Progress in Dynamic Adversarial Data Collection
&
Adventures in Multimodal Machine Learning

Douwe Kiela

Hugging Face Stanford University

DynaBench ML Commons UCL THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL Carnegie Mellon University The Alan Turing Institute

The University Of Sheffield. SFU SIMON FRASER UNIVERSITY JOHNS HOPKINS UNIVERSITY Stanford University FACEBOOK AI

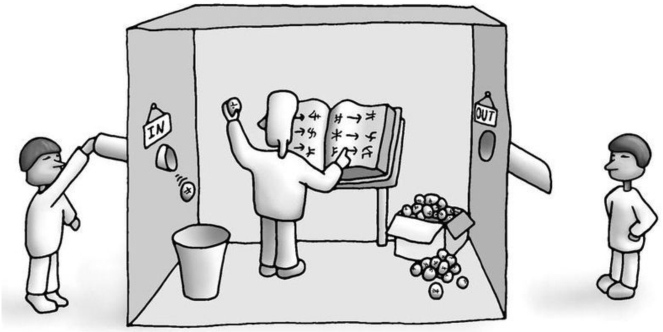
Outline

1. Dynabench
   a. **Overview**
   b. Common Objections & Misconceptions
2. Progress in Dynamic Adversarial Data Collection
   a. Humans and Models in Loops
   b. Dynamic Adversarial Training Data
3. Adventures in Multimodal ML
   a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
   b. Foundation Models: FLAVA

News / about me..

