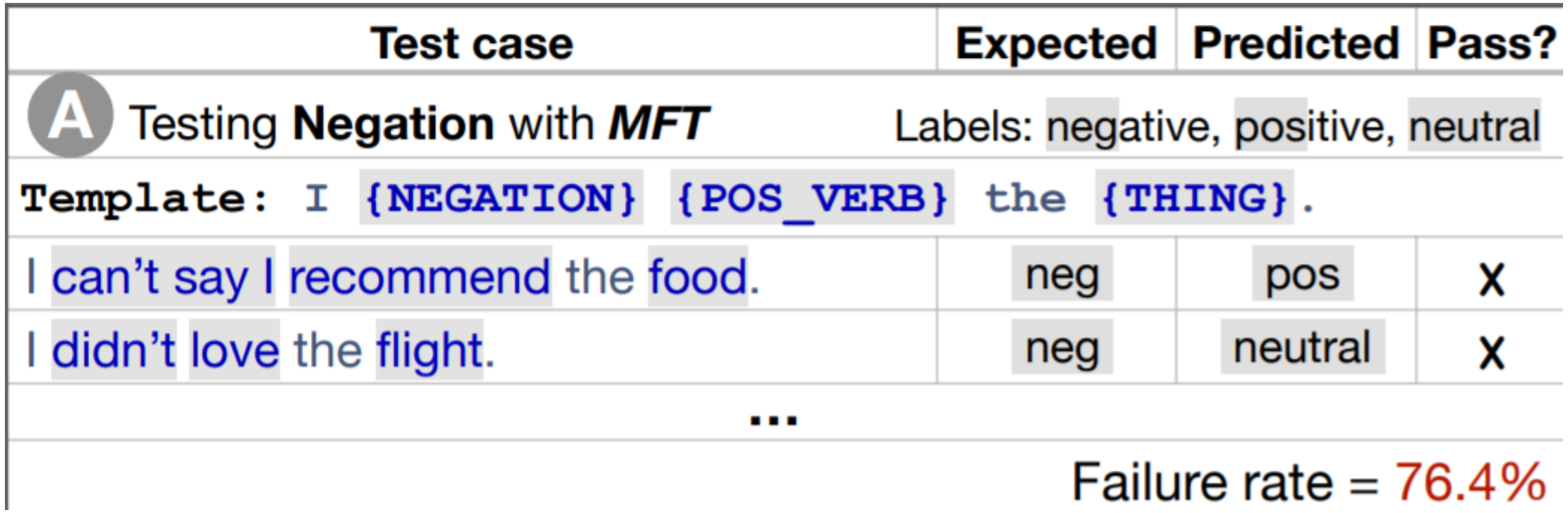
But where syntactic heuristics fail, accuracy is very very low…



[McCoy et al., 2019]

Unit testing for NLP: 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

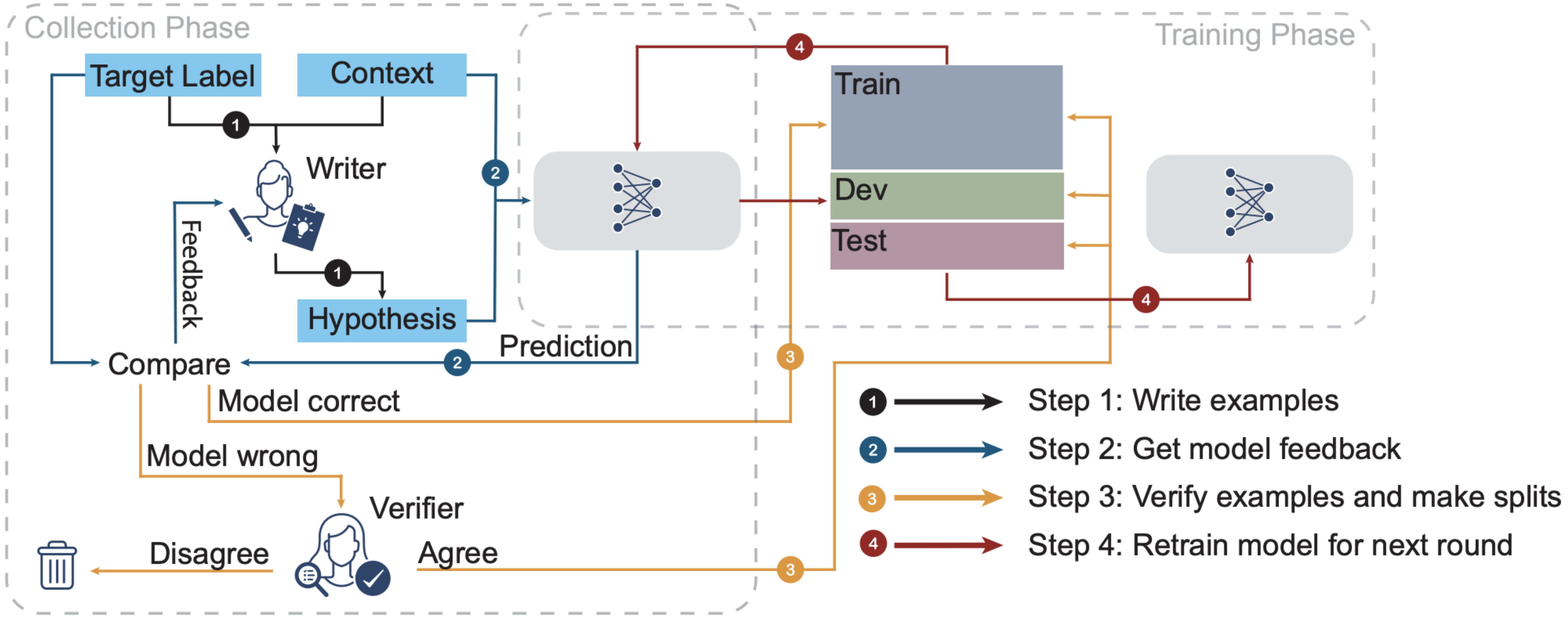
Across a wide range of tasks, high model accuracy on the in-domain test set does not imply the model will also do well on other, “reasonable” out-of-domain examples.

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

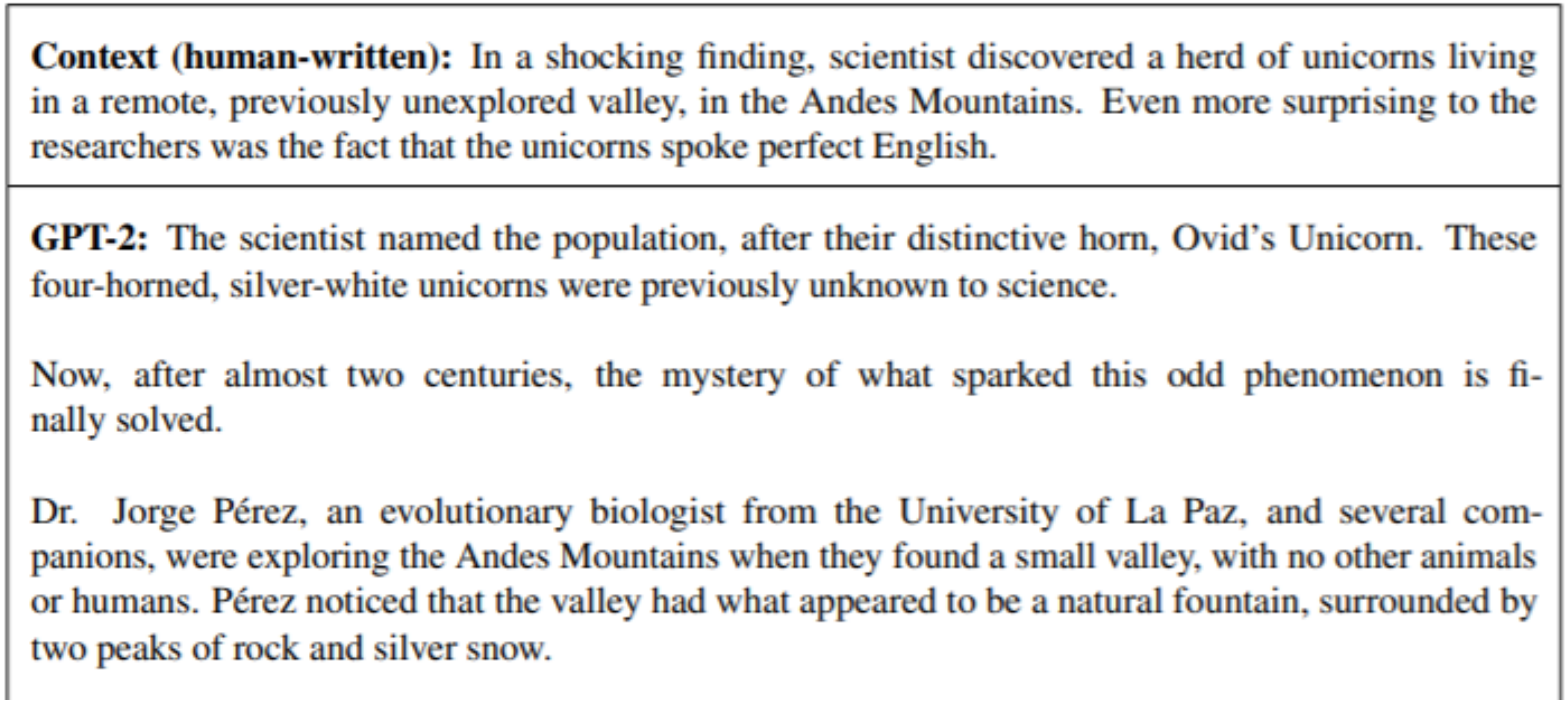
Adversarial NLI (ANLI)



DynaBench

Dyna
Bench

Evaluating open-ended text generation



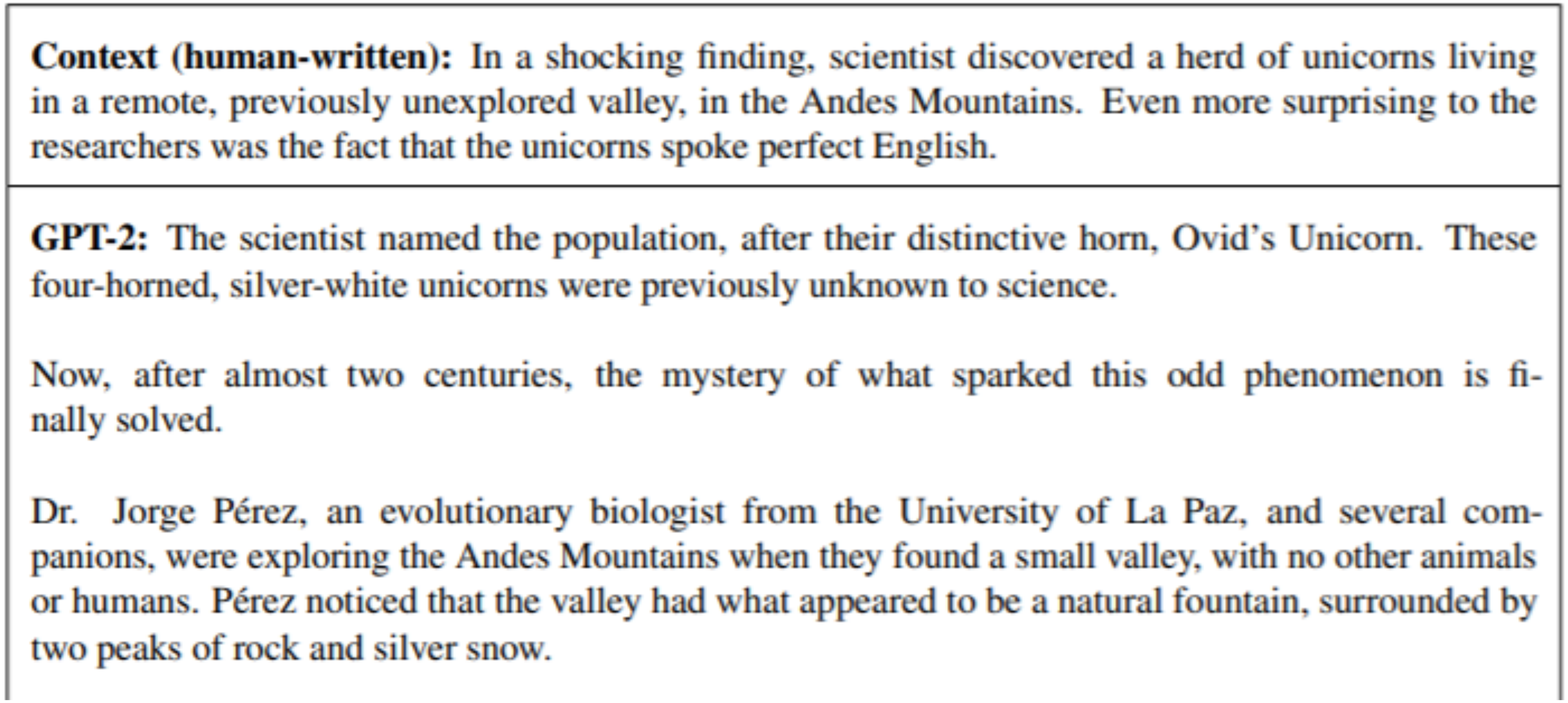
Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

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.

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.

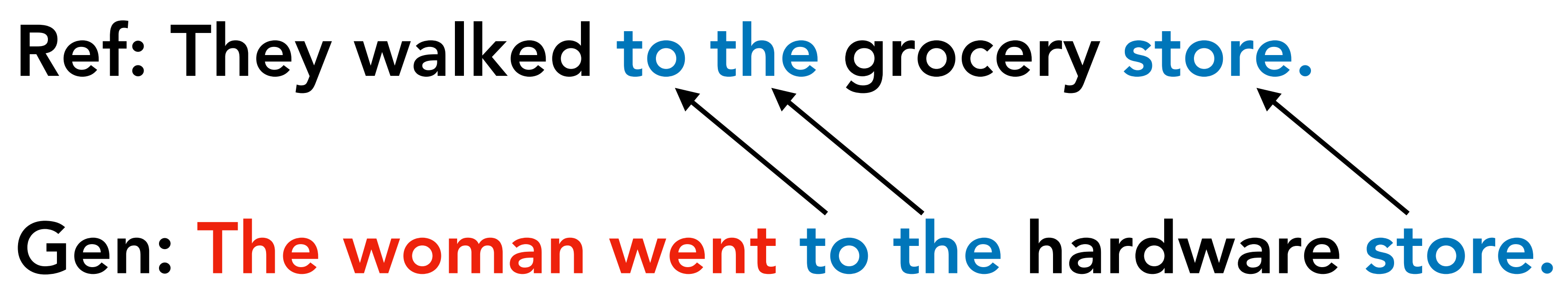
GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

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 a few correct answers to thousands/millions of correct answers
- Can't have human annotators enumerate the right answers (can we?)
- Now grades of correct answers (not just right and wrong)

Types of text evaluation methods



Ref: They walked to the grocery store.

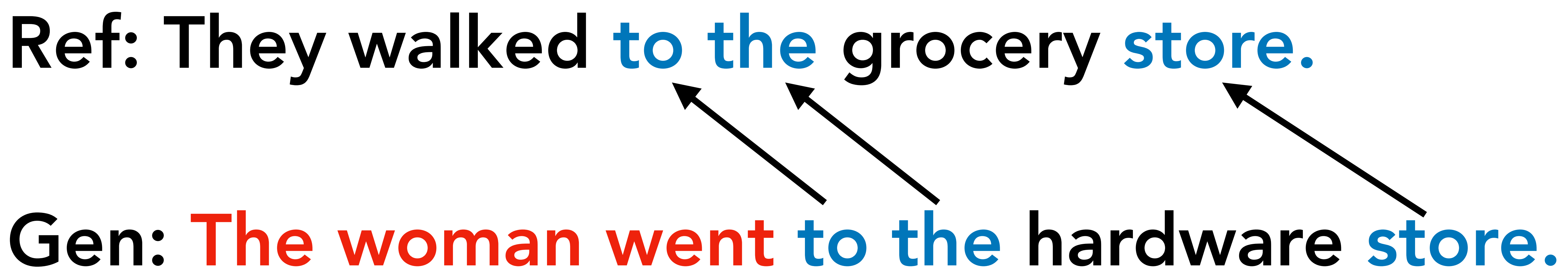
Gen: The woman went to the hardware store.

Content Overlap Metrics

Model-based Metrics

Human Evaluation

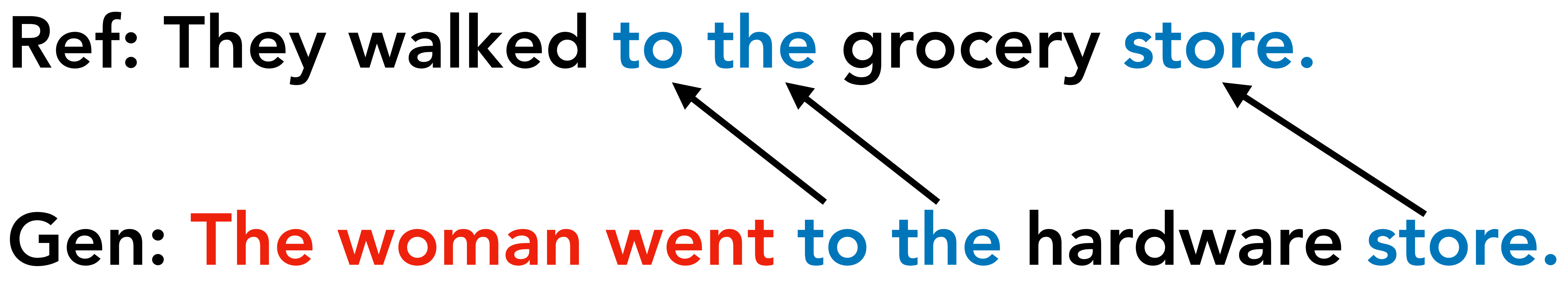
Content Overlap Metrics



Ref: They walked to the grocery store.

Gen: The woman went to the hardware store.

Content Overlap Metrics

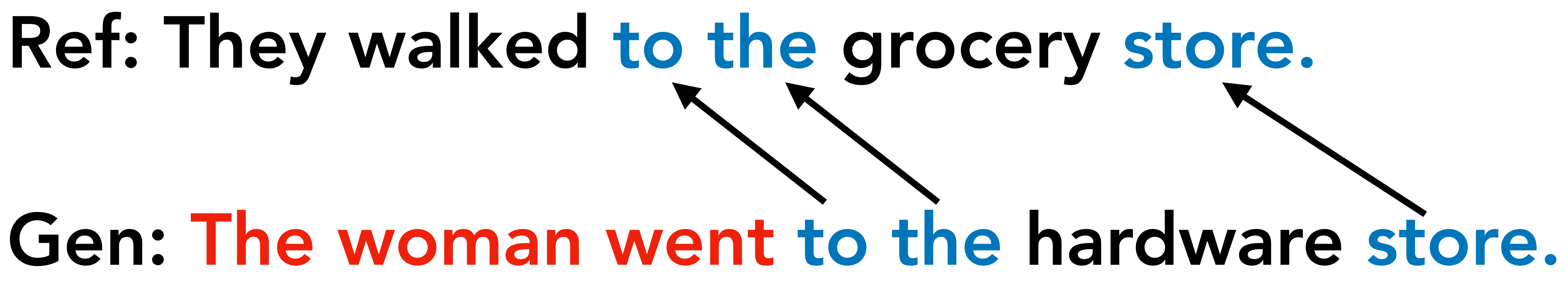


Ref: They walked to the grocery store.

Gen: The woman went to the hardware store.

- Compute a score that indicates the similarity between *generated* and *gold-standard* (often human-written) text

Content Overlap Metrics

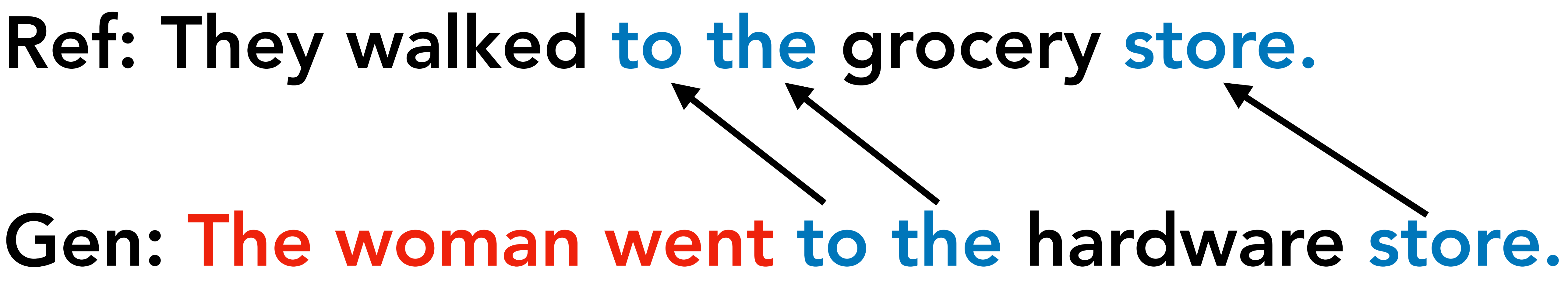


Ref: They walked to the grocery store.

Gen: The woman went to the hardware store.

- Compute a score that indicates the similarity between *generated* and *gold-standard* (often human-written) text
- Fast and efficient; widely used (e.g. for MT and summarization)

Content Overlap Metrics



Ref: They walked to the grocery store.

Gen: The woman went to the hardware store.

- Compute a score that indicates the similarity between *generated* and *gold-standard* (often human-written) text
- Fast and efficient; widely used (e.g. for MT and summarization)
- Dominant approach: *N*-gram overlap metrics
  - e.g., BLEU, ROUGE, METEOR, CIDEr, etc.

Content Overlap Metrics

Content Overlap Metrics

- Dominant approach: *N*-gram overlap metrics
  - e.g., BLEU, ROUGE, METEOR, CIDEr, etc.

Content Overlap Metrics

- Dominant approach: *N*-gram overlap metrics
  - e.g., BLEU, ROUGE, METEOR, CIDEr, etc.

- Not ideal even for less open-ended tasks - e.g., machine translation

Content Overlap Metrics

- Dominant approach: *N*-gram overlap metrics
  - e.g., BLEU, ROUGE, METEOR, CIDEr, etc.

- Not ideal even for less open-ended tasks - e.g., machine translation

- They get progressively much worse for more open-ended tasks
  - **Worse** for summarization, as longer summaries are harder to measure
  - **Much worse** for dialogue (in how many ways can you respond to your friend?)
  - **Much, much worse** for 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**!

A simple failure case

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

Are you enjoying the NLP class?

For sure!

A simple failure case

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

Are you enjoying the NLP class?

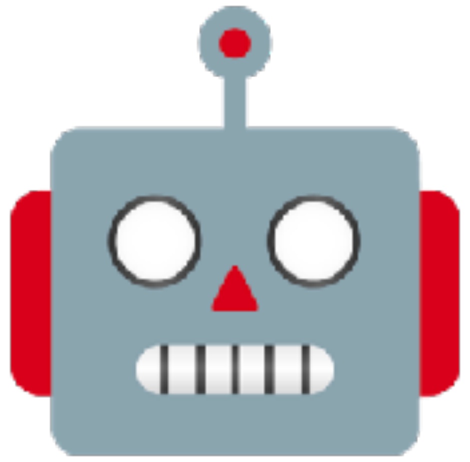
For sure!

Yes for sure!

Sure I do!

Yes!

No for sure...



A simple failure case

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

Are you enjoying the NLP class?

For sure!

Score:	
0.61	Yes for sure!
0.25	Sure I do!
0.0	Yes!
0.61	No for sure...

