Dynabench: Rethinking Benchmarking in NLP

ACL 2021 Workshop on Benchmarking: Past, Present and Future

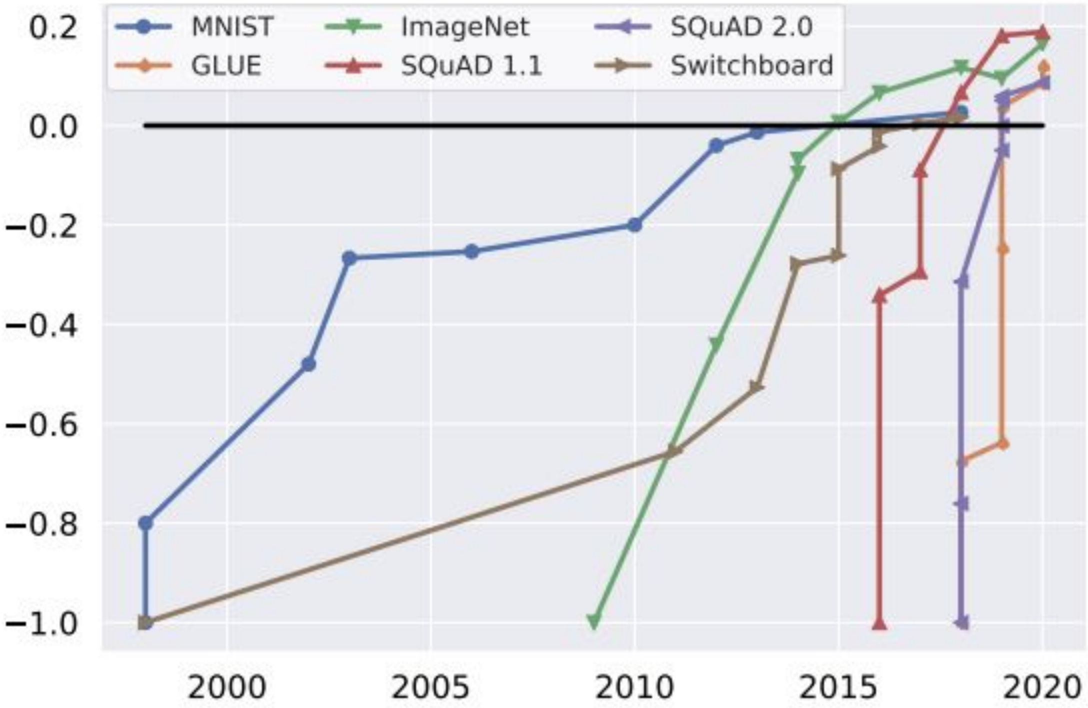
https://dynabench.org, @DynabenchAI

Douwe Kiela

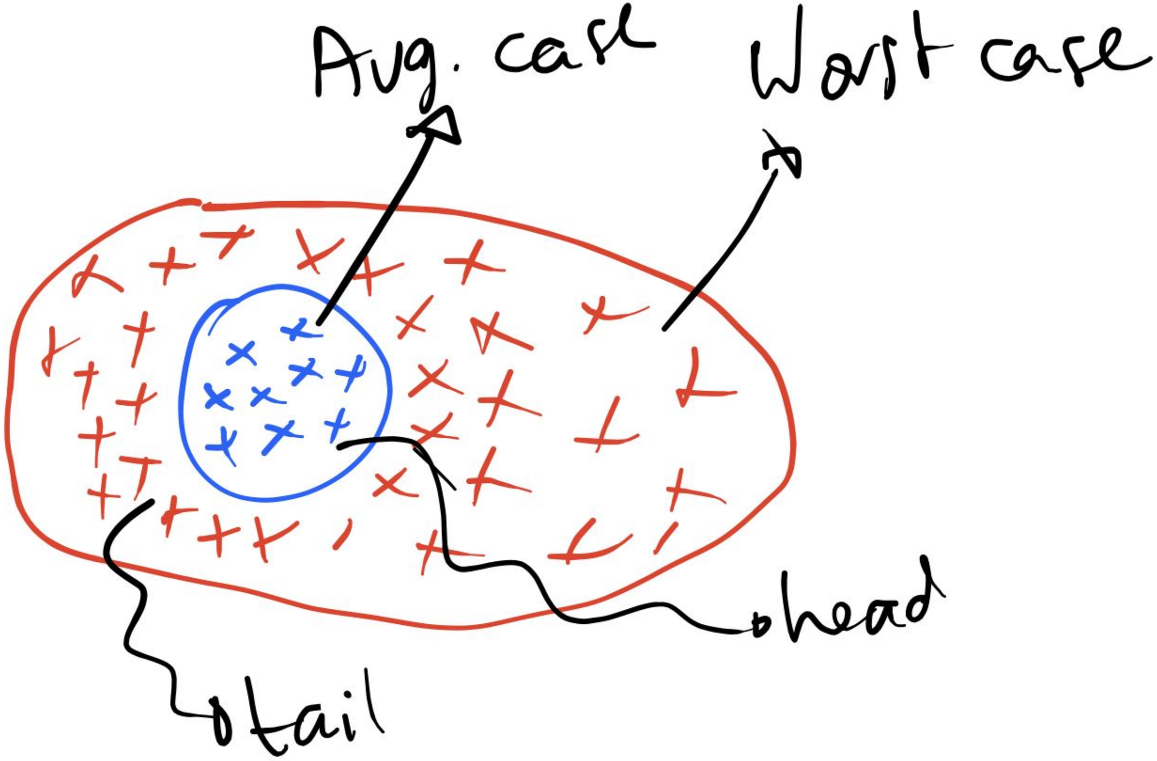
DynaBench

UCL
The University of North Carolina at Chapel Hill
Carnegie Mellon University
The Alan Turing Institute
SFU Simon Fraser University
Johns Hopkins University
Stanford University
FACEBOOK AI