Meta AI → Stanford University

Meta AI → ML Commons

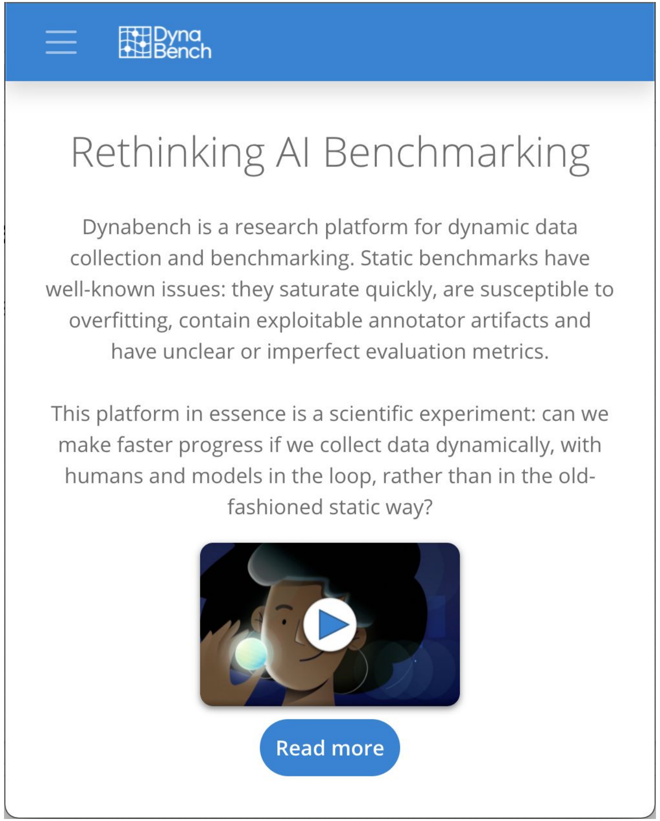


Rethinking benchmarking in AI

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,

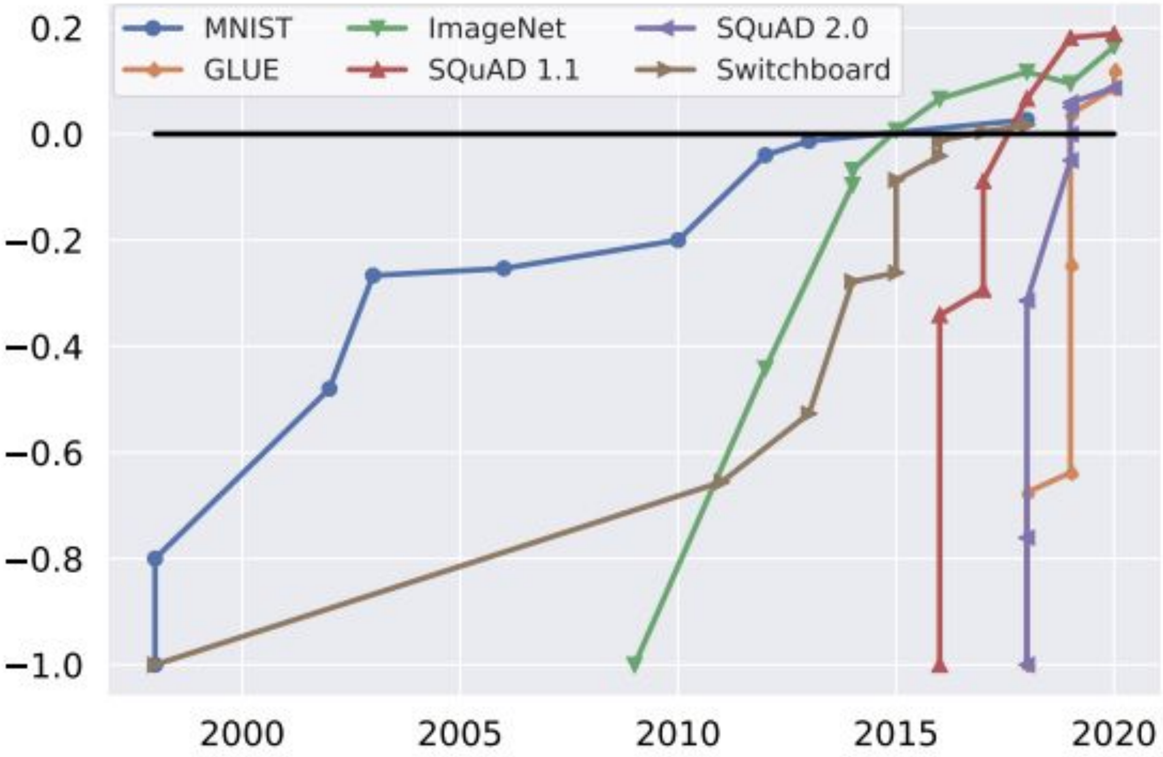


Rethinking AI Benchmarking

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?

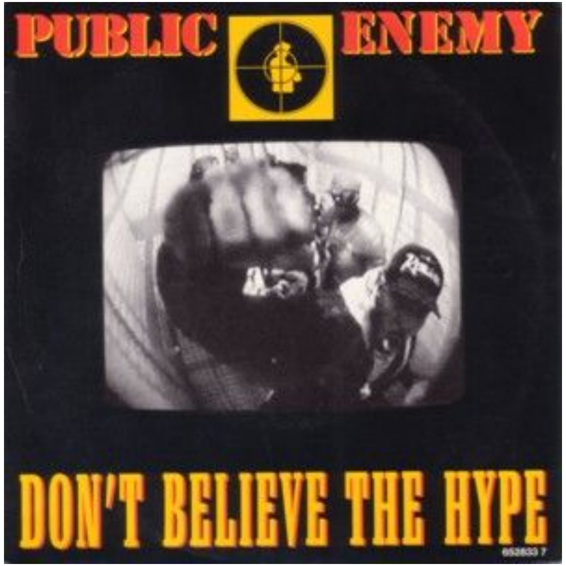
Benchmark saturation in NLP



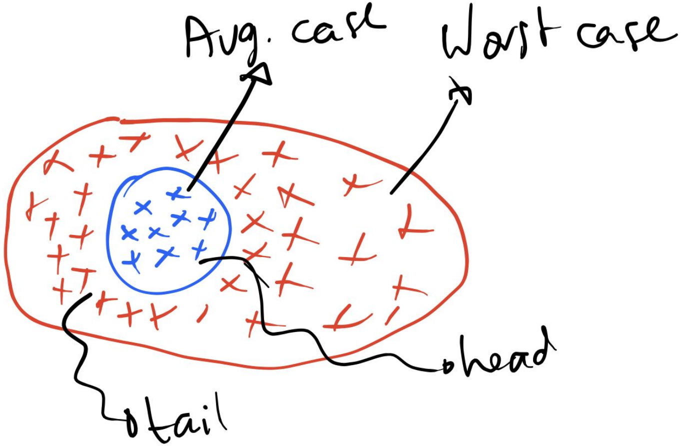
What is our goal? What is language?

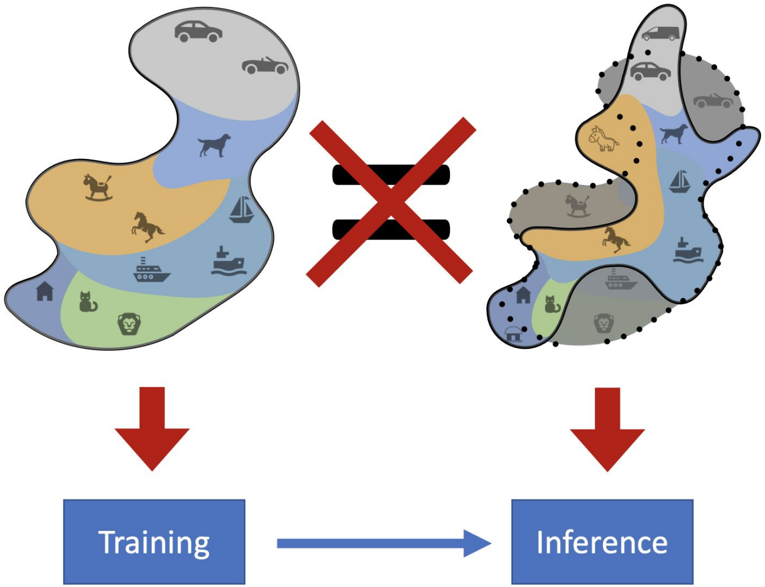
Do believe the hype: we're decent (but not great) at (some) i.i.d. problems when we have enough data and don't care about the worst case.

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.