A simple failure case

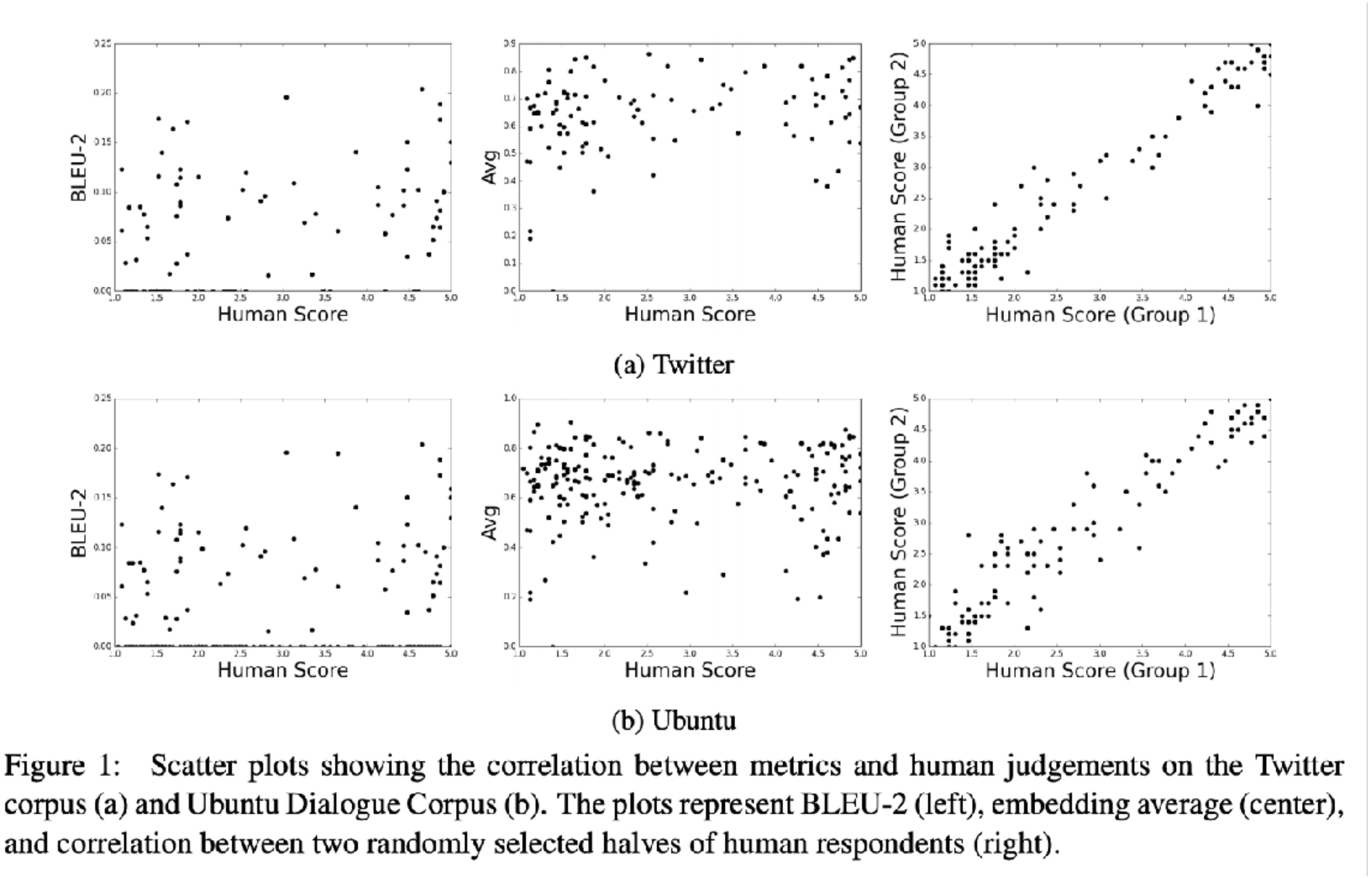
- *N*-gram overlap metrics have no concept of **semantic relatedness**!

Are you enjoying the NLP class?

For sure!

	Score:	
	0.61	Yes for sure!
	0.25	Sure I do!
False negative	0.0	Yes!
False positive	0.61	No for sure...

A more comprehensive failure analysis

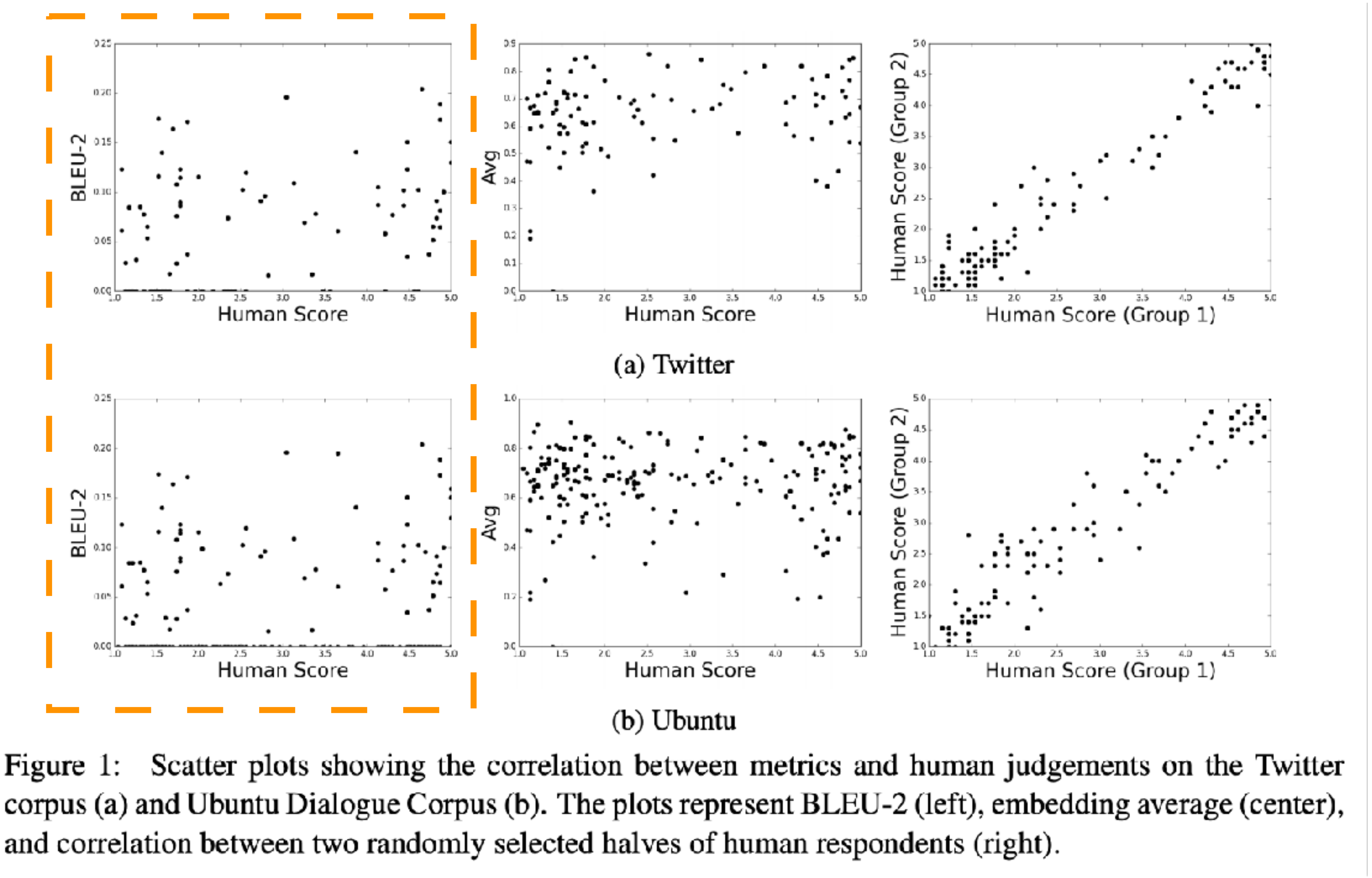


(a) Twitter

(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).

A more comprehensive failure analysis



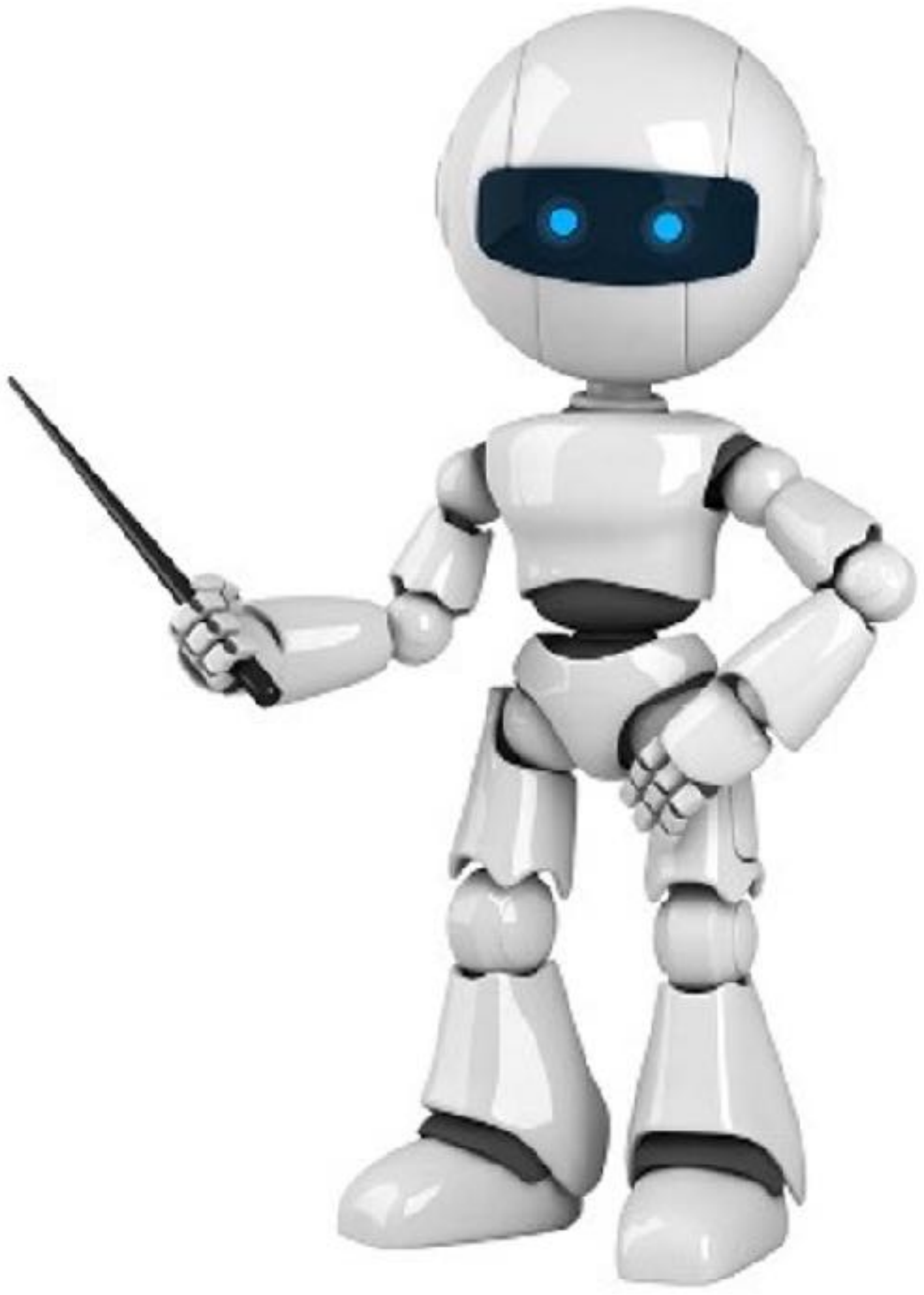
(a) Twitter

(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).

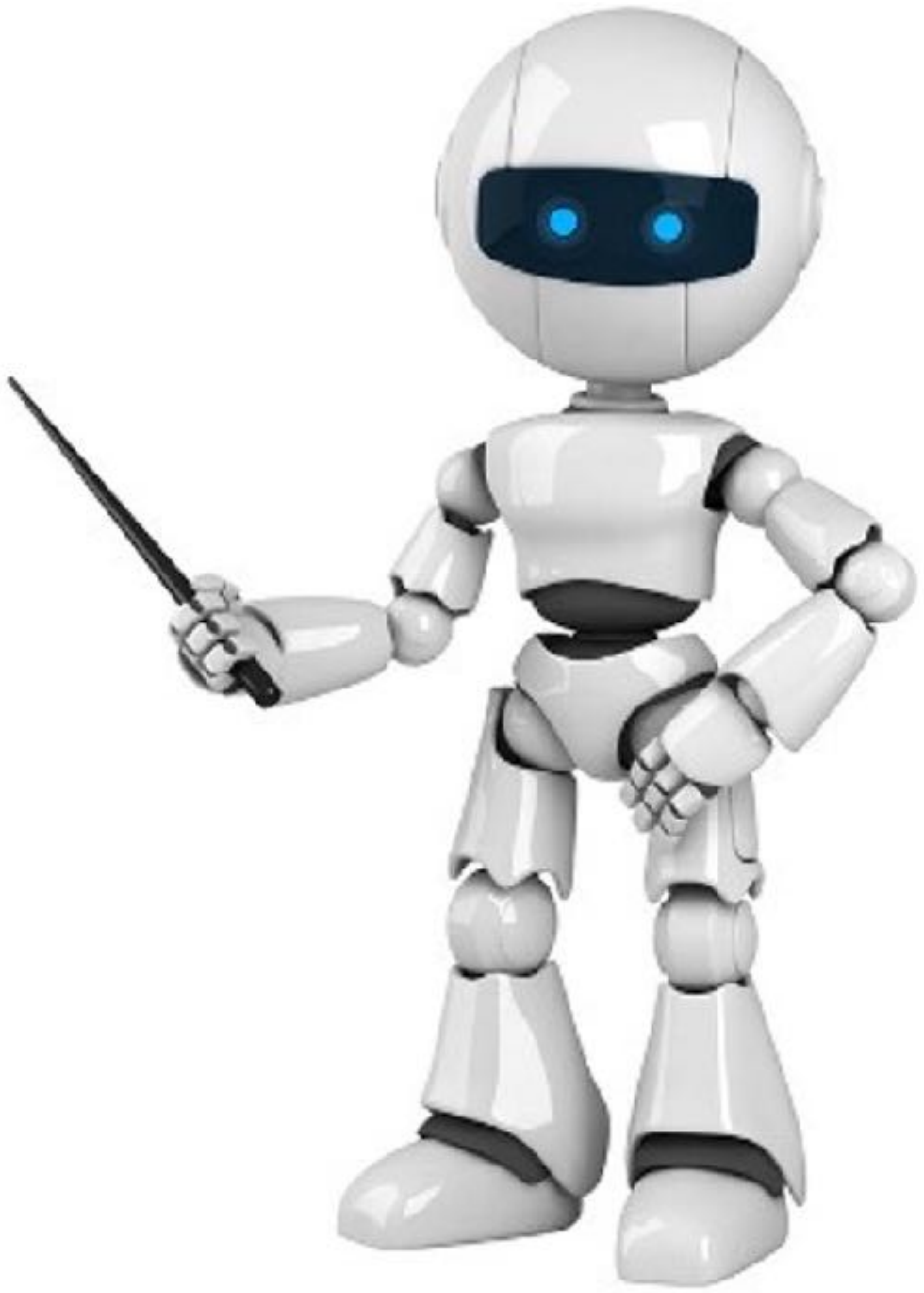
- Higher *n-gram overlap* does not imply higher ***human score***.

Model-based metrics to capture more semantics



Model-based metrics to capture more semantics

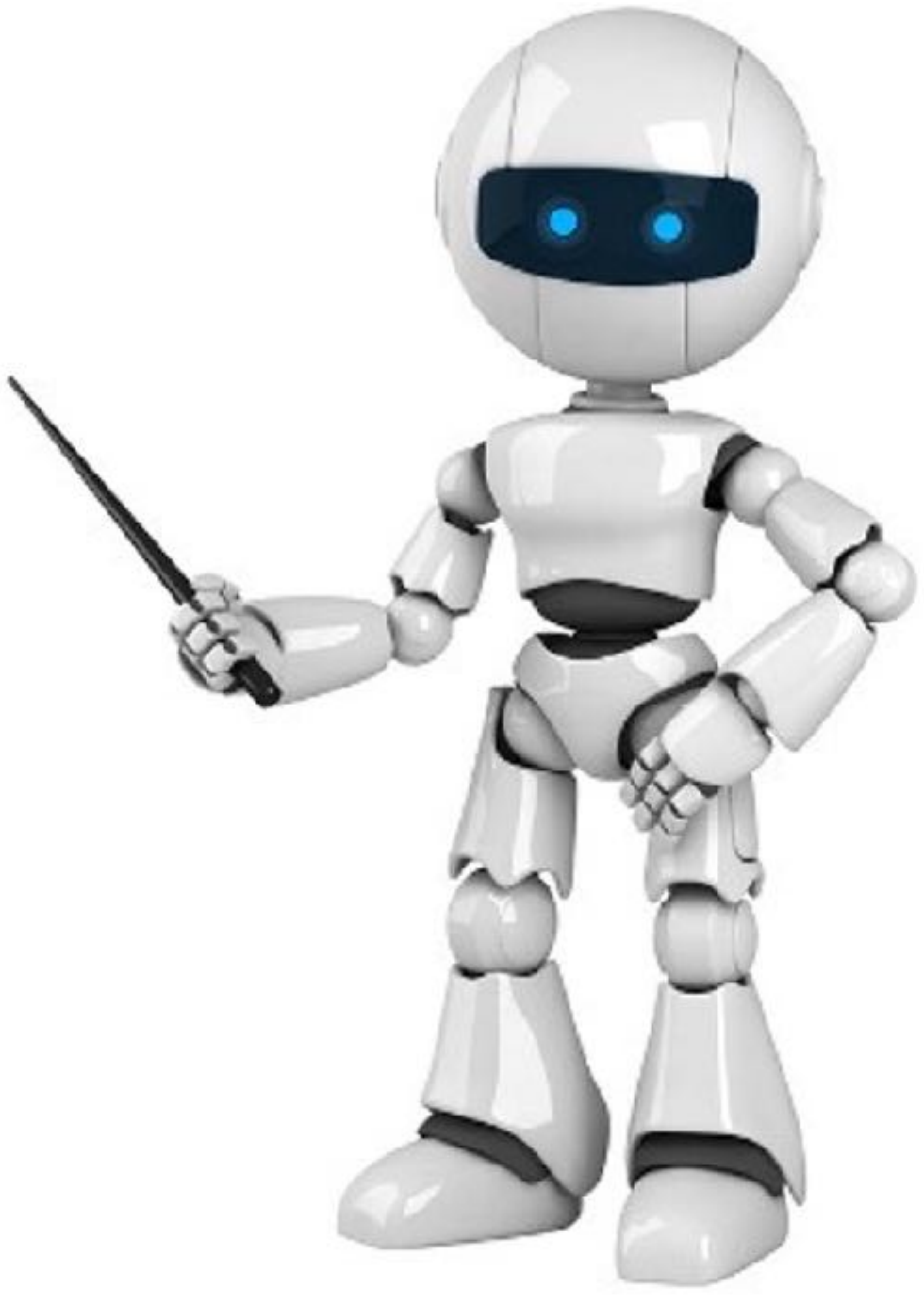
- Use learned representation of words and sentences to compute semantic similarity between generated and reference texts



Model-based metrics to capture more semantics

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

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

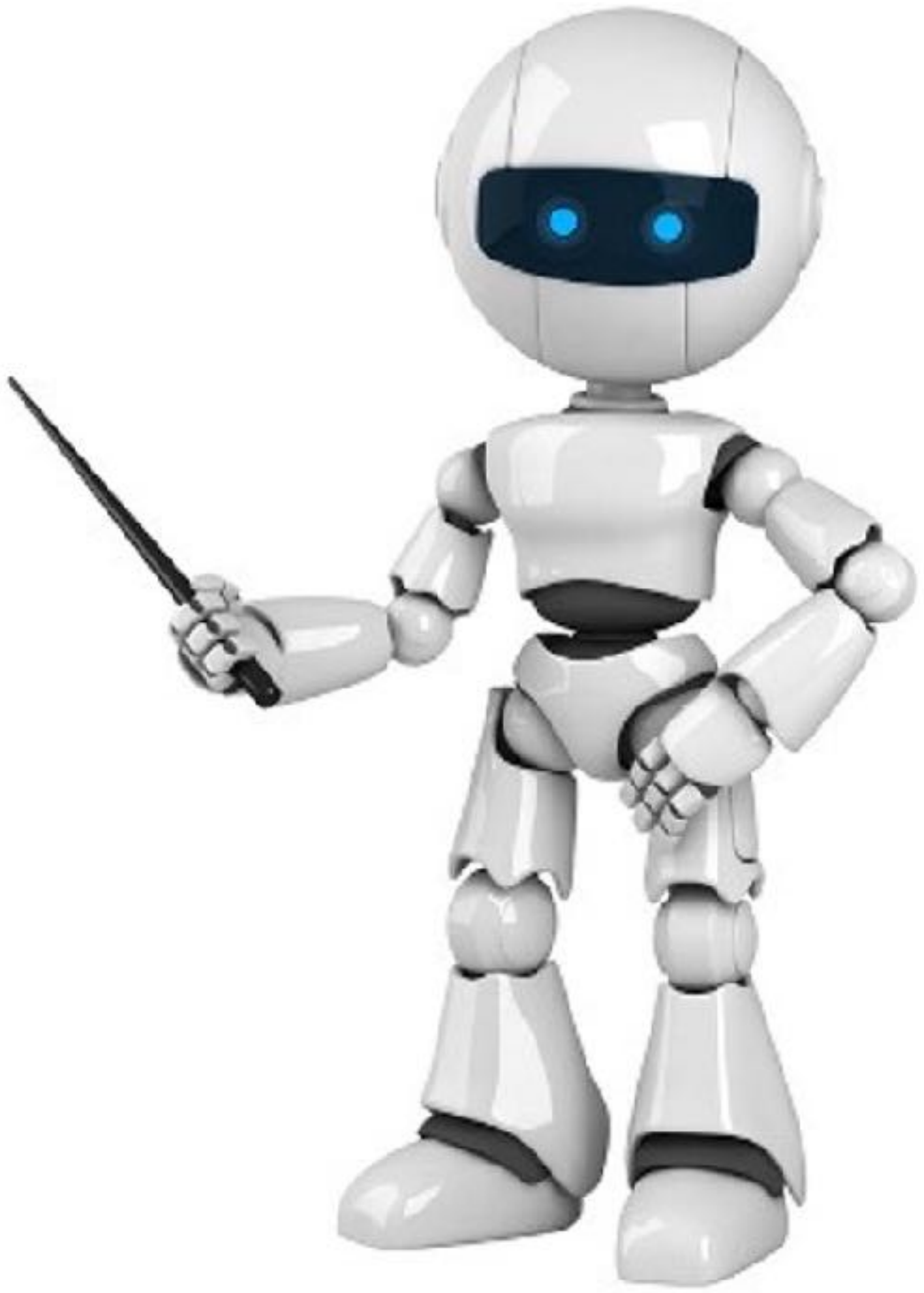


Model-based metrics to capture more semantics

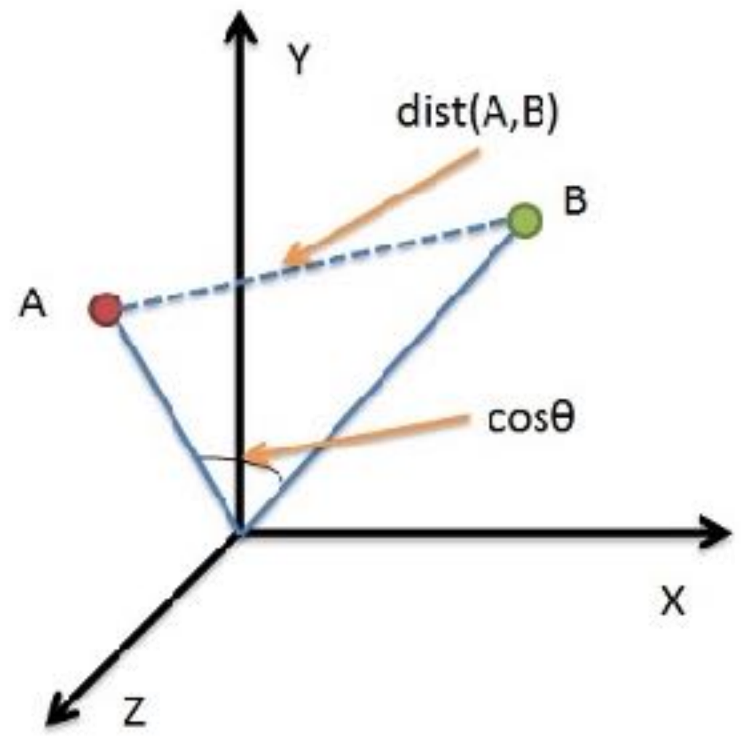
- Use learned representation of words and sentences to compute semantic similarity between generated and reference texts

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

- Even though embeddings are pre-trained, distance metrics used to measure similarity can be fixed.



