Why Adversarially-Collected Test Sets Don't Work as Benchmarks

ML² Machine Learning for Language

ANTHROP\C

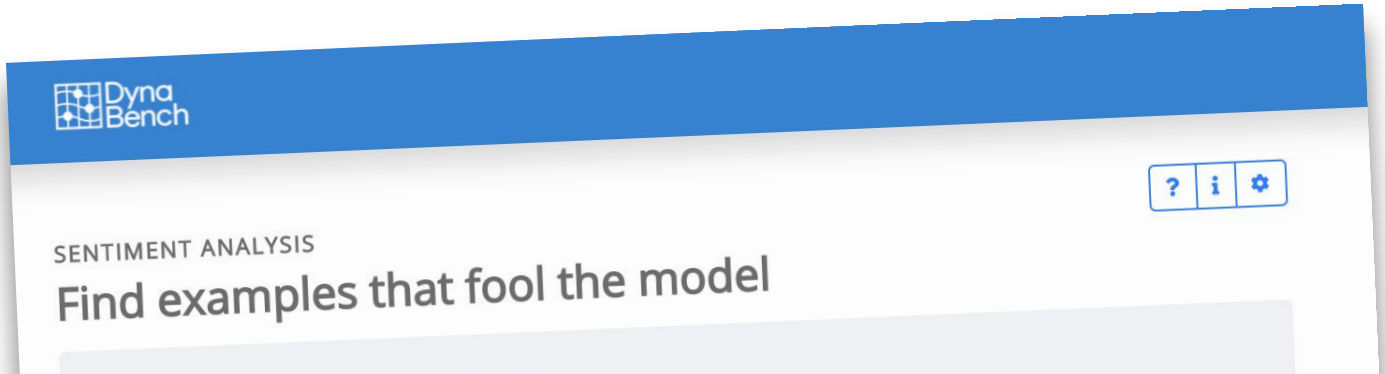
Sam Bowman
@sleepinyourhat

Adversarial Data Collection

Adversarial data collection (ADC) in this talk:

> *The practice of building datasets entirely out of examples on which a specific system fails.*

Bartolo et al. TACL ‘20;
Kiela et al. NAACL ‘21;
(Le Bras et al. ICML ‘20)



🔎 tl;dr

- ADC seems promising as a way of collecting training data.
- ADC seems promising as a way of analyzing model behavior.

🔎 tl;dr

- ADC seems promising as a way of comparing the robustness of a known set of models.
- **ADC is unfixably broken as a way of creating benchmark test sets.**

🔎 tl;dr

Why?

- It’s obscuring problems with NLP evaluation rather than fixing them.
- It makes test sets that can’t measure the relative performance of models.
- It makes test sets that can’t measure the absolute performance of models.

🔎 tl;dr

What should we do instead?

- Use ADC-based analyses as part of test set *design*.
- Build hard test sets the slow, simple way.
- It’s okay if they’re smaller!

ADC obscures problems with NLP evaluation rather than fixing them.

Bowman & Dahl NAACL ‘21

🥅 The Goal

We want benchmarks that measure the degree to which models can perform some specific language task on some specific language variety and topic domain.

🔎 Validity

This includes:

- Comprehensive coverage of language variation.
- Test cases isolating all necessary task skills.
- No artifacts that let bad models score highly.

This is hard.

The Problem

Benchmarking for language understanding is broken.

Model	EM
Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831
FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**

	Score
T5 + Meena, Single Model (Meena Team - Google Brain)	90.4
DeBERTa / TuringNLRv4	90.3
SuperGLUE Human Baselines	89.8
T5	89.3

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%			
B Testing **NER** with ***INV***	Same pred. (inv) after removals / additions		
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			

🤷 Does ADC Help?

It looks like it helps!

- Because ADC guarantees that test sets will be hard for SotA models, it guarantees that those test sets won't *look* broken.

🤷 Does ADC Help?

…but it doesn’t.