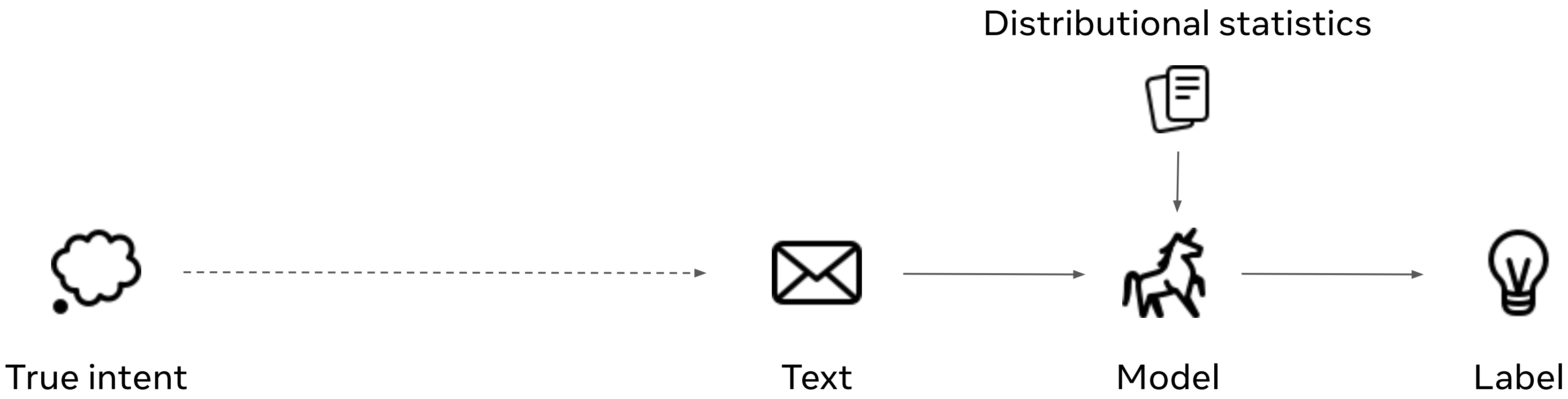
The ability to REALLY understand language





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

What we are doing..



Distributional statistics

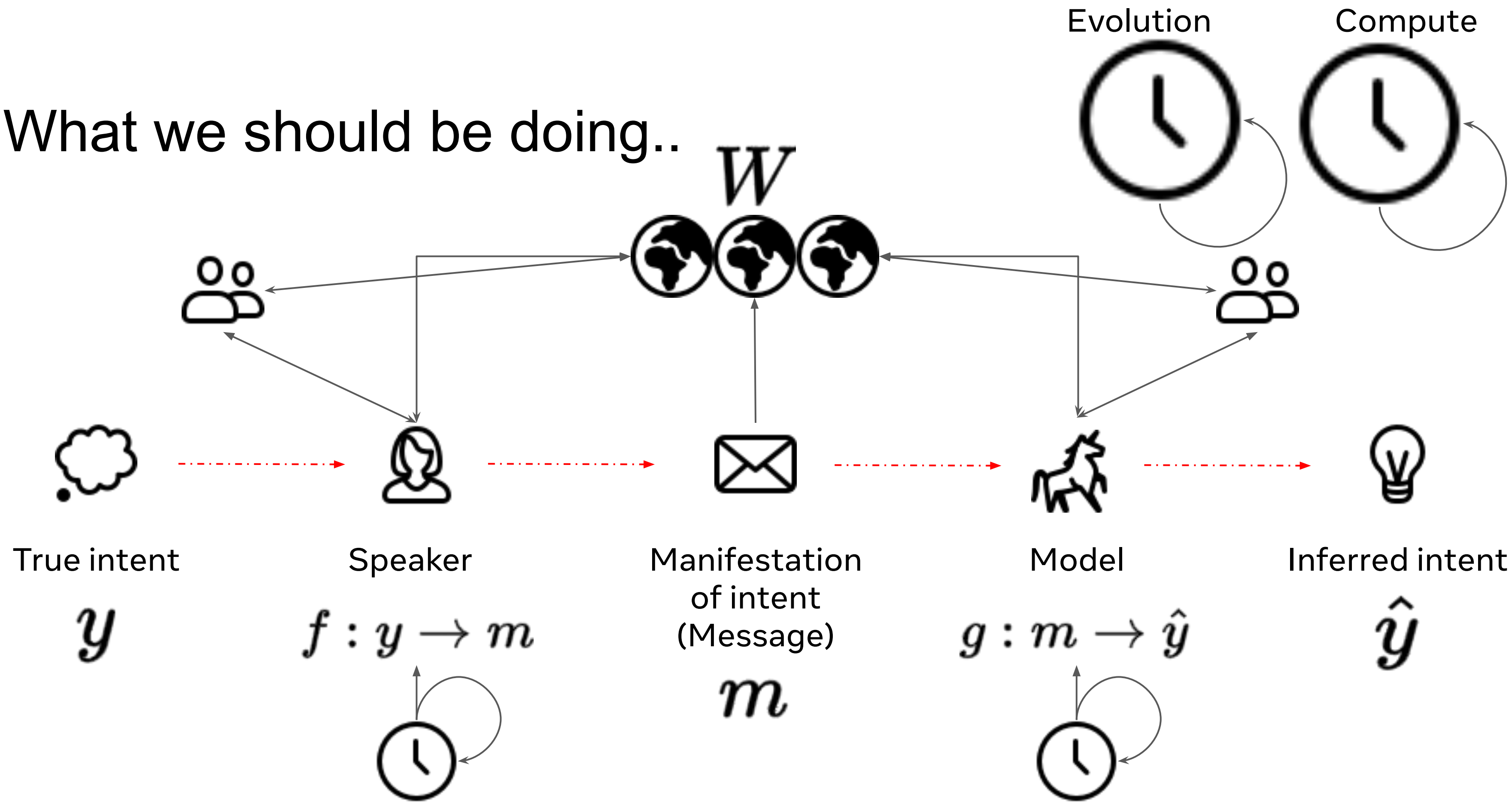
True intent

Text

Model

Label

What we should be doing..



