

Rethinking Benchmarking in AI

Evaluation-as-a-Service & Dynamic Adversarial Data Collection

AKBC 2022

Douwe Kiela

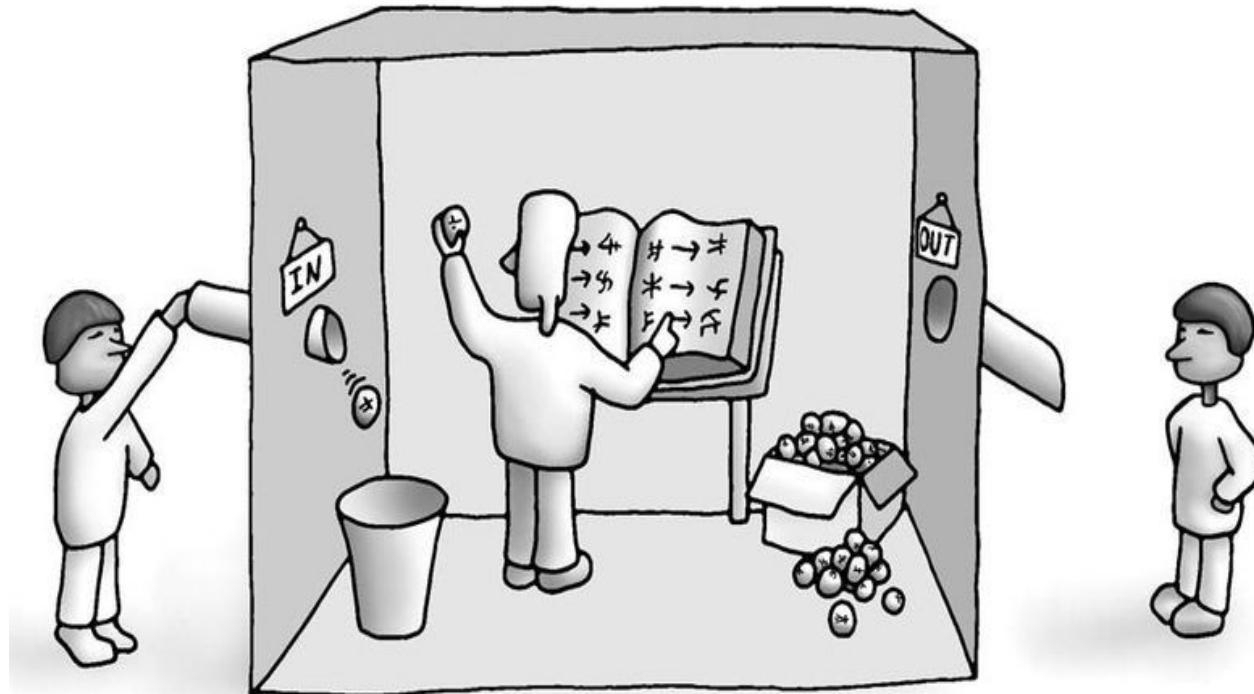


Hugging Face



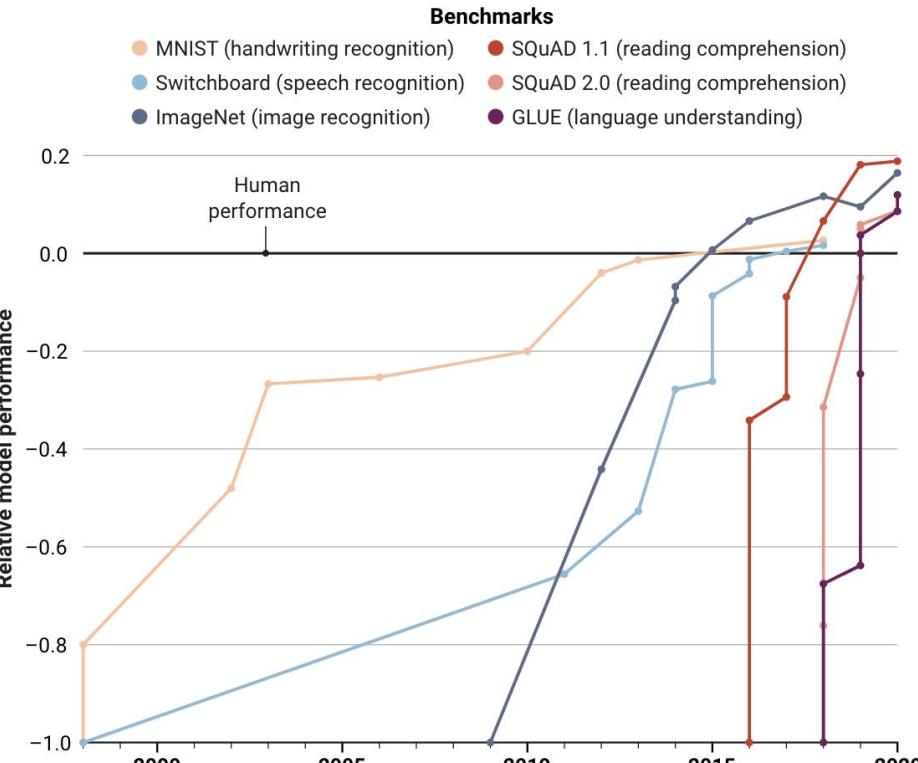
Stanford
University

Meaning in machines



Source: WikiCommons

Measuring progress



(GRAPHIC) K. FRANKLIN/SCIENCE; (DATA) D. KIELA ET AL., DYNABENCH: RETHINKING BENCHMARKING IN NLP. DOI:10.48550/ARXIV.2104.14337

TECHNOLOGY

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.



By Natasha Tiku

June 11, 2022 at 8:00 a.m. EDT



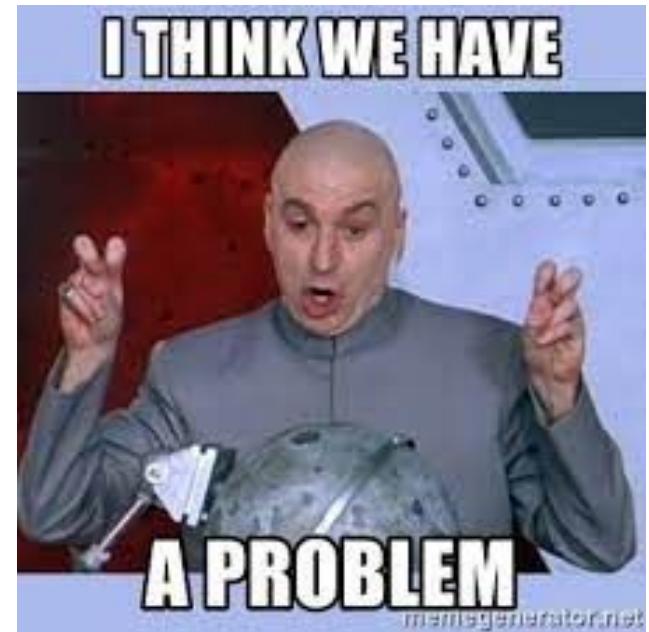
Google engineer Blake Lemoine. (Martin Klimek for The Washington Post)

Are we there yet?

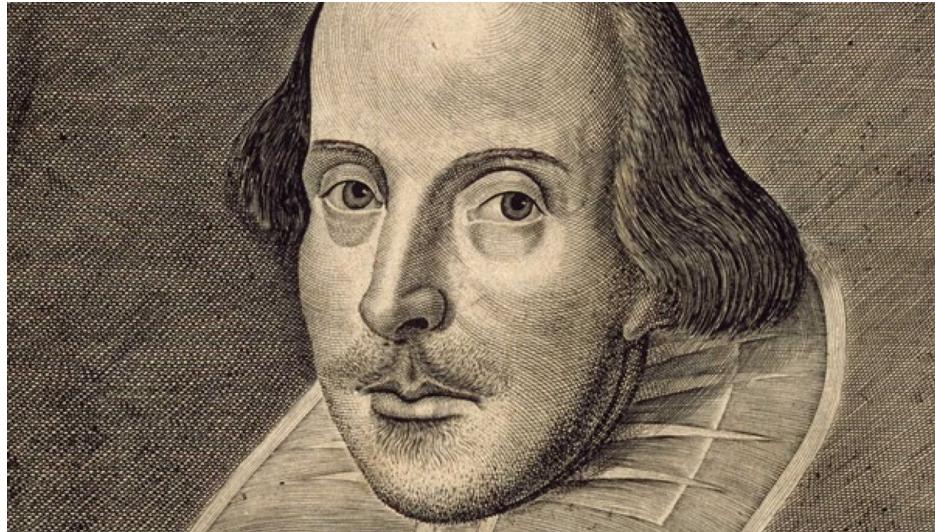
Progress on benchmarks has been remarkable.

Either:

- a) We are done; or
- b) We have a measurement problem.



