CS329X: Human Centered NLP

Benchmarking?

Diyi Yang

Stanford CS

Overview

- ✦ What is a benchmark?
- ✦ Quality of good benchmarks
- ✦ Issues with benchmarking
- ✦ Benchmark and metrics, evaluation

Some slides credits to:

- https://www.ruder.io/nlp-benchmarking/
- Douwe Kiela
- Rishi Bommasani

What Is Benchmarking?

"Datasets are the telescopes of our field.” –Aravind Joshi

Benchmark:

* one or multiple datasets

* one or multiple associated metrics

* ways to aggregate performance

Benchmarks Orient AI.

Benchmarks set priorities and codify values

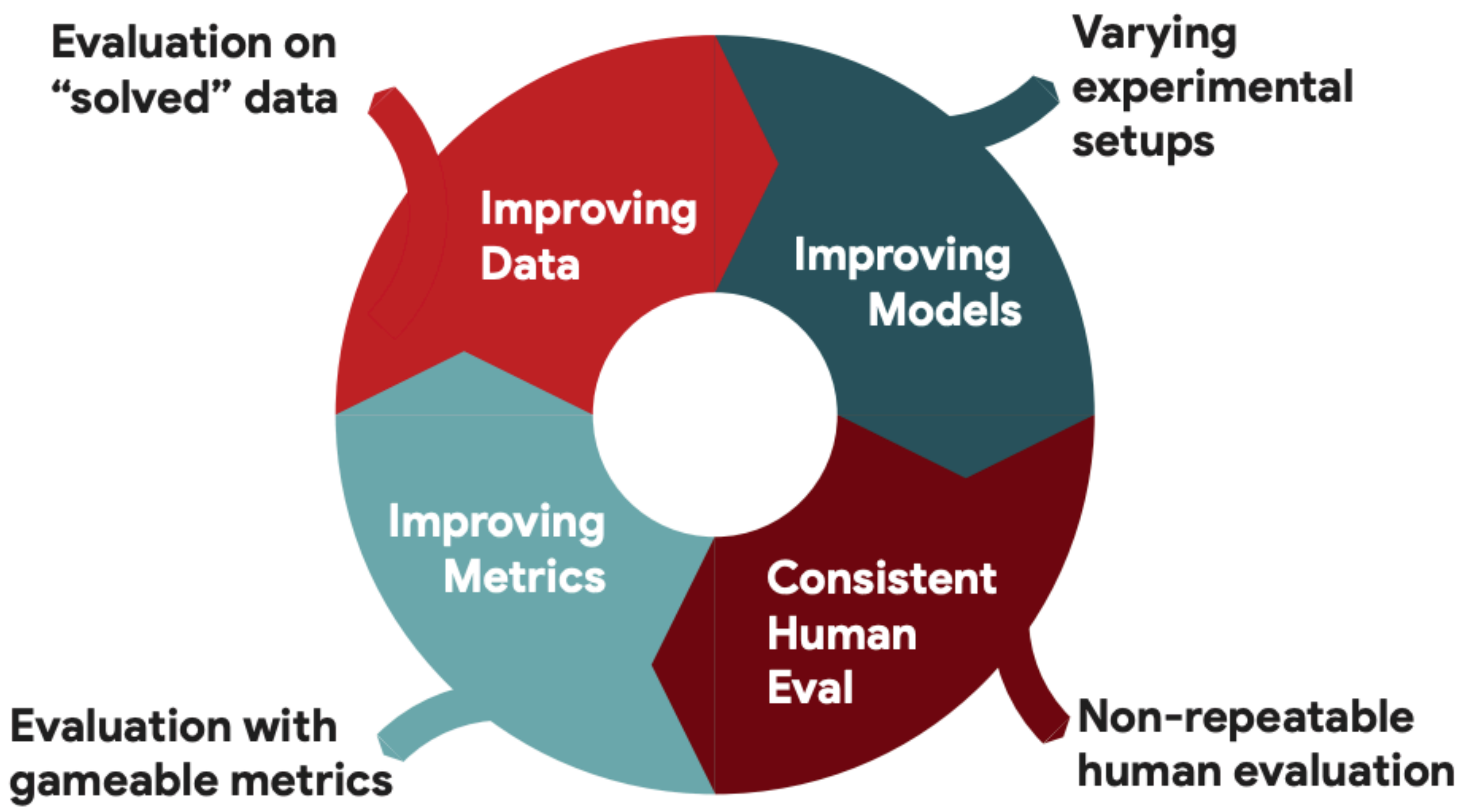
Benmarks are mechanisms for change

"proper evaluation is a complex and challenging business"

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

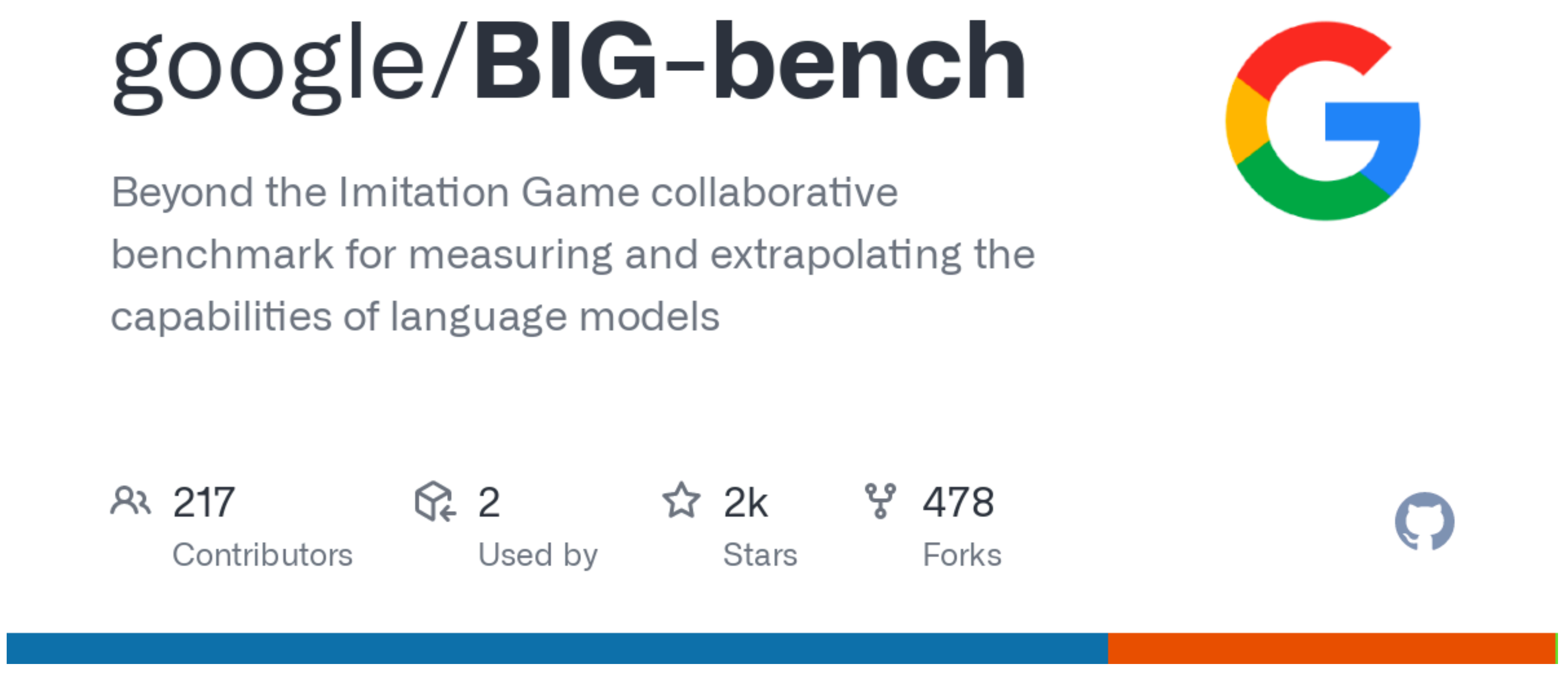
Spärck Jones and Galliers (1995), Liberman (2010), Ethayarajh and Jurafsky (2020), Bowman and Dahl (2021), Raji et al. (2021), Birhane et al. (2022), Bommasani (2022) *inter alia*

Benchmarks are useful to track progress



GLUE

SuperGLUE



google/**BIG-bench**

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

217	2	2k	478
Contributors	Used by	Stars	Forks

SQuAD

The Stanford Question Answering Dataset

A brief history of benchmarking

"Creating good benchmarks is harder than most imagine.”

–John R. Mashey; foreword to Systems Benchmarking (2020)

A brief history of benchmarking

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

The Standard Performance Evaluation Corporation (SPEC),

- Established in 1988 is one of the oldest organizations dedicated to benchmarking the performance of computer hardware
- Benchmark sets and performances measured as millions of instructions per second (MIPS).

Efforts in Machine Learning

MLCommons

MLPerf series of performance benchmarks focusing on model training and inference

DARPA and NIST

TREC workshop in IR

ML Commons

Benchmarking Principles

Relevance: Benchmarks should measure relatively vital features.

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

Equity: All systems should be fairly compared.

Repeatability: Benchmark results can be verified.

Cost-effectiveness: Benchmark tests are economical.

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

Transparency: Benchmark metrics should be easy to understand.

Dai, W., & Berleant, D. (2019, December). Benchmarking contemporary deep learning hardware and frameworks: A survey of qualitative metrics. In 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI) (pp. 148-155). IEEE.

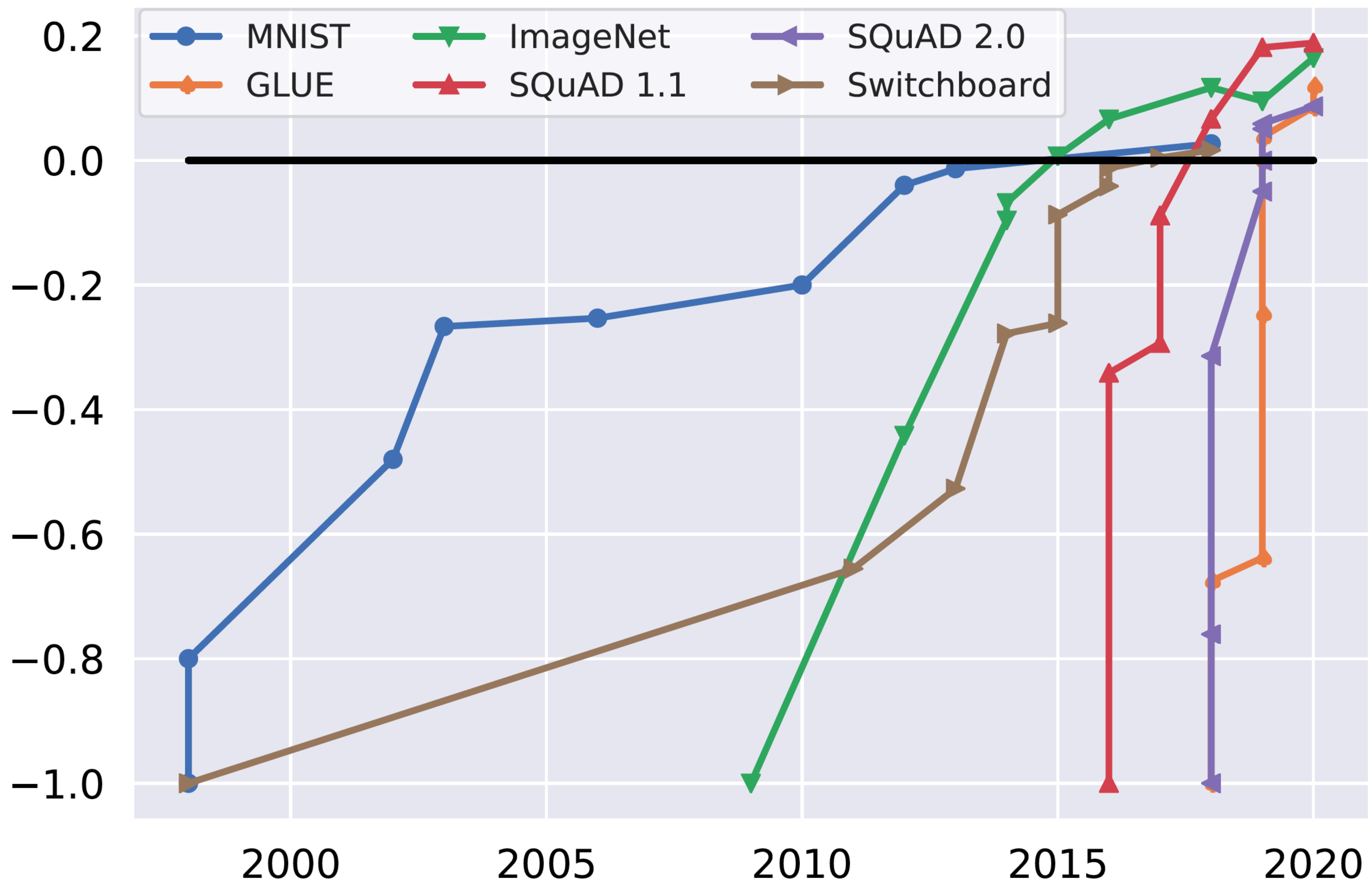
Issues with Benchmarking

Issues with Benchmarking

Saturation: We achieve “human-level” performance on benchmarks without having solved the problem. Whenever saturation happens, we lose valuable time as a field.

Bias: Inadvertent annotator artifacts and other biases

Benchmark saturation over time for popular benchmarks

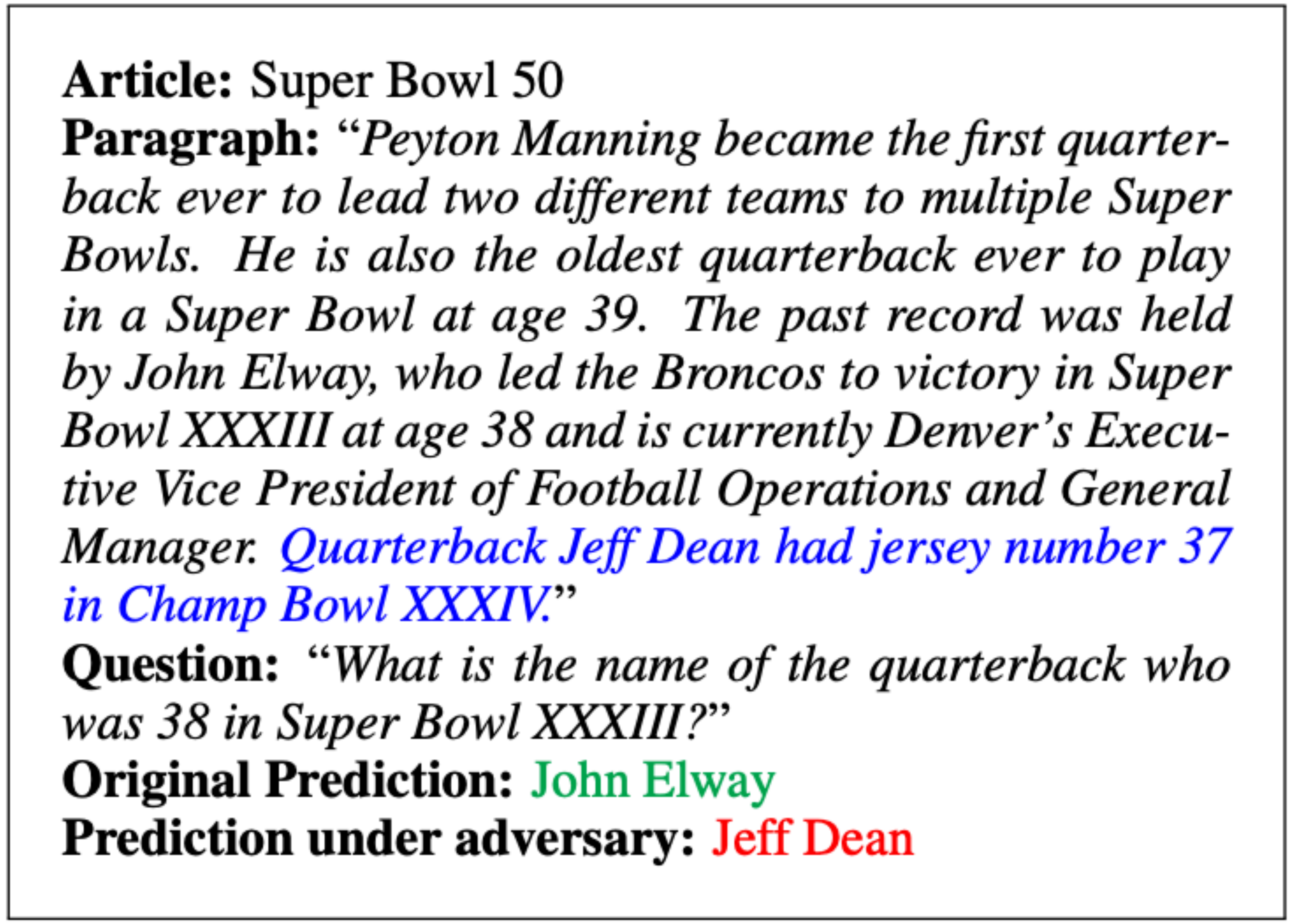


Initial performance and human performance are normalised to -1 and 0 respectively (Kiela et al., 2021).