Have we solved NLP?



The Ability to REALLY Understand a Language: Average vs Worst Case



Why this stuff matters

NLP models are being deployed with **real-world consequences** so we need to be very careful with what models we select for deployment.

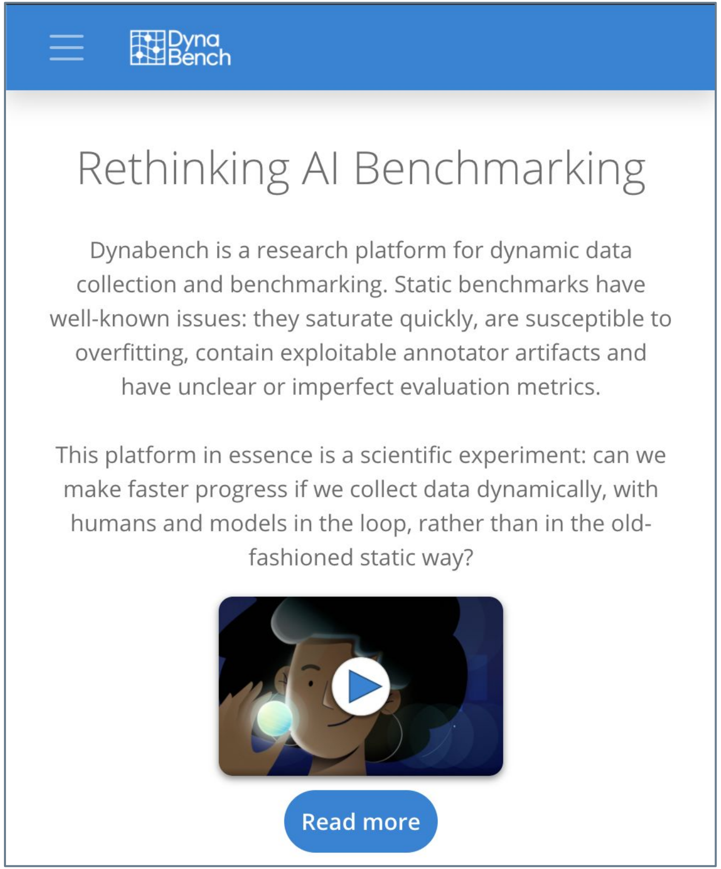
People anthropomorphize machines and are very bad at predicting when they’ll fail, so we shouldn’t expect them to correct for our systems’ weaknesses.