There is something rotten in the state of the art



Adversarial NLI: A New Benchmark for Natural Language Understanding

[Yixin Nie](#), [Adina Williams](#), [Emily Dinan](#), [Mohit Bansal](#), [Jason Weston](#), [Douwe Kiela](#)

We introduce a new large-scale NLI benchmark dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure. We show that training models on this new dataset leads to state-of-the-art performance on a variety of popular NLI benchmarks, while posing a more difficult challenge with its new test set. Our analysis sheds light on the shortcomings of current state-of-the-art models, and shows that non-expert annotators are successful at finding their weaknesses. The data collection method can be applied in a never-ending learning scenario, becoming a moving target for NLU, rather than a static benchmark that will quickly saturate.

Outline

- **Problems in AI evaluation**
- Solutions:
 - Evaluate and Evaluation on the Hub
 - Dynabench: A Platform for Rethinking Benchmarking in AI

An indictment: Problems in AI evaluation



Saturation

- AI systems “outperform humans” on benchmarks, but we know that that’s not really true in the real world.

Forbes

Are AI Systems About To Outperform Humans?

What! Machines are outperforming humans on reading comprehension

Guest

6 areas where artificial neural networks outperform humans

Roman Steinberg, uKit AI

December 8, 2017 4:10 PM



AI will be able to beat us at everything by 2060, say experts

Biases & Artifacts

- Datasets contain many inadvertent biases and annotator artifacts. Neural networks are especially good at picking up on those.

Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets

Mor Geva, Yoav Goldberg, Jonathan Berant

Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, Noah A. Smith

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy, Ellie Pavlick, Tal Linzen

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh

Universal Adversarial Triggers for Attacking and Analyzing NLP

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, Sameer Singh

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

JOYAB@MIT.EDU

Timnit Gebru
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

TIMNIT.GEBRU@MICROSOFT.COM

ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases

Pierre Stock, Moustapha Cisse

Does Object Recognition Work for Everyone?

Terrance DeVries* Ishan Misra* Changhan Wang* Laurens van der Maaten
Facebook AI Research

Reproducibility and backward compatibility

- Self-reported results cannot be trusted. Small implementational differences, even in the evaluation pipeline, may lead to very different results.

WIRED

GREGORY BARBER

BUSINESS SEP 16, 2019 7:00 AM

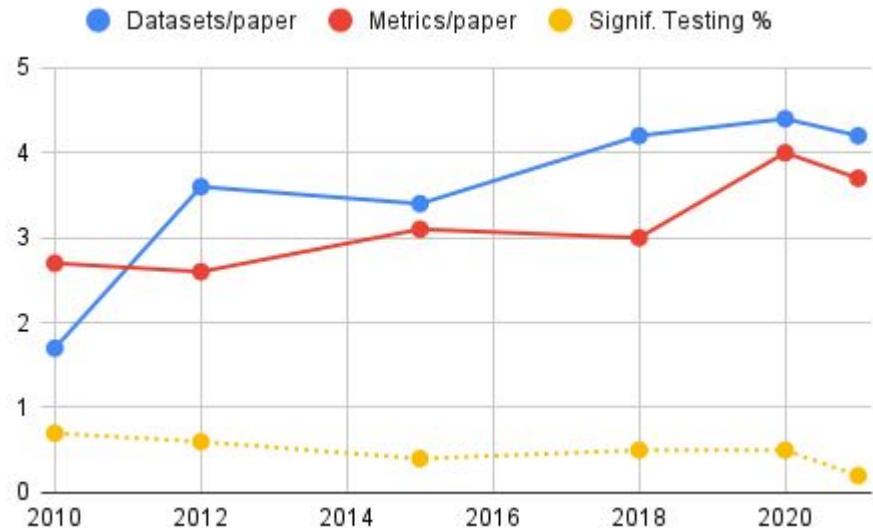
Artificial Intelligence Confronts a 'Reproducibility' Crisis

Machine-learning systems are black boxes even to the researchers that build them. That makes it hard for others to assess the results.

- Old models are not easily evaluated on new datasets and vice versa.

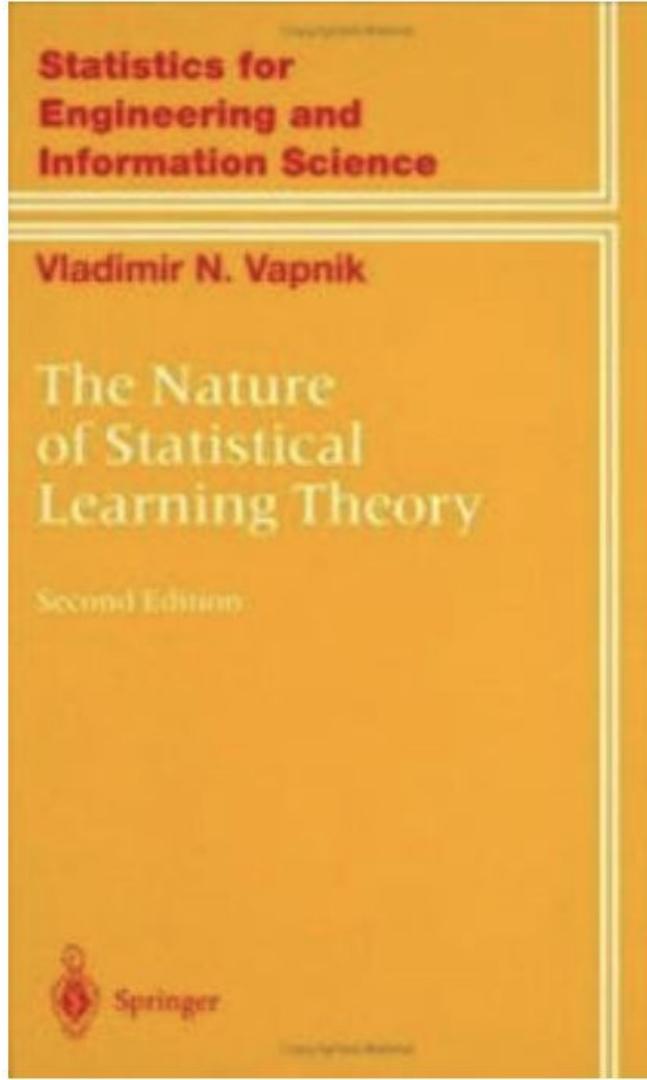
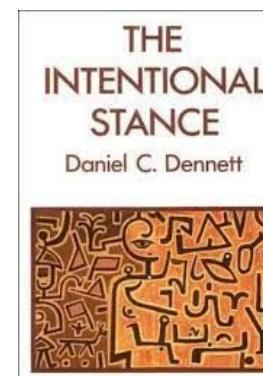
Ease of use and lack of best practices

- Proper system evaluation and comparison, at the scale of many models and many datasets, is unnecessarily cumbersome.
- Best practices are not well-established, despite massive growth.



Implicit assumptions

- Classic assumption:
 - We have a train set and an unseen test set
 - They are independent and identically distributed
- Modern day AI systems however are:
 - Pretrained on large data from different distribution
 - Prompted to “elicit” a certain behavior
- Lay-people interacting with AI will assume:
 - If it speaks language, it must be capable of strong generalization and have intentionality



Single-metric focus and leaderboard culture

- The community overfits on leaderboard performance
- Leaderboard metrics are based only on accuracy



Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9
2	Microsoft Alexander v-team	Turing NLP v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9
3	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6
4	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9
5	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2
6	DeBERTa Team - Microsoft	DeBERTa / TuringNLVR4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5
7	HFLFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5
+ 8	PING-AN Omni-Sinic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5
9	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5

Utility is in the Eye of the User: A Critique of NLP Leaderboards

Kawin Ethayarajh, Dan Jurafsky

Evaluation Examples are not Equally Informative: How should that change NLP Leaderboards?

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, Jordan Boyd-Graber

Alignment and measuring the right thing

- What we really care about: can this AI system successfully interact with humans for the task it is designed to do, in a way that it is aligned with their expectations and acts in their best interest.
- “Helpful, harmless, honest” (HHH)

A General Language Assistant as a Laboratory for Alignment

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Jared Kaplan

- That is not what we are currently measuring.

Outline

- Problems in AI evaluation
- Solutions:
 - Evaluate and Evaluation on the Hub
 - Dynabench: A Platform for Rethinking Benchmarking in AI



Evaluate & Evaluation on the Hub: Better Best Practices for Data and Model Measurements

**Leandro von Werra,* Lewis Tunstall,* Abhishek Thakur,* Sasha Luccioni,*
Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani,
Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško,
Albert Villanova, Quentin Lhoest, Julien Chaumond,
Margaret Mitchell, Alexander M. Rush, Thomas Wolf, Douwe Kiela
Hugging Face, Inc.**

Hugging Face

We are on a mission to democratize good machine learning, one commit at a time.



Hugging Face

[huggingface / transformers](#) Public

Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.