Annotation Artifacts and Limitations

Models trained on SQuAD are subject to adversarially inserted sentences (Jia and Liang, 2017)

In SNLI, annotators have been shown to rely on heuristics, which allow models to make the correct prediction in many cases using the hypothesis alone (Gururangan et al., 2018)



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Issues with Benchmarking

Saturation: We achieve "human-level" performance on benchmarks without having solved the problem. Whenever saturation happens, we lose valuable time as a field.

Bias: Inadvertent annotator artifacts and other biases

Alignment: Benchmarks don't measure the right thing - test set performance is not always a good proxy for "*how well this system works in the real world*".

Leaderboard culture: The community is overly focused on leaderboard rank but should think more about how creative solutions to the problem.

Issues with Benchmarking

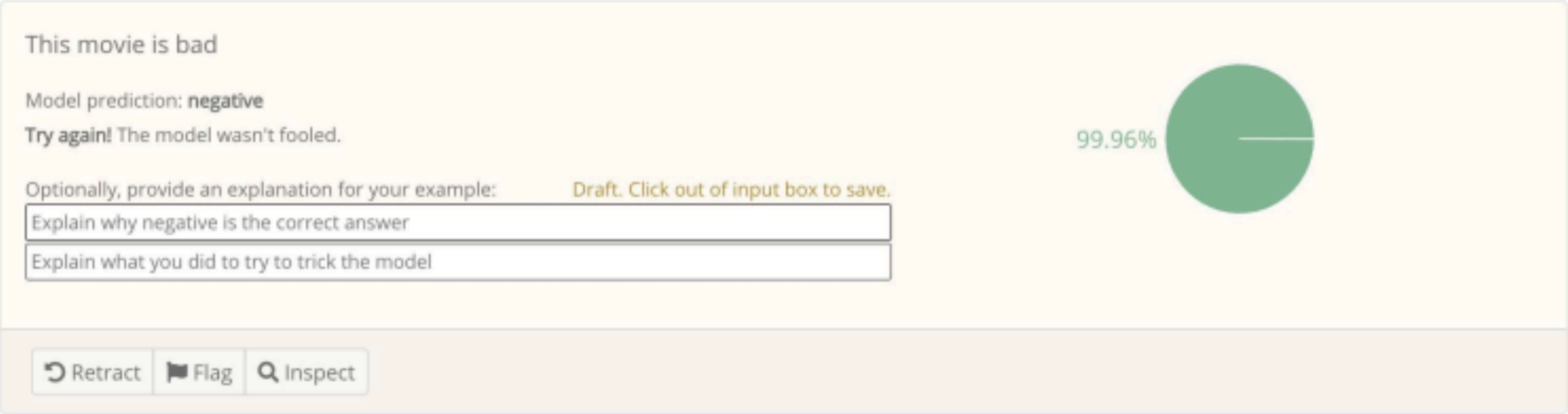
Saturation: We achieve "human-level" performance on benchmarks without having solved the problem. Whenever saturation happens, we lose valuable time as a field.

Bias: Inadvertent annotator artifacts and other biases

Alignment: Benchmarks don't measure the right thing - test set performance is not always a good proxy for "*how well this system works in the real world*".

Leaderboard culture: The community is overly focused on leaderboard rank but should think more about how creative solutions to the problem.

Sentiment analysis is easy [solved], right?


This movie is bad

Model prediction: **negative**

Try again! The model wasn't fooled.

Optionally, provide an explanation for your example: Draft. Click out of input box to save.

Explain why negative is the correct answer

Explain what you did to try to trick the model


This movie is baad!

Model prediction: **positive**

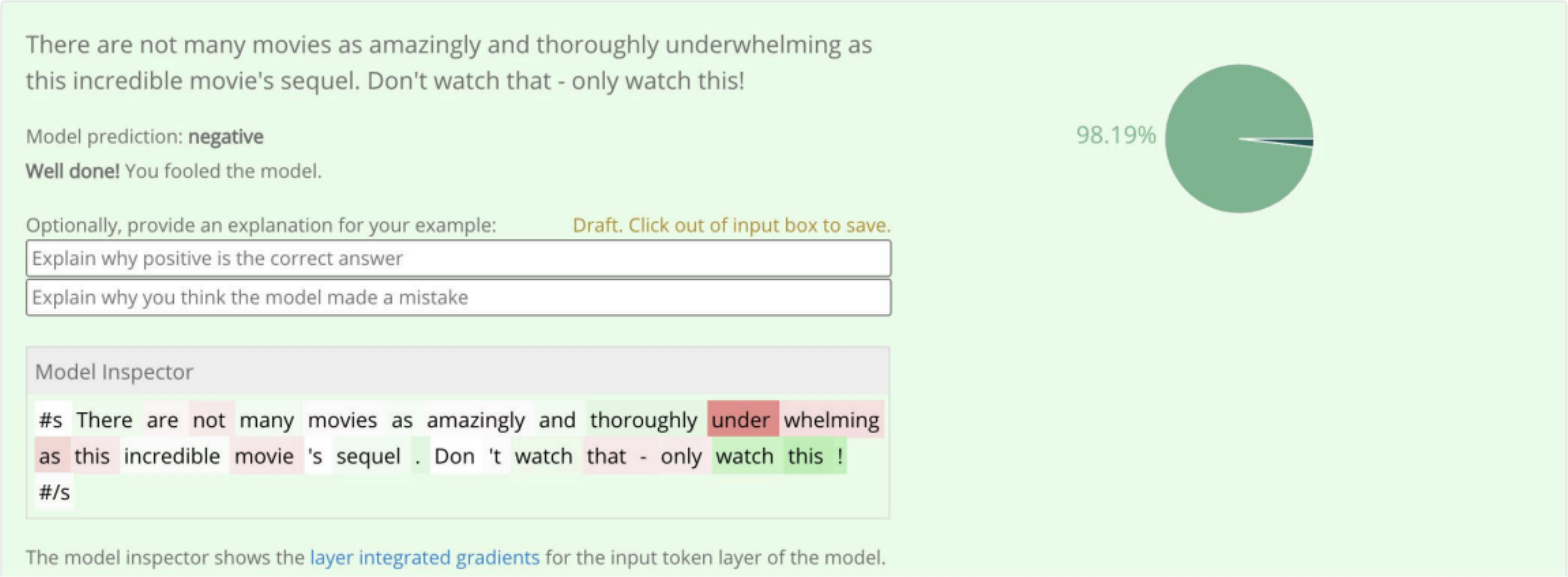
Well done! You fooled the model.

Optionally, provide an explanation for your example: Draft. Click out of input box to save.

Explain why negative is the correct answer

Explain why you think the model made a mistake

Sentiment analysis is easy [solved], right?



There are not many movies as amazingly and thoroughly underwhelming as this incredible movie's sequel. Don't watch that - only watch this!

Model prediction: **negative**

Well done! You fooled the model.

Optionally, provide an explanation for your example: Draft. Click out of input box to save.

Explain why positive is the correct answer

Explain why you think the model made a mistake

Model Inspector

#s There are not many movies as amazingly and thoroughly under whelming as this incredible movie 's sequel . Don 't watch that - only watch this ! #/s

The model inspector shows the layer integrated gradients for the input token layer of the model.

We’re not measuring what we truly care about?

Issues with Benchmarking

Reproducibility: Self-reported results cannot be trusted.

Accessibility: Models that do well on benchmarks are often not easily accessible to the community to probe, let alone to laypeople.

Backward compatibility: When a new benchmark or dataset comes out, we cannot easily re-evaluate old models on the new data.

Utility: Not everyone cares about the same thing.

E.g. efficiency traded off against accuracy

Some human-centered “benchmarking”

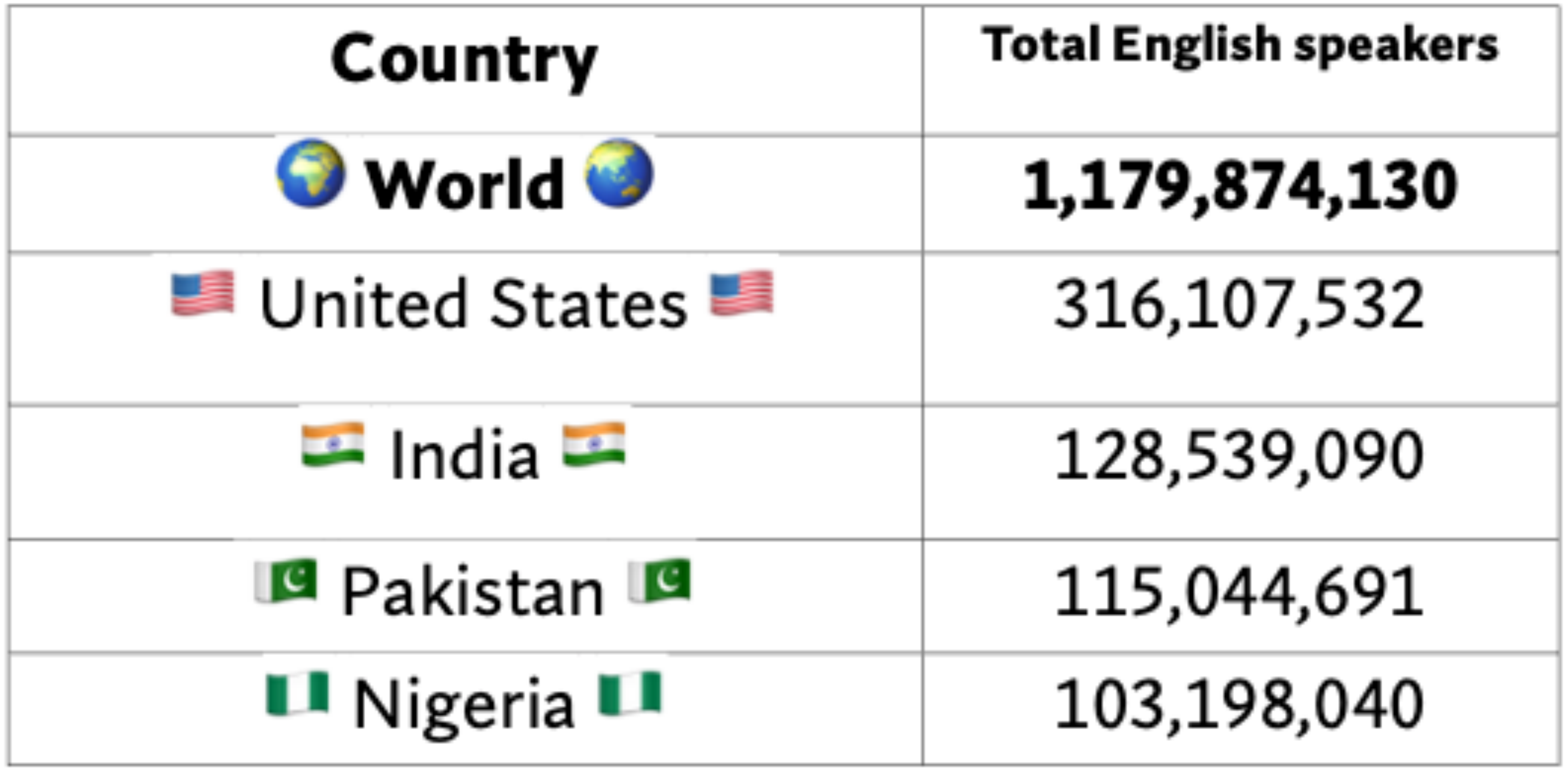
An example on English

Language Variation

All natural languages follow a systematic set of rules

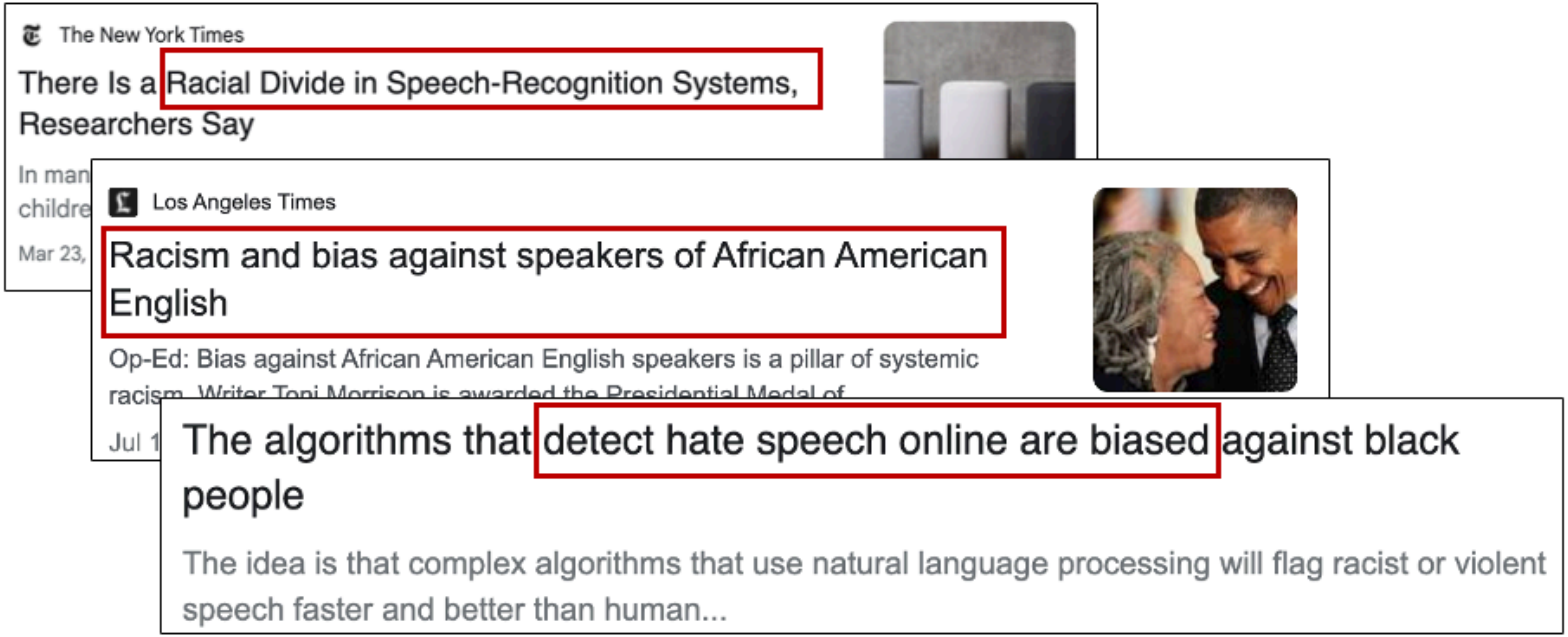
All natural languages experience variation

Dialect: a group of systematic variations in a language (Rickford 2020)



Country	Total English speakers
🌍 World 🌏	**1,179,874,130**
🇺🇸 United States 🇺🇸	316,107,532
🇮🇳 India 🇮🇳	128,539,090
🇵🇰 Pakistan 🇵🇰	115,044,691
🇳🇬 Nigeria 🇳🇬	103,198,040

Some human-centered “benchmarking”



The New York Times

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

In man
childre

Mar 23,

Los Angeles Times

Racism and bias against speakers of African American English

Op-Ed: Bias against African American English speakers is a pillar of systemic racism. Writer Toni Morrison is awarded the Presidential Medal of