True language understanding is about **strong generalization**, not i.i.d. average case generalization.

What is Dynabench?

A **research platform** for rethinking benchmarking:

- Try to **challenge existing notions** to see if we can make better progress.
- Can we **move beyond static** datasets, static leaderboards, static metrics?
- What happens when we collect data with **humans and models in the loop**?
- Can we **evaluate models along multiple axes**? And collect useful data while doing so?



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?

What are the problems we're trying to solve?

There are many well-known problems in 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 make benchmarks too easy.
- **Alignment**: Benchmarks don't measure the right thing - test set performance is not always a good proxy for "how well would this system work 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.
- **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.

What is the solution?

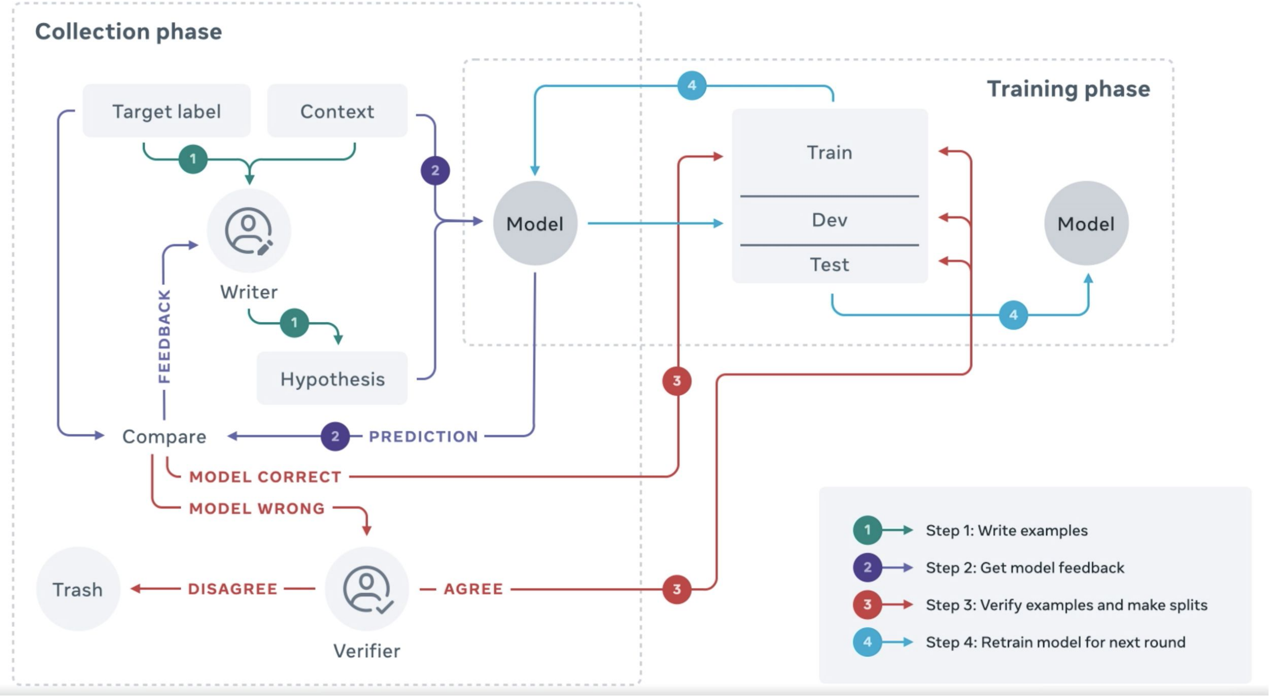
Nobody knows yet. The platform is dynamic and will keep evolving.