[huggingface.co/transformers](#)

Apache-2.0 license

70.6k stars 16.2k forks

[huggingface / datasets](#) Public

The largest hub of ready-to-use datasets for ML models with fast, easy-to-use and efficient data manipulation tools

[huggingface.co/docs/datasets](#)

Apache-2.0 license

14.1k stars 1.8k forks

Evaluate

- We distinguish between:
 - Metrics (e.g. “accuracy”)
 - Measurements (e.g. “length”)
 - Comparisons (e.g. “mcnemar”)
- Open source, standardized canonical implementations.



```
!pip install evaluate  
import evaluate
```



```
# General metrics  
evaluate.load("accuracy")
```



```
# Computer vision  
evaluate.load("mean_iou")
```



```
# NLP  
evaluate.load("bleu")
```



```
# Audio  
evaluate.load("wer")
```



```
# Information retrieval  
evaluate.load("trec_eval")
```



```
# Reinforcement learning  
evaluate.load("rl_reliability")
```

Metric/measurement/comparison cards

- Like “model cards” and “data sheets”, but for evaluation.
- Proper documentation is hugely important.
- Interactive widgets for easy intuition.

The screenshot shows a metric card for accuracy. At the top, there's a navigation bar with a space name, a status indicator (Running), and links for App, Files and versions, and Community. Below the navigation is the title "Metric: accuracy". A descriptive text explains that accuracy is the proportion of correct predictions among the total number of cases processed, with the formula $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$. It also defines TP (True positive), TN (True negative), FP (False positive), and FN (False negative). To the right of the text is a code snippet showing how to compute accuracy using Python's evaluate library. At the bottom, there's a table with two rows for "predictions" and "references", and a "Submit" button.

predictions	references
0	0
1	0

accuracy
{'accuracy': 0.5}

>>> accuracy_metric = evaluate.load("accuracy")
>>> results = accuracy_metric.compute(references=[0, 1], predictions=[0, 1])
>>> print(results)
{'accuracy': 1.0}

Clear Submit

Metric Card for Accuracy

Metric Description

Accuracy is the proportion of correct predictions among the total number of cases processed. It can be computed with: $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$ Where: TP: True positive TN: True negative FP: False positive FN: False negative

How to Use

At minimum, this metric requires predictions and references as inputs.

```
>>> accuracy_metric = evaluate.load("accuracy")
>>> results = accuracy_metric.compute(references=[0, 1], predictions=[0, 1])
>>> print(results)
{'accuracy': 1.0}
```

Limitations and Bias

This metric can be easily misleading, especially in the case of unbalanced classes. For example, a high accuracy might be because a model is doing well, but if the data is unbalanced, it might also be because the model is only accurately labeling the high-frequency class. In such cases, a more detailed analysis of the model's behavior, or the use of a different metric entirely, is necessary to determine how well the model is actually performing.

Sharing on the Hub

- Your own evaluation modules can easily be pushed to the HF Hub.
- Just like model papers open source models, and dataset papers open source datasets, evaluation papers can open source their methods for community usage.

```
evaluate-cli create "My Metric" module_type="metric"  
  
my_metric = evaluate.load("lvwerra/my_metric")
```

Evaluator

- Model developers use a Trainer abstraction for training logic.
- We offer an Evaluator abstraction for evaluation logic.
- The Evaluator also supports computing bootstrap confidence intervals, as well as compute throughput/efficiency statistics.

```
eval.compute(model_or_pipeline=pipe, data=data, metric=metric,
             label_mapping={"NEGATIVE": 0, "POSITIVE": 1},
             strategy="bootstrap", n_resamples=200)

>>> {'accuracy':
...     {
...         'confidence_interval': (0.906, 0.9406749892841922),
...         'standard_error': 0.00865213251082787,
...         'score': 0.923
...     }
... }
```

Data measurements tool

<https://huggingface.co/spaces/huggingface/data-measurements-tool>

This demo showcases the [dataset measures as we develop them](#). Right now this has a few pre-loaded datasets for which you can:

- view some general statistics about the text vocabulary, lengths, labels
- explore some distributional statistics to assess properties of the language
- view some comparison statistics and overview of the text distribution

The tool is in development, and will keep growing in utility and functionality 😊🚧

Comparison mode

Show text clusters

Choose dataset and field

-

Choose dataset to explore:

hate_speech18

▼

Choose configuration:

Data Measurements Tool

Showing: hate_speech18 - default - train - text

Dataset Description

+

General Text Statistics

+

Label Distribution

+

Text Lengths

+

Text Duplicates

+

Word Association: nPMI

+

Vocabulary Distribution: Zipf's Law Fit

+

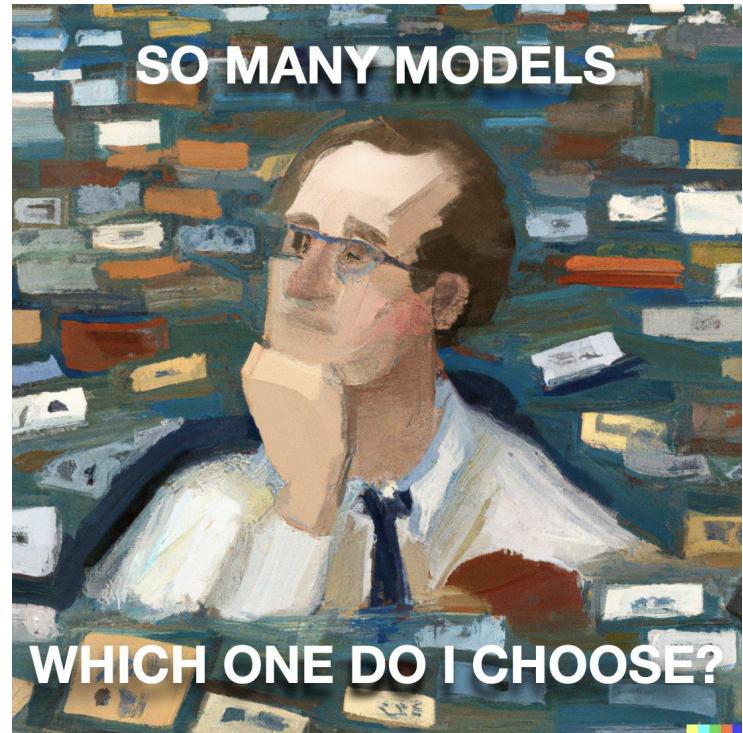
If we make it easy to follow best practices..

.. then people will follow them.

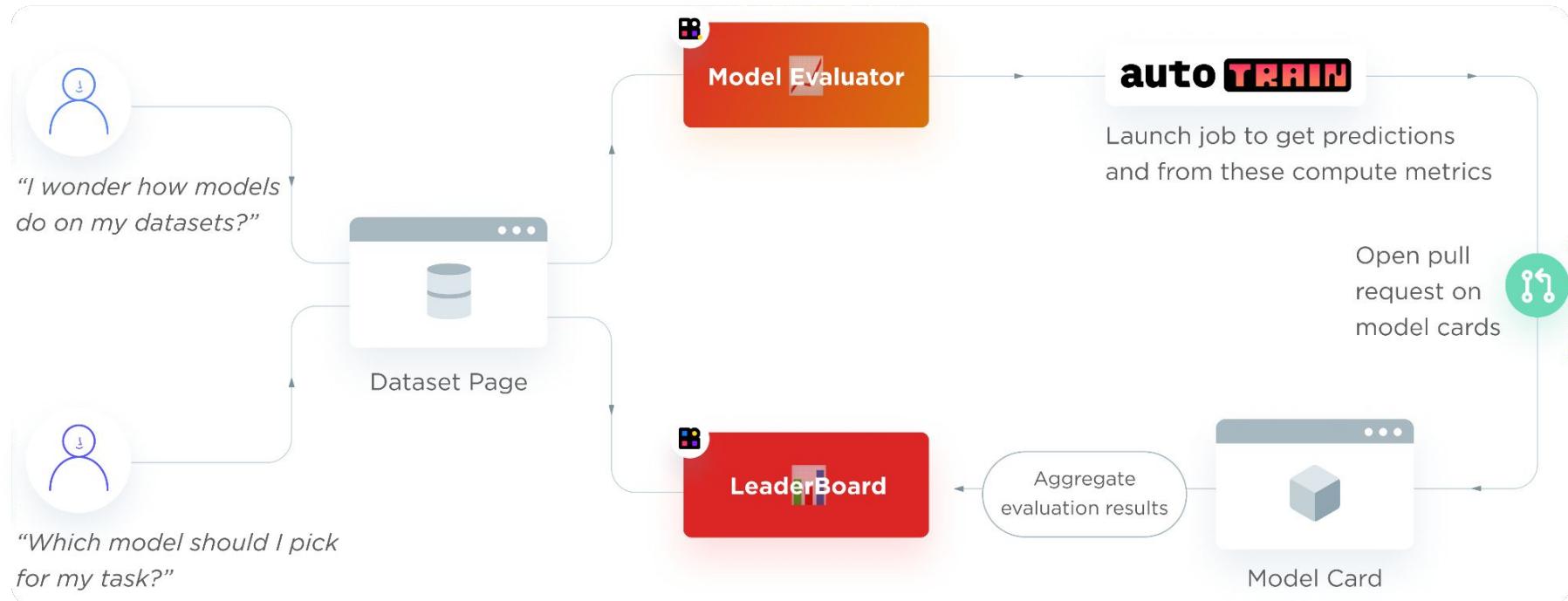
- How easy can we make it?