Evolution

Compute

W

True intent

y

Speaker

$f : y \rightarrow m$

Manifestation of intent (Message)

m

Model

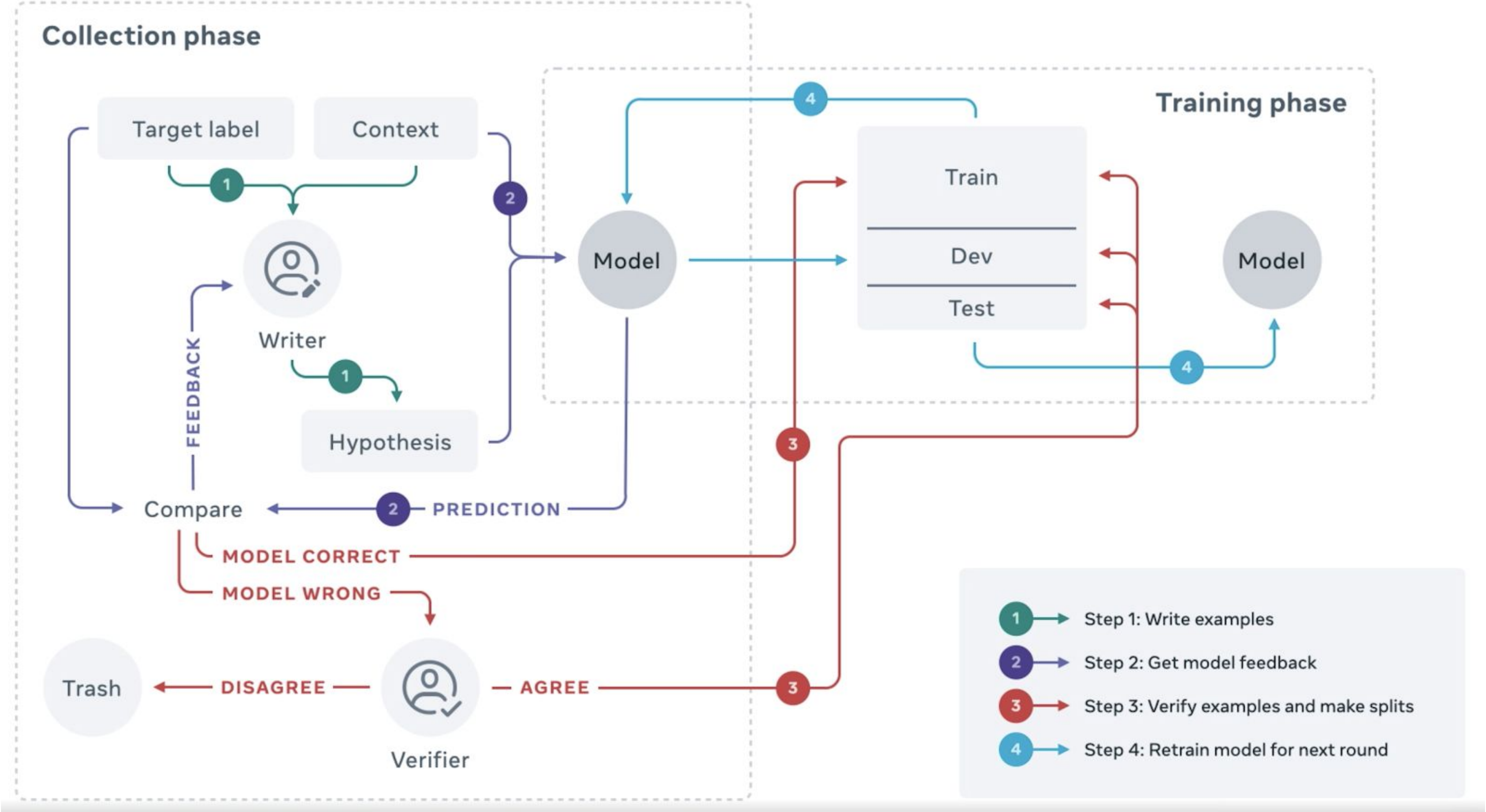
$g : m \rightarrow \hat{y}$

Inferred intent

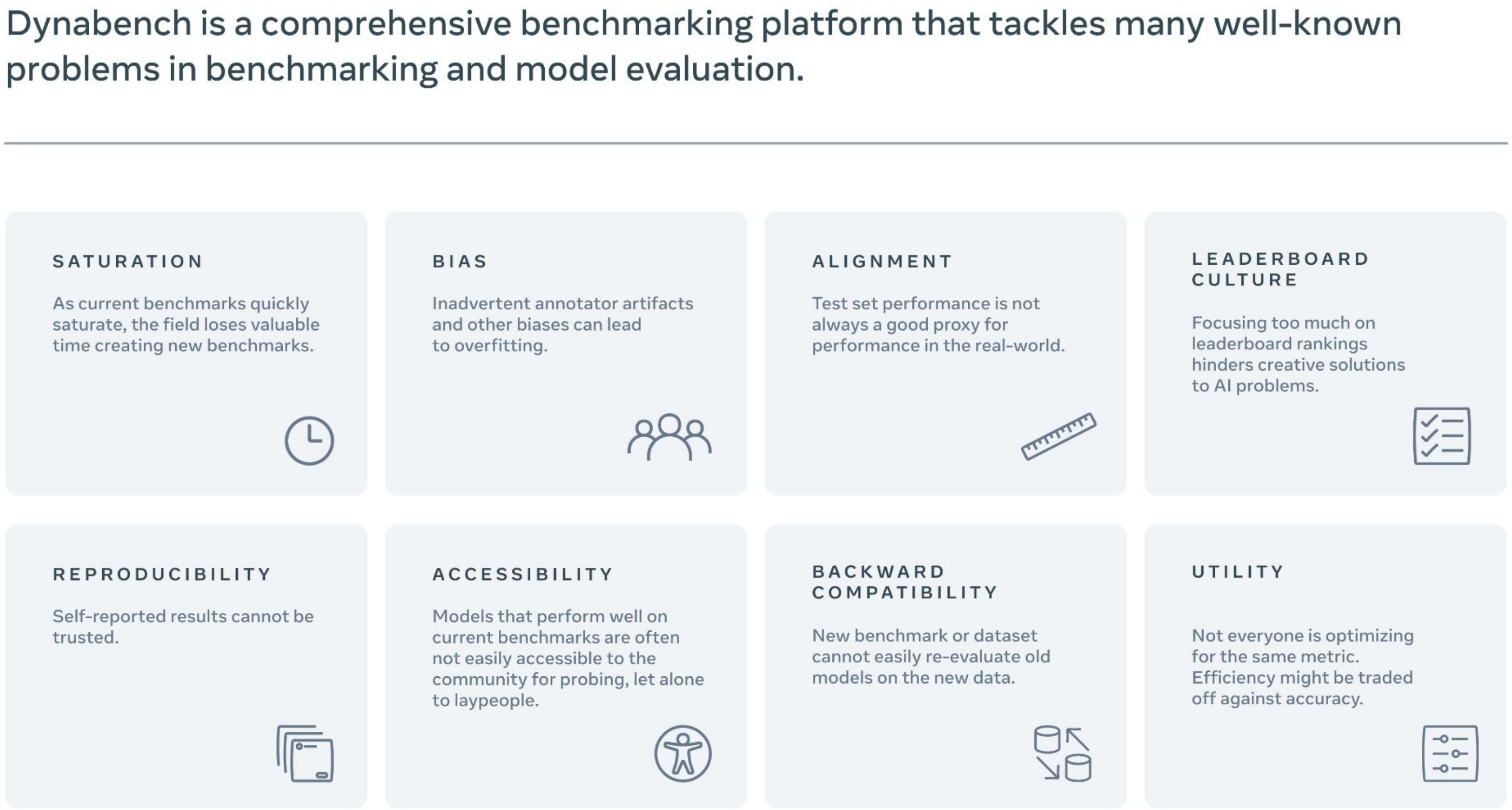
$\hat{y}$

Measuring not in the average case, but in the **worst case**.

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

Live demo

I was served rather the opposite of haute cuisine.

This restaurant was baad!

Summary

Language is about strong generalization. Humans expect this from other linguistic agents. But our systems might be right for the wrong reasons.

Evaluation is broken, we are haunted by Vapnik’s ghost but as a field are moving well beyond train-on-train-test-on-test in our deployments of AI (i.e., thanks to large scale pretraining and widely deployed systems).

We need to rethink this. Can we make cyclical progress and do more direct testing with humans in the loop?

Outline

1. Dynabench
    a. Overview
    b. Common Objections & Misconceptions
2. Progress in Dynamic Adversarial Data Collection
    a. Humans and Models in Loops
    b. Dynamic Adversarial Training Data
3. Adventures in Multimodal ML
    a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
    b. Foundation Models: FLAVA

Having all the answers

Dynabench is a research platform that changes over time. It’s not about having the right answers, it’s about changing the status quo: improving benchmarking will require experimentation.

Adversarialness

Dynabench does not require adversarialness. It’s straightforward to collect data with no model in the loop, or with many models in the loop, in an adversarial, collaborative, or other setting.

Language only

Dynabench is not about NLP. It supports many modalities and has tasks in multimodality, vision and (soon) speech. Tasks are currently in English but we'd like to change that.

Cost

Model-in-the-loop data collection can be more expensive, because it requires more creativity. It’s still unclear how this cost trades off against data quality - it might be worth it. There are ways to drive the cost down.

Naturalness and distributional shift

Open question how natural the data is, or how natural it will be in the long term.

Being “fair” to the model in the loop

Dynabench aims to measure how well models would hold up if they were deployed “in the wild”. The vMER metric is “fair” in that sense for comparing models that were both put in the loop.

Test sets created with a specific model in the loop are not fair to that model. I don’t really care about that sort of fairness: a) we could simply put a diverse ensemble of models in the loop; b) we are unfair at a specific point in time, but there will be many more other models to come - this is about measuring these future models.

All rounds count

Dynabench collects data over many rounds. ALL previous rounds, including non-adversarial ones, should be used for testing. An NLI model should perform well on all NLI datasets, adversarial or non-adversarial.