Model-based metrics: Word distance functions

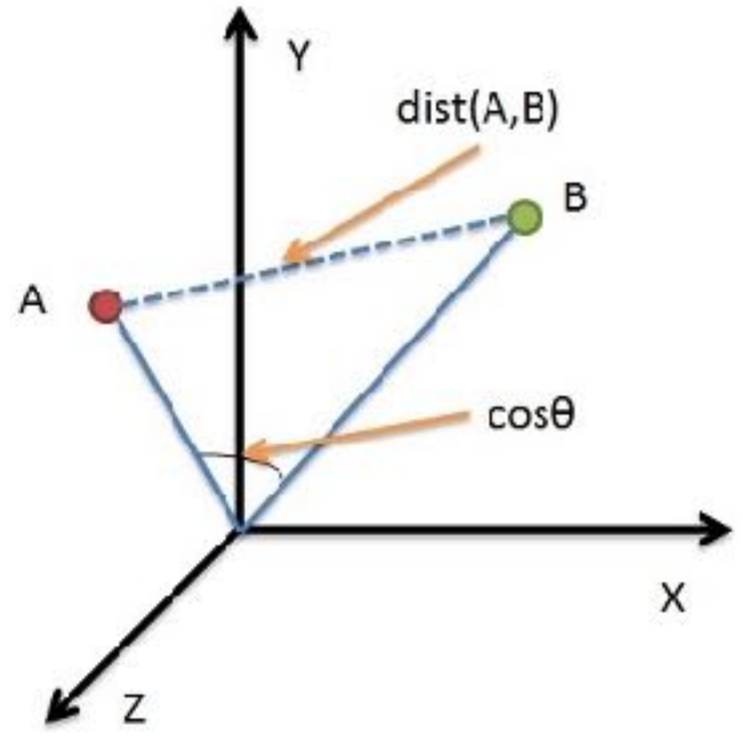


Vector Similarity

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)*

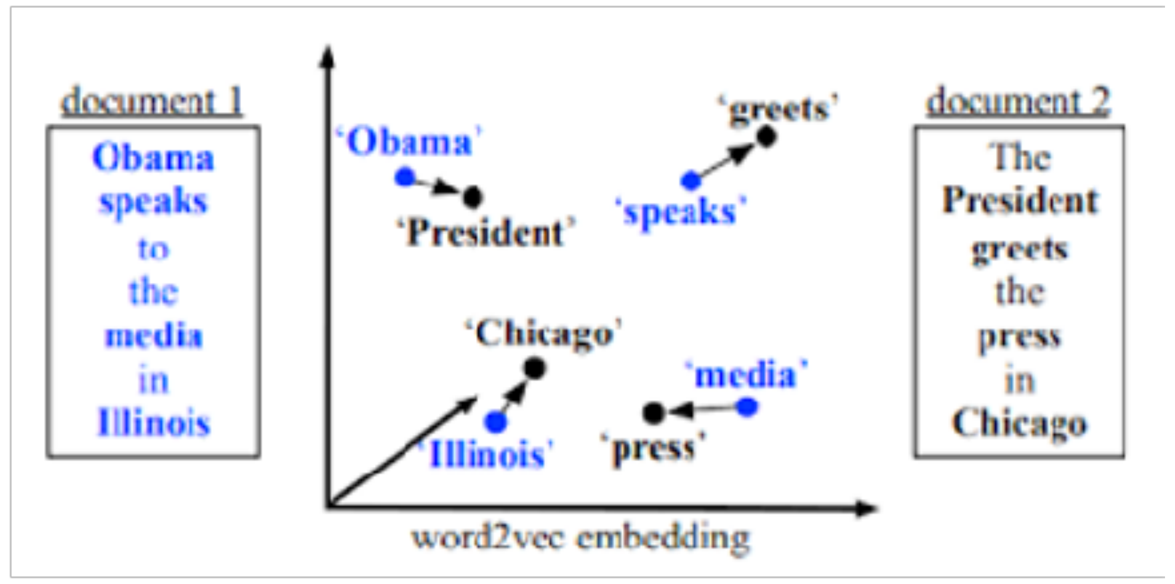
Model-based metrics: Word distance functions



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)*

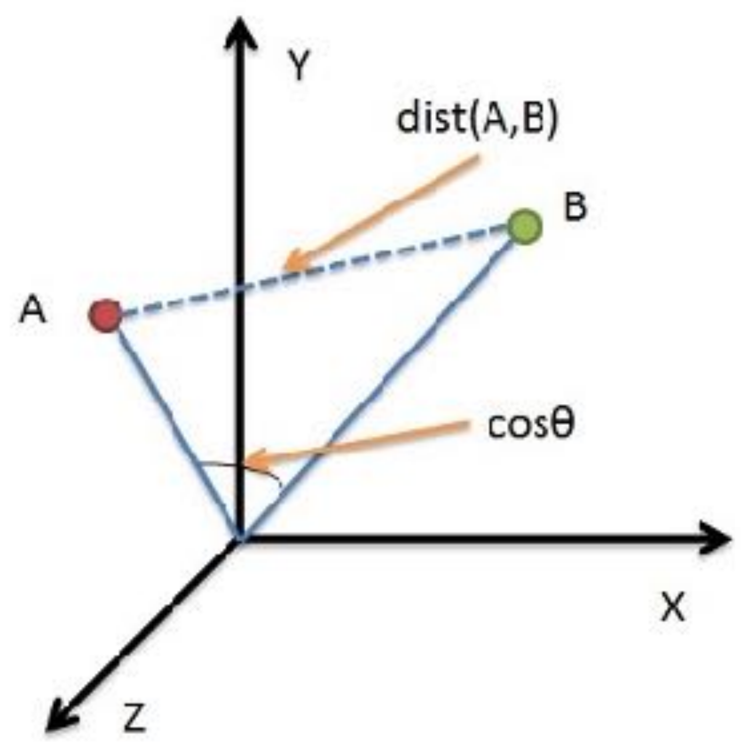


Word Mover's Distance

Measures the distance between two sequences using word embedding similarity matching.

- *(Kusner et al., 2015; Zhao et al., 2019)*

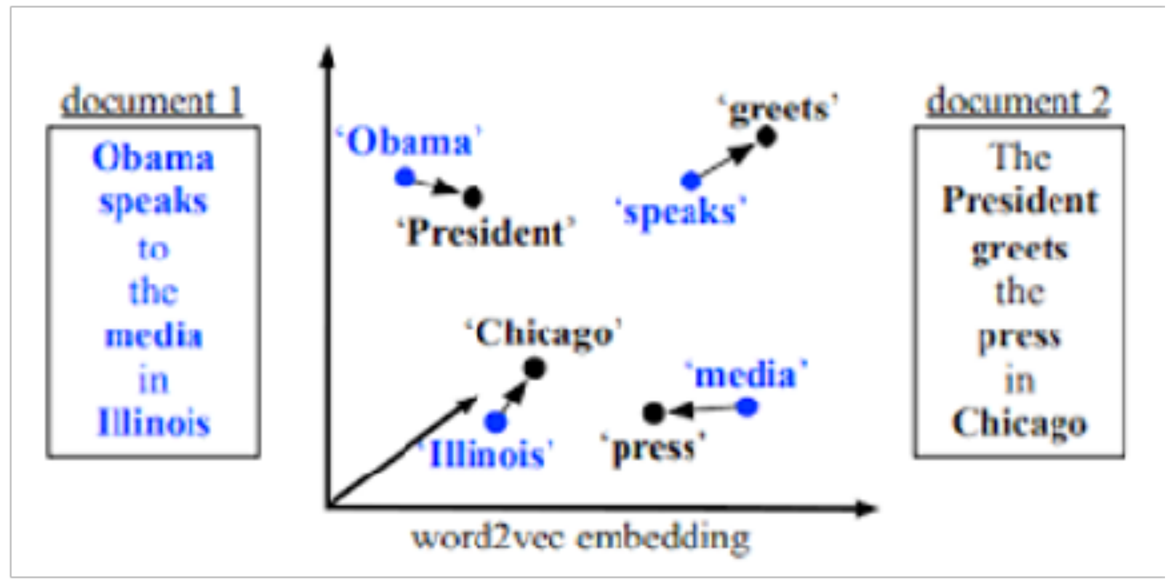
Model-based metrics: Word distance functions



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

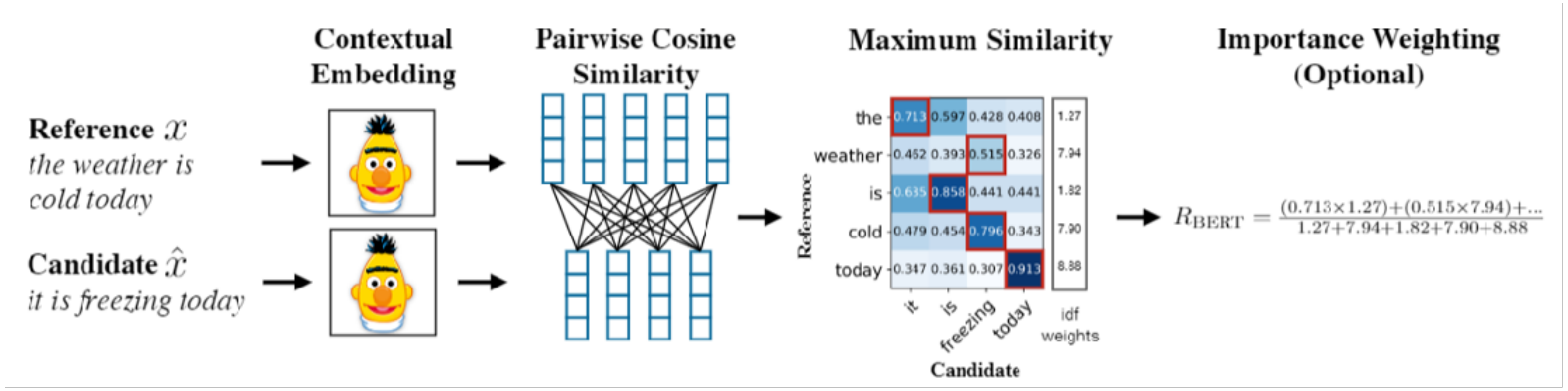
Measures the distance between two sequences using word embedding similarity matching.

- *(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., 2019)*

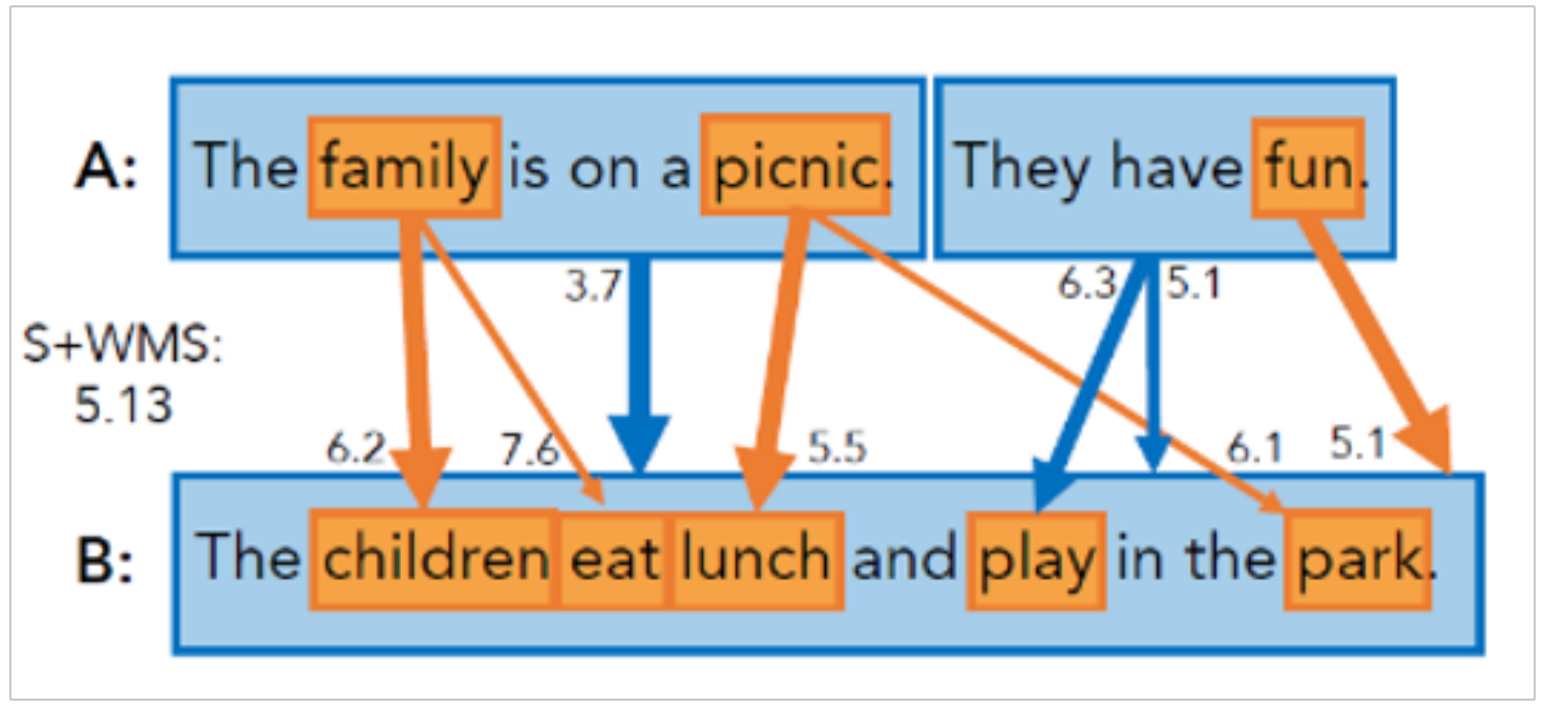


	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.452	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

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

Model-based metrics: Beyond word matching

Model-based metrics: Beyond word matching

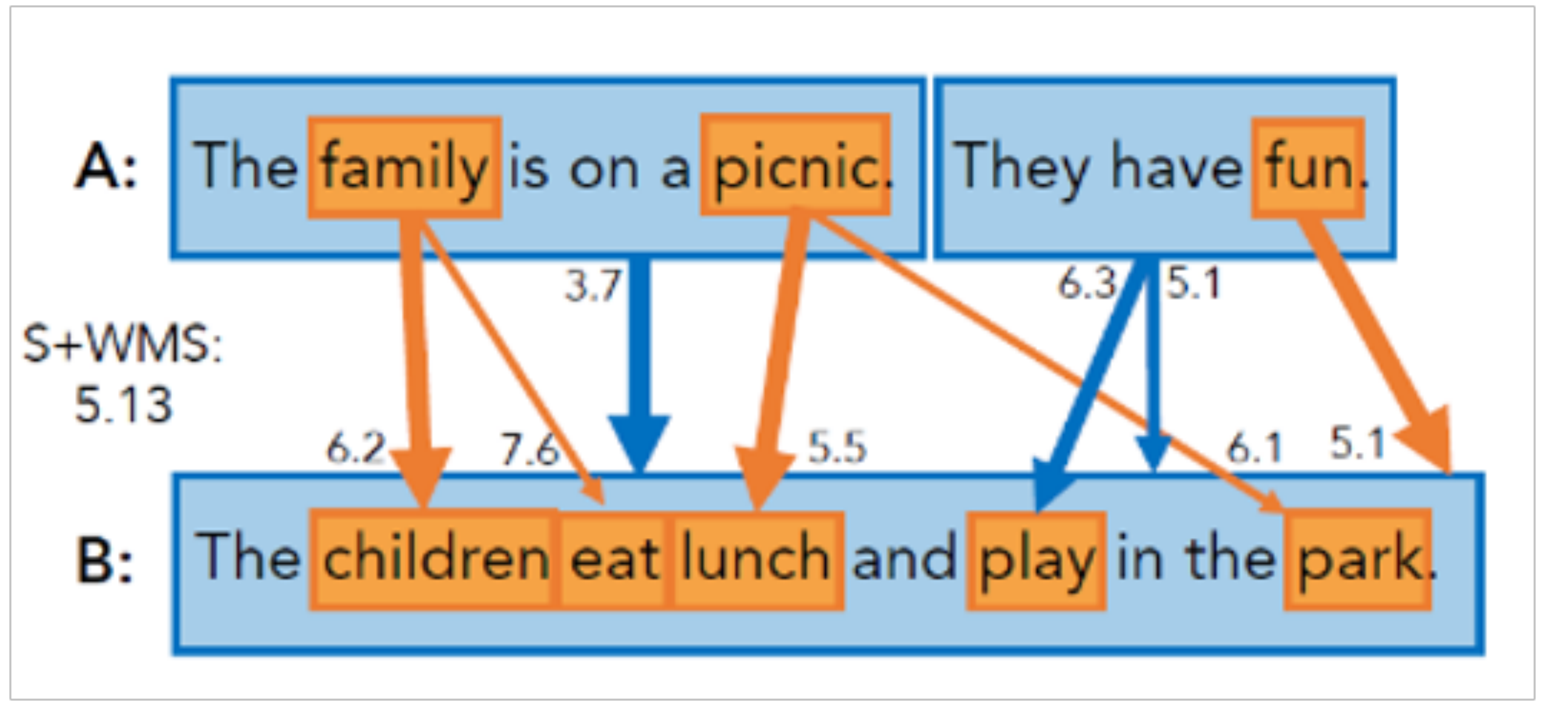

A: The family is on a picnic. They have fun.
S+WMS: 5.13
3.7
6.3
5.1
6.2
7.6
5.5
6.1
5.1
B: The children eat lunch and play in the park.

Sentence Mover's Similarity

Extends word mover's distance to multi-sentence level. Evaluates similarity using sentence embeddings from recurrent neural network representations.

- *(Clark et al., 2019)*

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 Mover's Similarity

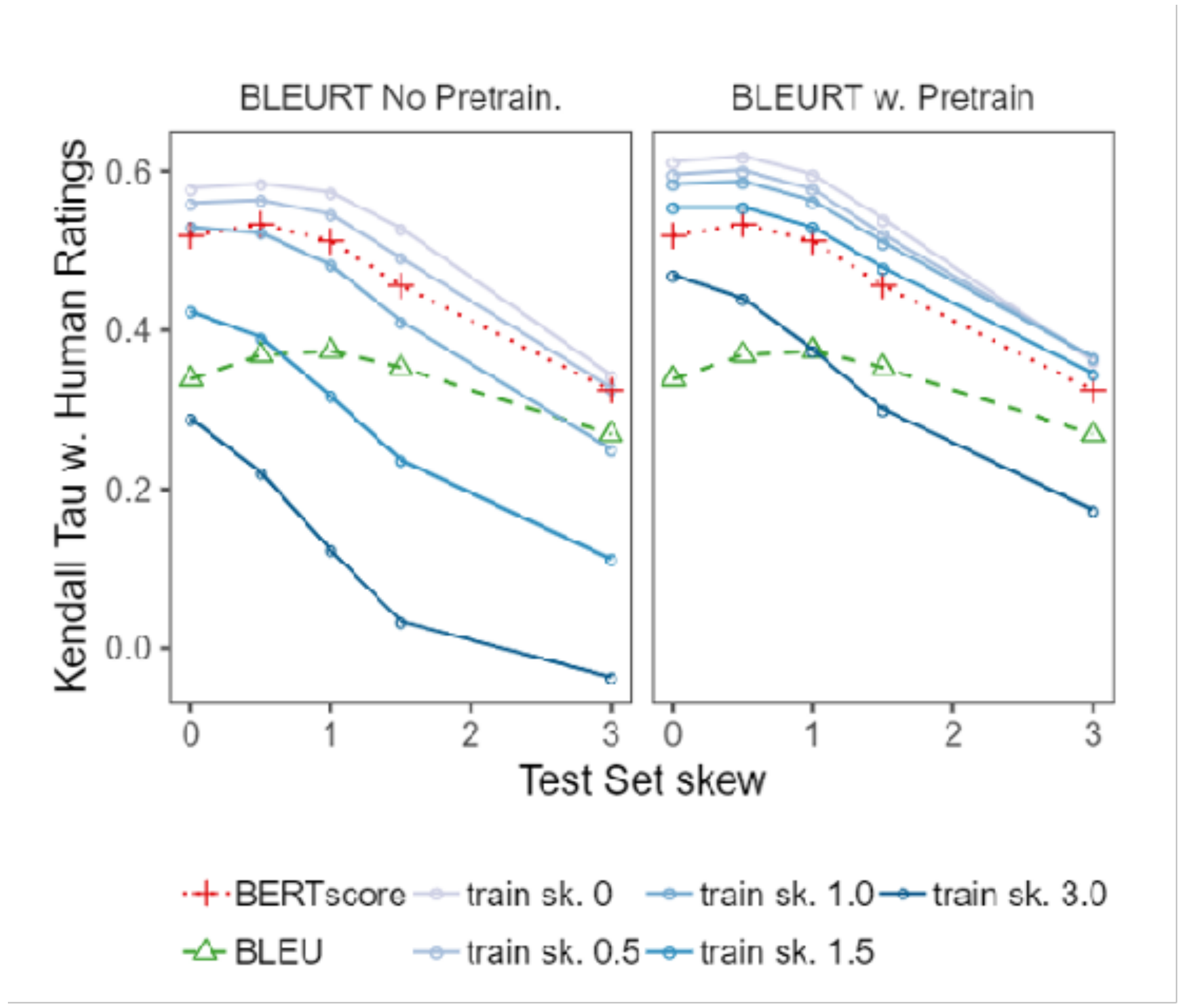
Extends word mover's distance to multi-sentence level. Evaluates similarity using sentence embeddings from recurrent neural network representations.

- *(Clark et al., 2019)*

BLEURT

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

- *(Sellam et al., 2020)*



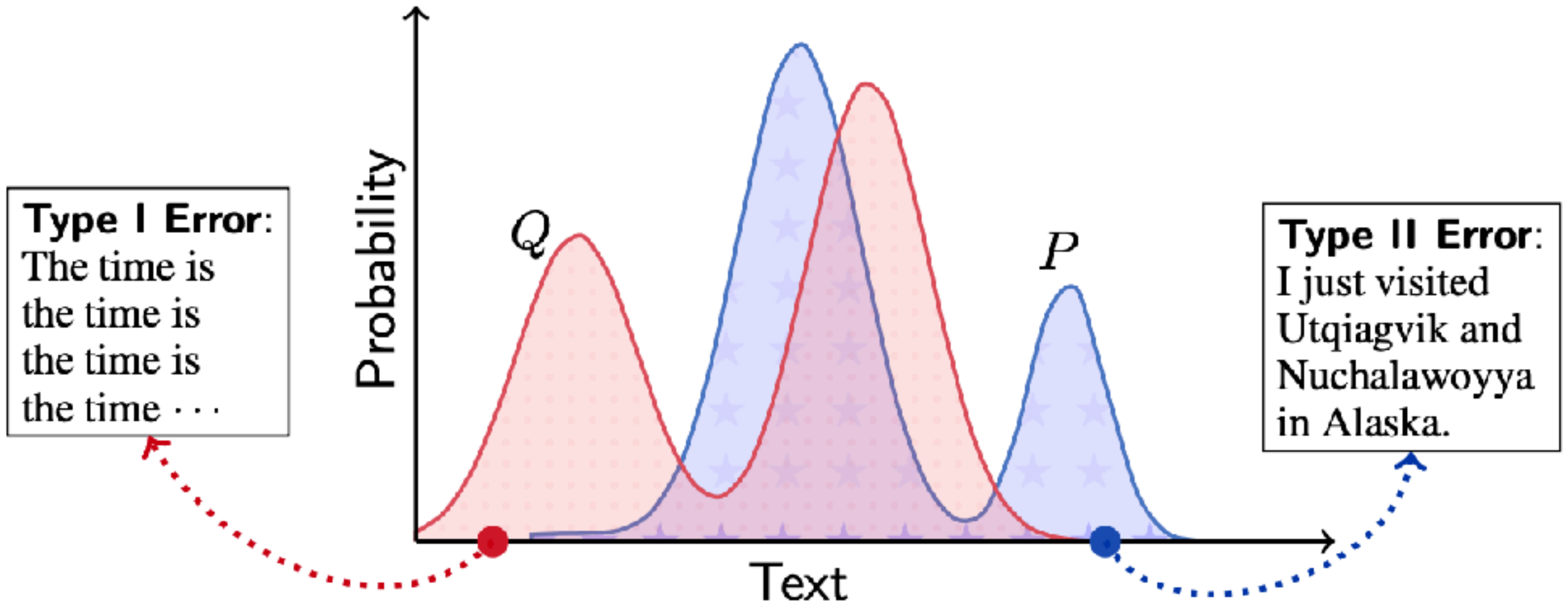
MAUVE: Beyond single sample matching

MAUVE: Beyond single sample matching

- In open-ended generation, comparing with a single reference may not say much. Can we instead compare the distribution of machine text vs. human text?