Evaluation on the Hub

- The Hugging Face Hub hosts models, datasets and evaluation methods.
- Can we automatically evaluate models on datasets using a given metric?
- Can we enable evaluation “at the click of a button”, **Evaluation-as-a-Service?**



How does it work?



Live demo

- Go to dataset page: https://huggingface.co/datasets/lewtun/dog_food
- Click on Evaluate model
- Trigger the evaluation job
- The result will appear on the dataset's HF Leaderboard

Free Evaluation of Very Large Language Models



Very Large Language Models
and how to evaluate them



without
code!



Outline

- Problems in AI evaluation
- Solutions:
 - Evaluate and Evaluation on the Hub
 - Dynabench: A Platform for Rethinking Benchmarking in AI



ML
Commons



UCL



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



The
Alan Turing
Institute

SFU

SIMON FRASER
UNIVERSITY

JOHNS HOPKINS
UNIVERSITY

Stanford
University

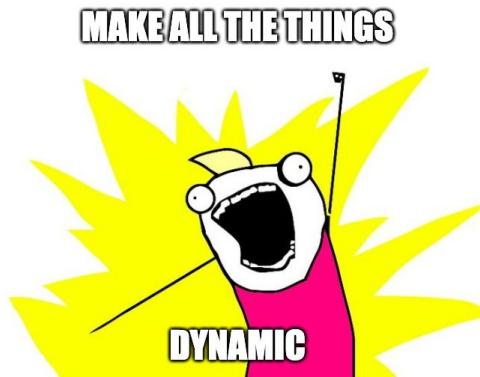
FACEBOOK AI

Rethinking benchmarking in AI

Dynabench (dynabench.org) is..

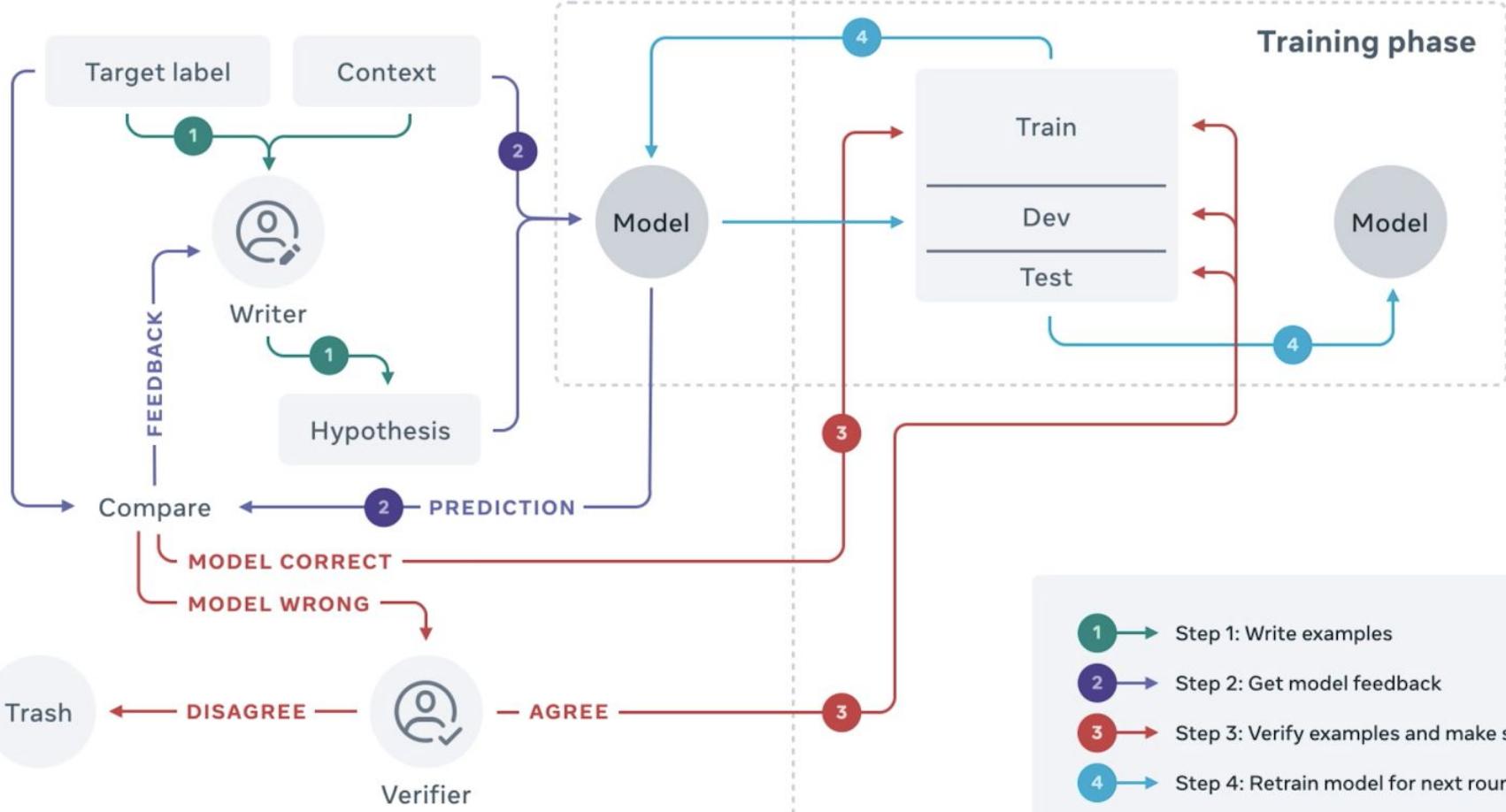
- A research platform.
- A community-based scientific experiment.
- An effort to challenge current benchmarking dogma and help push the boundaries of AI research.

As the name says,



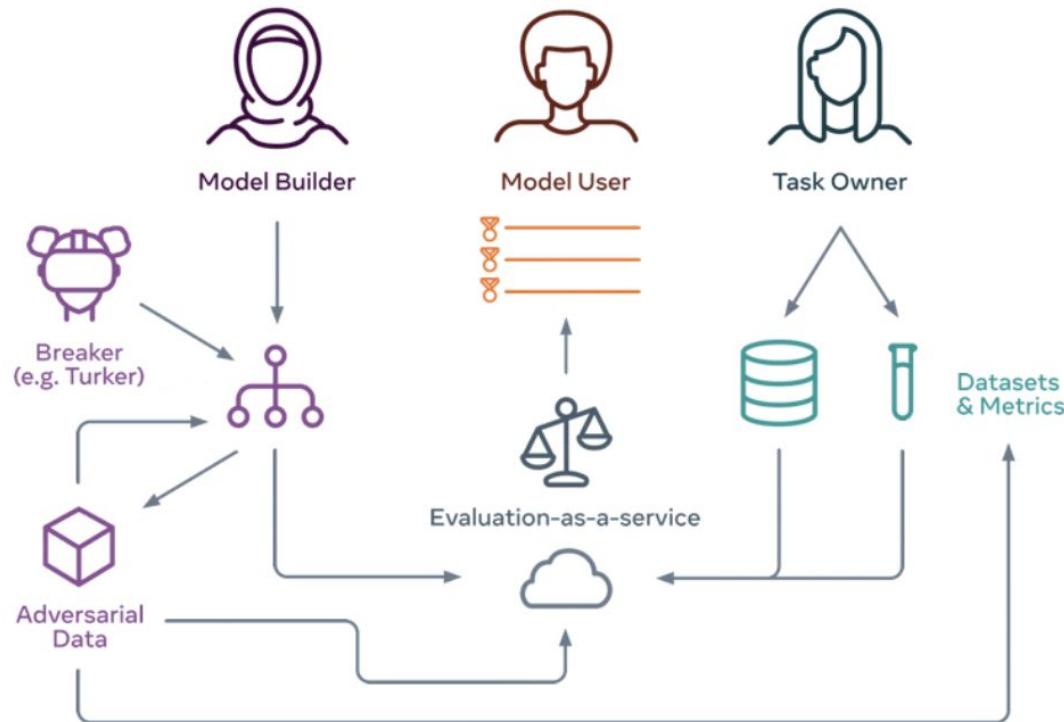
The screenshot shows a web page with a blue header containing the Dynabench logo and a menu icon. The main content area has a light gray background. At the top, the title "Rethinking AI Benchmarking" is displayed in a large, dark font. Below the title, a paragraph of text discusses the platform's purpose: "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." Further down, another paragraph states: "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?" At the bottom of the main content area is a blue button with a white play icon and the text "Read more".