Adversarial filtering

Crowdworkers are paid for every example they generate, including the ones that did not fool any model. Non-model-fooling data is generally not discarded, because it’s still useful. Different tasks/datasets filter data in different ways.

Adversarially-collected test sets

- "*the constraint that a specified system must fail on the test examples makes it difficult to infer much from absolute measures of test-set performance: As long as a model makes any errors at all on any possible inputs, then we expect it to be possible to construct an adversarial test set against the model, and we expect the model to achieve zero test accuracy on that test set*"
  - **In Dynabench we advocate for looking at many metrics, including the time it takes a crowdworker to fool a model and how many times a crowdworker needs to try before succeeding (vMER). Not just "did you fool" but also "how easy was it to fool".**
  - **The errors are not just any kind of error, they are things that humans easily get right and agree on. You need to look at the data itself! (So this reduces to the naturalness objection?)**
- "*We can further infer that any models that are sufficiently similar to the adversary should also perform very poorly on this test set, regardless of their ability*"
  - **What about the example of BERT performing below chance on WSC while deBERTa gets 96%? What does "sufficiently similar" mean and who gets to determine that?**
  - **What about "all models" or "a representative subset of all models"?**
- "*Neither of these observations would tell us anything non-trivial about the actual abilities of the models*" double neg => "*These observations only tell us trivial things about the ability of models*"
  - **That feels a bit strong?**

Absolute performance numbers

- “*Absolute performance numbers on adversarially-collected test sets are meaningless as measures of model capabilities*”

What makes a performance number meaningless? This seems to assume (again) that we haven’t looked at the actual data. If a human can easily get the right answer and humans (mostly) agree about a given answer, and the example is natural, why should performance on that example be meaningless?

If this is about naturalness of data, are Turker-collected free-form test sets guaranteed to be more natural?

The scientific process

“significant further work is needed to avoid catastrophe. This will be difficult to achieve without a clear accounting of the abilities and limitations of current and plausible near-future systems”

Exactly this! We should, as a field, work hard to develop a clearer picture of our current capabilities and fix measurement. If we are saturating benchmarks, while we know we have all these issues, something is wrong.

In other words: **be careful when you deploy a model and think about what you’re doing**. We want the world to realize that evaluation is something we should take more seriously. If we can measure better, we can make better progress. This will happen by building on previous work, in cycles of progress where benchmarks “saturate” and are replaced by better once. Science will do its job if we are open to new ideas.

Outline

1. Dynabench
    a. Overview
    b. Common Objections & Misconceptions
2. **Progress in Dynamic Adversarial Data Collection**
    a. Humans and Models in Loops
    b. Dynamic Adversarial Training Data
3. Adventures in Multimodal ML
    a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
    b. Foundation Models: FLAVA

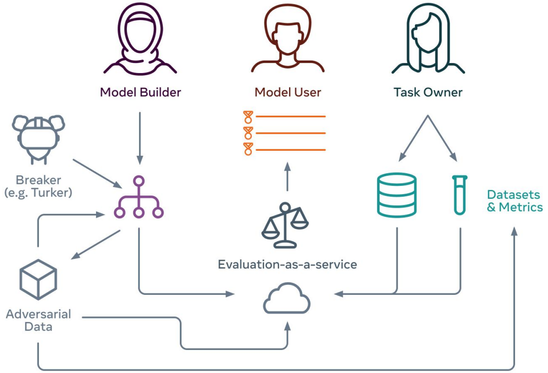
Broader research program

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. (2021). **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. (2021). **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. (2021). **Analyzing Dynamic Adversarial Training Data in the Limit**

flores

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. (2021). **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**

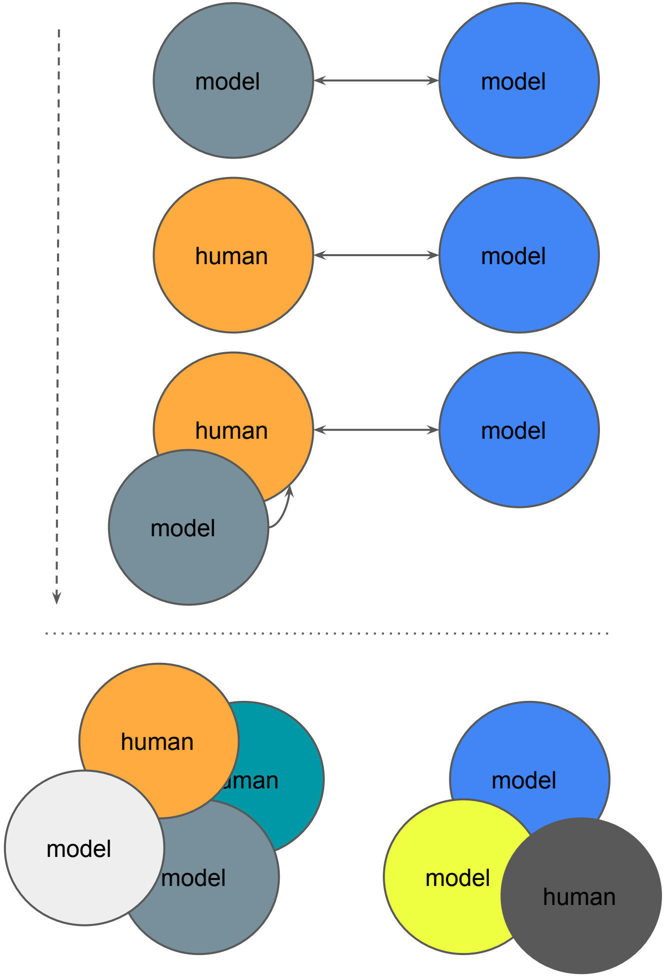
- Prasad et al. (Blackbox21). **To what extent do human explanations of model behavior align with actual behavior?**
- Ma, Ethayarajh, Thrush et al. (NeurIPS21). **Dynaboard: A Holistic Evaluation-As-A-Service Benchmarking Platform**
- Wenzek et al. (2021). **Findings of the WMT 2021 Shared Task on Large-Scale 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. (2021). **Analyzing Dynamic Adversarial Training Data in the Limit**