MAUVE: Beyond single sample matching

- In open-ended generation, comparing with a single reference may not say much.
  Can we instead compare the distribution of machine text vs. human text?
- **MAUVE** *(Pillutla et al., 2021)*
  - Computes the information divergence between the human text distribution P and the machine text distribution Q



Type I Error: The time is the time is the time is the time · · ·

Type II Error: I just visited Utqiagvik and Nuchalawoyya in Alaska.

MAUVE: Beyond single sample matching

- Divergence Curve

$$\mathcal{C}(P,Q) = \Big\{ \big(\exp(-c\,\mathrm{KL}(Q|R_\lambda)), \exp(-c\,\mathrm{KL}(P|R_\lambda)) \big) \;:\; R_\lambda = \lambda P + (1-\lambda)Q,\ \lambda \in (0,1) \Big\}$$

MAUVE: Beyond single sample matching

- Divergence Curve

$$\mathcal{C}(P, Q) = \Big\{ \big(\exp(-c\,\underline{\mathrm{KL}(Q|R_\lambda)}), \exp(-c\,\mathrm{KL}(P|R_\lambda)) \big) \;:\; R_\lambda = \lambda P + (1-\lambda) Q,\; \lambda \in (0,1) \Big\}$$

KL Divergence: Distance between two distributions Q and R_λ

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}$$

MAUVE: Beyond single sample matching

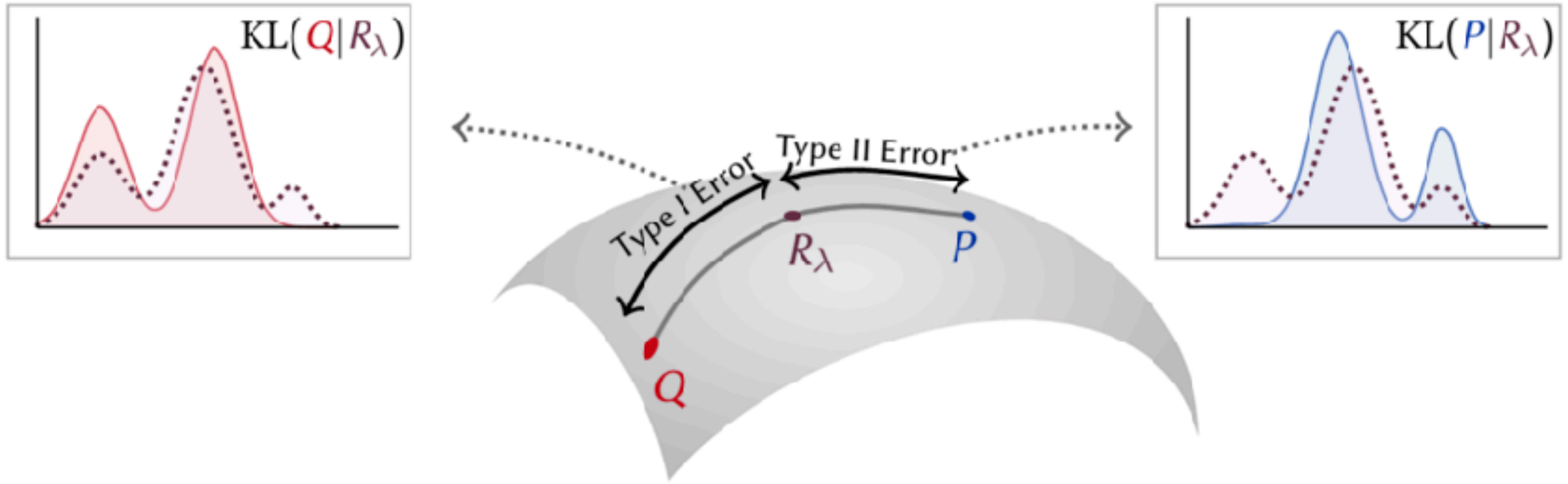
- Divergence Curve

$$\mathcal{C}(P,Q) = \Big\{ \big(\exp(-c\,\underline{\mathrm{KL}(Q|R_\lambda)}), \exp(-c\,\mathrm{KL}(P|R_\lambda)) \big) \;:\; \underline{R_\lambda = \lambda P + (1-\lambda) Q}, \; \lambda \in (0,1) \Big\}$$

KL Divergence: Distance between two distributions Q and R_λ

Interpolate between P and Q to draw a curve

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}$$



MAUVE: Beyond single sample matching

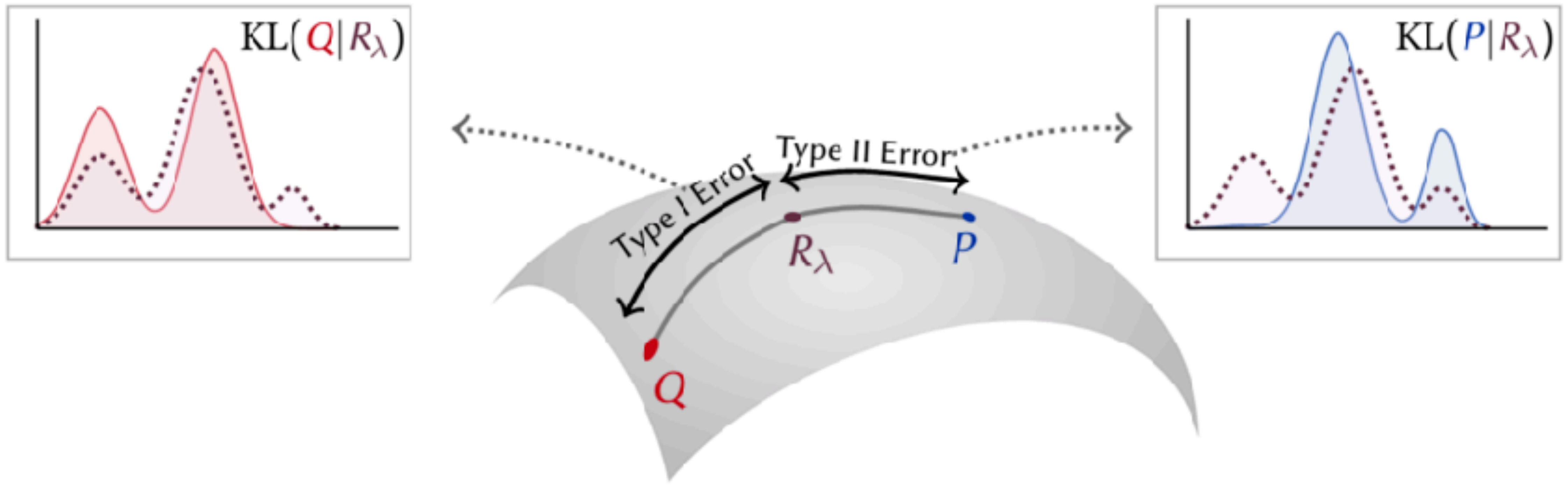
- Divergence Curve

$$\mathcal{C}(P,Q) = \Big\{ \big(\exp(-c\,\underline{\mathrm{KL}(Q|R_\lambda)}), \exp(-c\,\mathrm{KL}(P|R_\lambda)) \big) \;:\; \underline{R_\lambda = \lambda P + (1-\lambda)Q}, \; \lambda \in (0,1) \Big\}$$

KL Divergence: Distance between two distributions Q and R_λ

Interpolate between P and Q to draw a curve

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}$$



- $KL(P\,|\,Q)$ or $KL(Q\,|\,P)$ can be infinite, so measure errors softly using mixtures R_λ
- Draw a curve by varying the mixture weight λ: captures both type I / type 2 error!

MAUVE: Beyond single sample matching

- Divergence Curve

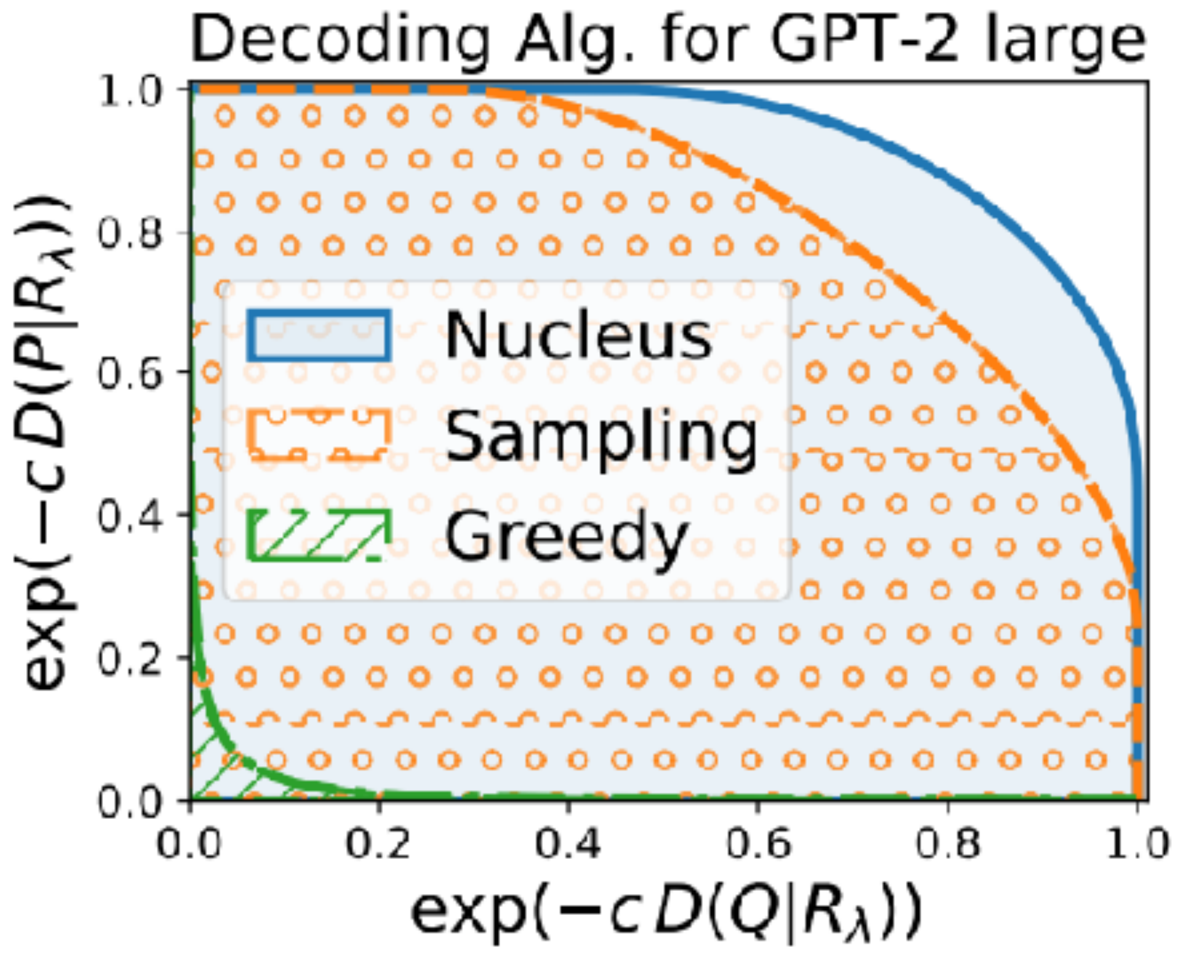
$$\mathcal{C}(P,Q) = \Big\{\big(\exp(-c\,\underline{\mathrm{KL}(Q|R_\lambda)}), \exp(-c\,\mathrm{KL}(P|R_\lambda))\big) \;:\; \underline{R_\lambda = \lambda P + (1-\lambda)Q,}\ \lambda \in (0,1)\Big\}$$

KL Divergence: Distance between two distributions Q and R_λ

Interpolate between P and Q to draw a curve

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}$$

- If P and Q are close, KL divergence will be lower, thus the divergence curve will be higher
- **MAUVE(P, Q)**: Area under the divergence curve (value in 0~1, **higher is better!**)



Nucleus sampling is better than naive sampling / greedy decoding.

MAUVE: Beyond single sample matching

- Problem: P and Q are distributions over all possible text!

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}$$

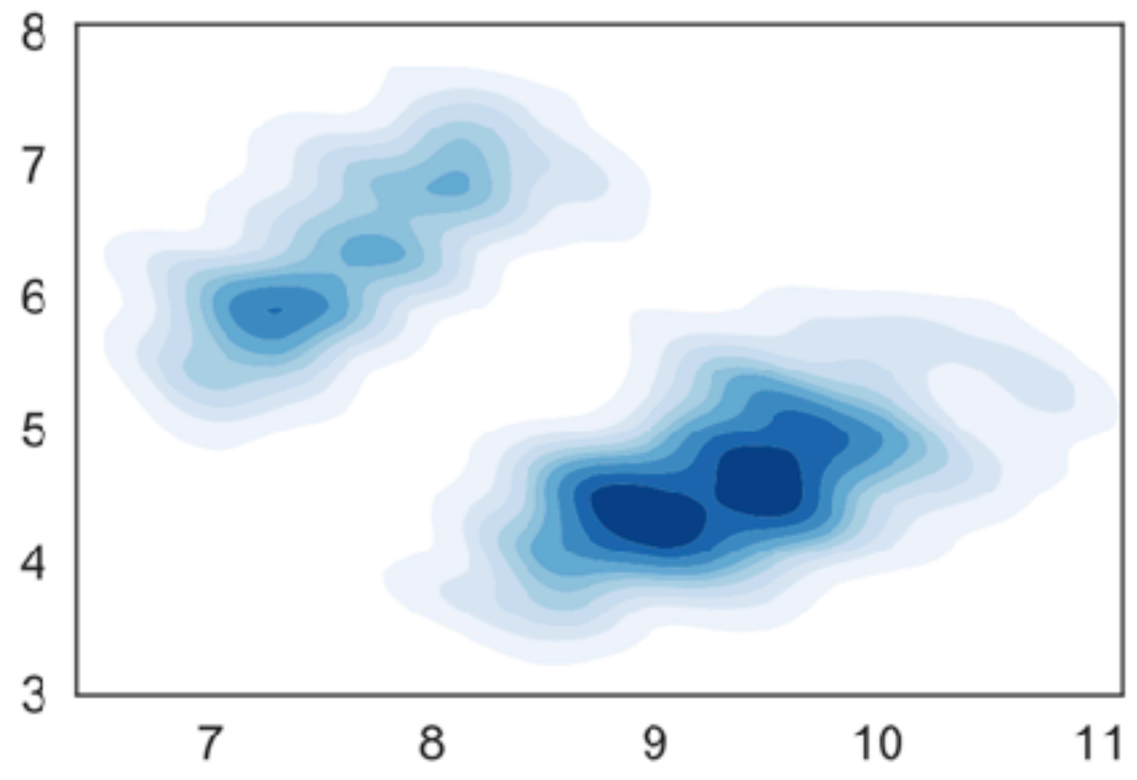
How do we compute the KL divergence?

MAUVE: Beyond single sample matching

- Problem: P and Q are distributions over all possible text!

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}$$

How do we compute the KL divergence?



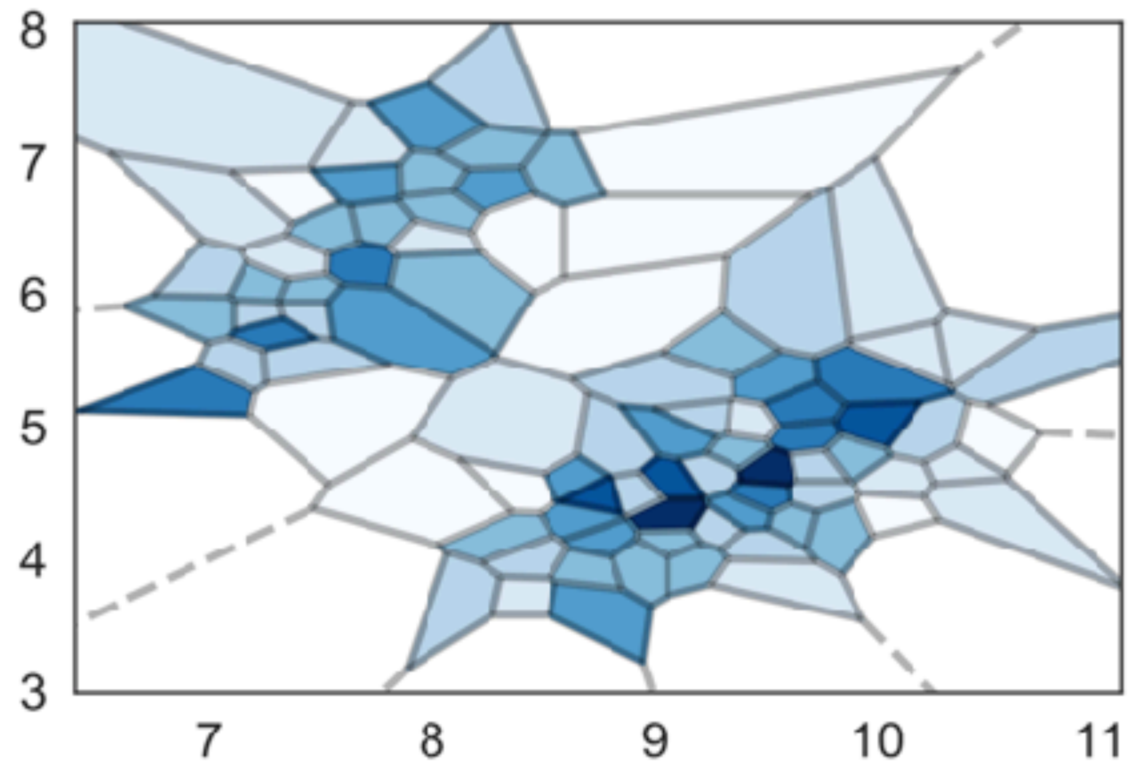
- Solution: Compute it over **quantized embedding distribution**

(1) Embed each sample x into latent space using e.g. GPT-2

(2) Quantize them into clusters

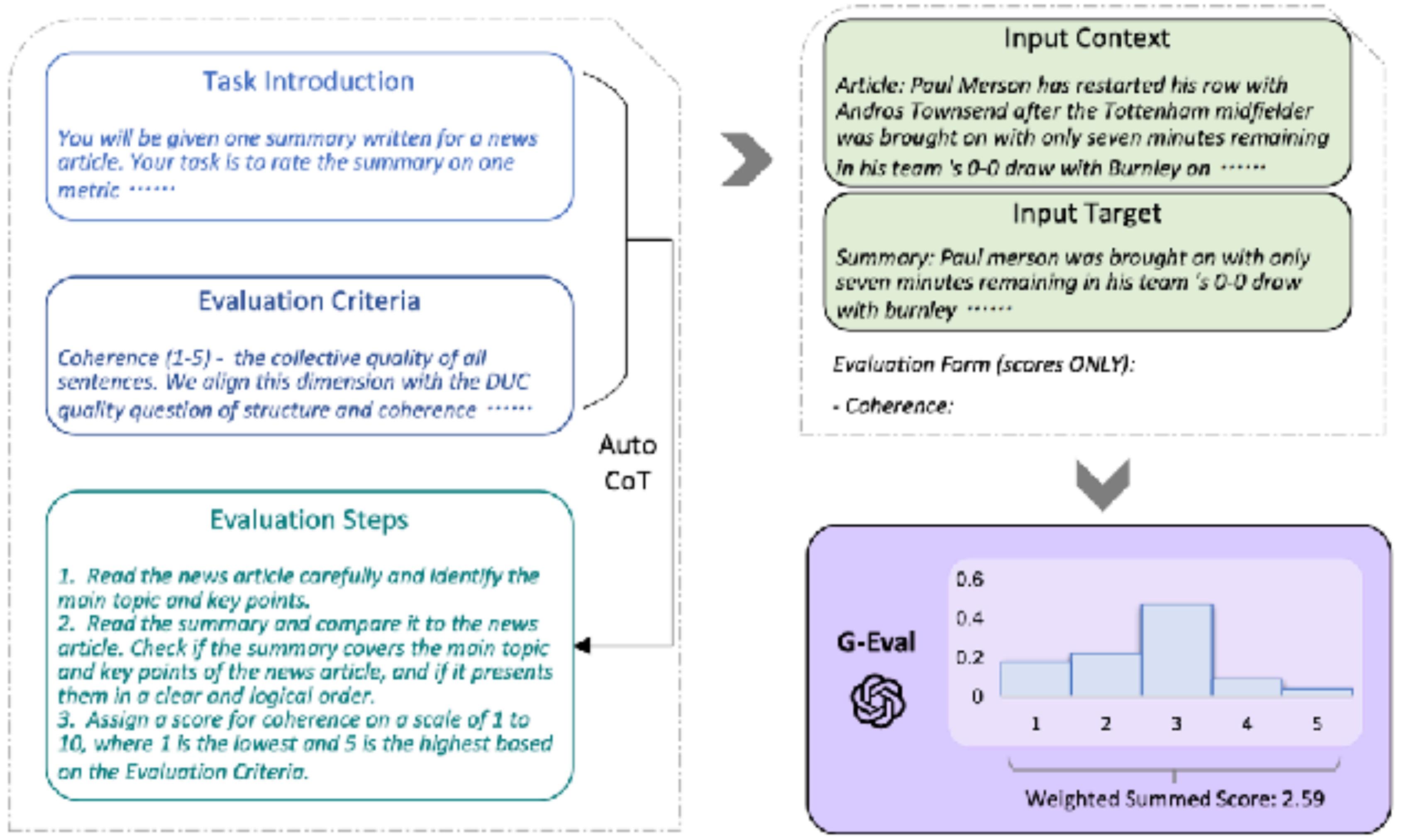
(3) Count cluster assignments to form histograms

Do (1) ~ (3) for both P and Q, now KL divergence is tractable 👍



Quantized Embedding Distribution

Model-based metrics: LLM as evaluator



Task Introduction

You will be given one summary written for a news article. Your task is to rate the summary on one metric ······

Evaluation Criteria

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence ······

Auto CoT

Evaluation Steps

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 10, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Input Context

Article: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on ······

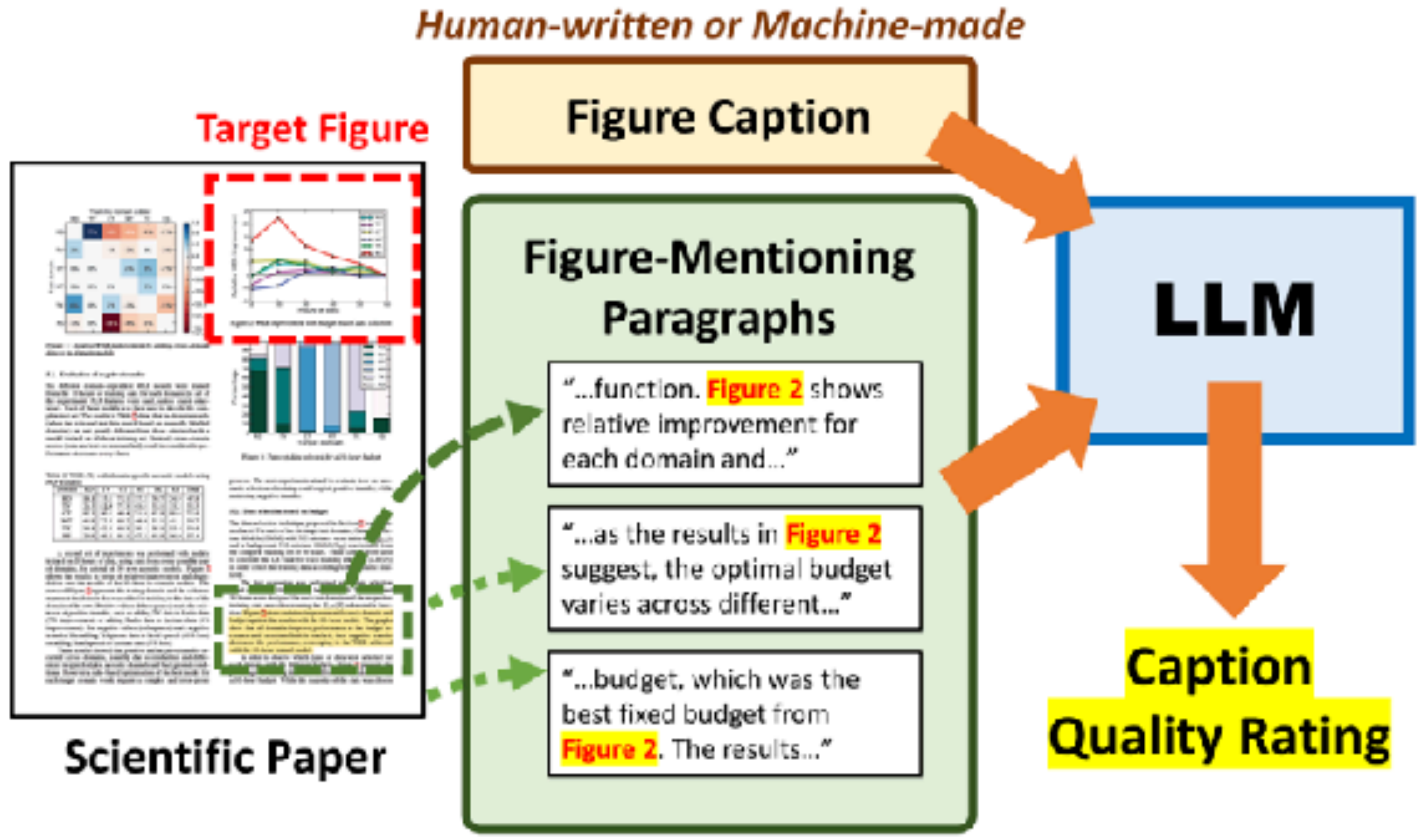
Input Target

Summary: Paul merson was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley ······

Evaluation Form (scores ONLY):

- Coherence:

Liu et al. 2023



"...function. Figure 2 shows relative improvement for each domain and..."