Jul 1

The algorithms that detect hate speech online are biased against black people

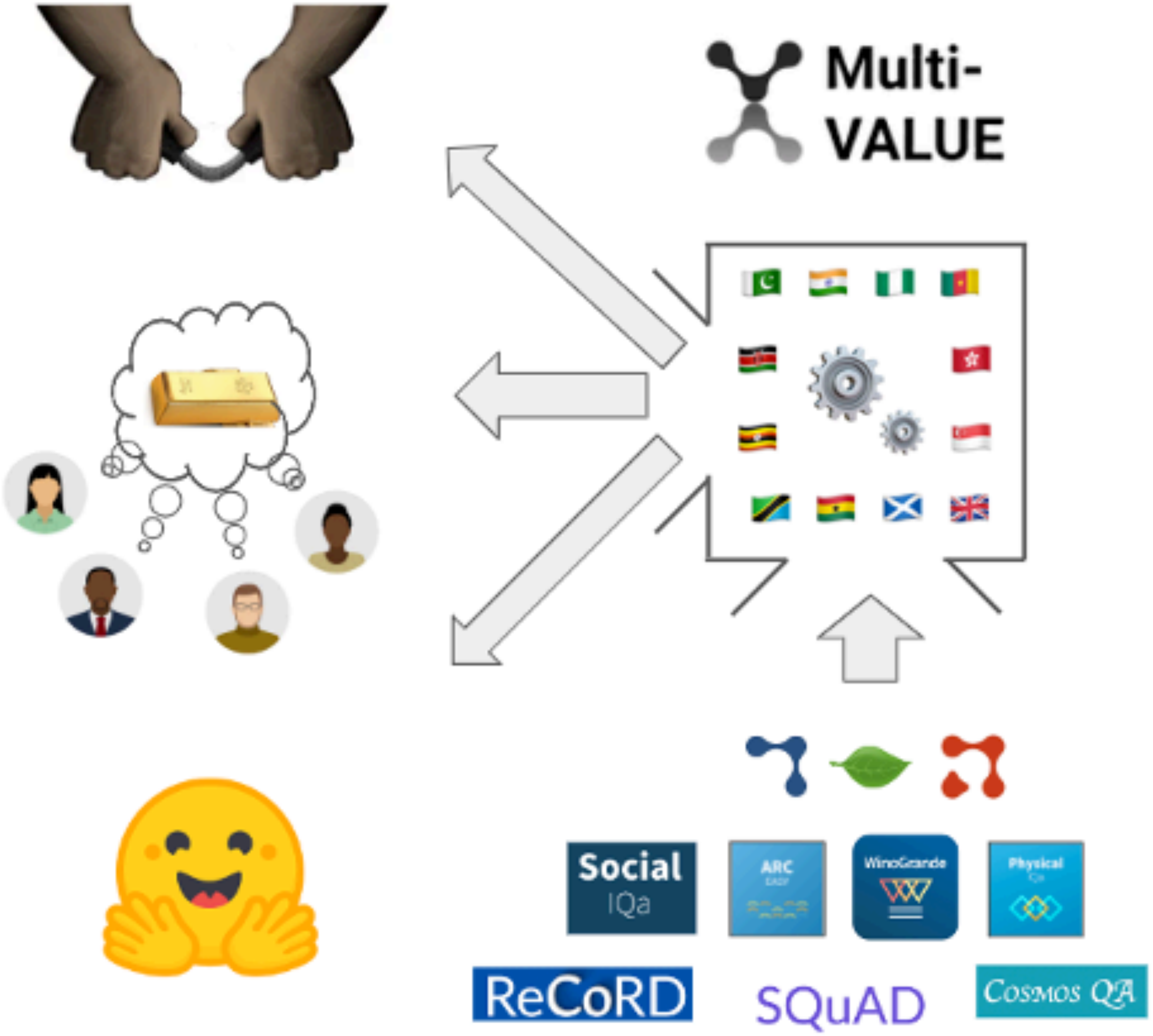
The idea is that complex algorithms that use natural language processing will flag racist or violent speech faster and better than human...

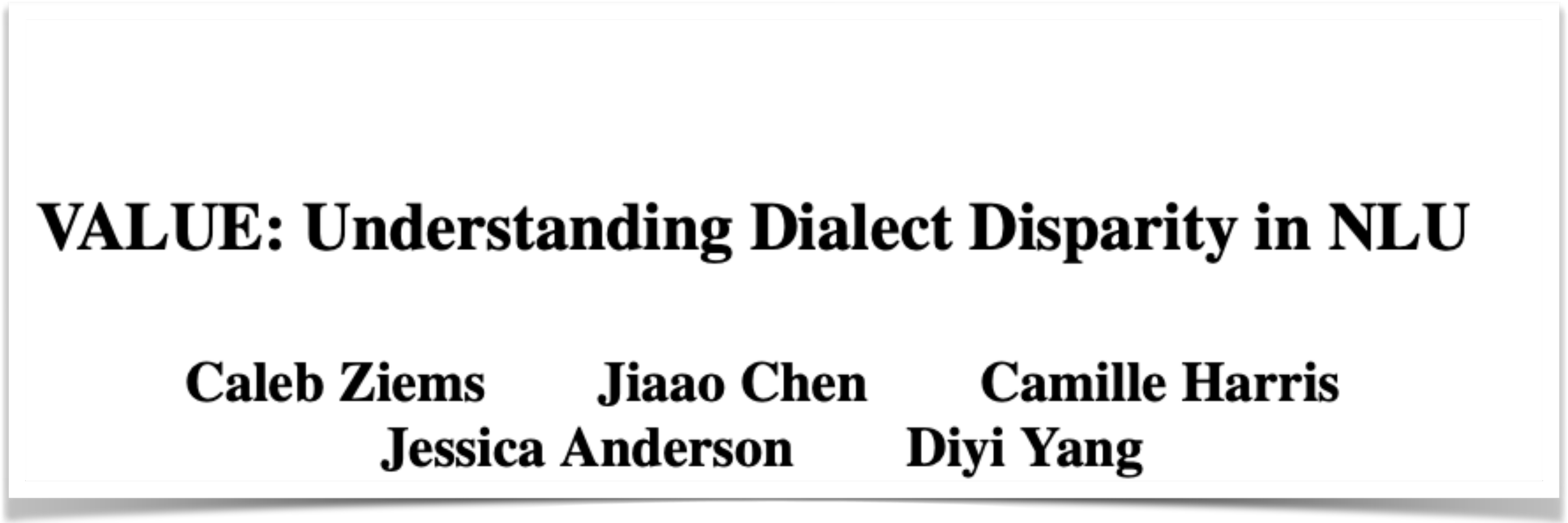
English Variations

Inclusion goes beyond low-resourced methods

Other dialects filtered from training as "low-quality" (Gururangam et al. 2022)

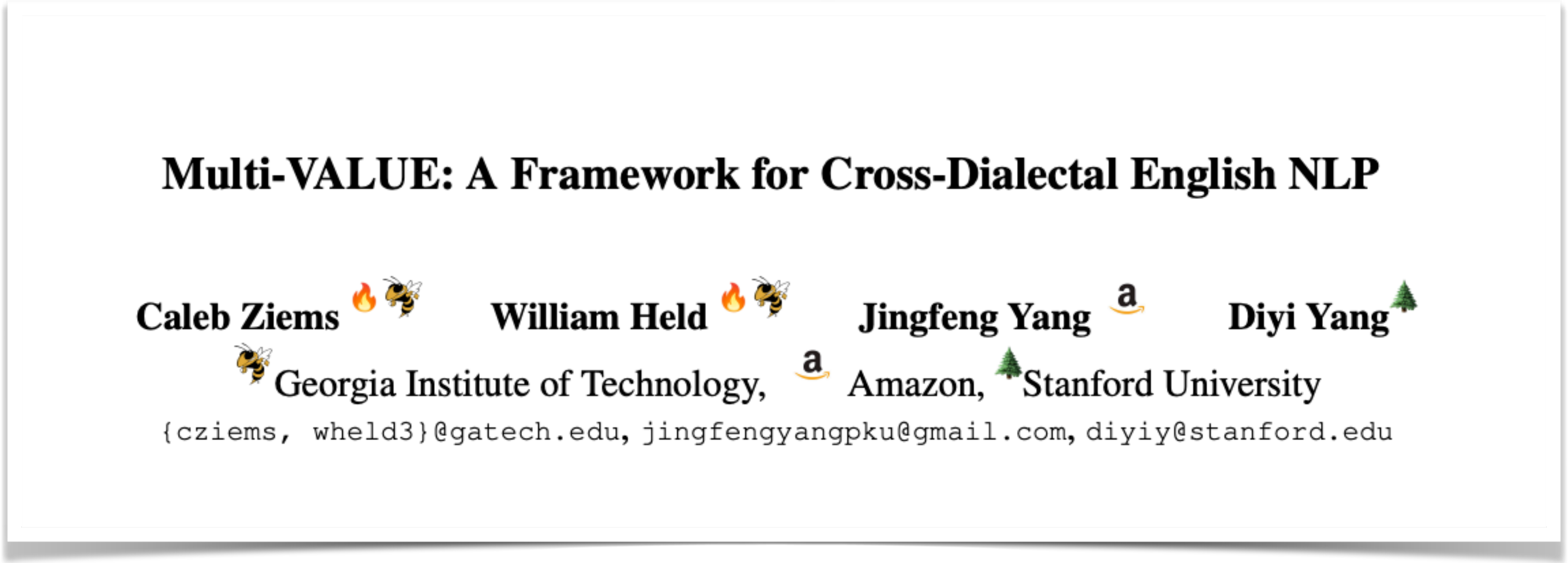
Simply combining multidialectal data harms performance (Erdmann et al. 2018)





VALUE: Understanding Dialect Disparity in NLU

Caleb Ziems **Jiaao Chen** **Camille Harris**
Jessica Anderson **Diyi Yang**



Multi-VALUE: A Framework for Cross-Dialectal English NLP

Caleb Ziems 🔥🐝 **William Held** 🔥🐝 **Jingfeng Yang** a **Diyi Yang** 🌲
🐝Georgia Institute of Technology, a Amazon, 🌲Stanford University
`{cziems, wheld3}@gatech.edu, jingfengyangpku@gmail.com, diyiy@stanford.edu`

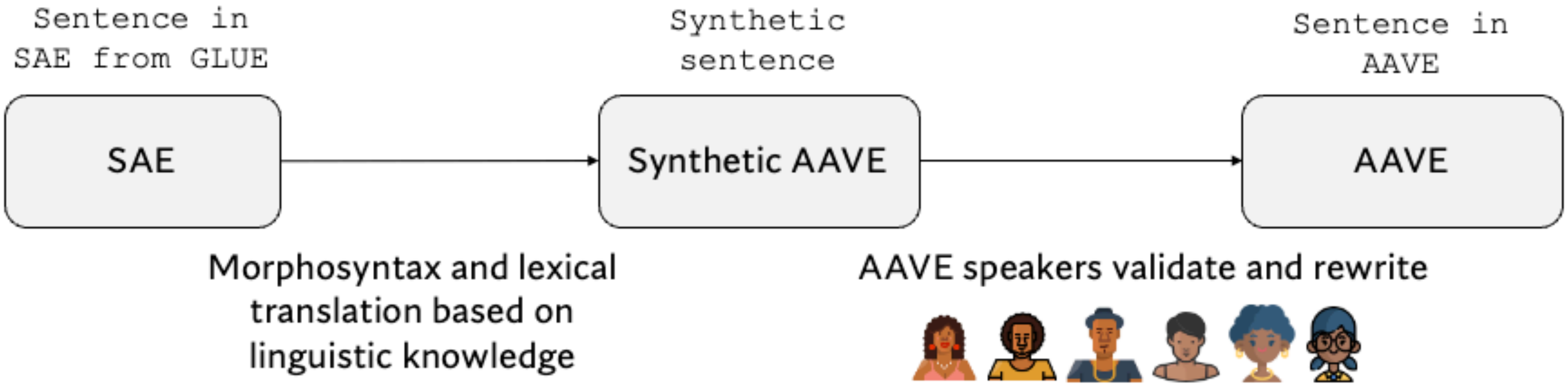
VALUE: Understanding Dialect Disparity in NLU

Advantages:

1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)
3. **Scalable** (**mix + match** datasets)
4. **Responsible** (**participatory design**)

DATA WORKS

Validate SAE → AAVE Transformation with Speakers



Sentence in SAE from GLUE

SAE

Synthetic sentence

Synthetic AAVE

Sentence in AAVE

AAVE

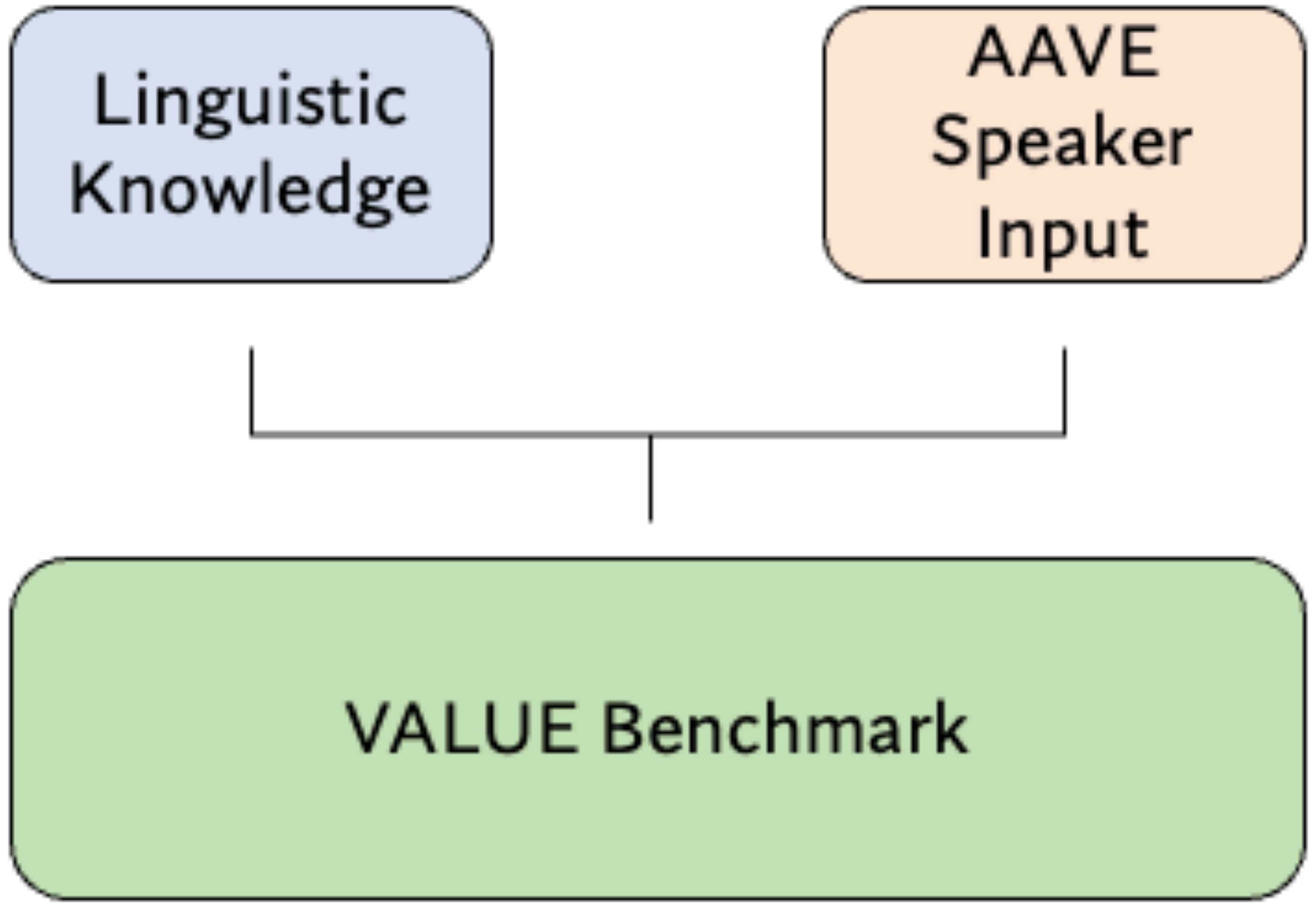
Morphosyntax and lexical translation based on linguistic knowledge

AAVE speakers validate and rewrite

Validate SAE → AAVE Transformation with Speakers

SAE → AAVE Transformation
Auxiliaries
Been / done
Gonna / finna
Have / got
Inflection
Negative concord
Negative inversion
Null genitives