- Making a dataset more difficult is distinct from making it more representative of the desired behavior.
  -
- Empowering *the adversary model* to define the test distribution removes a key point of leverage.

ADC obscures problems with NLP evaluation rather than fixing them.

Bowman & Dahl NAACL ‘21

ADC makes test sets that can’t measure the relative performance of models.

Phang et al. (under submission)

🥅 The Goal

One of the chief uses of benchmark test sets is to establish fair comparisons between different systems.

🥅 The Goal

In other words, the *ranking* of systems on the benchmark should reflect their relative ability on the task.

🥉 Ranking Artifacts

ADC introduces *ranking artifacts*:

> Patterns in model rankings on benchmarks that are predictable but not due to model ability.

🥉 Ranking Artifacts?

By design, if a model is tested on an adversarially-collected test set that was collected against that model, it will achieve zero accuracy…

…and sufficiently similar models will achieve low accuracy.

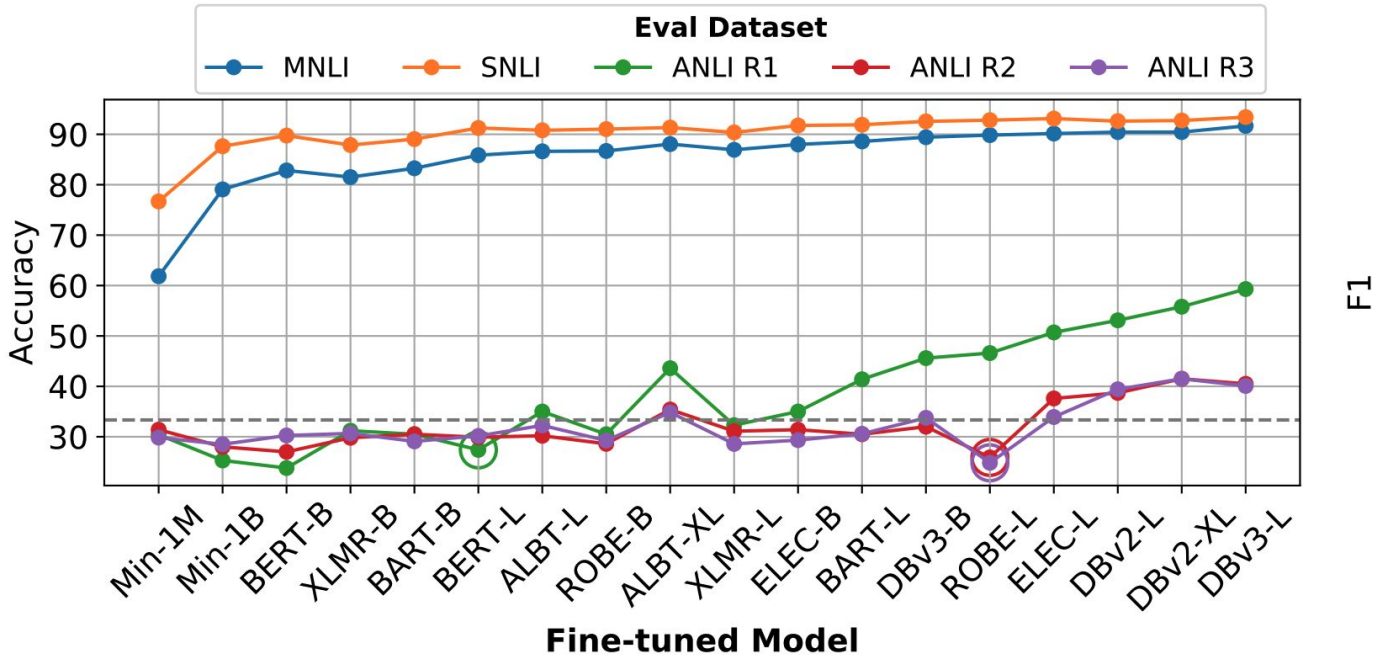
Ranking Artifacts

Model	Training Data	A1	A2	A3	ANLI
BERT	S,M[⋆1]	00.0	28.9	28.8	19.8
	+A1	44.2	32.6	29.3	35.0
	+A1+A2	57.3	45.2	33.4	44.6
	+A1+A2+A3	57.2	49.0	46.1	50.5
	S,M,F,ANLI	57.4	48.3	43.5	49.3
XLNet	S,M,F,ANLI	67.6	50.7	48.3	55.1
RoBERTa	S,M	47.6	25.4	22.1	31.1
	+F	54.0	24.2	22.4	32.8
	+F+A1[⋆2]	68.7	19.3	22.0	35.8
	+F+A1+A2[⋆3]	71.2	44.3	20.4	43.7
	S,M,F,ANLI	73.8	48.9	44.4	53.7

ANLI (ADC)

Nie et al. ACL ‘20

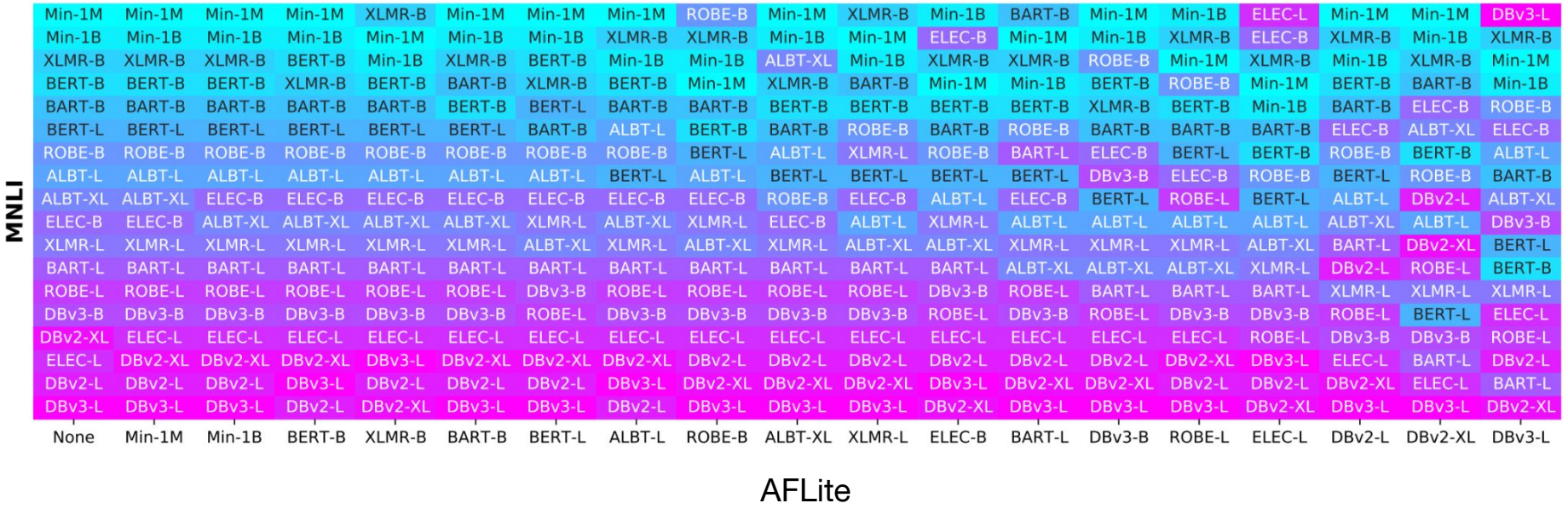
Ranking Artifacts



ANLI (ADC)

[Phang et al. (under submission)]

Ranking Artifacts



[Phang et al. (under submission)]

ADC makes test sets that can’t measure the relative performance of models.

Phang et al. (under submission)

ADC makes test sets that can’t measure the absolute performance of models.

🥅 The Goal

We want benchmarks that measure the degree to which models can perform some specific language task on some specific language variety and topic domain.