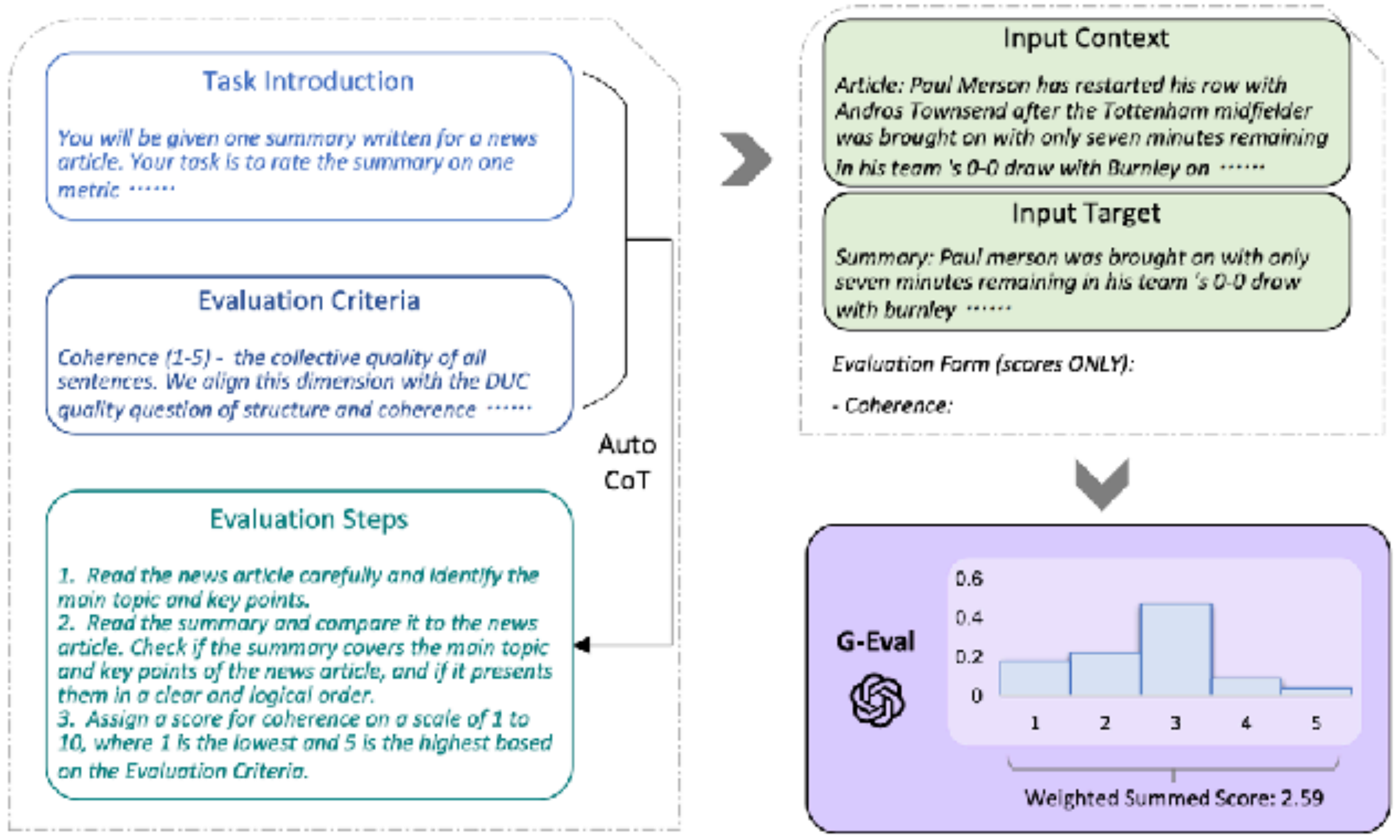
"...as the results in Figure 2 suggest, the optimal budget varies across different..."

"...budget, which was the best fixed budget from Figure 2. The results..."

Hsu et al. EMNLP Findings, 2023

Model-based metrics: LLM as evaluator

- Directly prompt LLM (GPT-4) to evaluate generated text.
  - Can be customized with evaluation criteria
  - (Often) better correlation with human evaluators than task-specific metrics (e.g. ROUGE)
  - (Often) is cheaper than human evaluation



You will be given one summary written for a news article. Your task is to rate the summary on one metric ……

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence ……

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 10, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

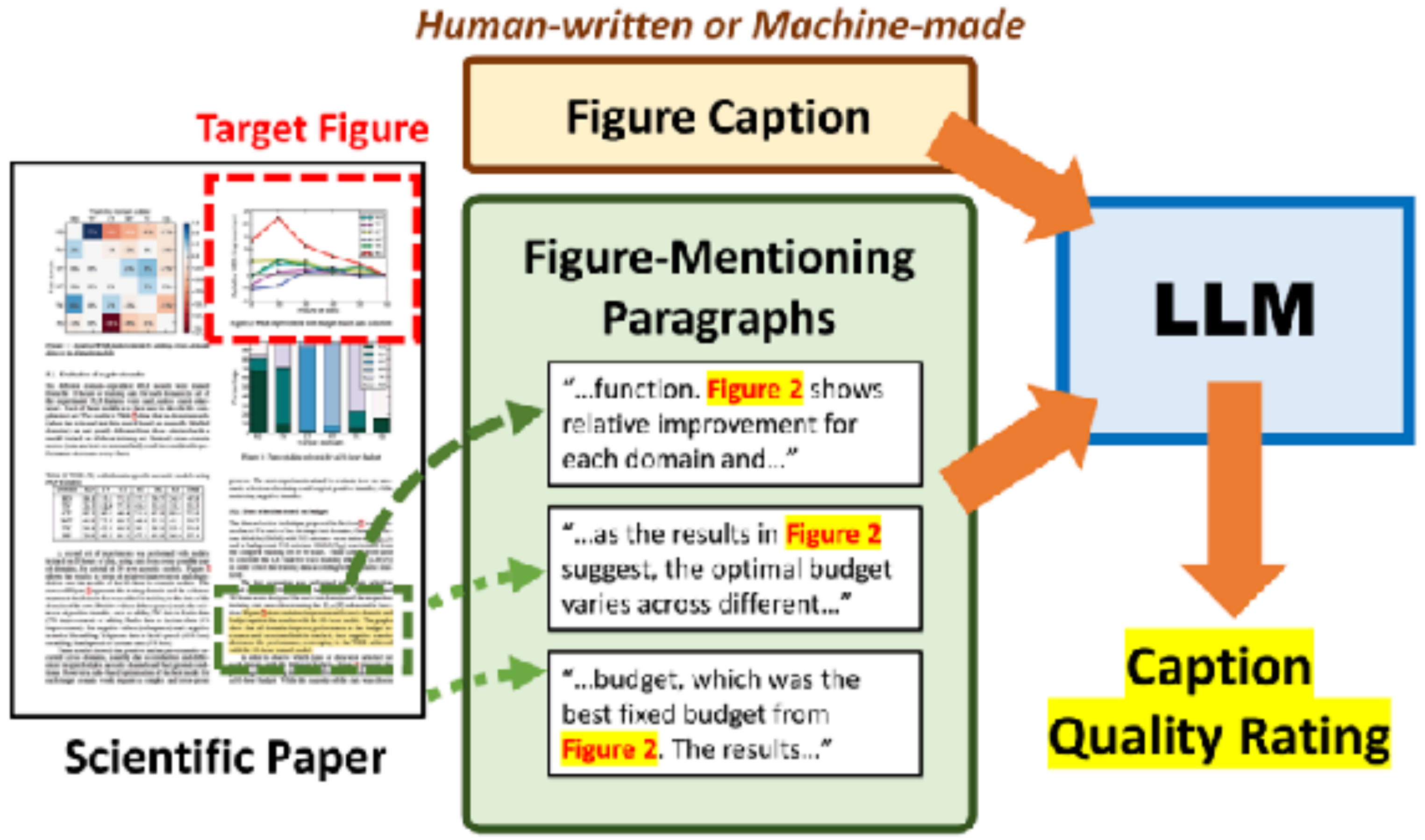
Article: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on ……

Summary: Paul merson was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley ……

Evaluation Form (scores ONLY):

- Coherence:

Liu et al. 2023



"...function. Figure 2 shows relative improvement for each domain and..."

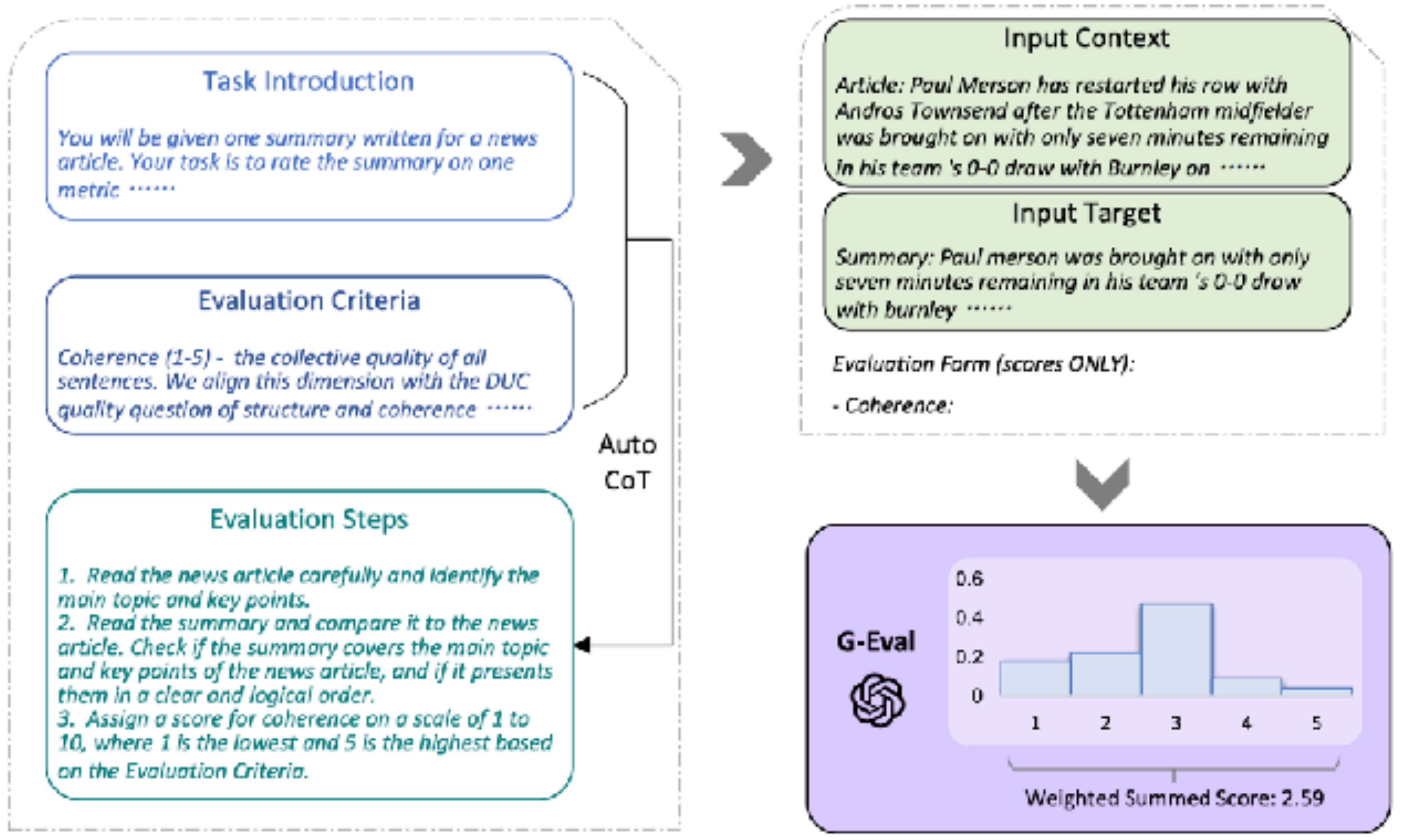
"...as the results in Figure 2 suggest, the optimal budget varies across different..."

"...budget, which was the best fixed budget from Figure 2. The results..."

Hsu et al. EMNLP Findings, 2023

Model-based metrics: LLM as evaluator

- Directly prompt LLM (GPT-4) to evaluate generated text.
  - Can be customized with evaluation criteria
  - (Often) better correlation with human evaluators than task-specific metrics (e.g. ROUGE)
  - (Often) is cheaper than human evaluation



You will be given one summary written for a news article. Your task is to rate the summary on one metric ……

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence ……

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 10, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Article: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on ……

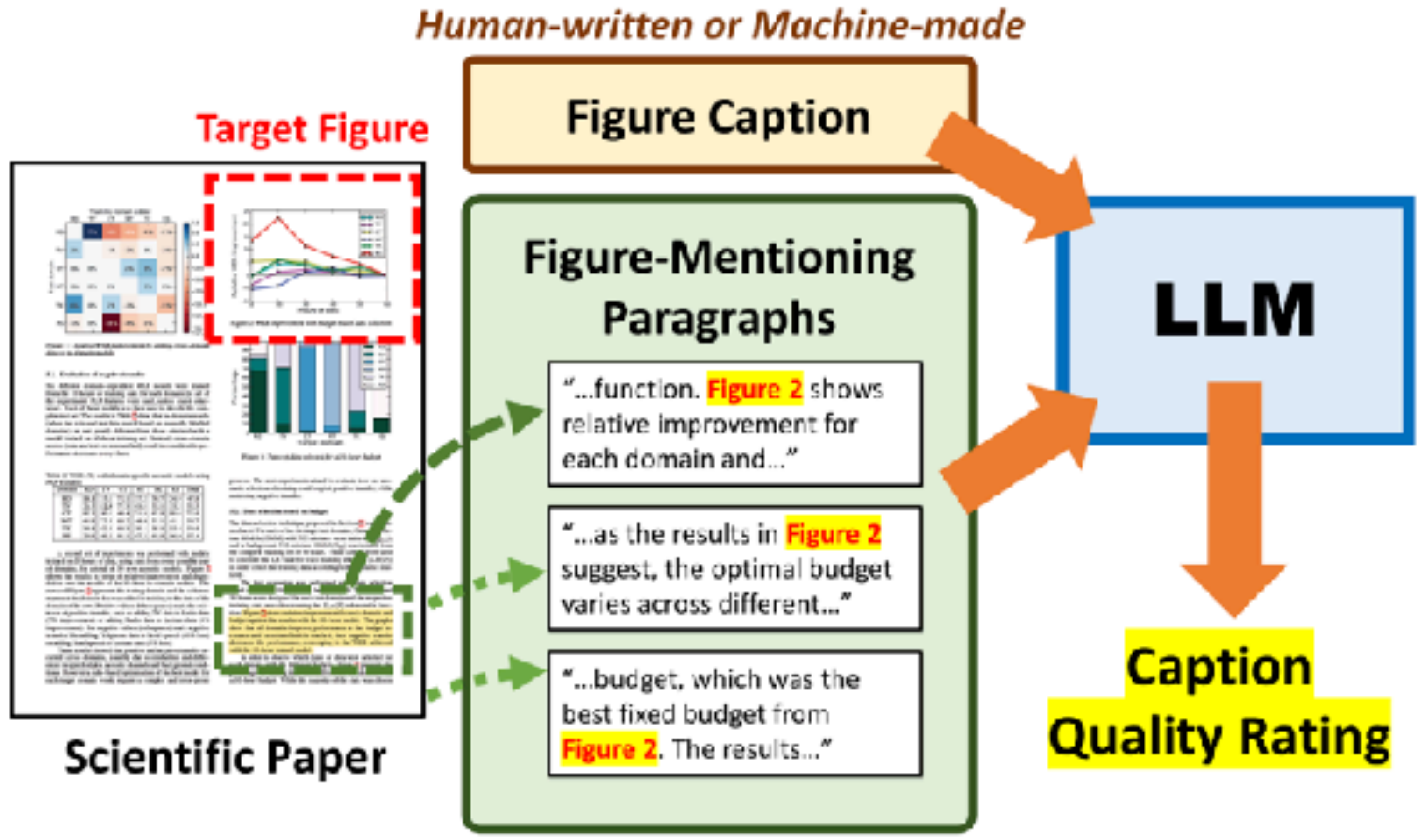
Summary: Paul merson was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley ……

Evaluation Form (scores ONLY):

- Coherence:

Liu et al. 2023

- Limitations
  - Brittleness: LLM evaluation can significantly vary when given different prompts!
  - Potential self-bias - LLMs may prefer what LLMs have generated…



"…function. Figure 2 shows relative improvement for each domain and…"

"…as the results in Figure 2 suggest, the optimal budget varies across different…"

"…budget, which was the best fixed budget from Figure 2. The results…"

Hsu et al. EMNLP Findings, 2023

Human Evaluations

Human Evaluations

- Automatic metrics fall short of matching human decisions

Human Evaluations

- Automatic metrics fall short of matching human decisions

- Most important form of evaluation for text generation systems

Human Evaluations

- Automatic metrics fall short of matching human decisions
- Most important form of evaluation for text generation systems
- Gold standard in developing new automatic metrics
  - Better automatic metrics will better correlate with human judgements!

Human Evaluations

- Sounds easy, but hard in practice: Ask humans to evaluate the quality of text

- Typical evaluation dimensions:
  - fluency
  - coherence / consistency
  - factuality and correctness
  - commonsense
  - style / formality
  - grammaticality
  - typicality
  - redundancy
  - …

Human Evaluations

- Sounds easy, but hard in practice: Ask humans to evaluate the quality of text

- Typical evaluation dimensions:
  - fluency
  - coherence / consistency
  - factuality and correctness
  - commonsense
  - style / formality
  - grammaticality
  - typicality
  - redundancy
  - ...

Note: Don't compare human evaluation scores across different studies

Even if they claim to evaluate on the same dimensions!

Human Evaluations

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

Human Evaluations

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

- Human judgements

Human Evaluations

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

- Human judgements
  - are inconsistent / irreproducible
  - can be illogical
  - can be misinterpreting your questionnaire
  - (cf. the venerable field of psychometrics)

Human Evaluations

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

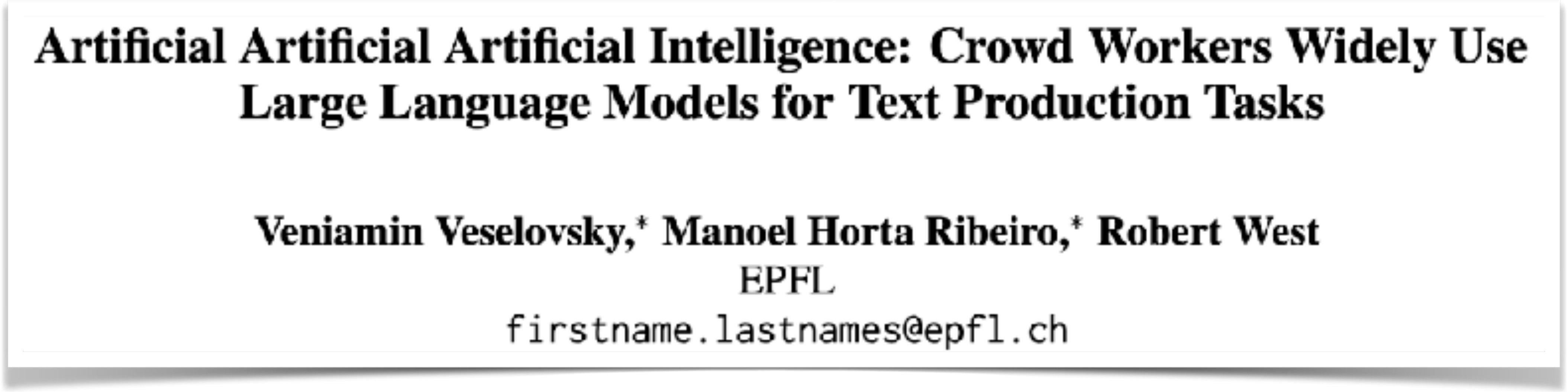
- Human judgements
  - are inconsistent / irreproducible
  - can be illogical
  - can be misinterpreting your questionnaire
  - (cf. the venerable field of psychometrics)
  - ...