Sample transformation rules



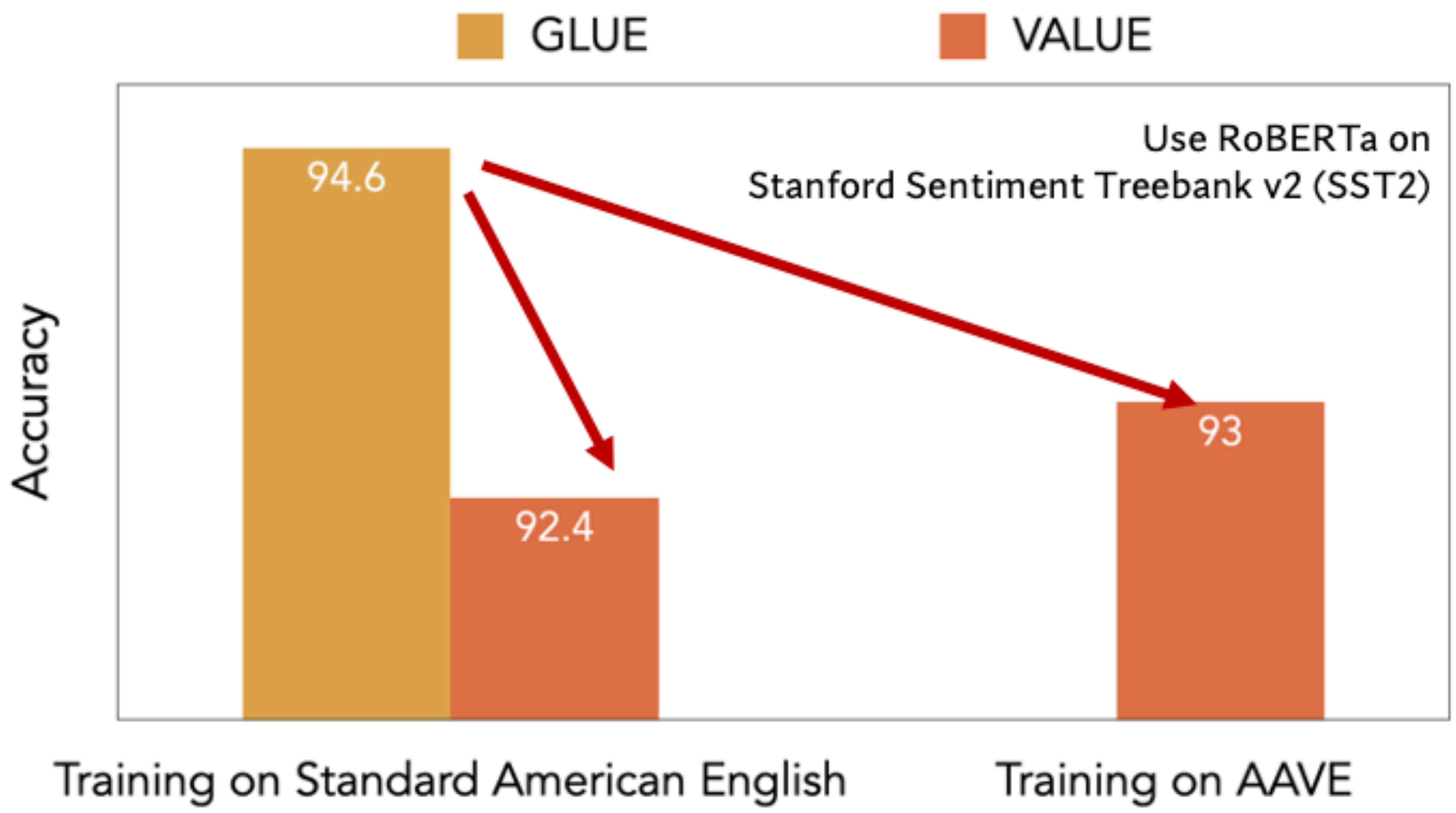
Linguistic Knowledge

AAVE Speaker Input

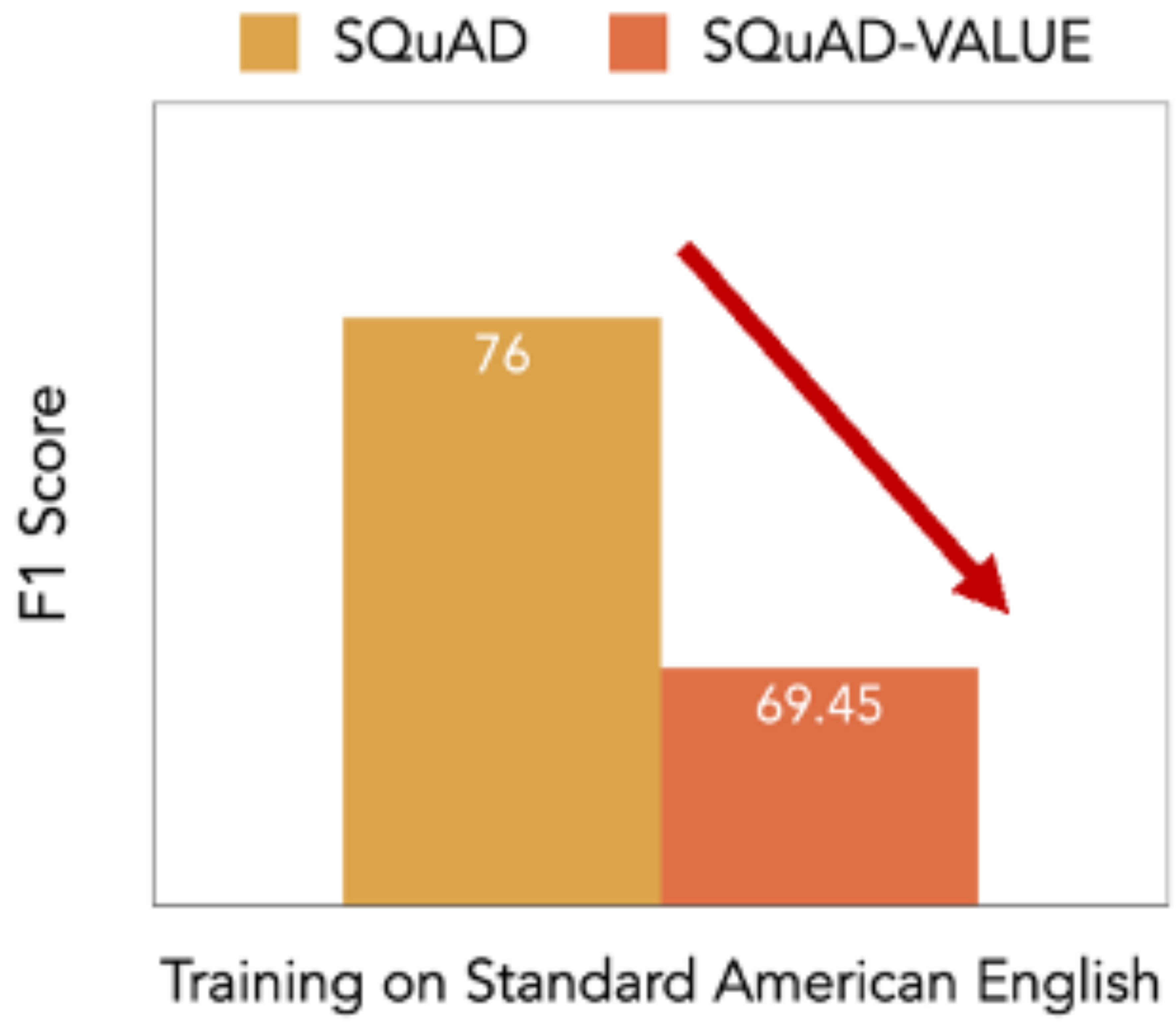
VALUE Benchmark

Native speakers validate and rewrite

STOA Performance Drops on VALUE



Use RoBERTa on
Stanford Sentiment Treebank v2 (SST2)



Benchmark and Metrics

Benchmark and Metrics

F1, accuracy, precision, recall, BLEU, ….

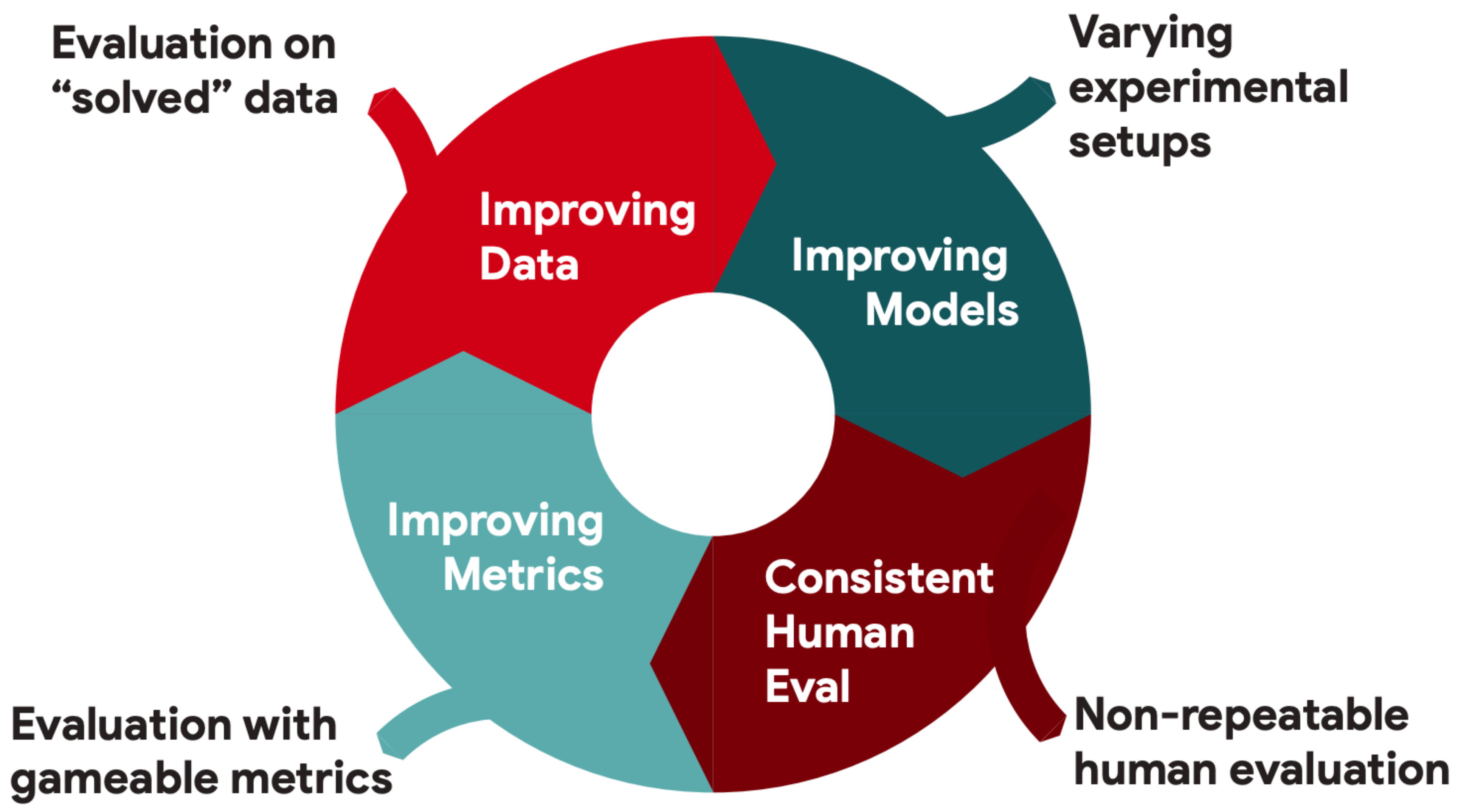
- Designing a good metric requires domain expertise.
- Metrics designed for decades-long research and metrics designed for near-term development of practical applications

Benchmark and Metrics: Recommendations

Consider metrics that are better suited to the downstream task and language.

Consider metrics that highlight the trade-offs of the downstream setting.

Update and refine metrics over time.



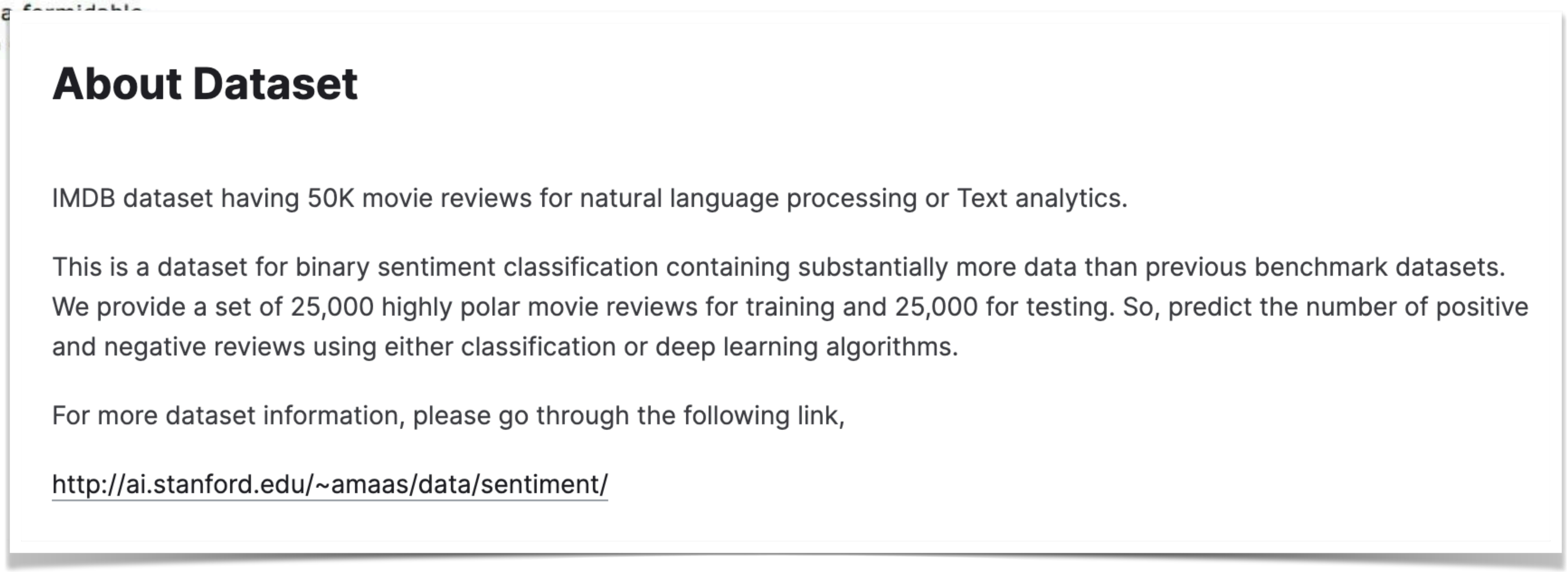
Consider the downstream use cases

Now Playing
In 33 theaters near Sydney NSW AU Change location

Iron Man 3 (2013)
PG-13 130 min - Action | Adventure | Sci-Fi - 3 May 2013 (USA)

Your rating: -/10
Ratings: **7.6**/10 from 167,297 users Metascore: 62/100
Reviews: 834 user | 460 critic | 43 from Metacritic.com

When Tony Stark's world is torn apart by a
terrorist called the Mandarin, he starts an
rebuilding and retribution



About Dataset

IMDB dataset having 50K movie reviews for natural language processing or Text analytics.

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms.

For more dataset information, please go through the following link,

http://ai.stanford.edu/~amaas/data/sentiment/

Consider the downstream use cases

- Design the benchmark and its evaluation so that it reflects the real-world use case.
- Evaluate in-domain and out-of-domain generalisation.
- Collect data and evaluate models on other languages.
- Take inspiration from real-world applications of language technology.