🥉 Ranking Artifacts Revisited

- By design, if a model is tested on an adversarially-collected test set that was collected against that model, it will achieve zero accuracy.
    - Sufficiently similar models will achieve low accuracy.
- True as long as the model makes *any* errors or debatable judgments on *any* possible inputs.
- So, possible to target *humans* for 0% accuracy, too!

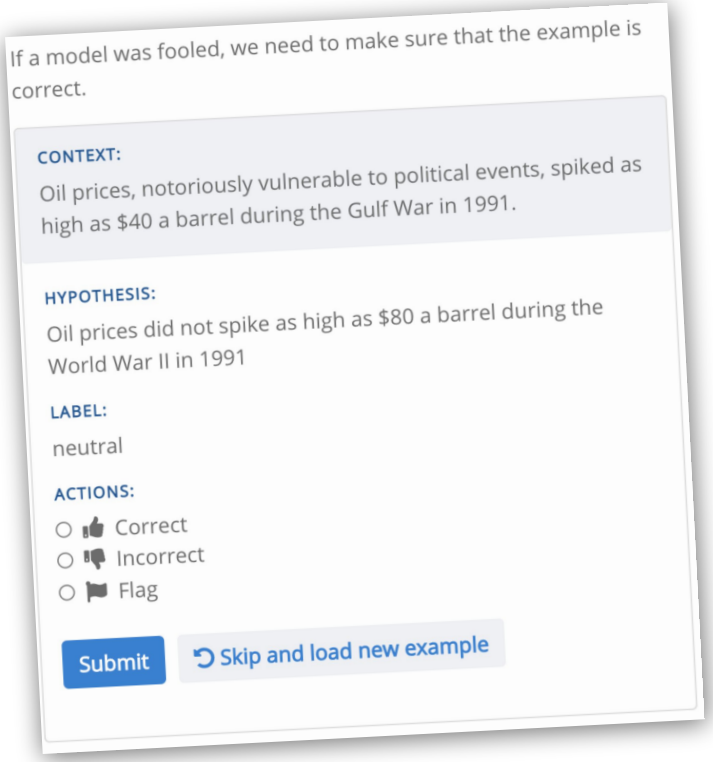
Ranking Artifacts Revisited

If our technique reports that some humans achieve **0%** competence at a language task, *absolute scores* originating from that technique aren't informative.

Absolute score on an adversarially-collected test set is meaningless as a measure of model performance.

🥉 Ranking Artifacts Revisited

Common DADC UIs make it relatively easy to accidentally skew subjective calls away from the target model:



[Wallace et al. ACL Findings '22]

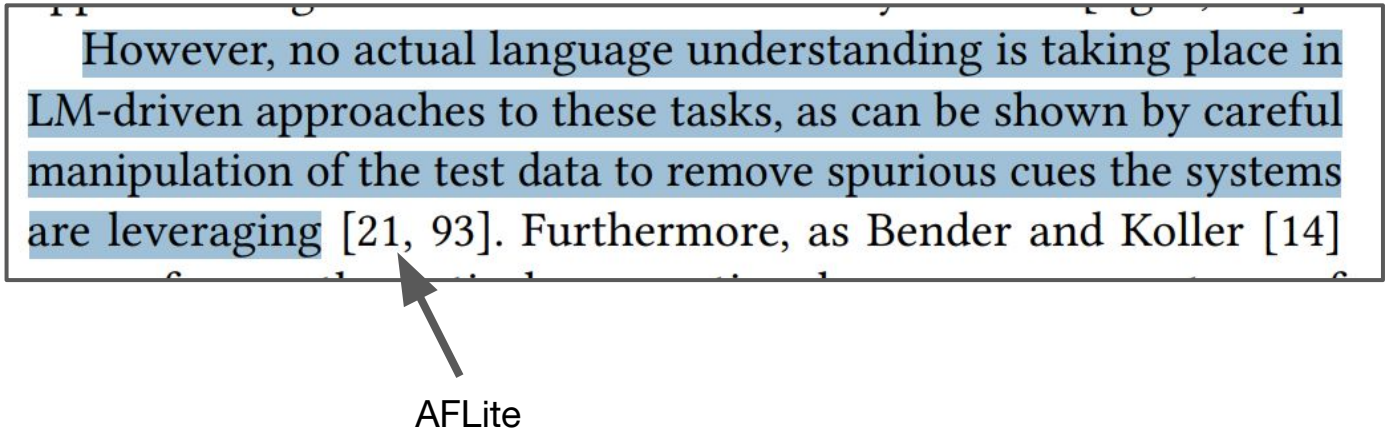
ADC makes test sets that can’t measure the absolute performance of models.