Humans and models in loops

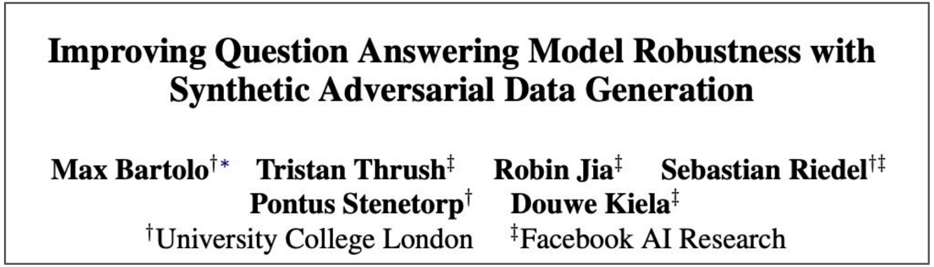
- Question 1:
    - Instead of human-adversarial, how much can we improve things by just being model-adversarial using human-adversarial data?
- Question 2:
    - Can generative (adversarial) models help humans fool discriminative models?

Work by **Max Bartolo** et al.



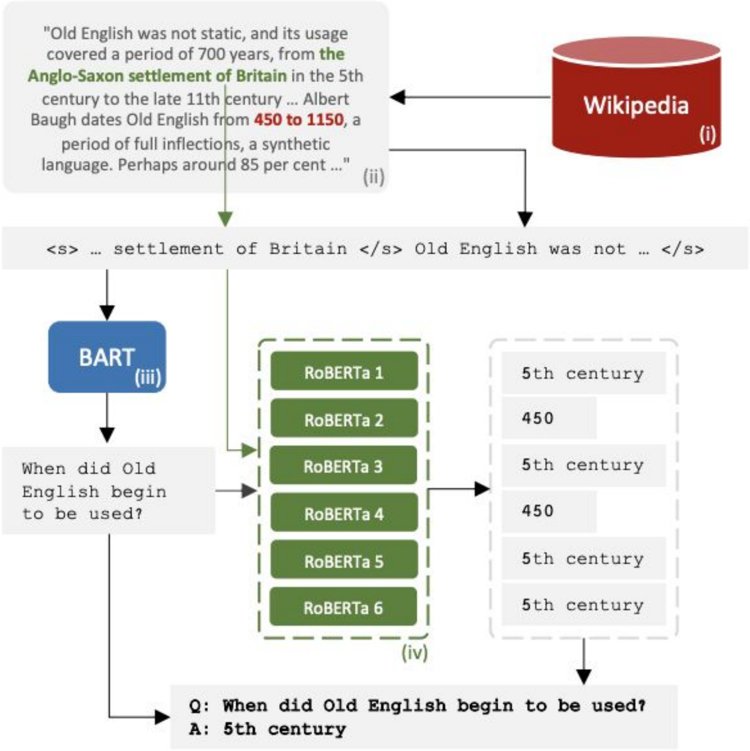
Improving QA robustness with synthetic adversarial data

- 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



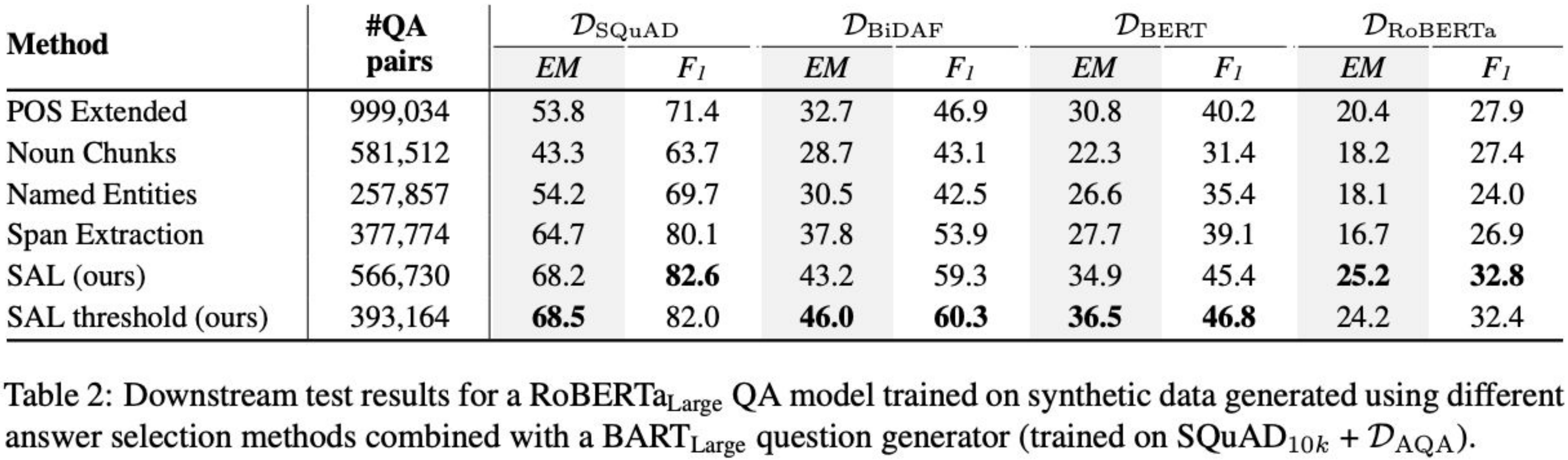
"Old English was not static, and its usage covered a period of 700 years, from the Anglo-Saxon settlement of Britain in the 5th century to the late 11th century … Albert Baugh dates Old English from 450 to 1150, a period of full inflections, a synthetic language. Perhaps around 85 per cent …"

```
<s> … settlement of Britain </s> Old English was not … </s>
```

```
When did Old
English begin
to be used?
```

```
Q: When did Old English begin to be used?
A: 5th century
```