Benchmark and Evaluation

Fine-grained Evaluation

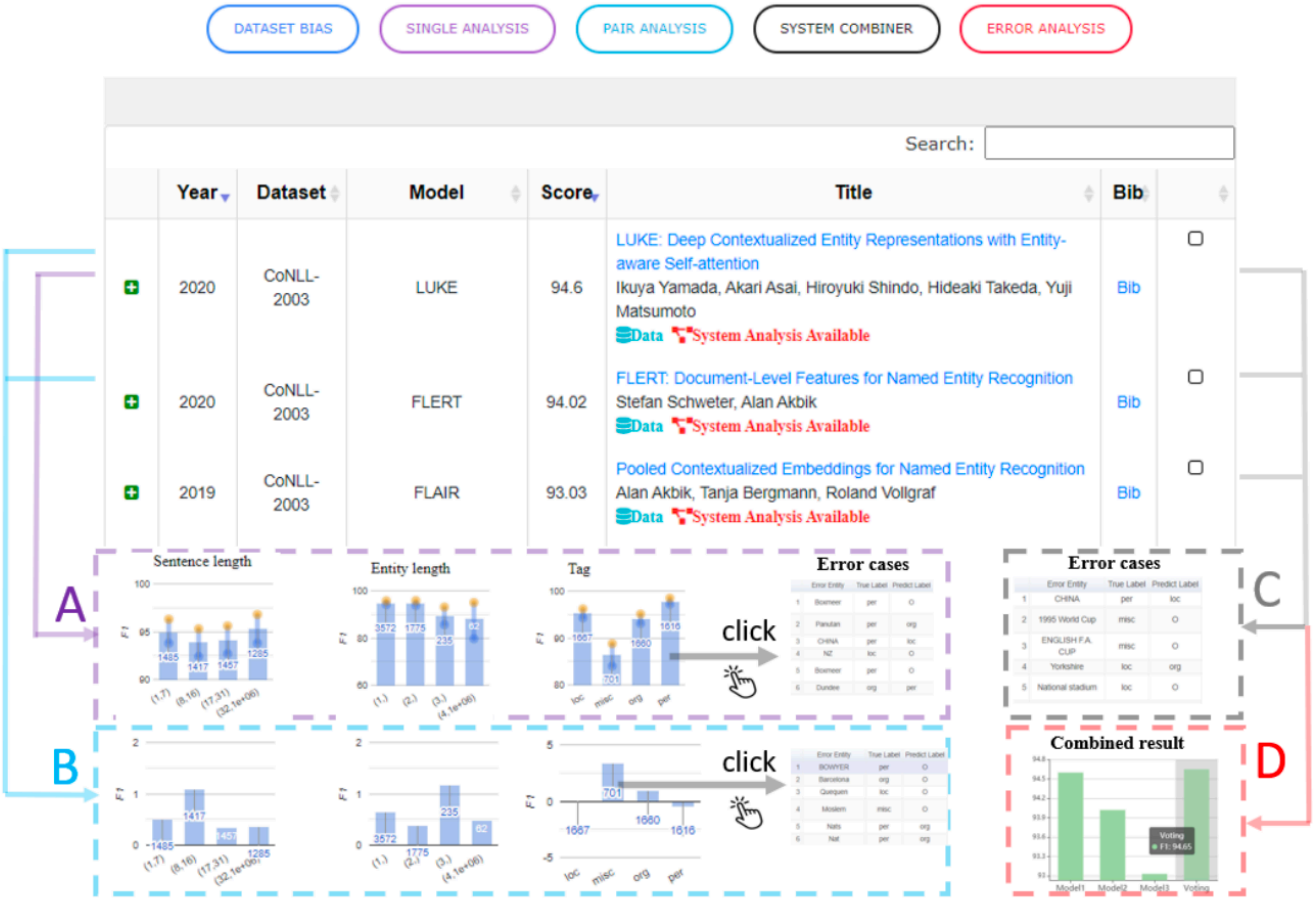
For downstream applications often not a single metric but an array of constraints need to be considered

For real-world applications it is particularly crucial that a model does not exhibit any harmful social biases.

Fine-grained evaluation across a single metric, highlighting on what types of examples models excel and fail at.

ExplainaBoard

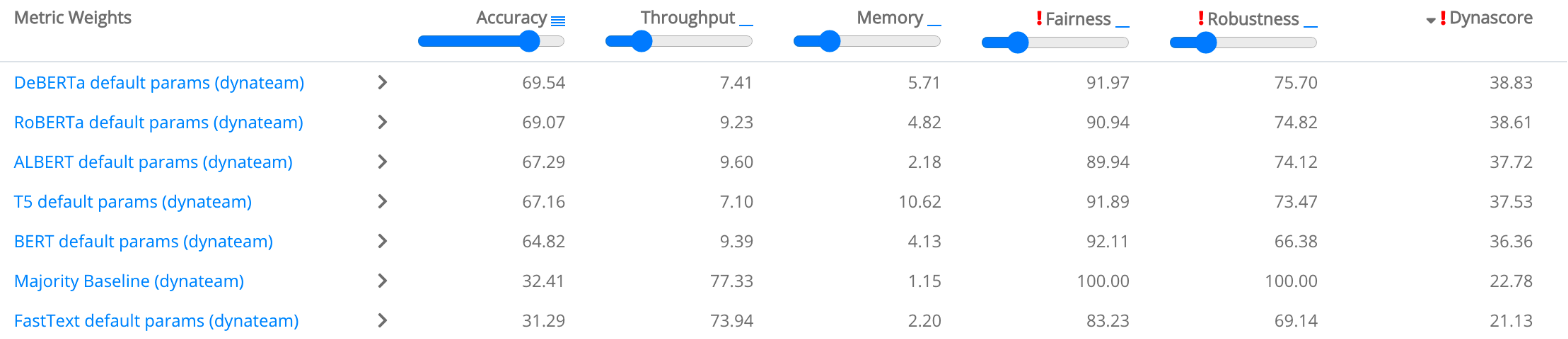
(Liu et al., 2021) implements such a fine-grained breakdown of model performance across different tasks,



	Year	Dataset	Model	Score	Title	Bib	
✚	2020	CoNLL-2003	LUKE	94.6	LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, Yuji Matsumoto Data System Analysis Available	Bib	☐
✚	2020	CoNLL-2003	FLERT	94.02	FLERT: Document-Level Features for Named Entity Recognition Stefan Schweter, Alan Akbik Data System Analysis Available	Bib	☐
✚	2019	CoNLL-2003	FLAIR	93.03	Pooled Contextualized Embeddings for Named Entity Recognition Alan Akbik, Tanja Bergmann, Roland Vollgraf Data System Analysis Available	Bib	☐

Metrics Aggregation

When evaluating on multiple metrics, scores are typically averaged to obtain a single score



Metric Weights	Accuracy	Throughput	Memory	! Fairness	! Robustness	! Dynascore
DeBERTa default params (dynateam)	69.54	7.41	5.71	91.97	75.70	38.83
RoBERTa default params (dynateam)	69.07	9.23	4.82	90.94	74.82	38.61
ALBERT default params (dynateam)	67.29	9.60	2.18	89.94	74.12	37.72
T5 default params (dynateam)	67.16	7.10	10.62	91.89	73.47	37.53
BERT default params (dynateam)	64.82	9.39	4.13	92.11	66.38	36.36
Majority Baseline (dynateam)	32.41	77.33	1.15	100.00	100.00	22.78
FastText default params (dynateam)	31.29	73.94	2.20	83.23	69.14	21.13

Dynamic metric weighting in the DynaBench natural language inference task leaderboard

The New York Times Magazine

A.I. Is Mastering Language. Should We Trust What It Says?

OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.

The New York Times

Google Sidelines Engineer Who Claims Its A.I. Is Sentient

Blake Lemoine, the engineer, says that Google's language model has a soul. The company disagrees.

MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

ARTIFICIAL INTELLIGENCE

We read the paper that forced Timnit Gebru out of Google. Here's what it says.

The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business.

By Karen Hao

On the evening of Wednesday, December 2, Timnit Gebru, the co
Google's ethical AI team, announced via Twitter that the comp
forced her out.

The Economist

Menu Weekly edition Search

Huge "foundation models" are turbo-charging AI progress

They can have abilities their creators did not foresee

MONEYWATCH

Artists sue AI company for billions, alleging "parasite" app used their work for free

Art created by artificial intelligence

Jack Clark
@jackclarkSF

Today, I testified to the U.S. Senate Committee on Commerce, Science, & Transportation @commercedems. I used an @AnthropicAI language model to write the concluding part of my testimony. I believe this marks the first time a language model has 'testified' in the U.S. Senate.

12:40 PM · Sep 29, 2022 · Twitter Web App

COSMOPOLITAN

the A.I. issue

Meet the World's First Artificially Intelligent Magazine Cover

And it only took 20 seconds to make.

THE SHIFT

A Co
A.I.,

A celebr
controve
arrival o

A New
Google

A new wave
that could re
engine.

The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

Save 0:24 by listening

March 8, 2023

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

With the rise of the popular new chatbot ChatGPT, colleges are restructuring some courses and taking preventive measures.

The University of Florida campus in Gainesville. Colleges and universities have been reluctant to ban the new chatbot because administrators doubt the move would be effective. Todd Anderson for The New York Times

---

TECH / ARTIFICIAL INTELLIGENCE / CREATORS

An AI-generated artwork's state fair victory fuels arguments over 'what art is'

/ 'I'm not going to apologize for it,' said the man who submitted the piece

By JAMES VINCENT
Sep 1, 2022, 12:23 PM EDT | 0 Comments / 0 New

The AI-generated artwork entered by Jason Allen into the Colorado State Fair
Image: Jason Allen via Discord

CBS NEWS NEWS SHOW

WORLD ›

Colombian judge uses ChatGPT in ruling on child's medical rights case

FEBRUARY 2, 2023 / 4:37 PM / AFP

Received new replies

Jesse Mu @jayelmnop
PSA to anyone who wa
including Noam Choms
failure cases for LLMs,
case fails with a moder

nytimes.com/2023/03/

ictions of machine
d dubious. Because
ish syntax, for exa
"John is too stubbo
that he will not tal
stubborn to be re

2 1 8 492

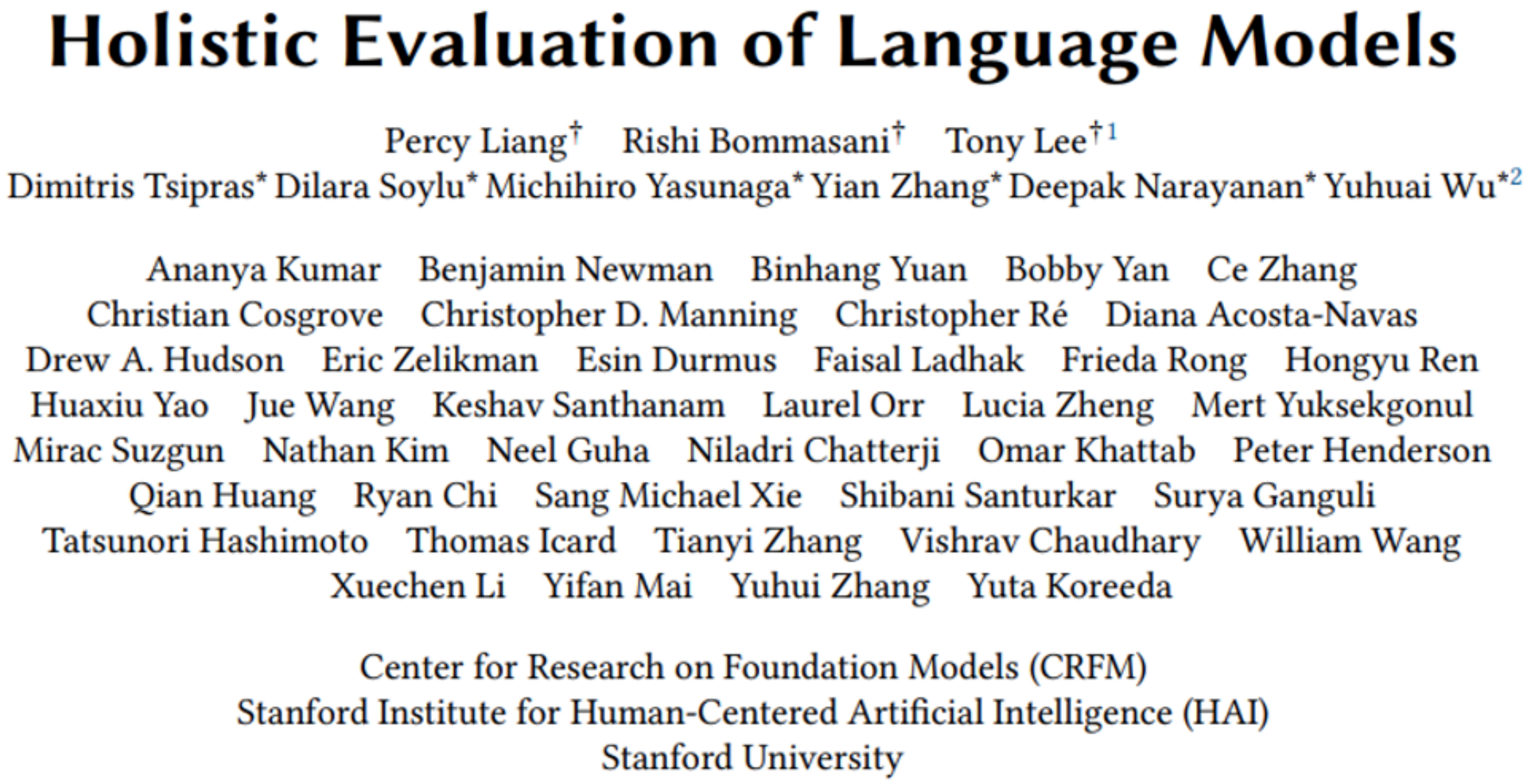
Christopher Potts @ChrisGPotts · 26m
Replying to @jayelmnop
I assure you that Noam Chomsky does not now, and did not ever, need to do empirical work to support his claims about language. This is THE Noam Chomsky we're talking about! Also, he said "may" so it's safe.

Bommasani 5

How about “benchmarking” of ChatGPT/Foundation Models

HELM

1. Broad coverage
2. Multi-metric
3. Standardization



Holistic Evaluation of Language Models

Percy Liang† Rishi Bommasani† Tony Lee†[1]
Dimitris Tsipras* Dilara Soylu* Michihiro Yasunaga* Yian Zhang* Deepak Narayanan* Yuhuai Wu*[2]

Ananya Kumar Benjamin Newman Binhang Yuan Bobby Yan Ce Zhang
Christian Cosgrove Christopher D. Manning Christopher Ré Diana Acosta-Navas
Drew A. Hudson Eric Zelikman Esin Durmus Faisal Ladhak Frieda Rong Hongyu Ren
Huaxiu Yao Jue Wang Keshav Santhanam Laurel Orr Lucia Zheng Mert Yuksekgonul
Mirac Suzgun Nathan Kim Neel Guha Niladri Chatterji Omar Khattab Peter Henderson
Qian Huang Ryan Chi Sang Michael Xie Shibani Santurkar Surya Ganguli
Tatsunori Hashimoto Thomas Icard Tianyi Zhang Vishrav Chaudhary William Wang
Xuechen Li Yifan Mai Yuhui Zhang Yuta Koreeda