Collection phase



- 1 → Step 1: Write examples
- 2 → Step 2: Get model feedback
- 3 → Step 3: Verify examples and make splits
- 4 → Step 4: Retrain model for next round

Dynabench roles



Live demo

<https://dynabench.org>

I was served rather the opposite of haute cuisine.

This restaurant was baad!

Broader research program

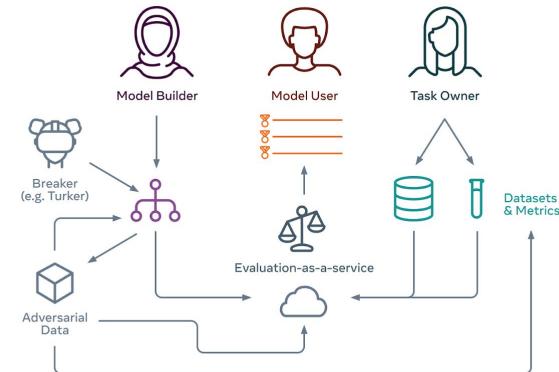
What happens when we put humans and models in loops?

Can we make faster progress? Can we make better measurements?

Can we have fewer biases and artifacts, and better robustness and alignment?

What are we still missing in our models?
What are the next challenges to solve?

How can we democratize model evaluation, help make research reproducible, learn from our mistakes as a community, and empower researchers?



Recent work out of the Dynabench team

- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP**
- Vidgen et al. (ACL21). **Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection**
- Potts et al. (ACL21). **DynaSent: A Dynamic Benchmark for Sentiment Analysis**
- Kirk et al. (2021). **Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate**
- Sheng & Singh et al. (NeurIPS21). **Human-Adversarial Visual Question Answering**
- Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
- Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
- Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**
- Thrush et al. (2022). **Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks**
- Bartolo et al. (EMNLP21). **Improving QA Model Robustness with Synthetic Adversarial Data Generation**
- Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
- Bartolo et al. (2022). **Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants**
- Wallace et al. (2022). **Analyzing Dynamic Adversarial Training Data in the Limit**

Recent work out of the Dynabench team

- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP** Dataset Papers
 - Vidgen et al. (ACL21). **Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection**
 - Potts et al. (ACL21). **DynaSent: A Dynamic Benchmark for Sentiment Analysis**
 - Kirk et al. (2021). **Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate**
 - Sheng & Singh et al. (NeurIPS21). **Human-Adversarial Visual Question Answering**
- Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
 - Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
 - Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**
 - Thrush et al. (2022). **Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks**
 - Bartolo et al. (EMNLP21). **Improving QA Model Robustness with Synthetic Adversarial Data Generation**
 - Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
 - Bartolo et al. (2022). **Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants**
 - Wallace et al. (2022). **Analyzing Dynamic Adversarial Training Data in the Limit**

Recent work out of the Dynabench team

- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP**
 - Vidgen et al. (ACL21). **Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection**
 - Potts et al. (ACL21). **DynaSent: A Dynamic Benchmark for Sentiment Analysis**
 - Kirk et al. (2021). **Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate**
 - Sheng & Singh et al. (NeurIPS21). **Human-Adversarial Visual Question Answering** Evaluation Papers
- Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
 - Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
 - Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**
 - Thrush et al. (2022). **Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks**
- Bartolo et al. (EMNLP21). **Improving QA Model Robustness with Synthetic Adversarial Data Generation**
 - Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
 - Bartolo et al. (2022). **Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants**
 - Wallace et al. (2022). **Analyzing Dynamic Adversarial Training Data in the Limit**

Recent work out of the Dynabench team

- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP**
- Vidgen et al. (ACL21). **Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection**
- Potts et al. (ACL21). **DynaSent: A Dynamic Benchmark for Sentiment Analysis**
- Kirk et al. (2021). **Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate**
- Sheng & Singh et al. (NeurIPS21). **Human-Adversarial Visual Question Answering**
- Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
- Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
- Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**
- Thrush et al. (2022). **Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks** [**Method Papers**](#)

- Bartolo et al. (EMNLP21). **Improving QA Model Robustness with Synthetic Adversarial Data Generation**
- Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
- Bartolo et al. (2022). **Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants**
- Wallace et al. (2022). **Analyzing Dynamic Adversarial Training Data in the Limit**

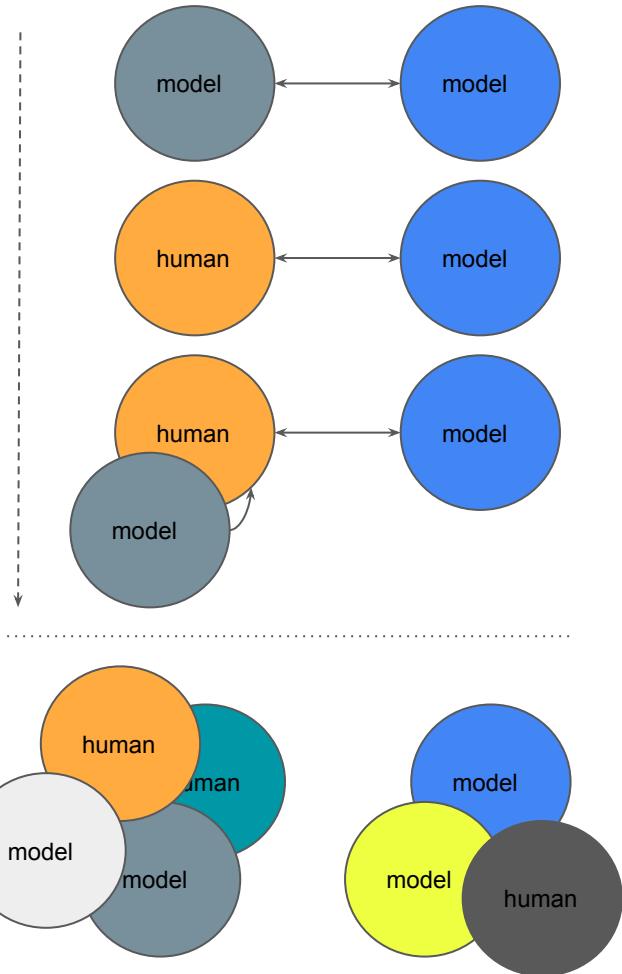
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP**
 - Vidgen et al. (ACL21). **Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection**
 - Potts et al. (ACL21). **DynaSent: A Dynamic Benchmark for Sentiment Analysis**
 - Kirk et al. (2021). **Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate**
 - Sheng & Singh et al. (NeurIPS21). **Human-Adversarial Visual Question Answering**
 - Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
 - Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
 - Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**
 - Thrush et al. (2022). **Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks**
- Bartolo et al. (EMNLP21). **Improving QA Model Robustness with Synthetic Adversarial Data Generation**
 - Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
 - Bartolo et al. (2022). **Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants**
 - Wallace et al. (2022). **Analyzing Dynamic Adversarial Training Data in the Limit**

Humans and models in loops

- Question 1:
 - Instead of human-adversarial, how much can we improve things by just being model-adversarial using human-adversarial data?
- Question 2:
 - Can generative (adversarial) models help humans fool discriminative models?

Work by **Max Bartolo et al.**

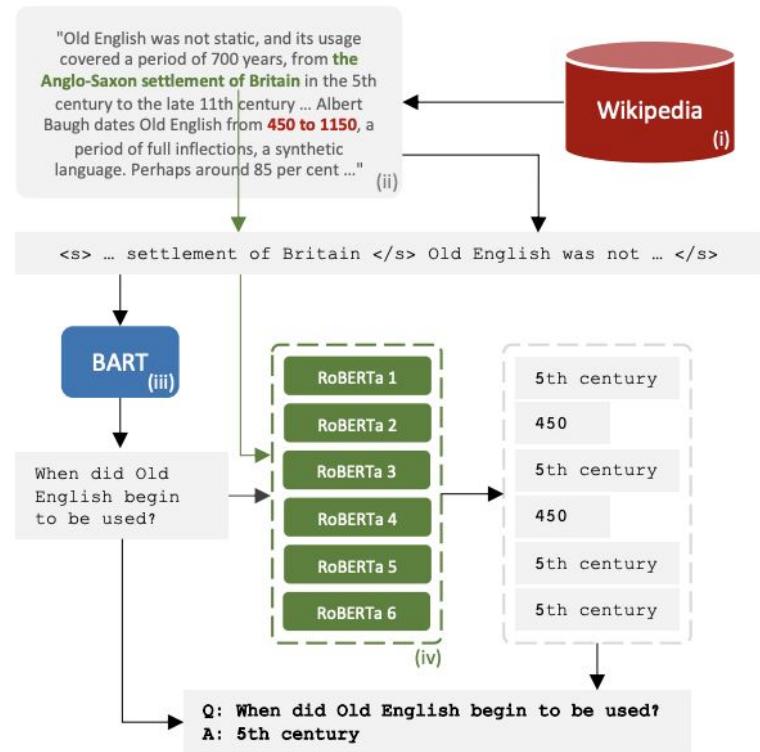


Improving QA robustness with synthetic adversarial data

- Pipeline:
 1. Passage selection
 2. Answer candidate selection
 3. Question generation
 4. Filtering and re-labeling
 5. Training a new QA model

Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation

Max Bartolo^{†*} Tristan Thrush[‡] Robin Jia[‡] Sebastian Riedel^{†‡}
Pontus Stenetorp[†] Douwe Kiela[‡]
[†]University College London [‡]Facebook AI Research



Findings

- Synthetic adversarial data derived from human-adversarial data **improves accuracy and robustness.**

Model	Training Data	\mathcal{D}_{BiDAF}		\mathcal{D}_{BERT}		$\mathcal{D}_{RoBERTa}$		mvMER*
		EM	F_1	EM	F_1	EM	F_1	
R _{SQuAD}	SQuAD	48.6 _{1.3}	64.2 _{1.5}	30.9 _{1.3}	43.3 _{1.7}	15.8 _{0.9}	26.4 _{1.3}	20.7%
R _{SQuAD+AQA}	↑ + AQA	59.6 _{0.5}	73.9 _{0.5}	54.8 _{0.7}	64.8 _{0.9}	41.7 _{0.6}	53.1 _{0.8}	17.6%
SynQA	↑ + SynQA _{SQuAD}	62.5 _{0.9}	76.0 _{1.0}	58.7 _{1.4}	68.3 _{1.4}	46.7 _{1.8}	58.0 _{1.8}	8.8%
SynQA _{Ext}	↑ + SynQA _{Ext}	62.7 _{0.6}	76.2 _{0.5}	59.0 _{0.7}	68.9 _{0.5}	46.8 _{0.5}	57.8 _{0.8}	12.3%

Model	MRQA in-domain													
	SQuAD		NewsQA		TriviaQA		SearchQA		HotpotQA		NQ		Avg	
	EM	F_1	EM	F_1										
R _{SQuAD}	84.1 _{1.3}	90.4 _{1.3}	41.0 _{1.2}	57.5 _{1.6}	60.2 _{0.7}	69.0 _{0.8}	16.0 _{1.8}	20.8 _{2.7}	53.6 _{0.8}	68.9 _{0.8}	40.5 _{2.7}	58.5 _{2.0}	49.2	60.9
R _{SQuAD+AQA}	84.4 _{1.0}	90.2 _{1.1}	41.7 _{1.6}	58.0 _{1.7}	62.7 _{0.4}	70.8 _{0.3}	20.6 _{2.9}	25.5 _{3.6}	56.3 _{1.1}	72.0 _{1.0}	54.4 _{0.5}	68.7 _{0.4}	53.3	64.2
SynQA	88.8 _{0.3}	94.3 _{0.2}	42.9 _{1.6}	60.0 _{1.4}	62.3 _{1.1}	70.2 _{1.1}	23.7 _{3.7}	29.5 _{4.4}	59.8 _{1.1}	75.3 _{1.0}	55.1 _{1.0}	68.7 _{0.8}	55.4	66.3
SynQA _{Ext}	89.0 _{0.3}	94.3 _{0.2}	46.2 _{0.9}	63.1 _{0.8}	58.1 _{1.8}	65.5 _{1.9}	28.7 _{3.2}	34.3 _{4.1}	59.6 _{0.6}	75.5 _{0.4}	55.3 _{1.1}	68.8 _{0.9}	56.2	66.9

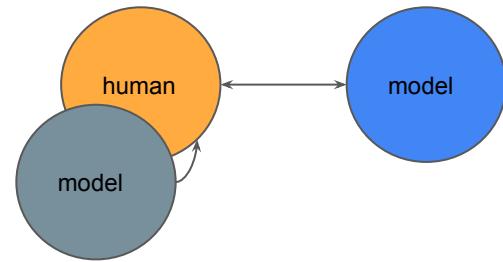
Model	MRQA out-of-domain													
	BioASQ		DROP		DuoRC		RACE		RelationExt.		TextbookQA		Avg	
	EM	F_1	EM	F_1										
R _{SQuAD}	53.2 _{1.1}	68.6 _{1.4}	39.8 _{2.6}	52.7 _{2.2}	49.3 _{0.7}	60.3 _{0.8}	35.1 _{1.0}	47.8 _{1.2}	74.1 _{3.0}	84.4 _{2.9}	35.0 _{3.8}	44.2 _{3.7}	47.7	59.7
R _{SQuAD+AQA}	54.6 _{1.2}	69.4 _{0.8}	59.8 _{1.3}	68.4 _{1.5}	51.8 _{1.1}	62.2 _{1.0}	38.4 _{0.9}	51.6 _{0.9}	75.4 _{2.3}	85.8 _{2.4}	40.1 _{3.1}	48.2 _{3.6}	53.3	64.3
SynQA	55.1 _{1.5}	68.7 _{1.2}	64.3 _{1.5}	72.5 _{1.7}	51.7 _{1.3}	62.1 _{0.9}	40.2 _{1.2}	54.2 _{1.3}	78.1 _{0.2}	87.8 _{0.2}	40.2 _{1.3}	49.2 _{1.5}	54.9	65.8
SynQA _{Ext}	54.9 _{1.3}	68.5 _{0.9}	64.9 _{1.1}	73.0 _{0.9}	48.8 _{1.2}	58.0 _{1.2}	38.6 _{0.4}	52.2 _{0.6}	78.9 _{0.4}	88.6 _{0.2}	41.4 _{1.1}	50.2 _{1.0}	54.6	65.1

SynQA outperforms alternatives

SynQA models are much harder to fool (i.e. more robust)

Empowering crowdworkers with generative assistants

- We know now that generative models trained on adversarial data can help make models more robust.
- Can we use those models to help humans fool models as “generative adversarial assistants”?
ModelS in the loop!
 - a. Adversarial data is expensive - can it be made cheaper?
 - b. Adversarial data can be noisy - can it be made higher quality?



Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants

Max Bartolo* Tristan Thrush† Sebastian Riedel‡*†

Pontus Stenetorp* Robin Jia†‡ Douwe Kiela‡

*UCL †USC ‡Facebook AI Research

Concrete example

A hole is classified by its par, meaning the number of strokes a skilled golfer should require to complete play of the hole. The minimum par of any hole is **3** because par always includes a stroke for the tee shot and **two** putts. Pars of 4 and 5 strokes are ubiquitous on golf courses; more rarely, a few courses feature par-6 and even par-7 holes. Strokes other than the tee shot and putts are expected to be made from the fairway; for example, a skilled golfer expects to reach the green on a par-4 hole in two strokes—one from the...



A: **two**



Q: How many putts are considered minimum to make par?



A: **3**

Standard (SDC) vs Adversarial (ADC) Data Collection

Adversary-in-the-loop?	Median time per example			Time per model-fooling ex				Domain generalization	
	t (s)	vMER (%)	t/vMFE (s)	SQuAD _{dev}	$\mathcal{D}_{\text{BiDAF}}$	$\mathcal{D}_{\text{BERT}}$	$\mathcal{D}_{\text{RoBERTa}}$	MRQA	
✗	56.3 _{23.6}	0.63	11274	45.4	14.7	9.2	8.8	25.2	
✓	61.2 _{27.4}	1.62	4863	82.0	44.4	29.2	22.4	53.8	

Validated model error rate

Standard

Adversarial QA

↓

↓

↓

↓

↓

↓

↑

↑

↑

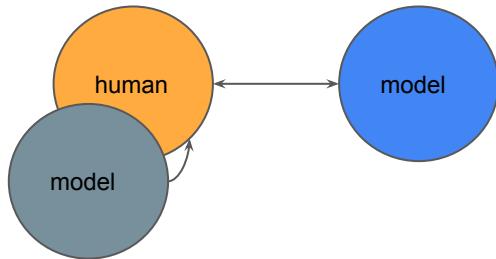
Improving ADC further

- If you do “answer prompting” where you don’t force annotators to pick the answer but suggest one, ADC gets even faster and much higher quality.
- Starting point, traditional data collection: vMER=0.63 with t=56.3
- End point, ADC with GAA: vMER=6.08 with t=43.8

GAA Training	Sampling	t (s)	vMER (%)	t/vMFE (s)	SQuAD _{dev}	$\mathcal{D}_{\text{BiDAF}}$	$\mathcal{D}_{\text{BERT}}$	$\mathcal{D}_{\text{RoBERTa}}$	MRQA
AdversarialQA	<i>Likelihood</i>	49.9 _{29.9}	6.08	1086	78.2	44.0	33.7	26.2	52.0
AdversarialQA	<i>Adversarial</i>	43.8 _{22.1}	2.22	2587	79.9	44.2	30.6	23.6	52.1
AdversarialQA	<i>Uncertainty</i>	50.9 _{23.5}	4.04	1667	80.4	42.8	28.8	22.1	51.1
Combined	<i>Likelihood</i>	49.0 _{23.0}	2.72	2510	79.6	42.7	31.1	23.8	50.2
Combined	<i>Adversarial</i>	65.2 _{30.9}	4.41	2042	80.2	44.7	31.5	24.8	53.0
Combined	<i>Uncertainty</i>	54.1 _{22.0}	2.94	2740	81.1	44.8	27.9	23.8	51.2

A “new paradigm”?