We think part of the solution is:

- Support multiple tasks over multiple rounds
- Community effort, with task experts as “owners” calling the shots
- Dynamic data collection with model(s) in the loop
- Dynamic leaderboards that rely on Evaluation-as-a-Service

In the rest of the talk, I’ll go over some of the crucial components.

Human and model in the loop

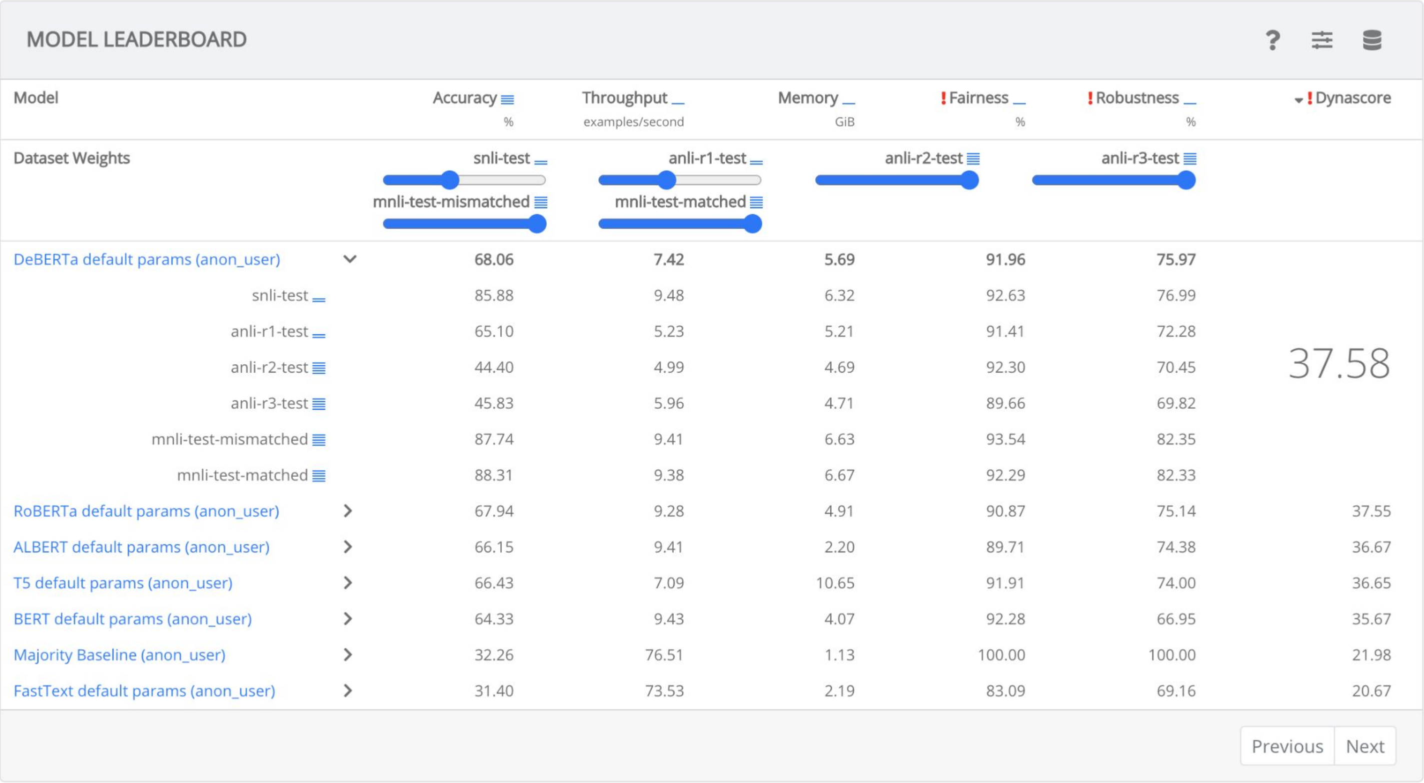


Model upload and multi-metric eval

Problem with data collection: what about backward compatibility? With our **evaluation cloud**, we can try to **address many existing problems** with current benchmarks:

- For every model, we can **record the validated model error rate** (vMER), i.e., how easily would this model get fooled if it was deployed?
- We can **evaluate models along multiple axes**, and check for efficiency, but also fairness and robustness, i.e., our systems should work just as well for James as for Jamal.
- We can have **task owners choose** what metrics they want to use, what datasets they want to evaluate on, and how they want to aggregate results.
- New **SOTA models can be put in the loop** for new rounds of data collection, and we can evaluate old models on new data.