Detour: Underclaiming

Bowman ACL ‘22

ADC and Underclaiming

If results on ADC test sets are misrepresented as capturing absolute performance, they can feed into unjustified negative messages about the current state of the art:



ADC and Underclaiming

This phenomenon, *underclaiming*, is increasingly common, and it's important that we learn to avoid it.

Three Reasons Underclaiming Is Dangerous

Bowman ACL ‘22

 The Health of the Field

- We like to think of NLP as a scientific field.
- This means not accepting claims without good evidence.

Managing Current Impacts

- Underclaiming can be superficially appealing here:
    - Arguing that systems don’t work should discourage their deployment, limiting the harms from biased or untrustworthy systems.
- But this approach backfires:
    - If operators of deployed systems realize that they can’t trust our assessments of system ability, they might not listen to any of our other concerns.

🔮 Managing Future Impacts

- We *seem* to be making progress, and it's reasonable to expect that NLP technology will eventually get good.

- Many of the most important impacts from NLP deployments depend on systems *working very well*.

  🥳: worldwide access to excellent education, medical advice, legal services, …

  😬: abrupt mass unemployment, mass misinformation/surveillance, potential catastrophic risks, ...

🔮 Managing Future Impacts

- To manage these impacts, we'll need to start the relevant technical work and policy work long before the impacts start to arrive.

- Widespread underclaiming makes it hard for the NLP community to take these issues seriously.

👷 What Should We Do Instead?

There’s No Easy Fix

Evaluating language understanding in machines for some task requires careful thinking about language, machines, and the task.

👷 What Should We Do Instead?

Collect data the hard, slow, boring way:

- Figure out what phenomena and domains will be informative to study.
- Hire careful workers to collect a representative sample of those phenomena in those domains.
- Thoroughly validate those examples.

👷 What Should We Do Instead?

This is slower, but not necessarily prohibitive:

- Large-scale pretraining means that benchmarks no longer need to come with large training sets…
- …and a big decrease in the importance of hyperparameter tuning makes it safer to launch benchmarks with small test sets.

👷 What Should We Do Instead?

Room for creativity here:

- Use DADC to identify phenomena to study (cf. ANLIzing ANLI)

- Use DADC where *unqualified humans* are the adversary (cf. QuALITY)

ADC is valuable.

ADC does not produce usable test sets.

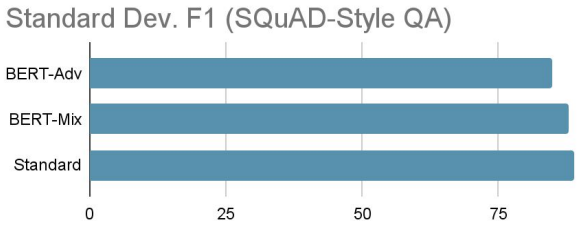
…but we don’t need it to.

Fin

Thanks to Schmidt Futures, the US NSF, Apple, Samsung, Intuit, and Google for funding. See papers for project details.

🤷 Does ADC Help?

- Empirically, ADC data can get arbitrarily far from the task under study…



On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study

Divyansh Kaushik[†], **Douwe Kiela**[‡], **Zachary C. Lipton**[†], **Wen-tau Yih**[‡]
[†] Carnegie Mellon University; [‡] Facebook AI Research
{dkaushik,zlipton}@cmu.edu, {dkiela,scottyih}@fb.com

rial data collection (ADC)