- ModelS in LoopS:
 - Yes, we can collect much higher quality data than static data using this method.
 - Yes, we can collect higher quality data than regular human-and-model-in-the-loop.
 - Yes, we can do so at a cost that is much lower than human-and-model-in-the-loop, matching standard data collection.



Douwe Kiela @douwekiela · Jan 16

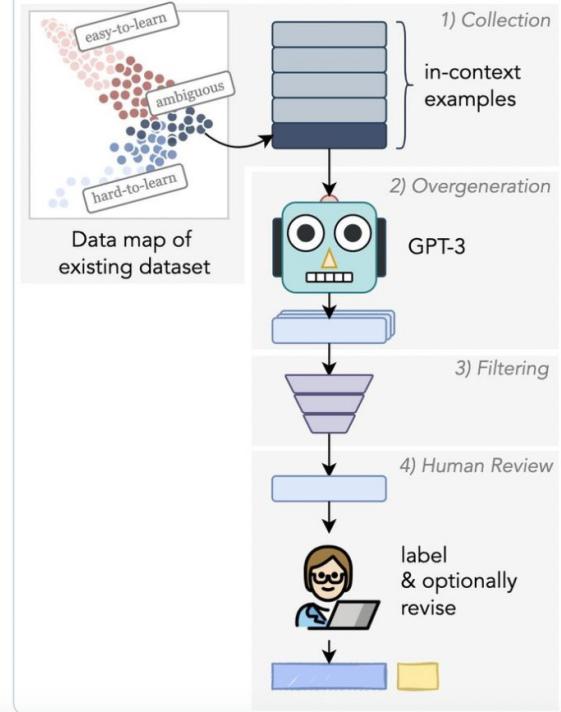
I wish Einstein had done this, it would have been so much easier. "A New Paradigm for Brownian Motion", or "A New (Special/General) Paradigm for Relativity".

Alisa Liu @alisawuffles · Jan 15

We introduce a new paradigm for dataset creation based on human 🧑‍💻 and machine 🤖 collaboration, which brings together the generative strength of LMs and the evaluative strength of humans. And we collect 🎉 WaNLI, a dataset of 108K NLI examples! 📈

Paper: [swabhs.com/assets/pdf/wan...](https://swabhs.com/assets/pdf/wanli.pdf)

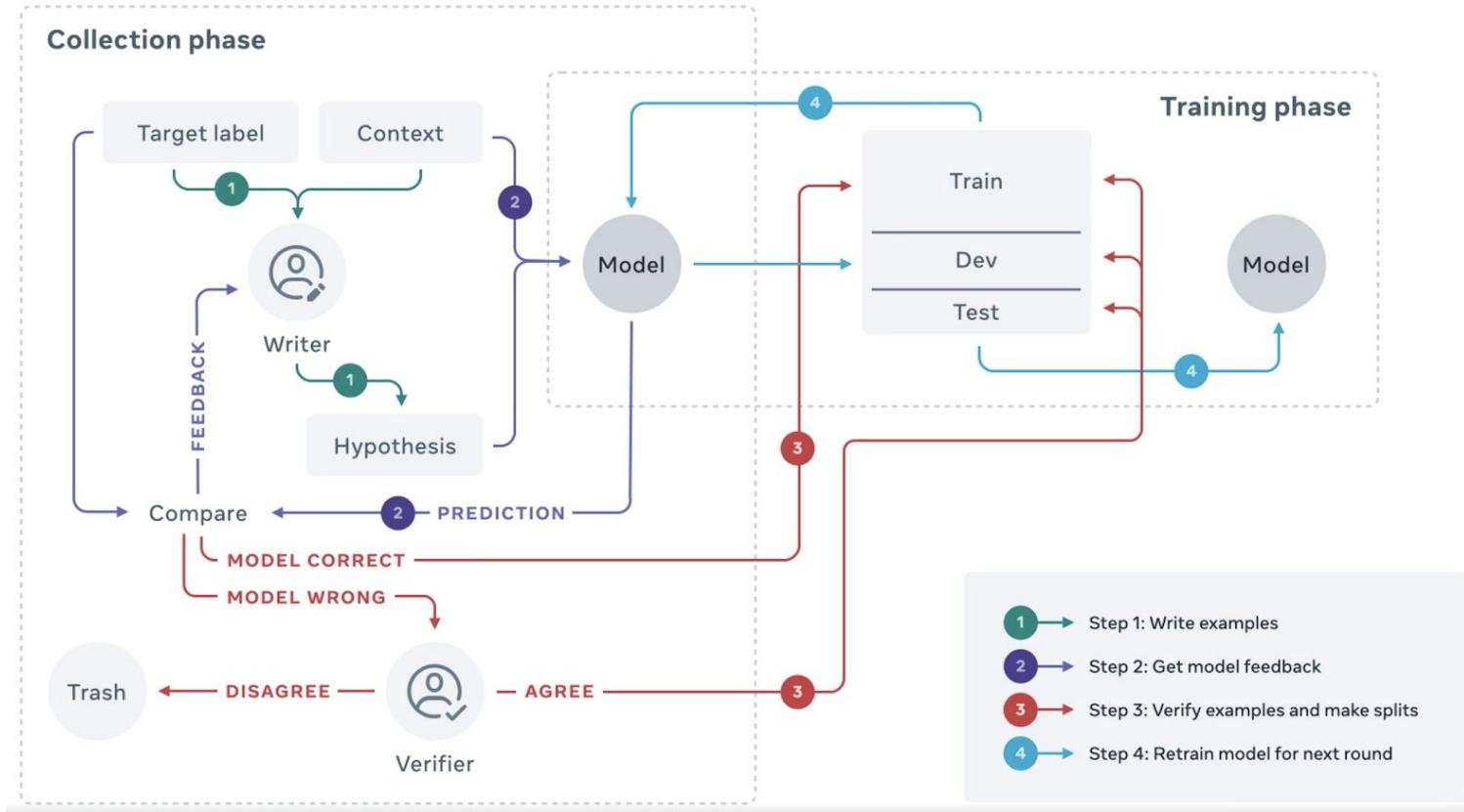
Show this thread



Recent work out of the Dynabench team

- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP**
- Vidgen et al. (ACL21). **Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection**
- Potts et al. (ACL21). **DynaSent: A Dynamic Benchmark for Sentiment Analysis**
- Kirk et al. (2021). **Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate**
- Sheng & Singh et al. (NeurIPS21). **Human-Adversarial Visual Question Answering**
- Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
- Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
- Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**
- Thrush et al. (2022). **Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks**
- Bartolo et al. (EMNLP21). **Improving QA Model Robustness with Synthetic Adversarial Data Generation**
- Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
- Bartolo et al. (2022). **Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants**
- Wallace et al. (2022). **Analyzing Dynamic Adversarial Training Data in the Limit**

Dynamic adversarial data collection in the limit



Experimental setup

- Starting point: Roberta trained on “All NLI” (MNLI+SNLI+FEVER)
- We hand-construct an expert-curated test set covering a wide range of NLI phenomena.
- We do DADC for 20 rounds (ANLI only did 3).
- We select 10 contexts so that:
 - a. We can afford collecting many rounds of data
 - b. We have some hope of achieving saturation
 - c. We have a broad range of phenomena
 - d. We can create a wide-coverage test set

Analyzing Dynamic Adversarial Training Data in the Limit

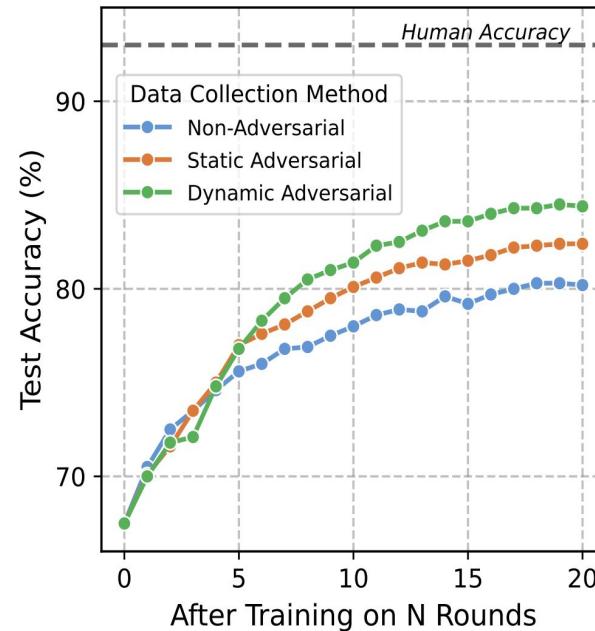
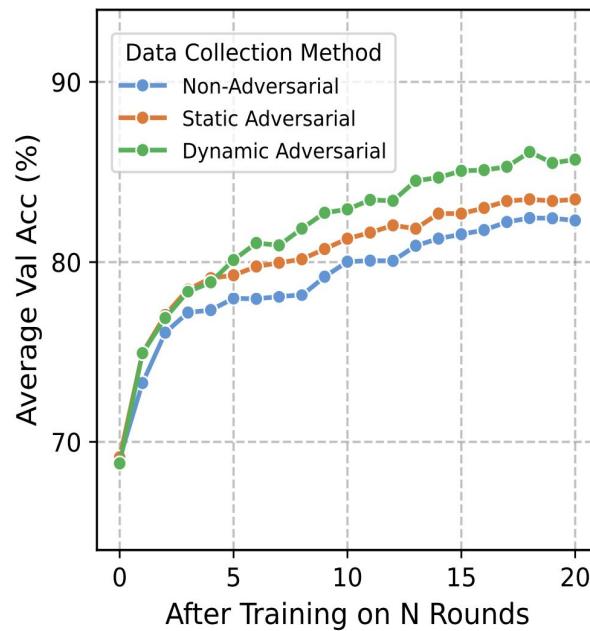
Work by **Eric Wallace et al.**