Human Evaluations

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

- Human judgements
  - are inconsistent / irreproducible
  - can be illogical
  - can be misinterpreting your questionnaire
  - (cf. the venerable field of psychometrics)
  - ...
  - and recently, use of LLMs by crowd-source workers 🙄

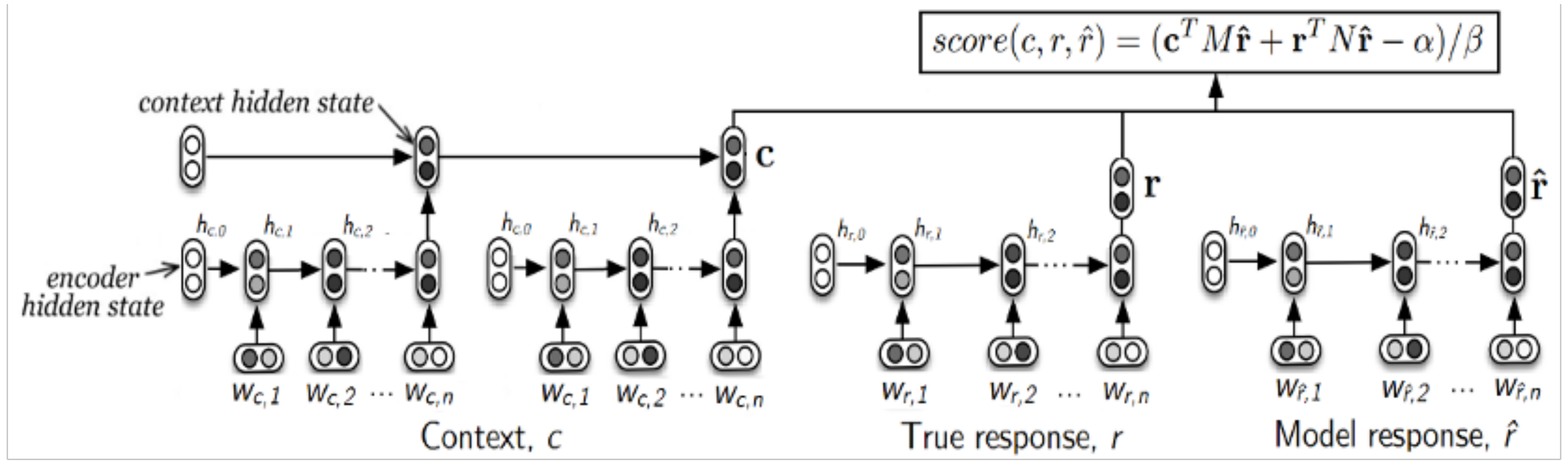
(Veselovsky et al., 2023)



Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks

Veniamin Veselovsky,* Manoel Horta Ribeiro,* Robert West
EPFL
firstname.lastnames@epfl.ch

Learning metrics from humans

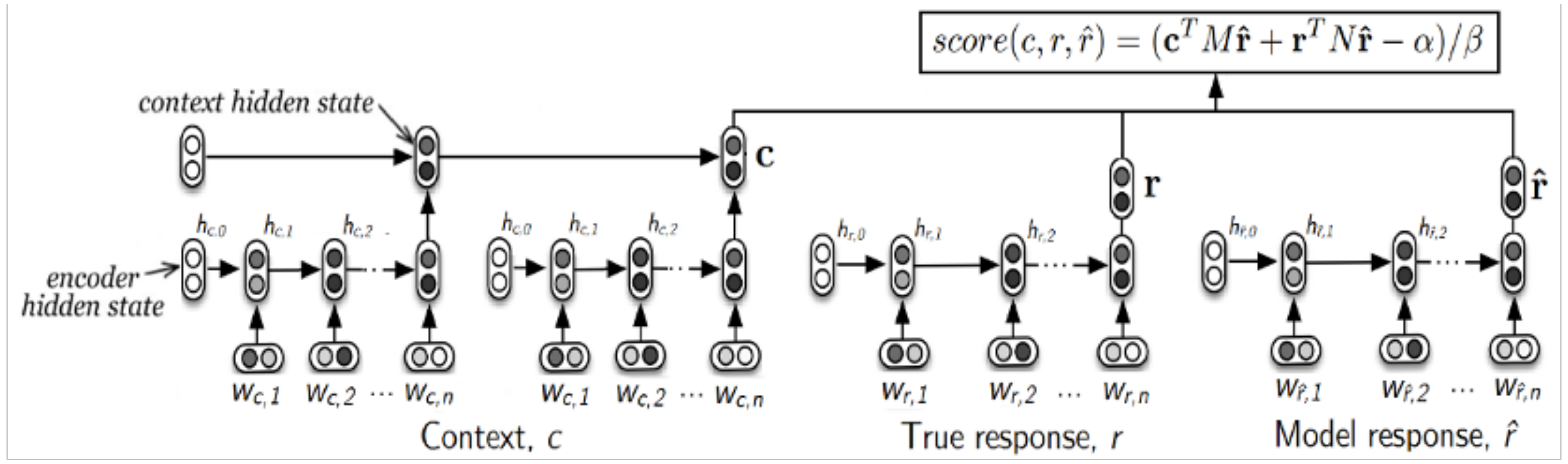

$score(c, r, \hat{r}) = (\mathbf{c}^T M \hat{\mathbf{r}} + \mathbf{r}^T N \hat{\mathbf{r}} - \alpha)/\beta$
context hidden state
encoder hidden state
$\mathbf{c}$
$\mathbf{r}$
$\hat{\mathbf{r}}$
$h_{c,0}$ $h_{c,1}$ $h_{c,2}$
$h_{r,0}$ $h_{r,1}$ $h_{r,2}$
$h_{\hat{r},0}$ $h_{\hat{r},1}$ $h_{\hat{r},2}$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

ADEM

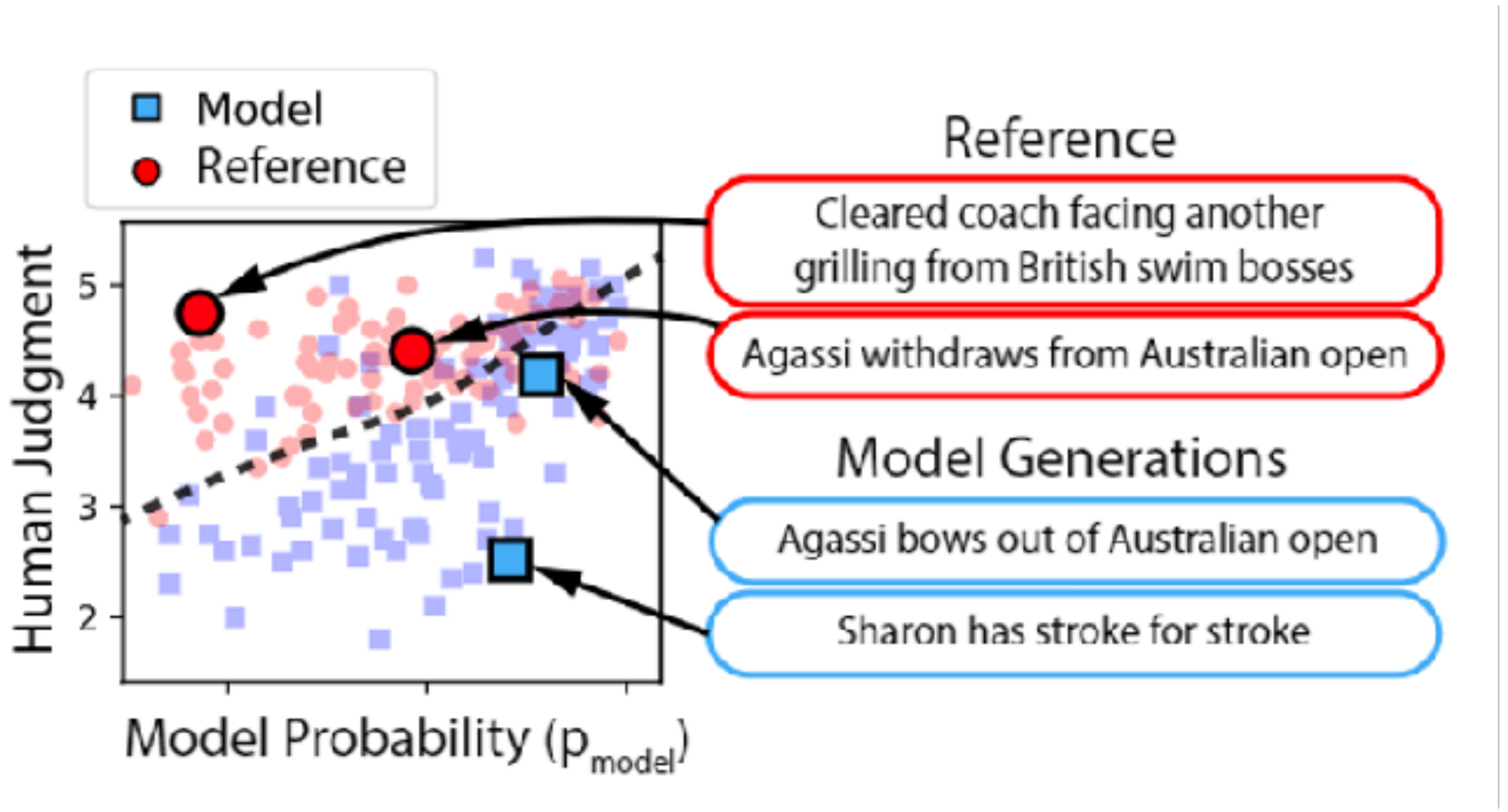
A learned metric from human judgments for dialog system evaluation in a chatbot setting

- *(Lowe et al., 2017)*

Learning metrics from humans



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

VS

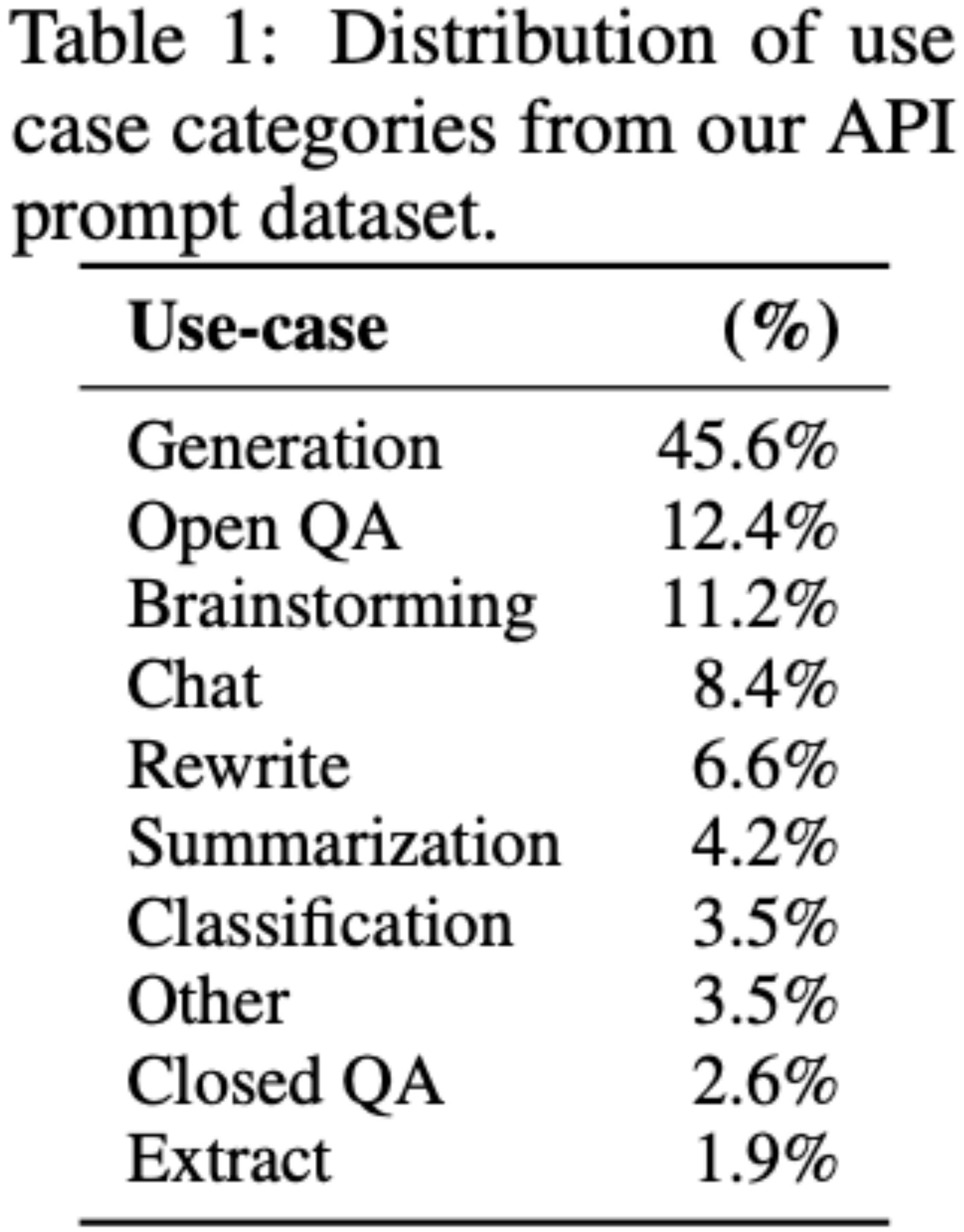
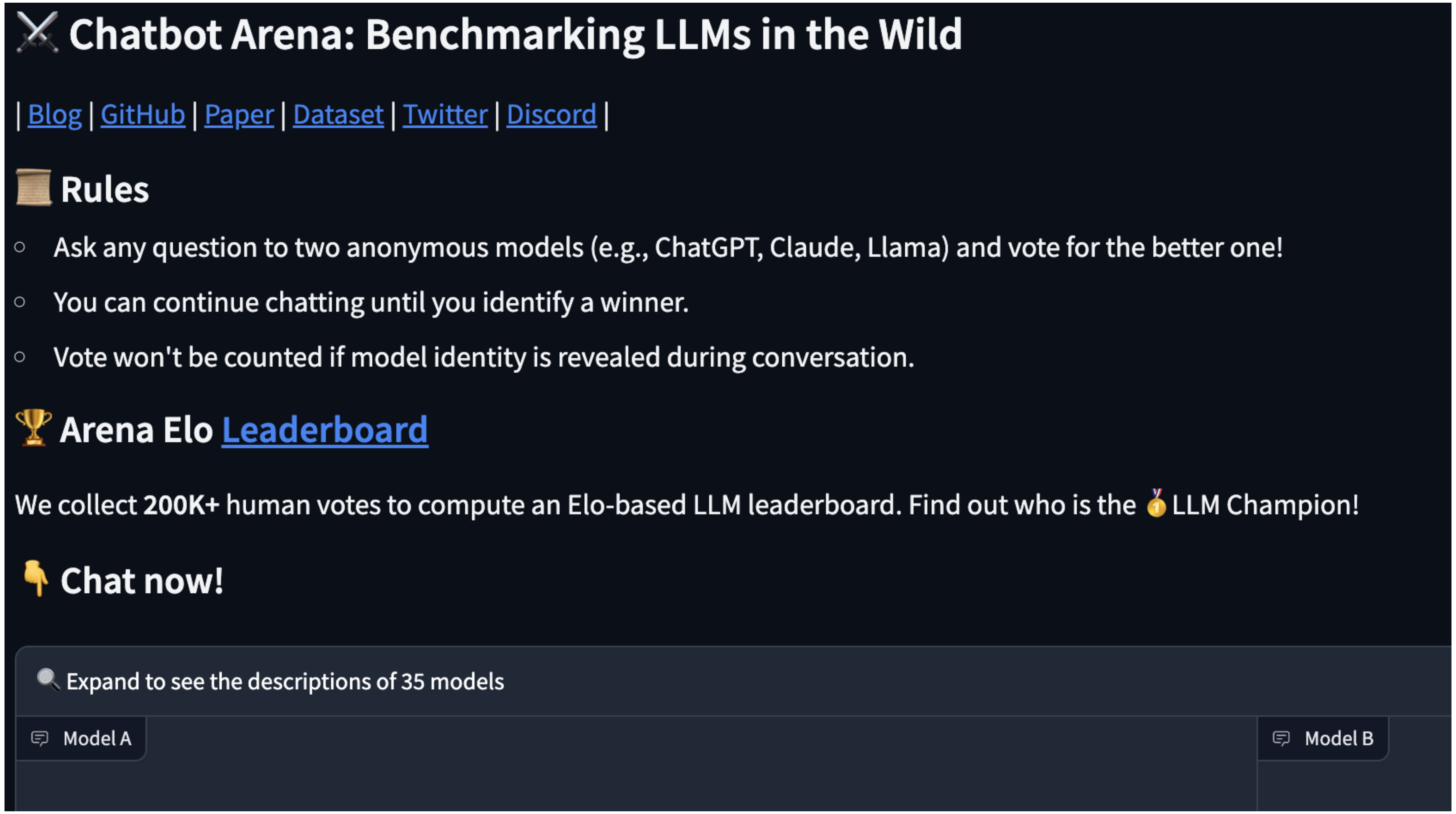


Table 1: Distribution of use case categories from our API prompt dataset.

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!

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

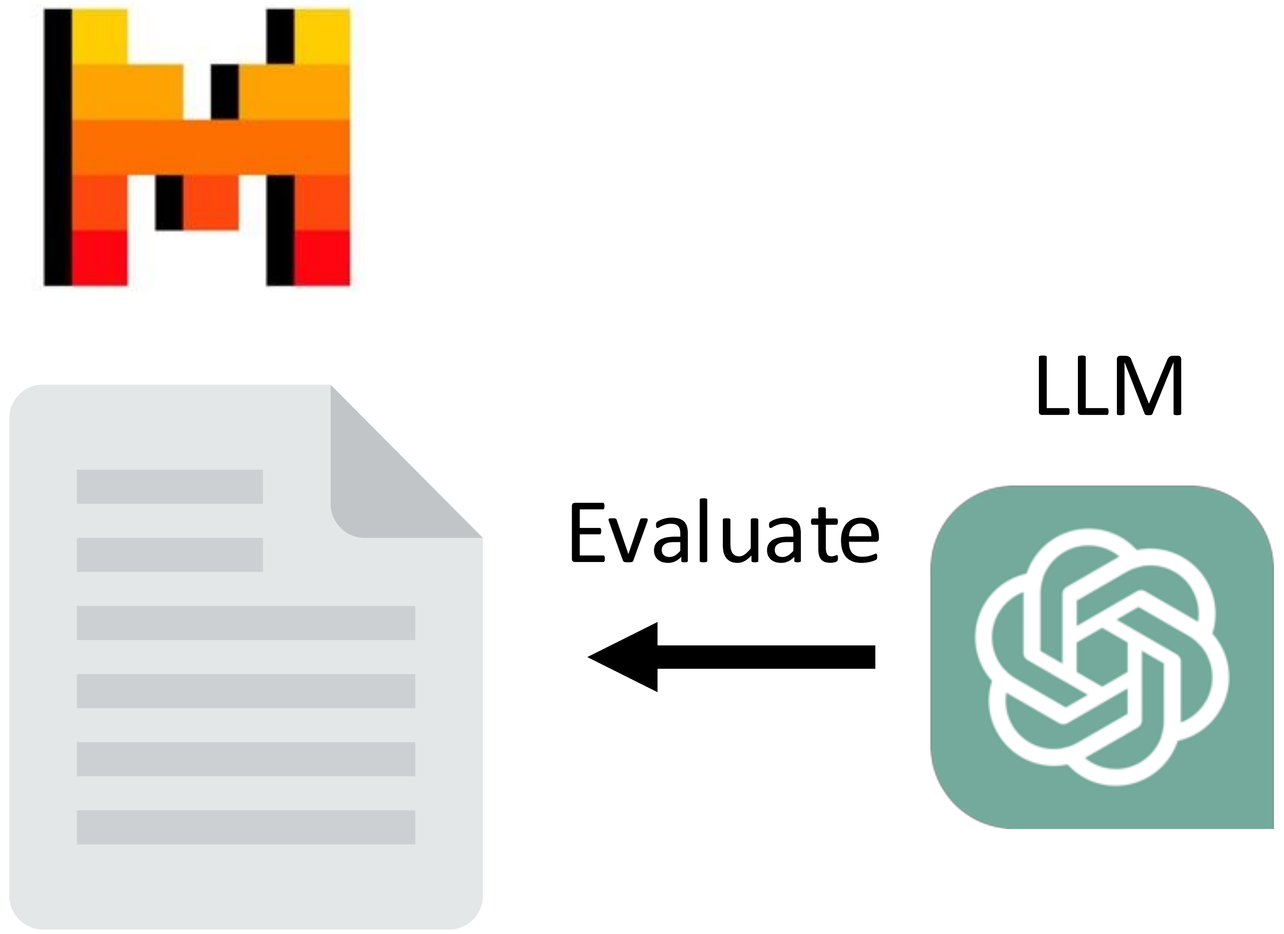
What’s missing from side-by-side human evaluation?

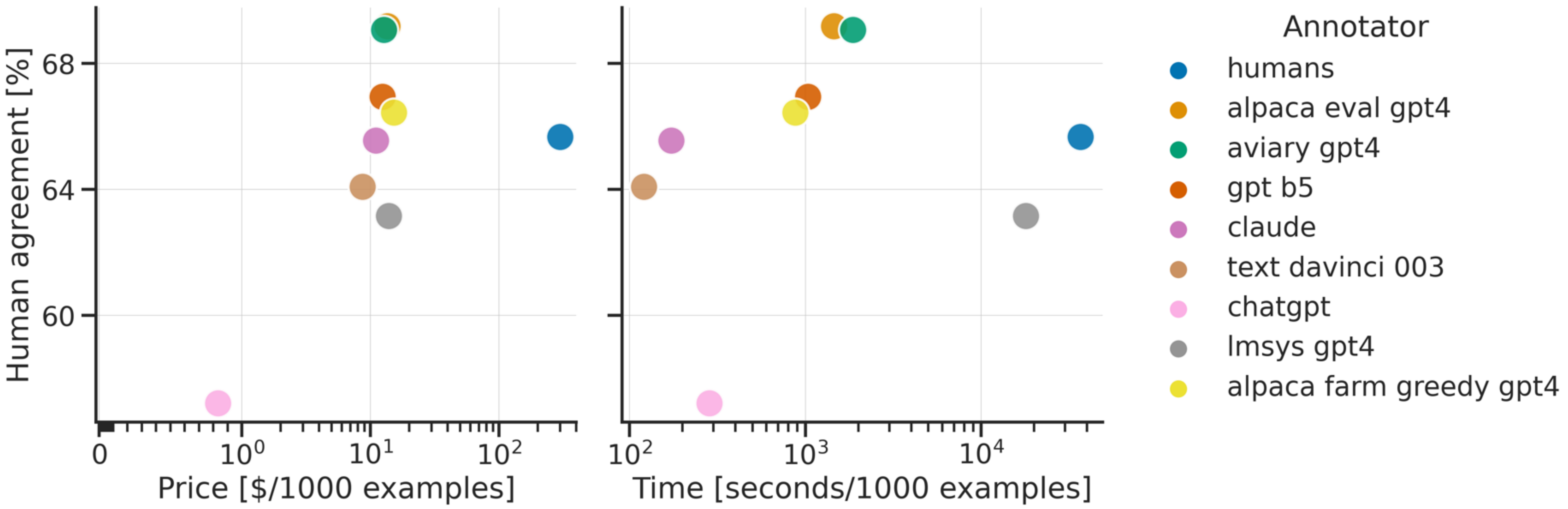
What’s missing from side-by-side human evaluation?