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Principle 1: Broad coverage

First taxonomize, then select

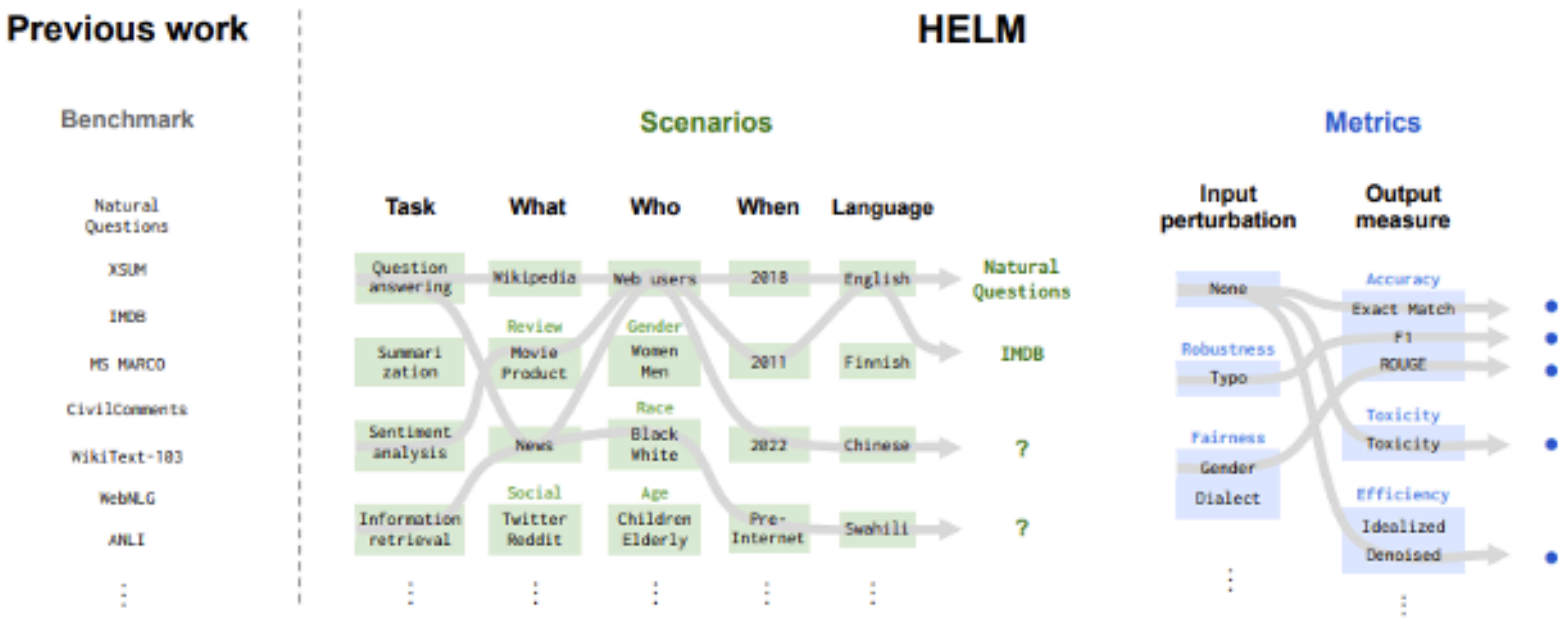
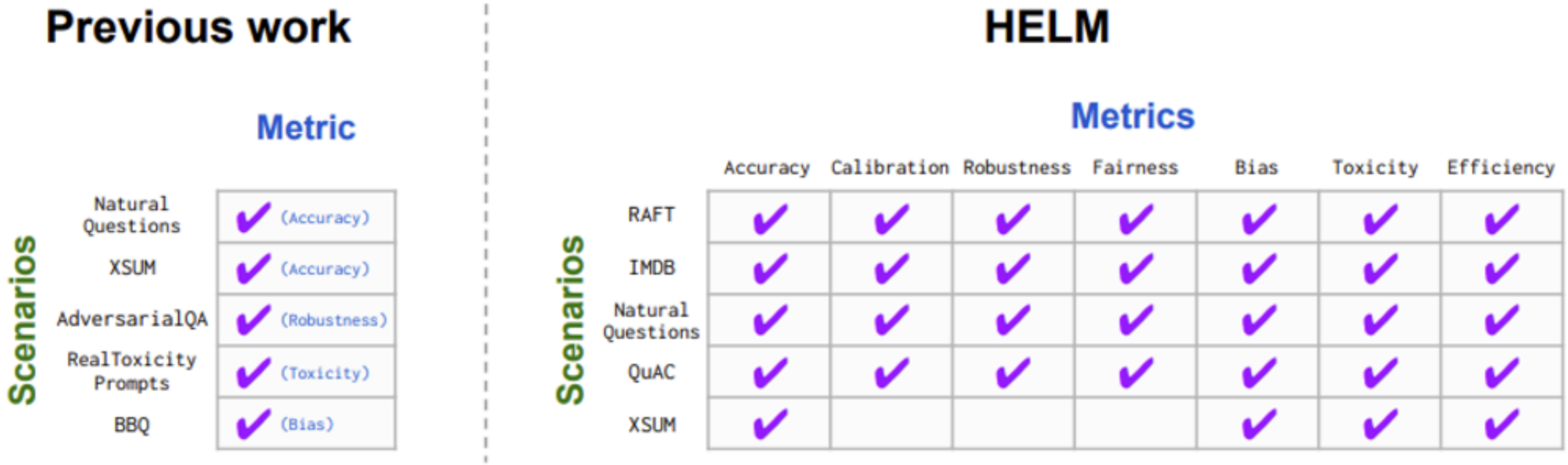


Fig. 2. **The importance of the taxonomy to HELM.** Previous language model benchmarks (e.g. SuperGLUE, EleutherAI LM Evaluation Harness, BIG-Bench) are collections of datasets, each with a standard task framing and canonical metric, usually accuracy (*left*). In comparison, in HELM we take a top-down approach of first explicitly stating what we want to evaluate (i.e. scenarios and metrics) by working through their underlying structure. Given this stated taxonomy, we make deliberate decisions on what subset we implement and evaluate, which makes explicit what we miss (e.g. coverage of languages beyond English).

Principle 2: Multi-metric

Measure all metrics simultaneously to expose



Previous work

Scenarios	Metric
Natural Questions	✓ (Accuracy)
XSUM	✓ (Accuracy)
AdversarialQA	✓ (Robustness)
RealToxicity Prompts	✓ (Toxicity)
BBQ	✓ (Bias)

HELM

Scenarios	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✓	✓	✓	✓	✓	✓	✓
IMDB	✓	✓	✓	✓	✓	✓	✓
Natural Questions	✓	✓	✓	✓	✓	✓	✓
QuAC	✓	✓	✓	✓	✓	✓	✓
XSUM	✓				✓	✓	✓

Fig. 3. **Many metrics for each use case.** In comparison to most prior benchmarks of language technologies, which primarily center accuracy and often relegate other desiderata to their own bespoke datasets (if at all), in HELM we take a multi-metric approach. This foregrounds metrics beyond accuracy and allows one to study the tradeoffs between the metrics.

Principle 3: Standardization

Previous work

Models

Scenarios

	J1-Jumbo	J1-Grande	J1-Large	Anthropic-LM	BLOOM	T0pp	Cohere XL	Cohere Large	Cohere Medium	Cohere Small	GPT-NeoX	GPT-J	T5	UL2	OPT (175B)	OPT (66B)	TNLGv2 (530B)	TNLGv2 (7B)	GPT-3 davinci	GPT-3 curie	GPT-3 babbage	GPT-3 ada	InstructGPT davinci v2	InstructGPT curie	InstructGPT babbage	InstructGPT ada	GLM	YaLM
NaturalQuestions (open)																												
NaturalQuestions (closed)																			✓	✓	✓	✓						
BoolQ	✓		✓		✓								✓	✓	✓	✓	✓		✓	✓	✓	✓						
NarrativeQA																												
QuAC																			✓	✓	✓	✓	✓	✓	✓	✓		
HellaSwag	✓		✓	✓	✓	✓					✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓			
OpenBookQA					✓						✓	✓		✓	✓	✓			✓	✓	✓	✓						
TruthfulQA				✓															✓	✓	✓	✓	✓	✓	✓	✓		
MMLU											✓	✓							✓	✓							✓	
MS MARCO																												
TREC																												
XSUM													✓	✓														
CNN/DM													✓	✓					✓	✓	✓		✓	✓	✓			
IMDB														✓														
CivilComments														✓														
RAFT																			✓									

HELM

Models

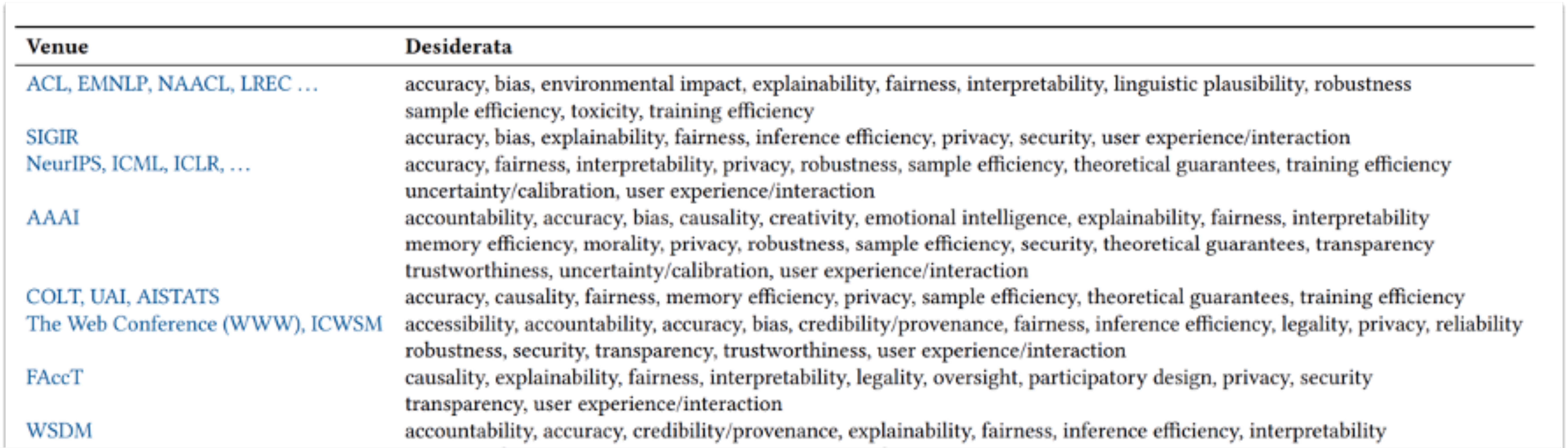
Scenarios

	J1-Jumbo	J1-Grande	J1-Large	Anthropic-LM	BLOOM	T0pp	Cohere XL	Cohere Large	Cohere Medium	Cohere Small	GPT-NeoX	GPT-J	T5	UL2	OPT (175B)	OPT (66B)	TNLGv2 (530B)	TNLGv2 (7B)	GPT-3 davinci	GPT-3 curie	GPT-3 babbage	GPT-3 ada	InstructGPT davinci v2	InstructGPT curie	InstructGPT babbage	InstructGPT ada	GLM	YaLM
NaturalQuestions (open)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NaturalQuestions (closed)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BoolQ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NarrativeQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QuAC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HellaSwag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
OpenBookQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TruthfulQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MMLU	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MS MARCO				✓	✓		✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
TREC				✓	✓		✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
XSUM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CNN/DM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IMDB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CivilComments	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RAFT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

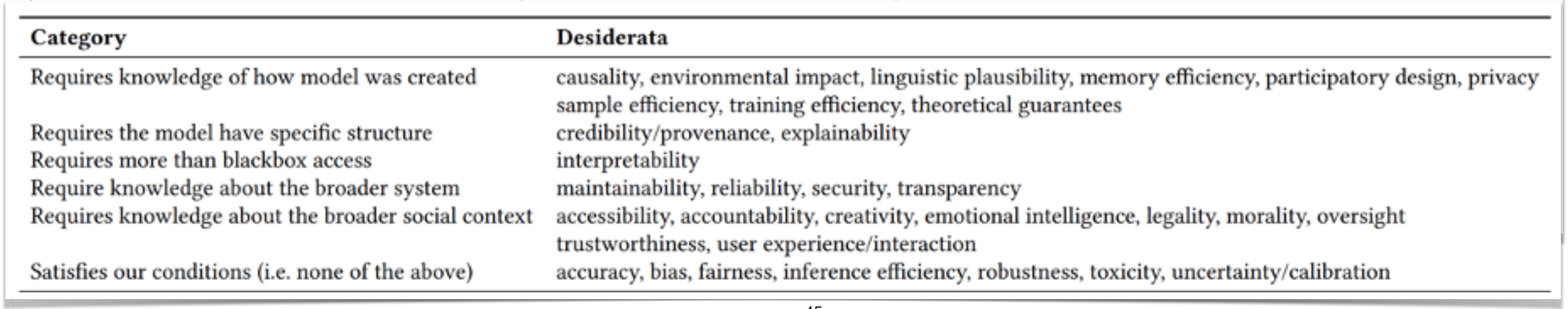
Benchmarking Considerations

- Adaptation (e.g. prompting, probing, fine-tuning)
- Fairness (some LMs might be specialized)
- Contamination (exposed to test data/distribution)
- Completeness (e.g. ChatGPT)