Dynamic leaderboards



MODEL LEADERBOARD

Model	Accuracy ≣ %	Throughput _ examples/second	Memory _ GiB	! Fairness _ %	! Robustness _ %	▾ ! Dynascore
Dataset Weights	snli-test _ mnli-test-mismatched ≣	anli-r1-test _ mnli-test-matched ≣	anli-r2-test ≣	anli-r3-test ≣		
DeBERTa default params (anon_user)	68.06	7.42	5.69	91.96	75.97	37.58
snli-test _	85.88	9.48	6.32	92.63	76.99	
anli-r1-test _	65.10	5.23	5.21	91.41	72.28	
anli-r2-test ≣	44.40	4.99	4.69	92.30	70.45	
anli-r3-test ≣	45.83	5.96	4.71	89.66	69.82	
mnli-test-mismatched ≣	87.74	9.41	6.63	93.54	82.35	
mnli-test-matched ≣	88.31	9.38	6.67	92.29	82.33	
RoBERTa default params (anon_user)	67.94	9.28	4.91	90.87	75.14	37.55
ALBERT default params (anon_user)	66.15	9.41	2.20	89.71	74.38	36.67
T5 default params (anon_user)	66.43	7.09	10.65	91.91	74.00	36.65
BERT default params (anon_user)	64.33	9.43	4.07	92.28	66.95	35.67
Majority Baseline (anon_user)	32.26	76.51	1.13	100.00	100.00	21.98
FastText default params (anon_user)	31.40	73.53	2.19	83.09	69.16	20.67

What have we done so far?

- Datasets created:
  - DynaSent (Sentiment), LFTW (Hate speech), Human-Adversarial VQA
- Papers published:
  - 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**
  - Kaushik et al. (ACL21). **On the Efficacy of Adversarial Data Collection for Question Answering**
  - Bartolo et al. (2021). **Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation**
  - Sheng & Singh et al. (2021). **Human-Adversarial Visual Question Answering**
  - Ma, Ethayarajh, Thrush et al. (2021). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
  - .. and many more in the pipeline ..
- Challenges enabled:
  - Flores101 large-scale multilingual machine translation @ WMT
- Raw examples collected:
  - 400k+

Demo

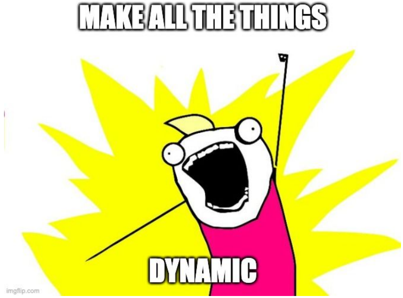
- Quick live demo

Research questions

- How do we deal with distributional shift?
- How do we combine i.i.d with non-i.i.d? What does i.i.d. even mean in the case of language?
- How useful is adversarial data for training purposes?
- What better metrics can we come up with?
- How can we empower annotators using models in the loop?

Conclusion

- If we want to deploy NLP models in the wild, they should work beyond the average case.
- We can use the vMER to assess how good models really are.
- This process can be repeated over many rounds, yielding useful training data.
- We need to move beyond accuracy towards a more holistic approach.
- We're doing continuous testing in the worst case, where a test set is a "known unknown", we also want to capture the "unknown unknowns".



Thanks for listening

- We would love your help! Please join our community!
    - Come up with new model-fooling examples
    - Contribute new models
    - Add new features or metrics
    - Help improve the code base
    - Start a new task
- We are hiring: interns, pre-, post-docs, please reach out!

Dynabench.org
@DynabenchAI