- **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

Use an LM to lower costs

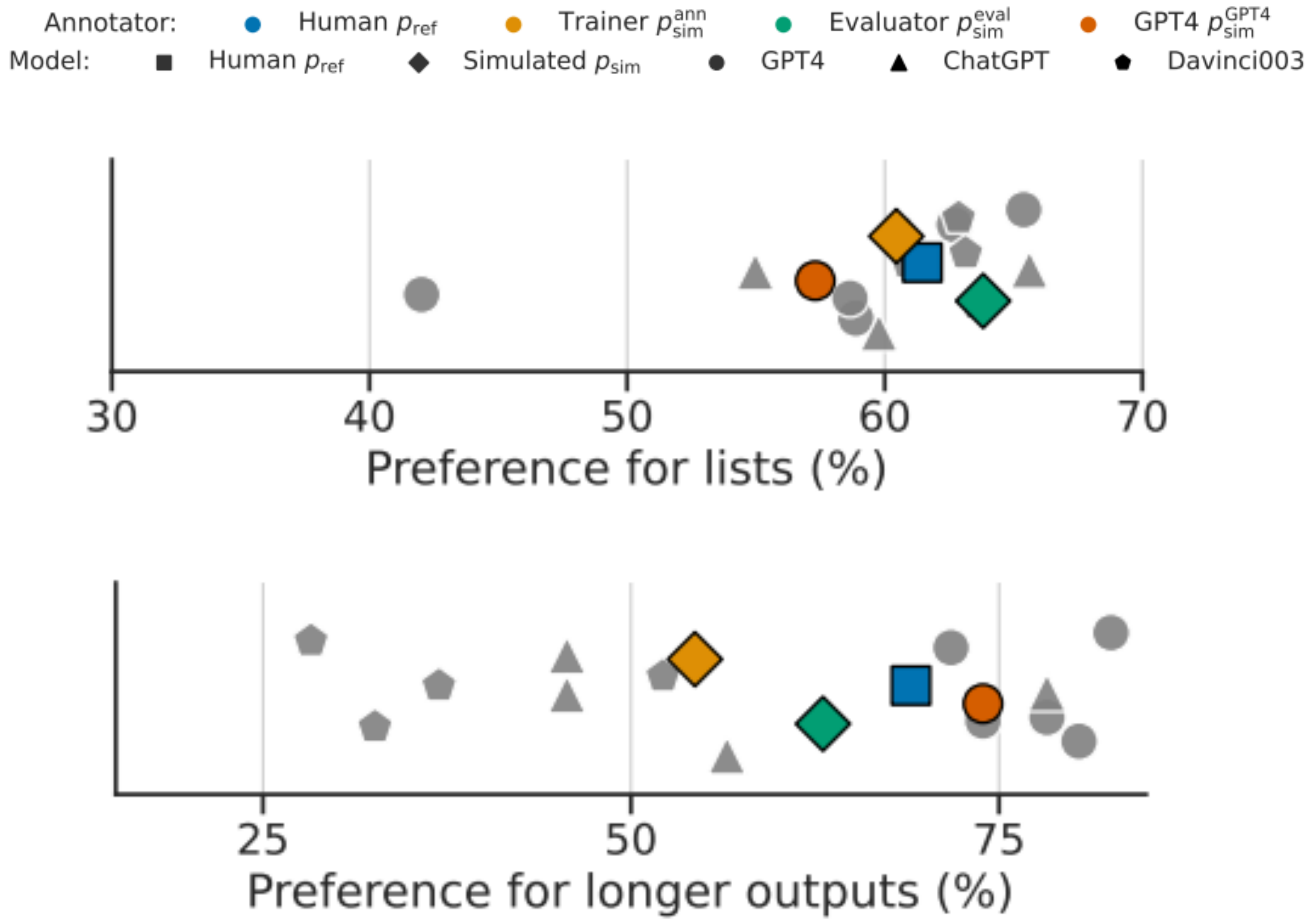




- Use a LM as a reference free evaluator
    - Surprisingly high correlations with human

- Common versions: MT-bench, AlpacaEval, ‘GPT-as-judge’

Potential pitfalls with LLMs as evaluators



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

External 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 […]

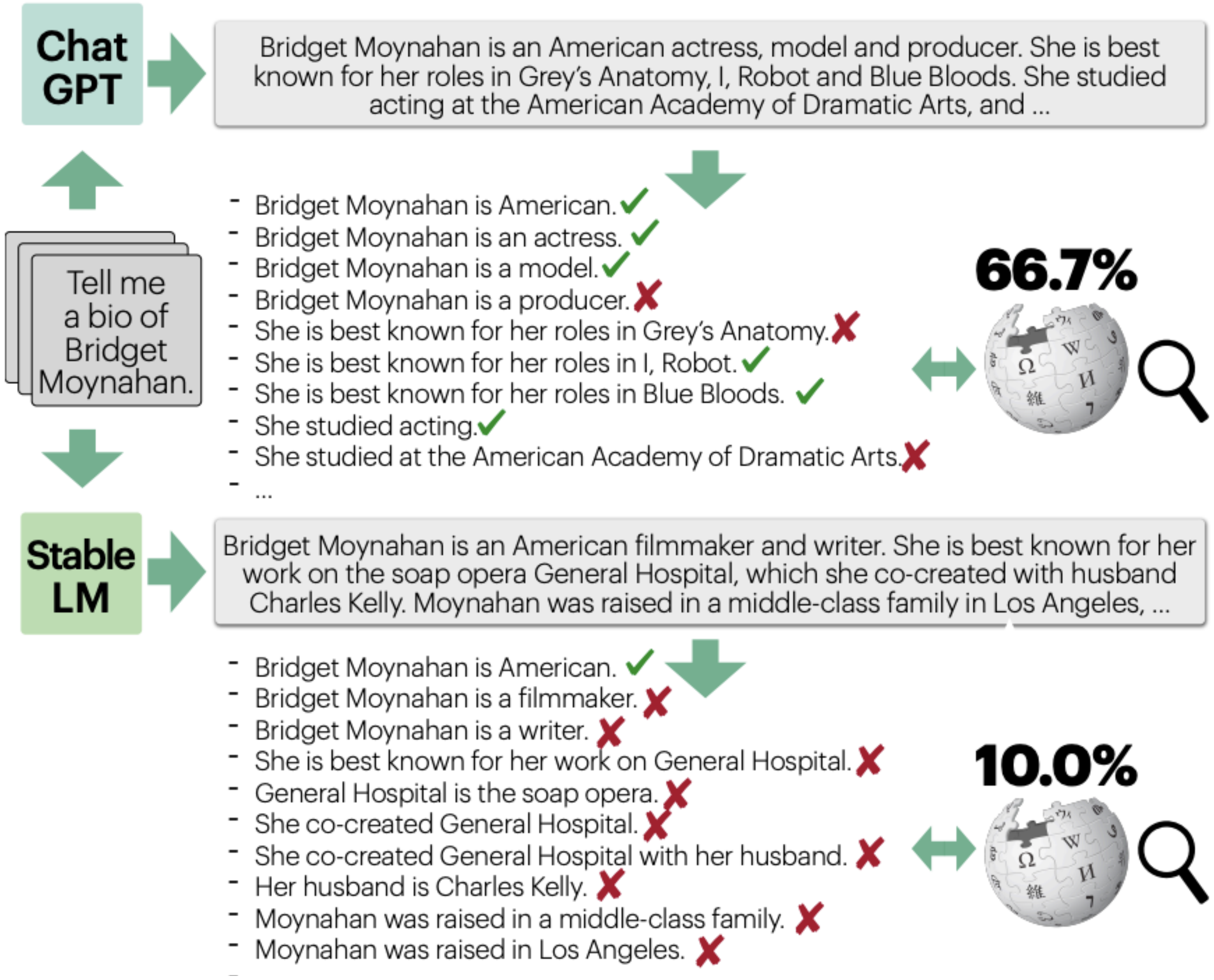
Other tasks: Long-form expository writing

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



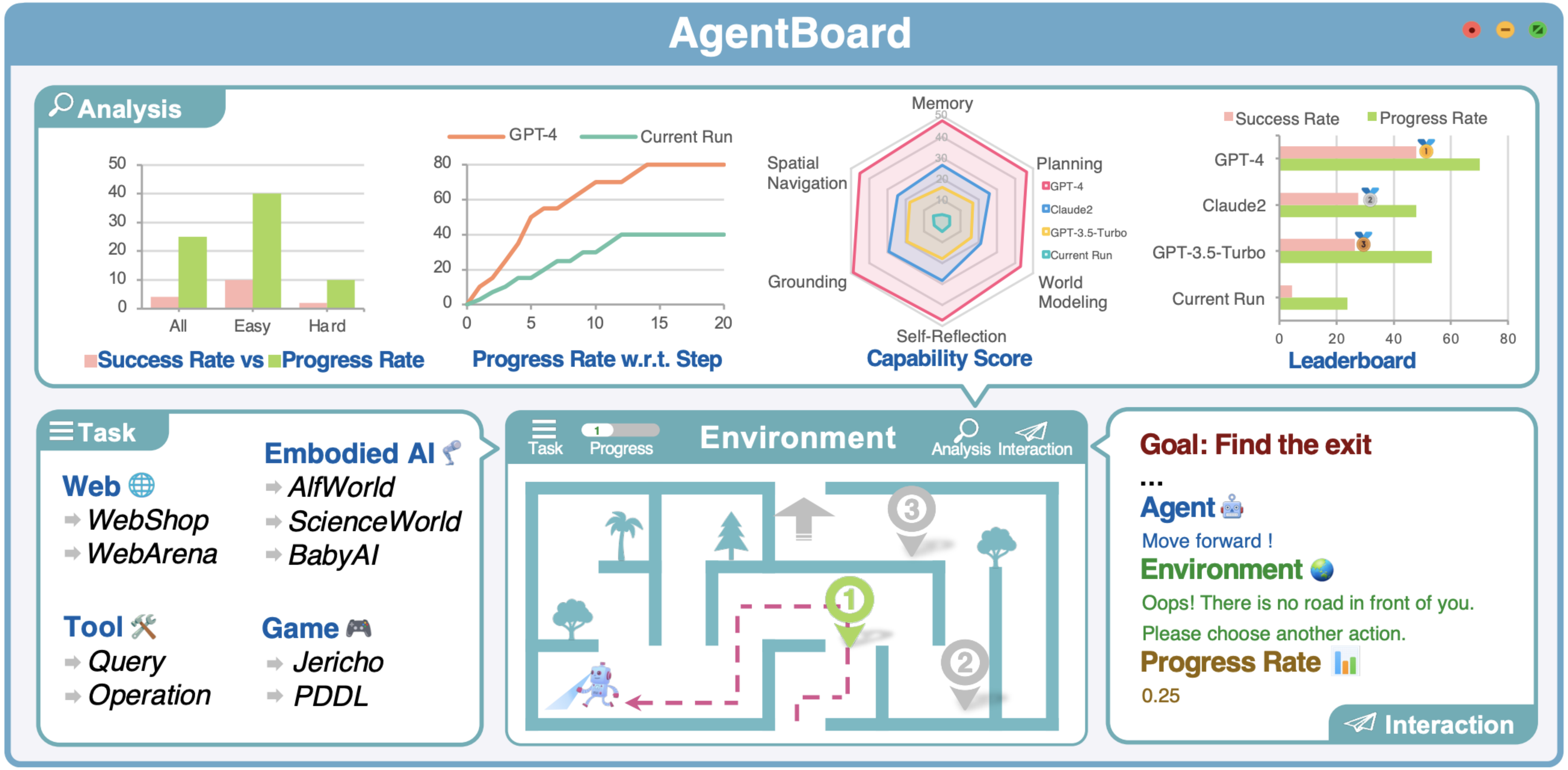
Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

- 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. ✗
- ...

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, ...

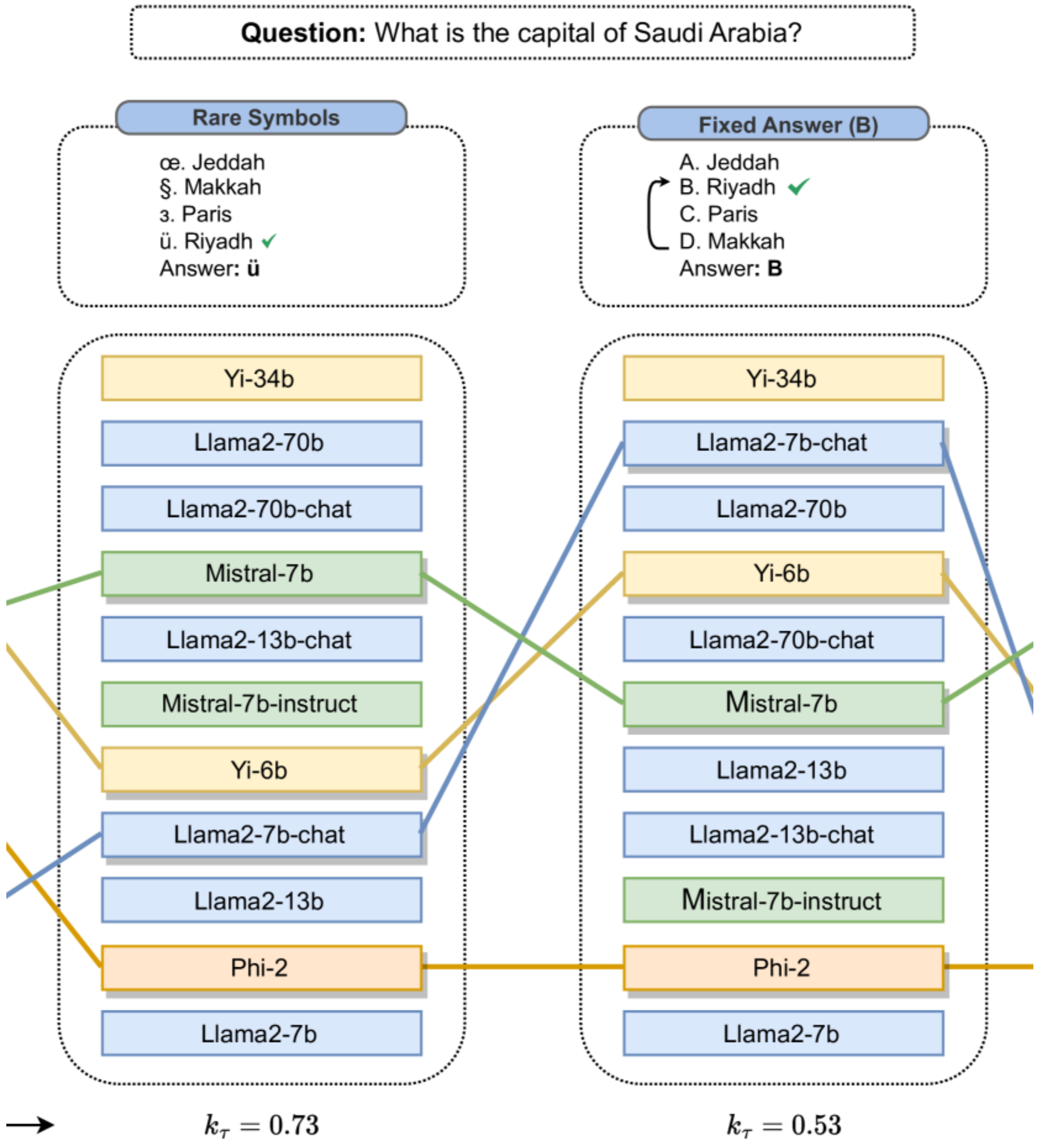
- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband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

Other tasks: Agents



- 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)

Open problems: Threats to reliable evaluations



Question: What is the capital of Saudi Arabia?

Rare Symbols

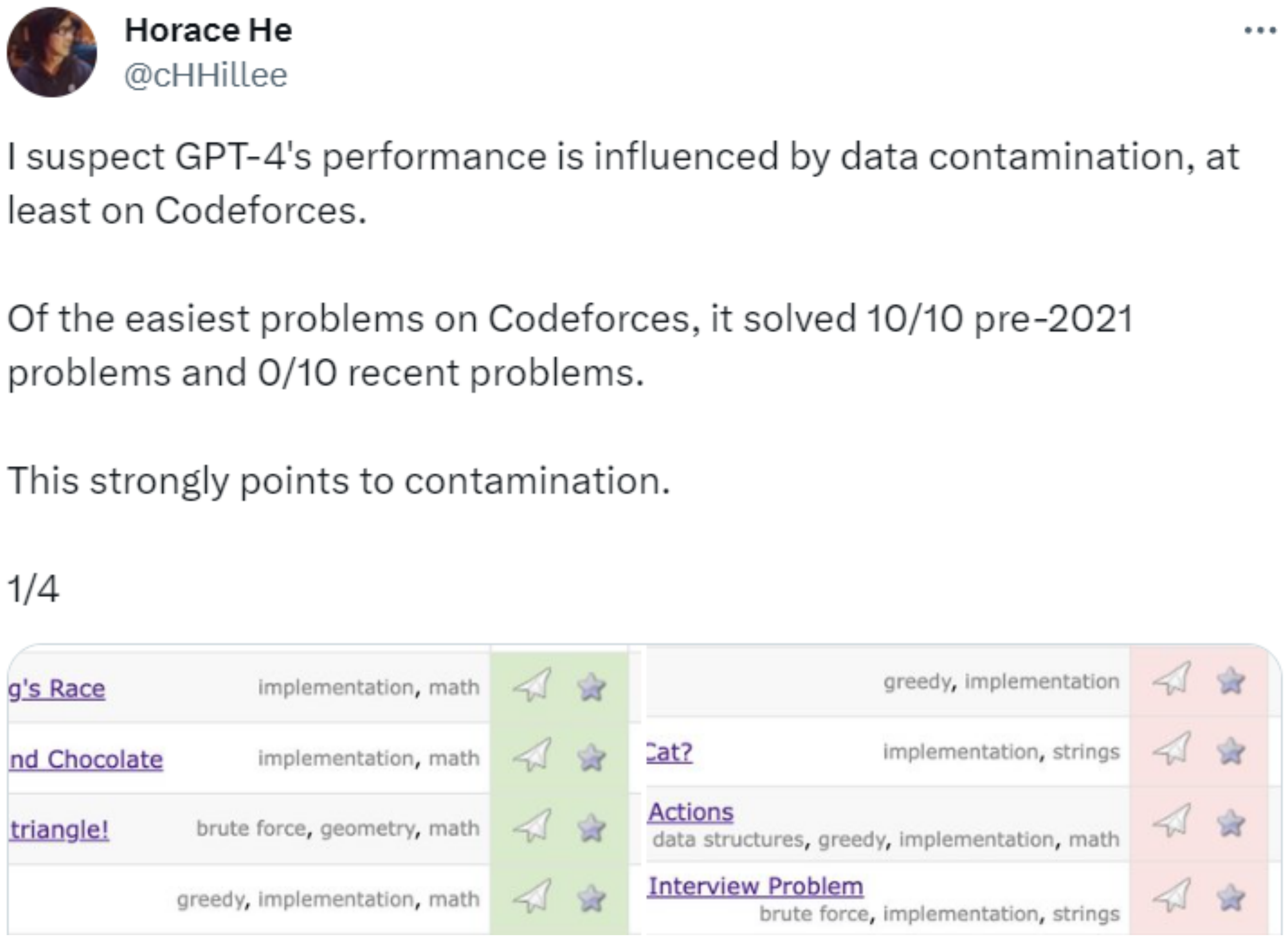
œ. Jeddah
§. Makkah
з. Paris
ü. Riyadh ✓
Answer: **ü**

Fixed Answer (B)

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

[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.

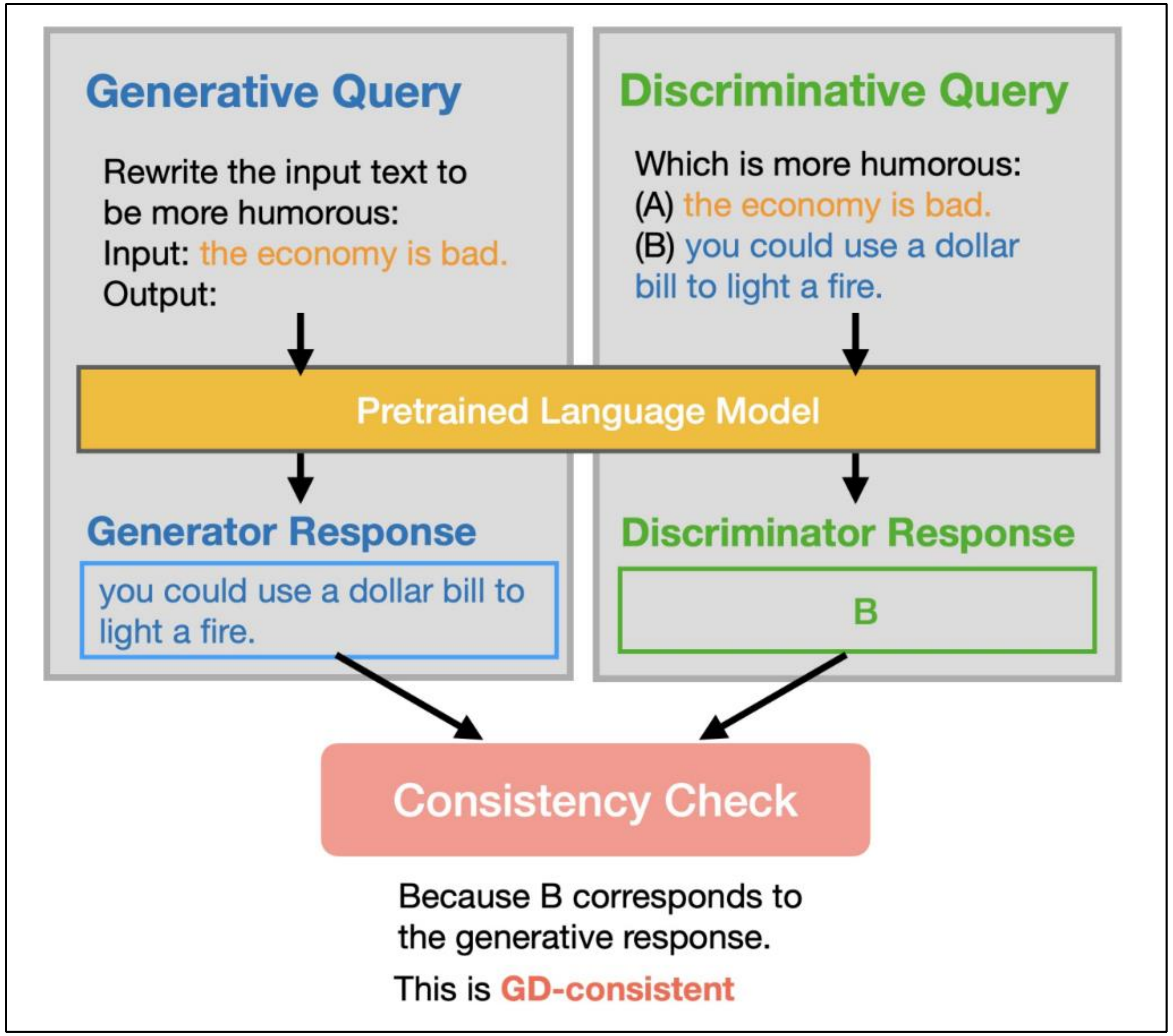
This strongly points to contamination.