Eric Wallace^{1*} Adina Williams^{2†} Robin Jia^{2,3†} Douwe Kiela^{2†}

¹UC Berkeley ²Facebook AI Research ³USC

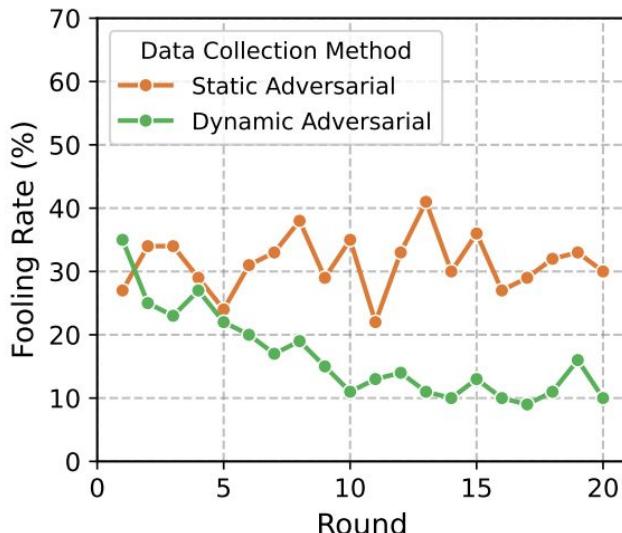
Findings: A virtuous cycle

Promising results when exploring Dynamic Adversarial Data Collection in the limit:



Findings: Diversity is key

- DADC data is more diverse, more complex and has fewer artifacts.
- DADC models gets stronger over time.



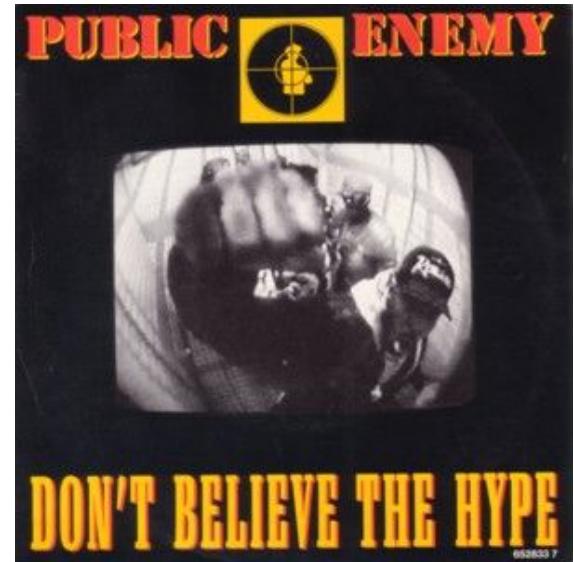
	No Model	Static Model	Dynamic Model
<i>Diversity</i>			
Unique Unigrams	4.0k	4.2k	4.3k
Unique Bigrams	23.3k	24.8k	25.6k
Inter-example Sim.	41.2	41.9	39.5
<i>Complexity</i>			
Syntax	2.0	2.1	2.3
Reading Level	4.9	5.4	5.9
Length	10.1	10.9	12.1
<i>Artifacts</i>			
Hypo-only Acc %	75.4	69.3	69.7
Overlap Entail %	54.2	49.2	47.3

Method take-aways

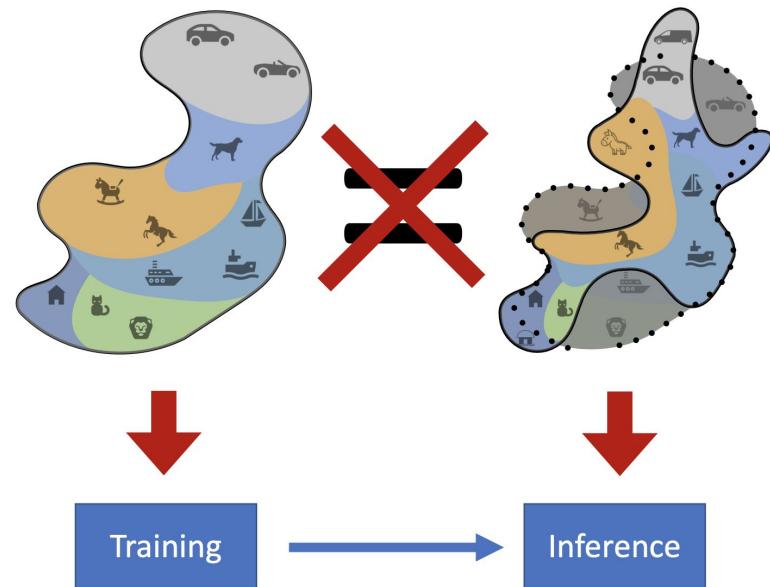
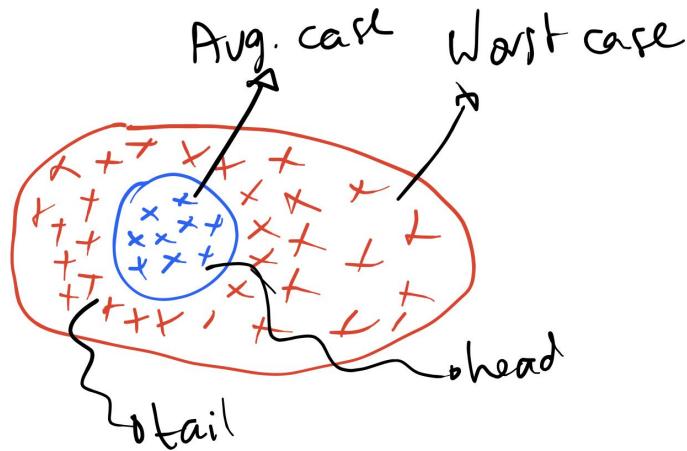
- If DADC gets you better training and testing data faster, why do traditional static crowdworker data collection? Humans and models in the loop!
- Further work needed on many questions, including:
 - a. How (un)natural is adversarial data and how much does that matter?
 - b. How does dynamic adversarial data collection relate to active learning and continual learning?
 - c. Can we incorporate knowledge about the model in the loop in our optimization procedures?
 - d. Exploring ensembles in the loop, different scoring functions, etc.

What is our goal? What is language?

- **Do believe the hype:** we're decent (but not great) at (some) i.i.d. problems when we have enough data and don't care about the worst case.
- **Don't believe the hype:** we are FAR from truly general language understanding that encompasses all of language's recursive, structured, generative, productive, and creative nature.



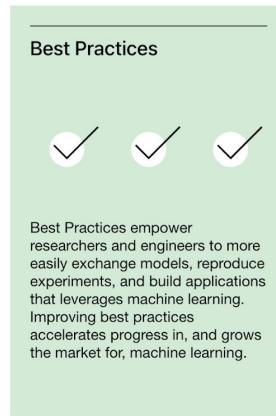
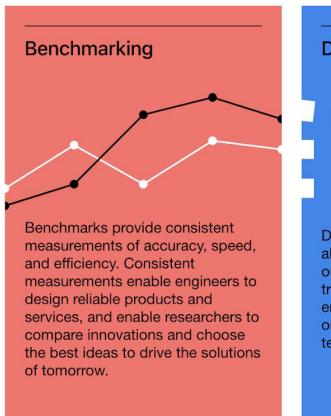
The ability to REALLY understand language



(Madry, 2018; <https://adversarial-ml-tutorial.org/>)

Teaming up with ML Commons and DataPerf

MLCommons aims to answer the needs of the nascent machine learning industry through open, collaborative engineering in three areas:



“Everyone wants to do the model work, not the data work”:
Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
[nithyasamba,kapania,hhighfill,dakrong,pkp.loraa]@google.com
Google Research
Mountain View, CA

Thanks!

Thank you to my many collaborators on these projects (they deserve the credit, I'm just the conduit here).

And thank you for listening!