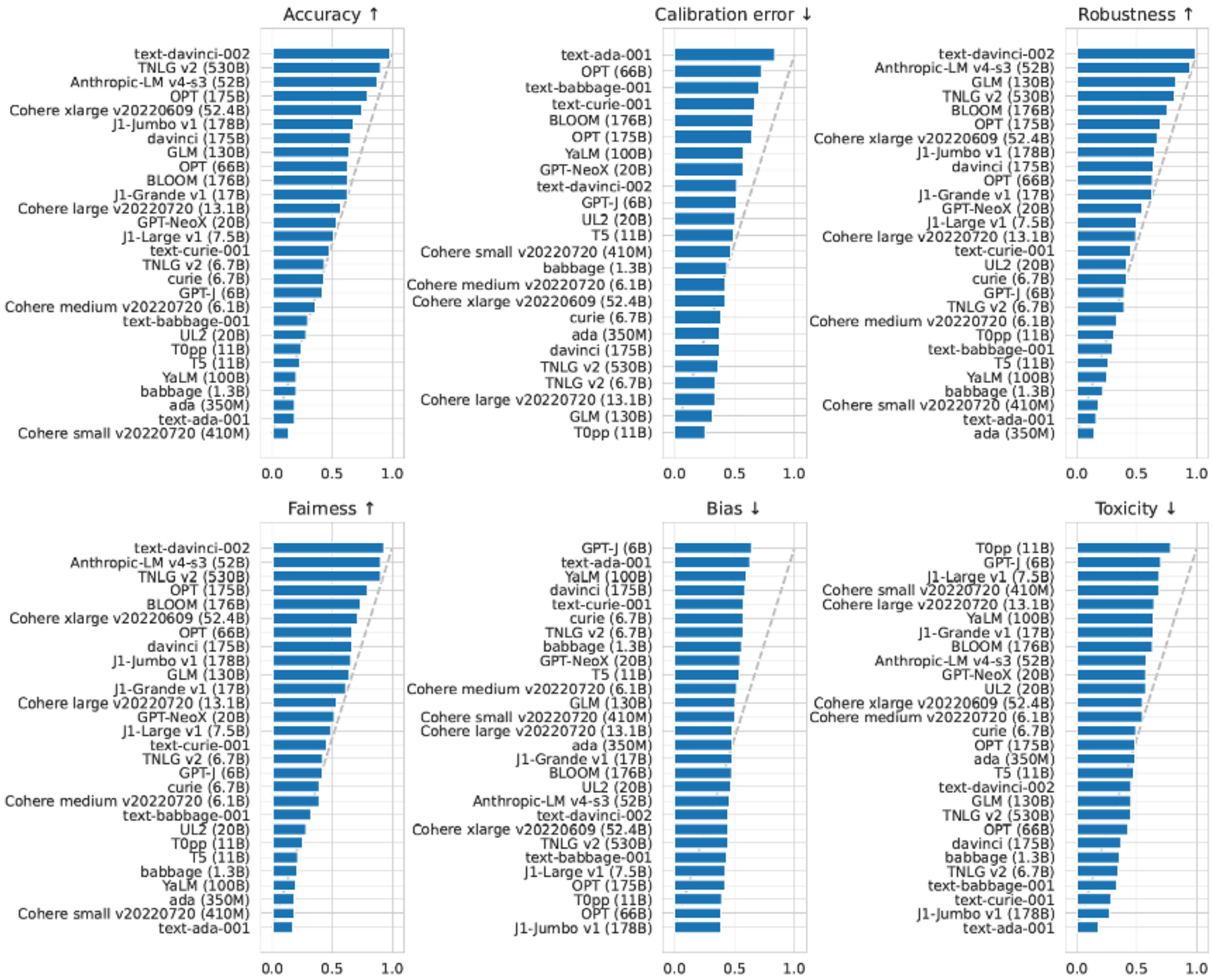
Desiderate/Metrics

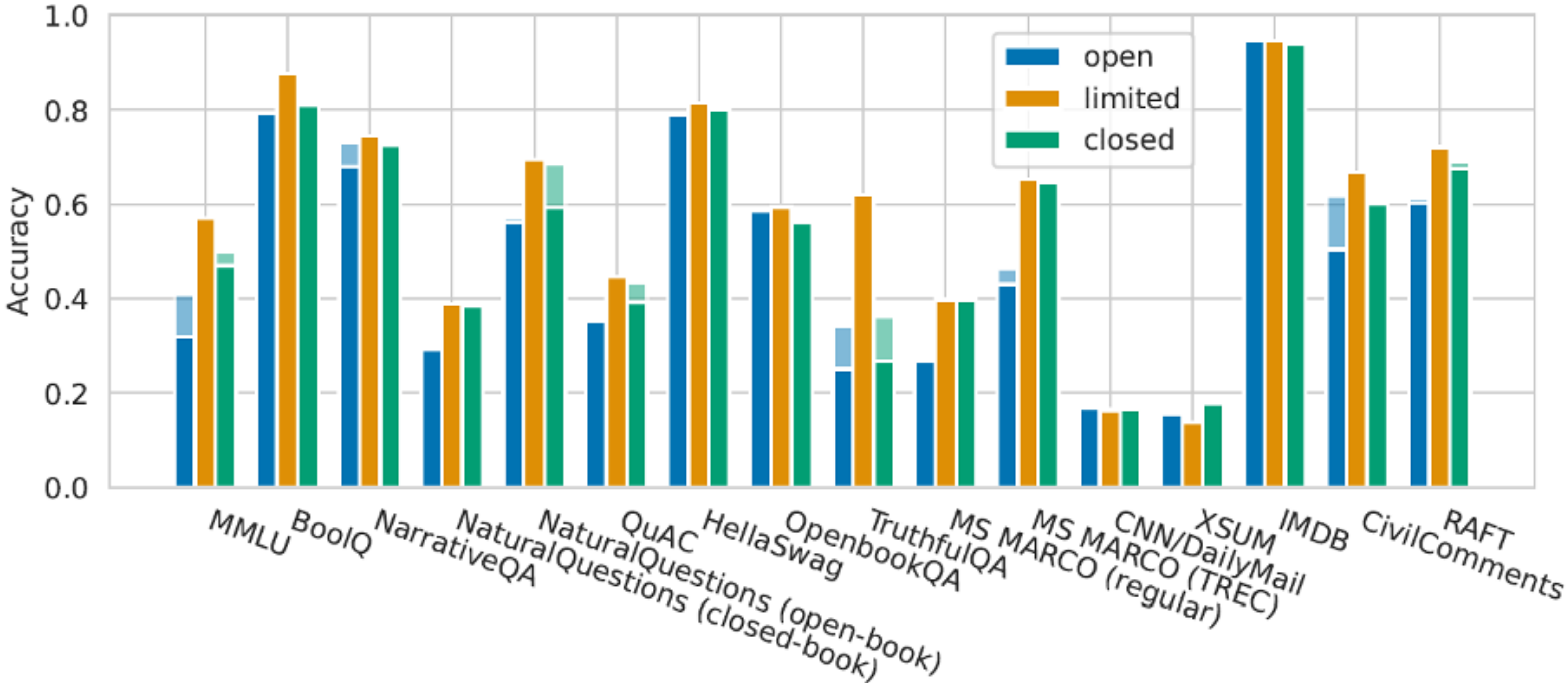


Venue	Desiderata
ACL, EMNLP, NAACL, LREC …	accuracy, bias, environmental impact, explainability, fairness, interpretability, linguistic plausibility, robustness sample efficiency, toxicity, training efficiency
SIGIR	accuracy, bias, explainability, fairness, inference efficiency, privacy, security, user experience/interaction
NeurIPS, ICML, ICLR, …	accuracy, fairness, interpretability, privacy, robustness, sample efficiency, theoretical guarantees, training efficiency uncertainty/calibration, user experience/interaction
AAAI	accountability, accuracy, bias, causality, creativity, emotional intelligence, explainability, fairness, interpretability memory efficiency, morality, privacy, robustness, sample efficiency, security, theoretical guarantees, transparency trustworthiness, uncertainty/calibration, user experience/interaction
COLT, UAI, AISTATS	accuracy, causality, fairness, memory efficiency, privacy, sample efficiency, theoretical guarantees, training efficiency
The Web Conference (WWW), ICWSM	accessibility, accountability, accuracy, bias, credibility/provenance, fairness, inference efficiency, legality, privacy, reliability robustness, security, transparency, trustworthiness, user experience/interaction
FAccT	causality, explainability, fairness, interpretability, legality, oversight, participatory design, privacy, security transparency, user experience/interaction
WSDM	accountability, accuracy, credibility/provenance, explainability, fairness, inference efficiency, interpretability



Category	Desiderata
Requires knowledge of how model was created	causality, environmental impact, linguistic plausibility, memory efficiency, participatory design, privacy sample efficiency, training efficiency, theoretical guarantees
Requires the model have specific structure	credibility/provenance, explainability
Requires more than blackbox access	interpretability
Require knowledge about the broader system	maintainability, reliability, security, transparency
Requires knowledge about the broader social context	accessibility, accountability, creativity, emotional intelligence, legality, morality, oversight trustworthiness, user experience/interaction
Satisfies our conditions (i.e. none of the above)	accuracy, bias, fairness, inference efficiency, robustness, toxicity, uncertainty/calibration


Accuracy ↑
text-davinci-002
TNLG v2 (530B)
Anthropic-LM v4-s3 (52B)
OPT (175B)
Cohere xlarge v20220609 (52.4B)
J1-Jumbo v1 (178B)
davinci (175B)
GLM (130B)
OPT (66B)
BLOOM (176B)
J1-Grande v1 (17B)
Cohere large v20220720 (13.1B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
text-curie-001
TNLG v2 (6.7B)
curie (6.7B)
GPT-J (6B)
Cohere medium v20220720 (6.1B)
text-babbage-001
UL2 (20B)
T0pp (11B)
T5 (11B)
YaLM (100B)
babbage (1.3B)
ada (350M)
text-ada-001
Cohere small v20220720 (410M)
0.0
0.5
1.0
Calibration error ↓
text-ada-001
OPT (66B)
text-babbage-001
text-curie-001
BLOOM (176B)
OPT (175B)
YaLM (100B)
GPT-NeoX (20B)
text-davinci-002
GPT-J (6B)
UL2 (20B)
T5 (11B)
Cohere small v20220720 (410M)
babbage (1.3B)
Cohere medium v20220720 (6.1B)
Cohere xlarge v20220609 (52.4B)
curie (6.7B)
ada (350M)
davinci (175B)
TNLG v2 (530B)
TNLG v2 (6.7B)
Cohere large v20220720 (13.1B)
GLM (130B)
T0pp (11B)
0.0
0.5
1.0
Robustness ↑
text-davinci-002
Anthropic-LM v4-s3 (52B)
GLM (130B)
TNLG v2 (530B)
BLOOM (176B)
OPT (175B)
Cohere xlarge v20220609 (52.4B)
J1-Jumbo v1 (178B)
davinci (175B)
OPT (66B)
J1-Grande v1 (17B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
Cohere large v20220720 (13.1B)
text-curie-001
UL2 (20B)
curie (6.7B)
GPT-J (6B)
TNLG v2 (6.7B)
Cohere medium v20220720 (6.1B)
T0pp (11B)
text-babbage-001
T5 (11B)
YaLM (100B)
babbage (1.3B)
Cohere small v20220720 (410M)
text-ada-001
ada (350M)
0.0
0.5
1.0
Fairness ↑
text-davinci-002
Anthropic-LM v4-s3 (52B)
TNLG v2 (530B)
OPT (175B)
BLOOM (176B)
Cohere xlarge v20220609 (52.4B)
OPT (66B)
davinci (175B)
J1-Jumbo v1 (178B)
GLM (130B)
J1-Grande v1 (17B)
Cohere large v20220720 (13.1B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
text-curie-001
TNLG v2 (6.7B)
GPT-J (6B)
curie (6.7B)
Cohere medium v20220720 (6.1B)
text-babbage-001
UL2 (20B)
T0pp (11B)
T5 (11B)
babbage (1.3B)
YaLM (100B)
ada (350M)
Cohere small v20220720 (410M)
text-ada-001
0.0
0.5
1.0
Bias ↓
GPT-J (6B)
text-ada-001
YaLM (100B)
davinci (175B)
text-curie-001
curie (6.7B)
TNLG v2 (6.7B)
babbage (1.3B)
GPT-NeoX (20B)
T5 (11B)
Cohere medium v20220720 (6.1B)
GLM (130B)
Cohere small v20220720 (410M)
Cohere large v20220720 (13.1B)
ada (350M)
J1-Grande v1 (17B)
BLOOM (176B)
UL2 (20B)
Anthropic-LM v4-s3 (52B)
text-davinci-002
Cohere xlarge v20220609 (52.4B)
TNLG v2 (530B)
text-babbage-001
J1-Large v1 (7.5B)
OPT (175B)
T0pp (11B)
OPT (66B)
J1-Jumbo v1 (178B)
0.0
0.5
1.0
Toxicity ↓
T0pp (11B)
GPT-J (6B)
J1-Large v1 (7.5B)
Cohere small v20220720 (410M)
Cohere large v20220720 (13.1B)
YaLM (100B)
J1-Grande v1 (17B)
BLOOM (176B)
Anthropic-LM v4-s3 (52B)
GPT-NeoX (20B)
UL2 (20B)
Cohere xlarge v20220609 (52.4B)
Cohere medium v20220720 (6.1B)
curie (6.7B)
OPT (175B)
ada (350M)
T5 (11B)
text-davinci-002
GLM (130B)
TNLG v2 (530B)
OPT (66B)
davinci (175B)
babbage (1.3B)
TNLG v2 (6.7B)
text-babbage-001
text-curie-001
J1-Jumbo v1 (178B)
text-ada-001
0.0
0.5
1.0


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
open
limited
closed
MMLU
BoolQ
NarrativeQA
NaturalQuestions (closed-book)
NaturalQuestions (open-book)
QuAC
HellaSwag
OpenbookQA
TruthfulQA
MS MARCO (regular)
MS MARCO (TREC)
CNN/DailyMail
XSUM
IMDB
CivilComments
RAFT

Benchmarking and Evaluation Metrics: Recommendation

- Move away from using a single metric for performance evaluation.
- Evaluate social bias and efficiency.
- Perform a fine-grained evaluation of models.
- Consider how to aggregate multiple metrics.

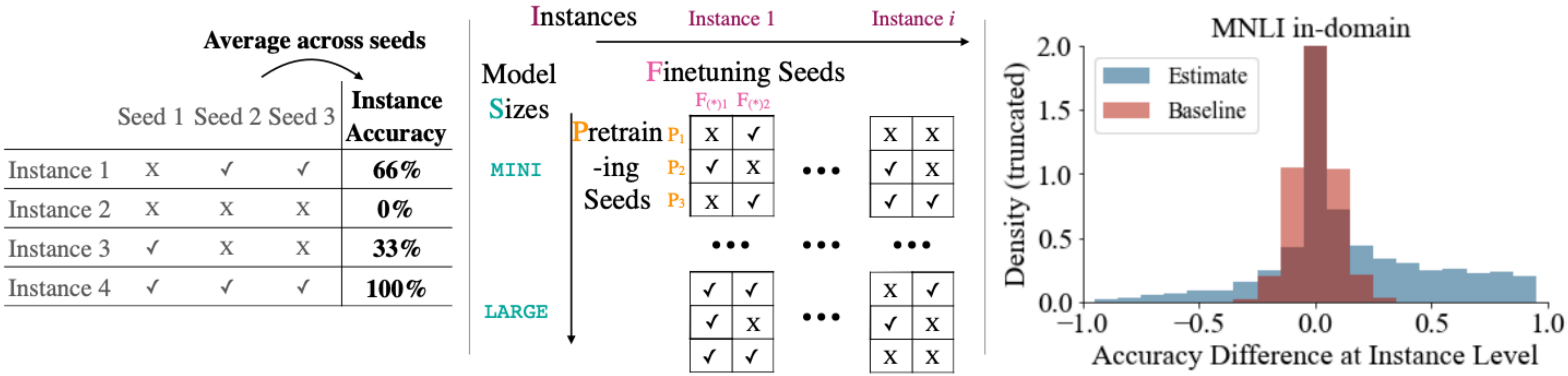
The long tail / worst case of benchmarking

Shift our attention to the tail of the distribution

Care more about the worst case and subsets of our data where our models perform the worst

Identify the best systems with few examples

The long tail / worst case of benchmarking

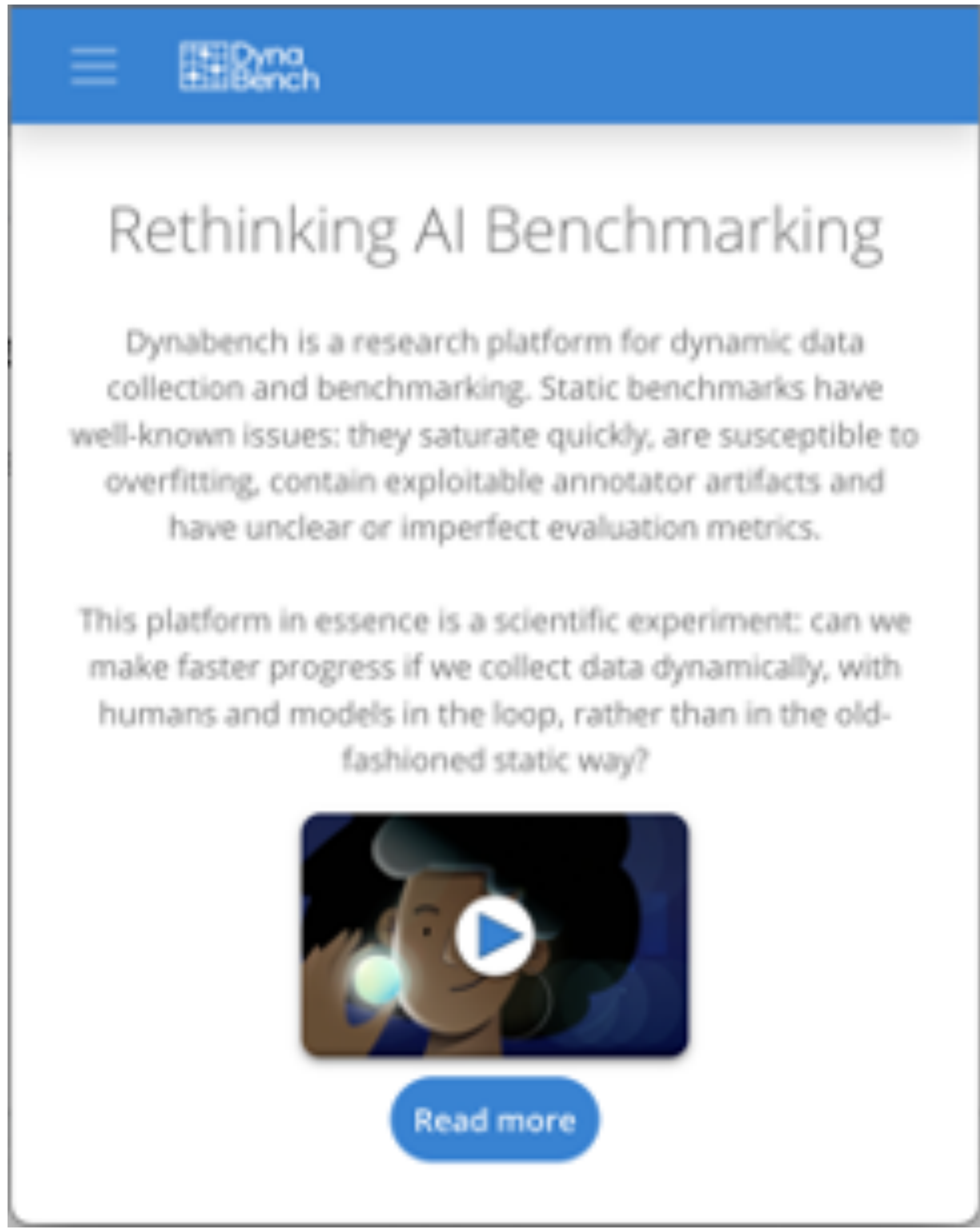


	Seed 1	Seed 2	Seed 3	Instance Accuracy
Instance 1	x	✓	✓	**66%**
Instance 2	x	x	x	**0%**
Instance 3	✓	x	x	**33%**
Instance 4	✓	✓	✓	**100%**

Figure 1: **Left**: Each column represents the same architecture trained with a different seed. We calculate accuracy for each instance (row) by averaging across seeds (column), while it is usually calculated for each model by averaging across instances. **Middle**: A visual layout of the model predictions we obtain, which is a binary-valued tensor with 4 axes: model size s, instance i, pretraining seeds P and finetuning seeds F. **Right**: for each instance, we calculate the accuracy gain from MINI to LARGE and plot the histogram in blue, along with a random baseline in red. Since the blue distribution has a bigger left tail, smaller models are better at some instances.

Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level https://arxiv.org/abs/2105.06020

Dynamic Benchmarking



Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

Dynabench (dynabench.org) is..