1/4

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

Contamination

Prompt-sensitivity and inconsistency

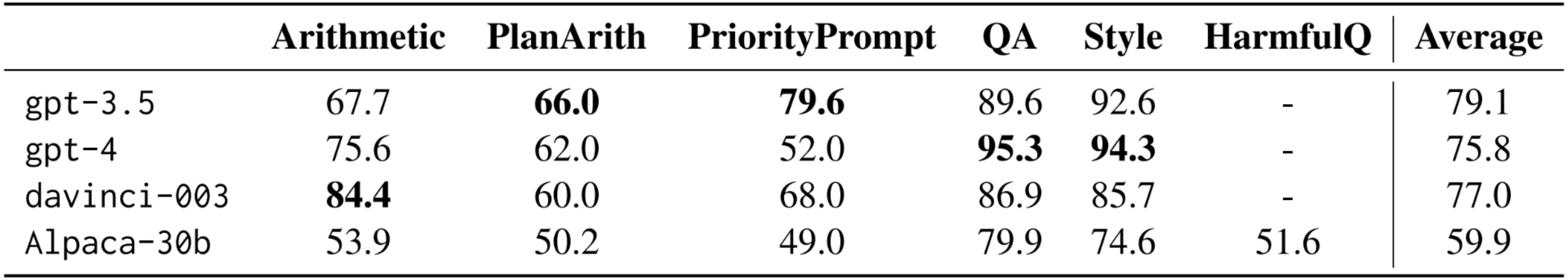

Generative Query
Rewrite the input text to be more humorous:
Input: the economy is bad.
Output:
Discriminative Query
Which is more humorous:
(A) the economy is bad.
(B) you could use a dollar bill to light a fire.
Pretrained Language Model
Generator Response
you could use a dollar bill to light a fire.
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
```

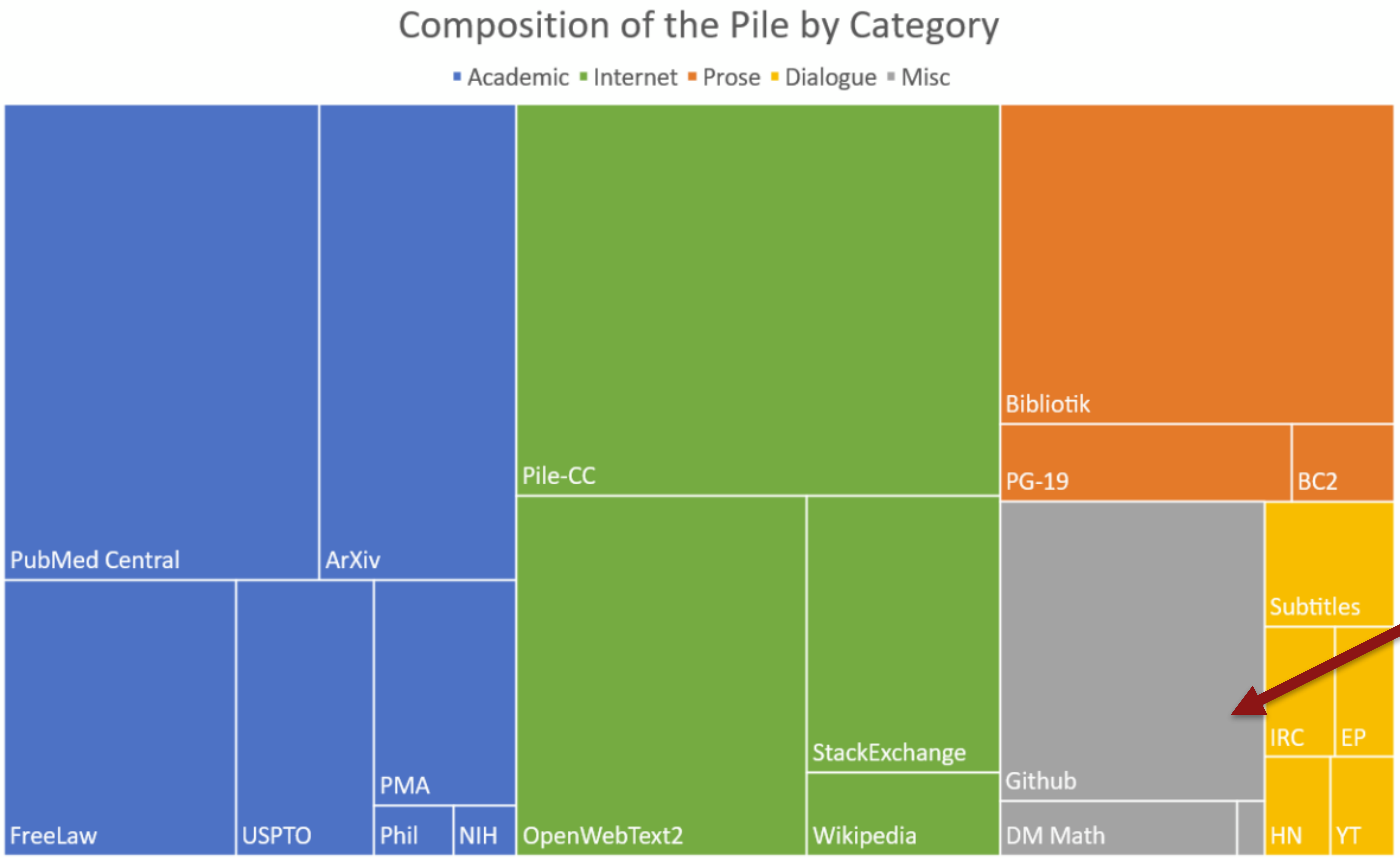
Prompt-sensitivity and inconsistency



	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

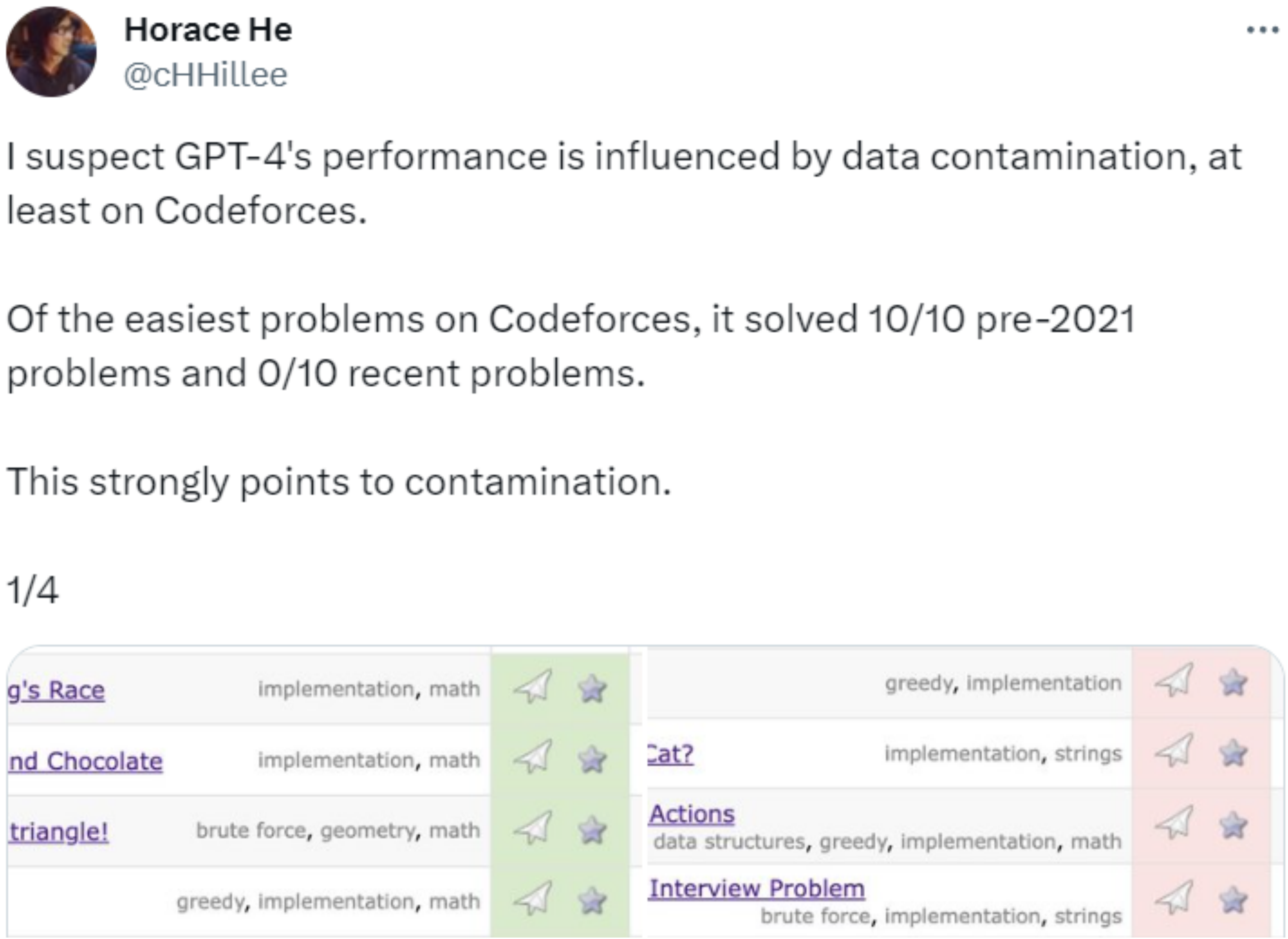
What’s in the training data of your LLM?



.. But maybe your test set is in here?

CODEFORCES
Sponsored by TON

Benchmarks are hard to trust for closed models



Horace He
@cHHillee

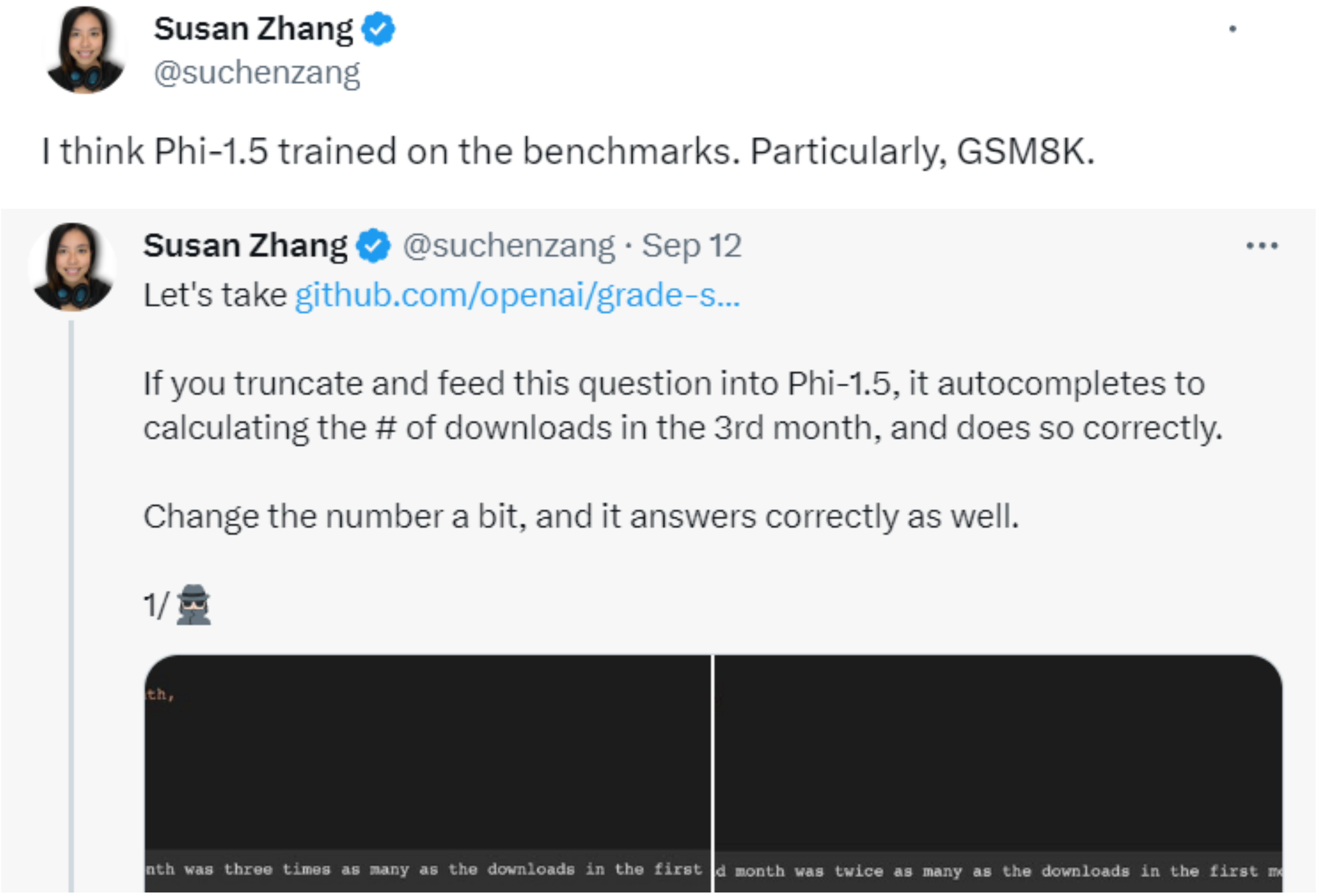
I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

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	Cat?	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/🕵️

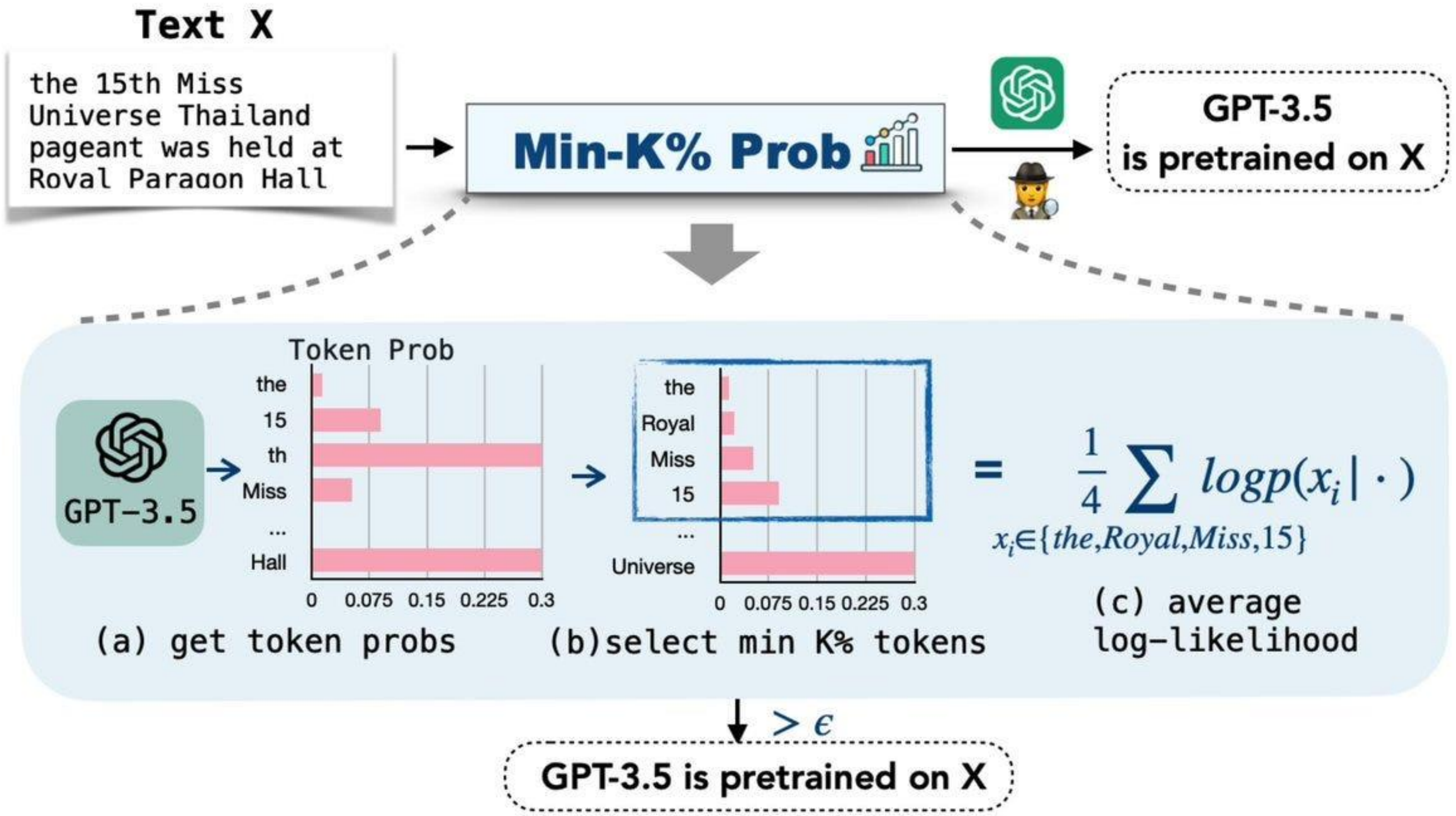
```
th,
nth was three times as many as the downloads in the first
```

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

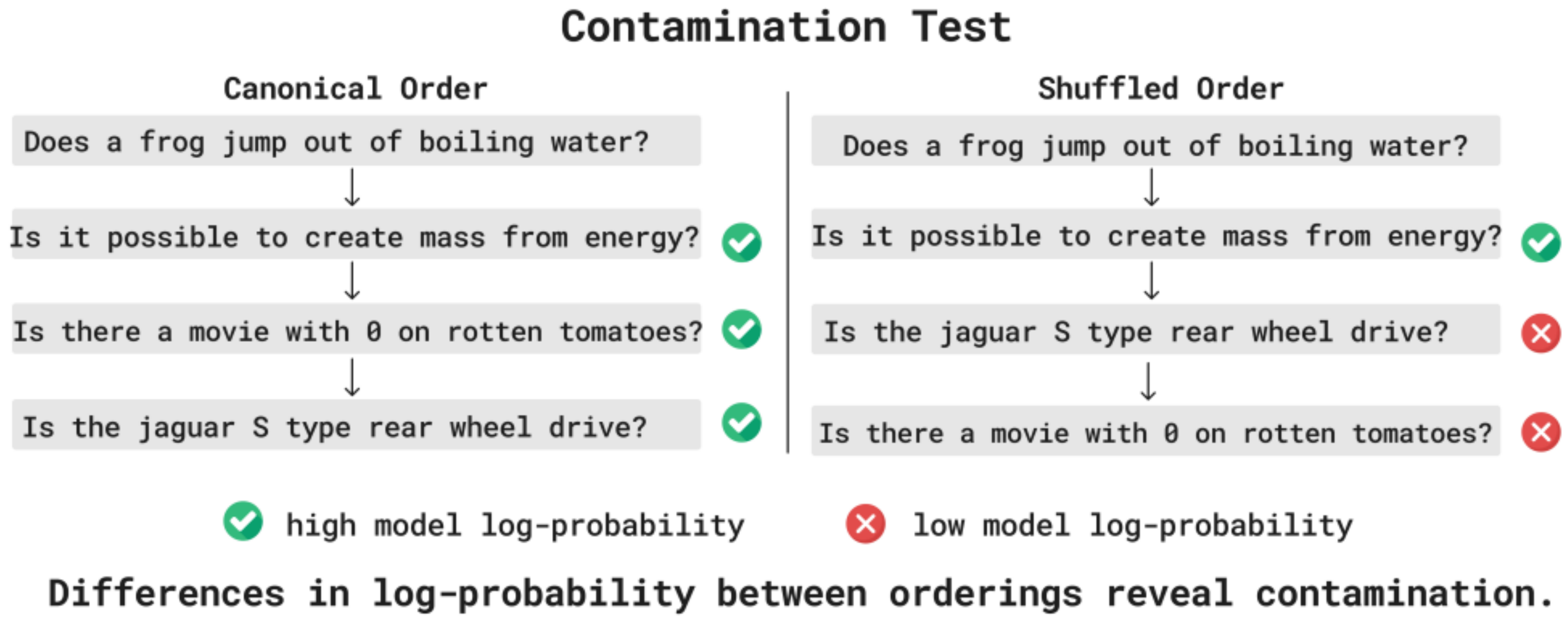
Closed models + pretraining: hard to know that benchmarks are truly 'new'

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
$= \frac{1}{4}\sum_{x_i \in \{the, Royal, Miss, 15\}} logp(x_i | \cdot)$
(a) get token probs
(b)select min K% tokens
(c) average log-likelihood
$> \epsilon$
GPT-3.5 is pretrained on X

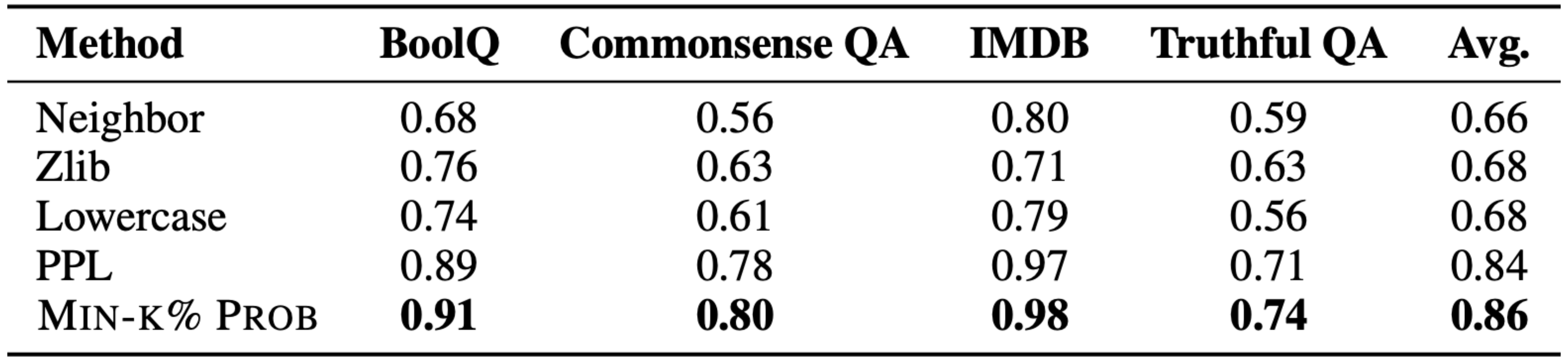
Exchangeability test


Contamination 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? ✗
✓ high model log-probability
✗ low model log-probability
Differences in log-probability between orderings reveal contamination.

- Detect if models trained on a benchmark by checking if probabilities are 'too high' (what is too high?). Often heuristic.
- Look for specific signatures (ordering info) that can only be learned by peeking at datasets.

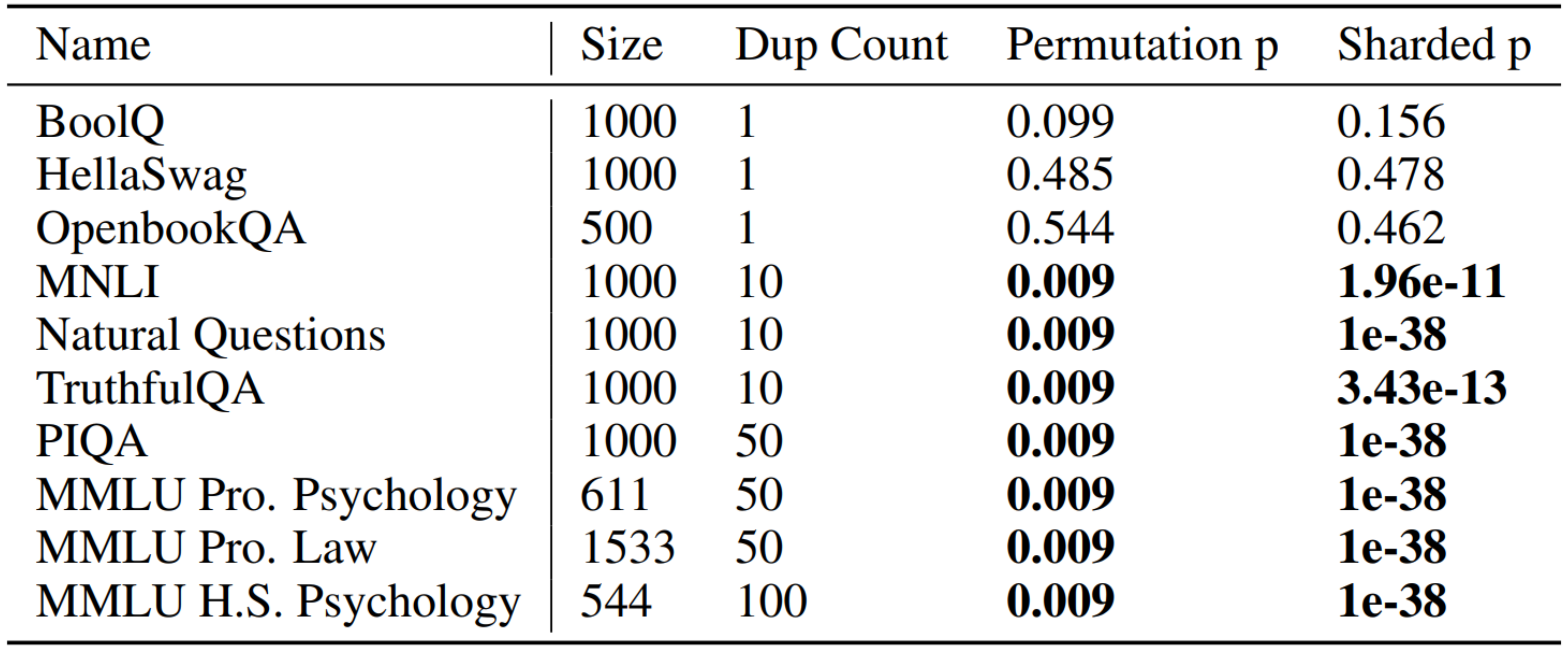
Min-k-prob and other detectors

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

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're not good enough on their own

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're not good enough on their own

- *Model-based metrics* can be more correlated with human judgment, but often are not interpretable

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're not good enough on their own

- *Model-based metrics* can be more correlated with human judgment, but often are not interpretable

- Human judgments are critical
  - But remember humans are inconsistent!

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're not good enough on their own

- *Model-based metrics* can be more correlated with human judgment, but often are not interpretable

- Human judgments are critical
  - But remember humans are inconsistent!

- Challenges
  - Consistency: Does the evaluation ignore nuisance variation?
  - Contamination: Can we trust the numbers?

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're not good enough on their own

- *Model-based metrics* can be more correlated with human judgment, but often are not interpretable

- Human judgments are critical
  - But remember humans are inconsistent!

- Challenges
  - Consistency: Does the evaluation ignore nuisance variation?
  - Contamination: Can we trust the numbers?

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're not good enough on their own

- *Model-based metrics* can be more correlated with human judgment, but often are not interpretable

- Human judgments are critical
  - But remember humans are inconsistent!

- Challenges
  - Consistency: Does the evaluation ignore nuisance variation?
  - Contamination: Can we trust the numbers?

- In many cases, the best judge of output quality is **YOU**!
  - **Look at the actual generations - don't just rely on numbers.**
  - **Publicly release large samples of outputs from your system!**