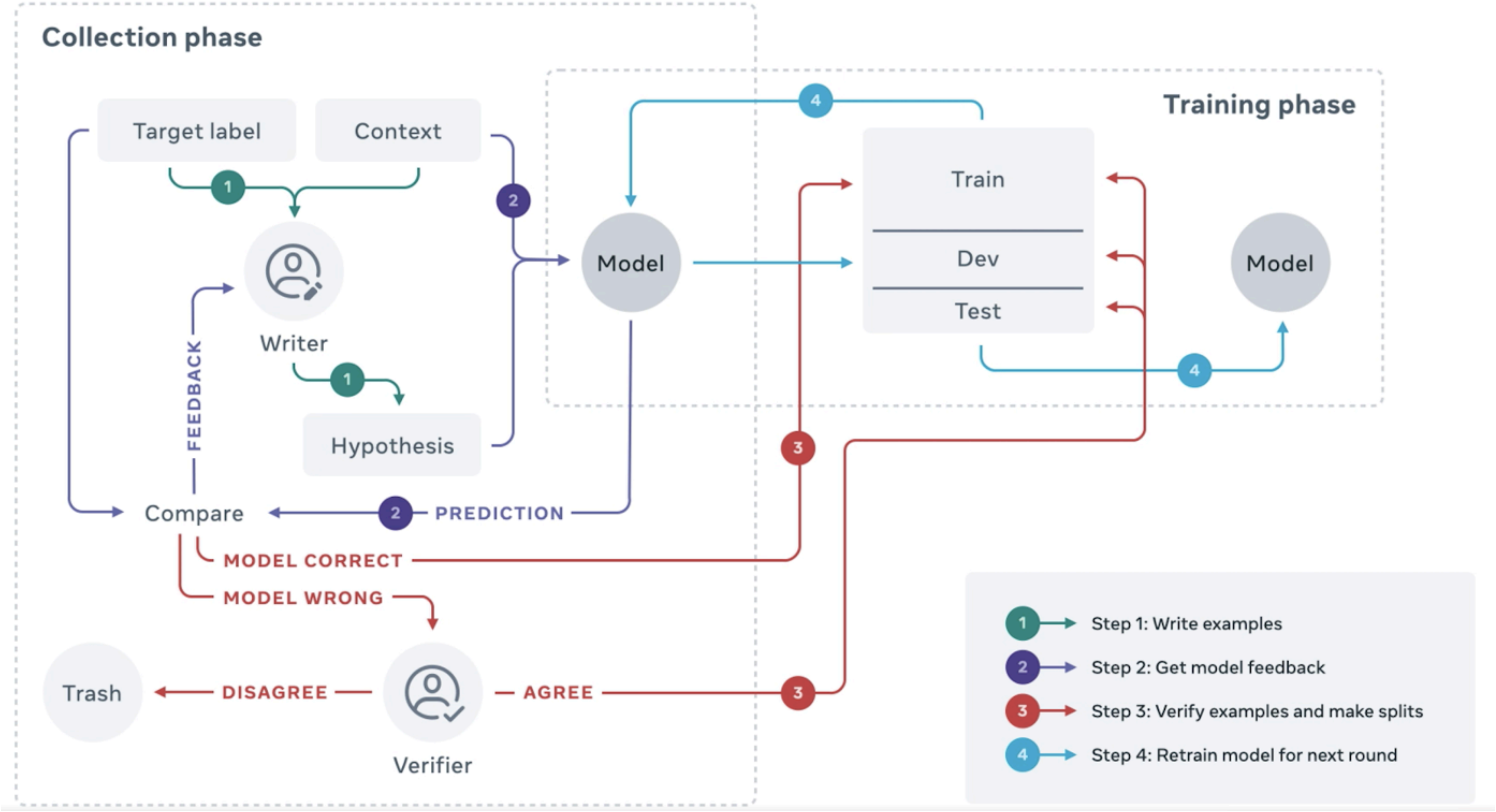
A research platform.

A community-based scientific experiment.

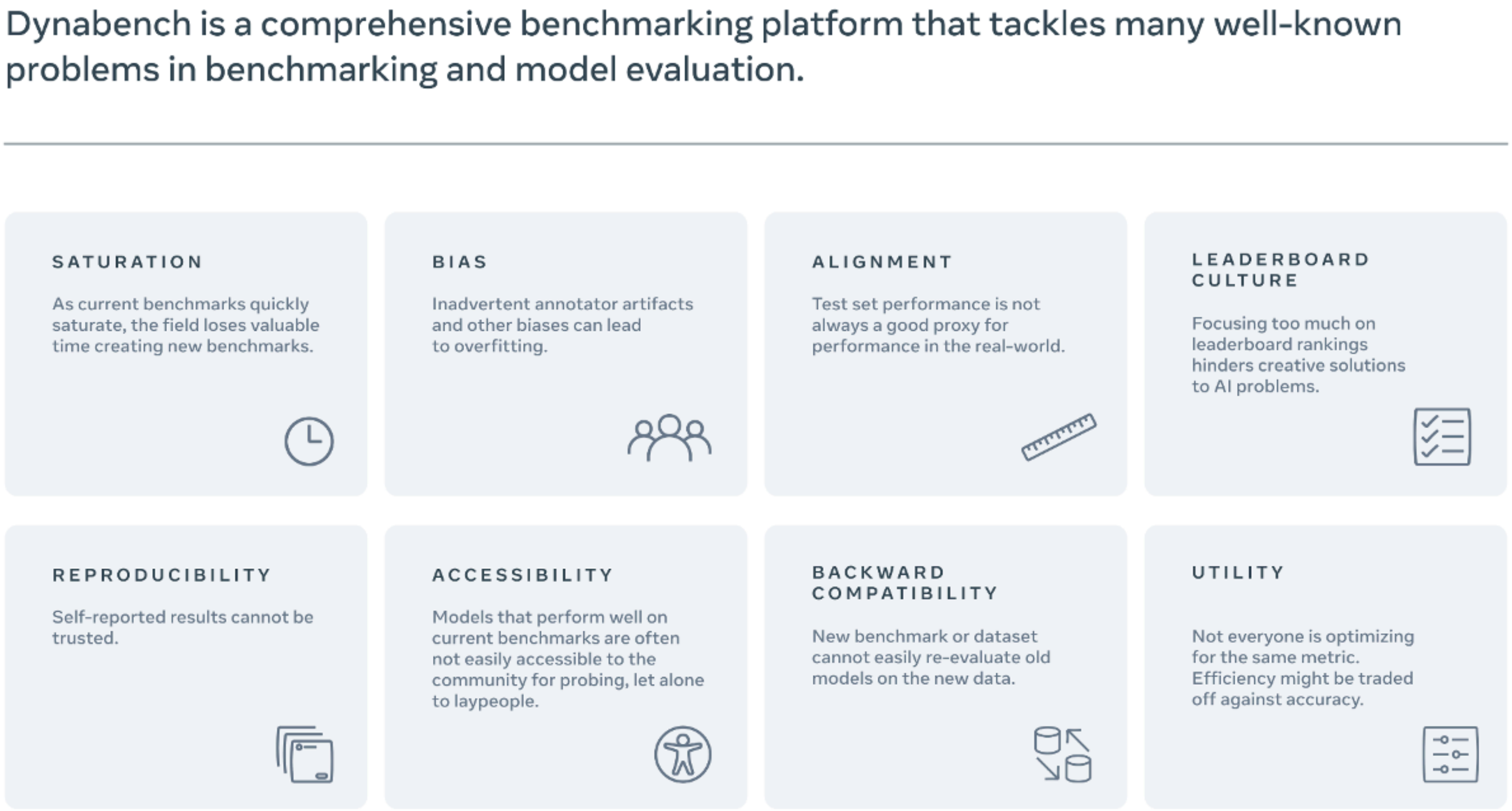
An effort to challenge current benchmarking dogma and help push the boundaries of AI research.

As the name says,

Dynamic adversarial data collection (ANLI; Nie et al. 2019)



Dynabench Goals



Dynabench is a comprehensive benchmarking platform that tackles many well-known problems in benchmarking and model evaluation.

SATURATION

As current benchmarks quickly saturate, the field loses valuable time creating new benchmarks.

BIAS

Inadvertent annotator artifacts and other biases can lead to overfitting.

ALIGNMENT

Test set performance is not always a good proxy for performance in the real-world.

LEADERBOARD CULTURE

Focusing too much on leaderboard rankings hinders creative solutions to AI problems.

REPRODUCIBILITY

Self-reported results cannot be trusted.

ACCESSIBILITY

Models that perform well on current benchmarks are often not easily accessible to the community for probing, let alone to laypeople.

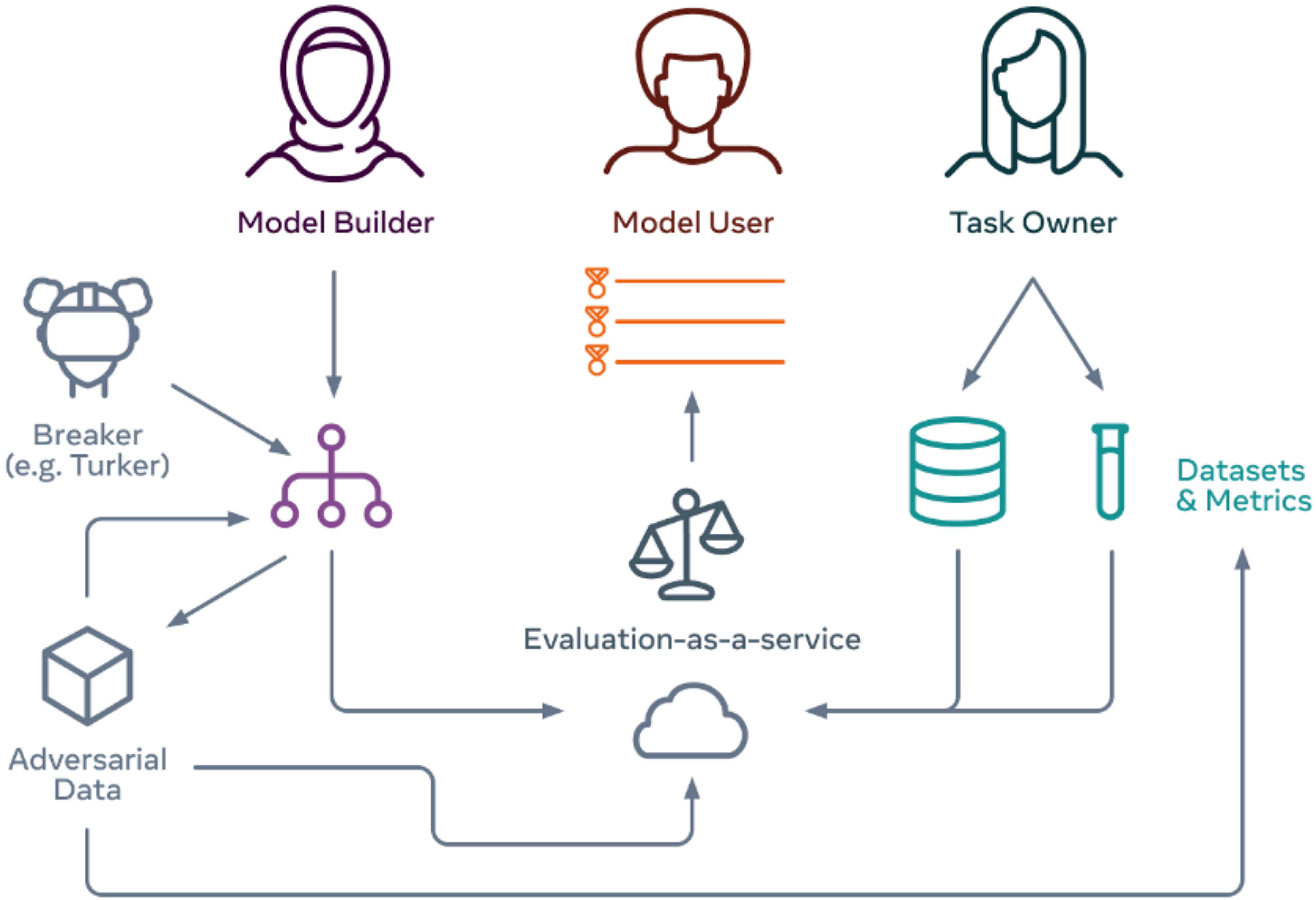
BACKWARD COMPATIBILITY

New benchmark or dataset cannot easily re-evaluate old models on the new data.

UTILITY

Not everyone is optimizing for the same metric. Efficiency might be traded off against accuracy.

Dynabench Roles



Large-scale continuous evaluation

"When a measure becomes a target, it ceases to be a good measure." –Goodhart's law

GEM (Gehrmann et al., 2021), which explicitly aims to be a 'living' benchmark, generally include around 10–15 different tasks.

BIG-Bench, a recent collaborative benchmark for language model probing

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

Generative AI Agents

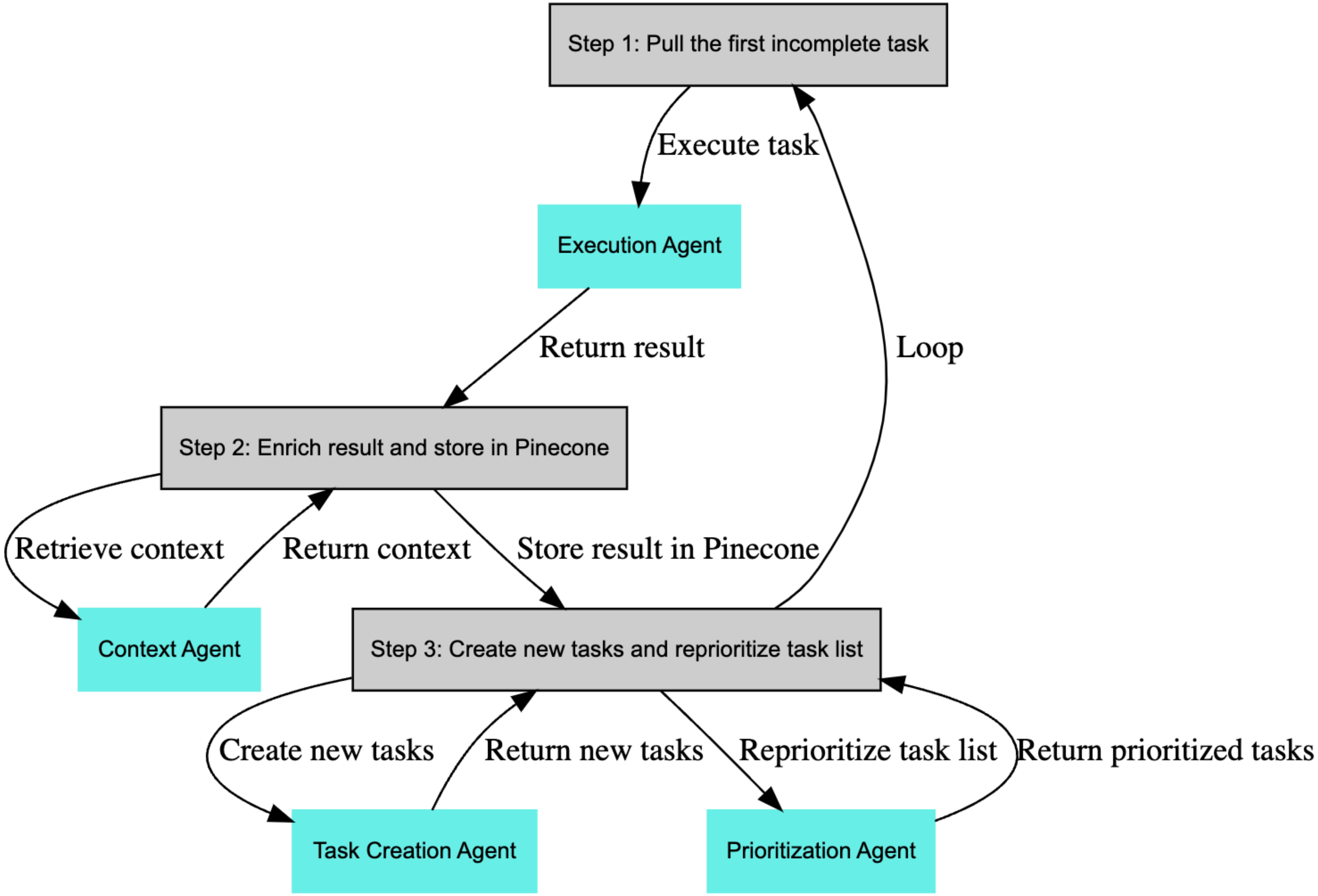
[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?

[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?

https://arxiv.org/pdf/2304.03442.pdf

BabyAGI

https://github.com/yoheinakajima/babyagi



Step 1: Pull the first incomplete task

Execute task

Execution Agent

Return result

Loop

Step 2: Enrich result and store in Pinecone

Retrieve context

Return context

Store result in Pinecone

Context Agent

Step 3: Create new tasks and reprioritize task list

Create new tasks

Return new tasks

Reprioritize task list

Return prioritized tasks

Task Creation Agent

Prioritization Agent

If benchmark is helping us reach a goal, what is that goal today?