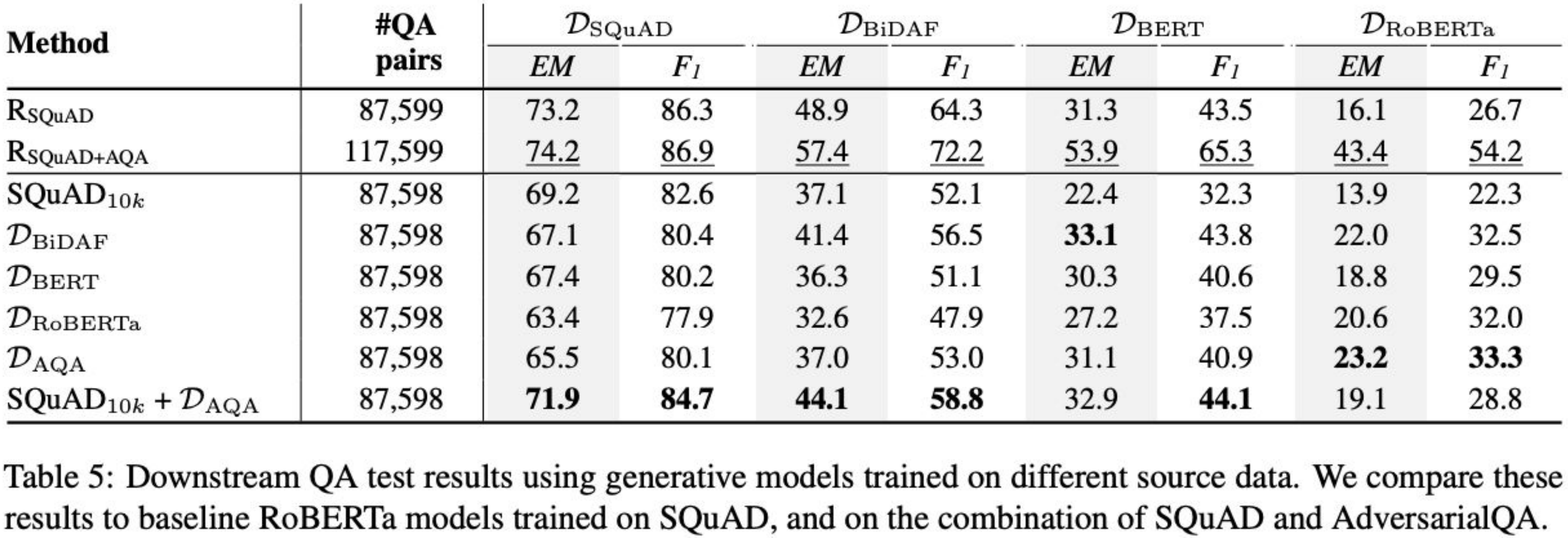
Step 2: Answer selection



Method	#QA pairs	$\mathcal{D}_{\text{SQuAD}}$		$\mathcal{D}_{\text{BiDAF}}$		$\mathcal{D}_{\text{BERT}}$		$\mathcal{D}_{\text{RoBERTa}}$	
		EM	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1
POS Extended	999,034	53.8	71.4	32.7	46.9	30.8	40.2	20.4	27.9
Noun Chunks	581,512	43.3	63.7	28.7	43.1	22.3	31.4	18.2	27.4
Named Entities	257,857	54.2	69.7	30.5	42.5	26.6	35.4	18.1	24.0
Span Extraction	377,774	64.7	80.1	37.8	53.9	27.7	39.1	16.7	26.9
SAL (ours)	566,730	68.2	**82.6**	43.2	59.3	34.9	45.4	**25.2**	**32.8**
SAL threshold (ours)	393,164	**68.5**	82.0	**46.0**	**60.3**	**36.5**	**46.8**	24.2	32.4

Table 2: Downstream test results for a $\text{RoBERTa}_{\text{Large}}$ QA model trained on synthetic data generated using different answer selection methods combined with a $\text{BART}_{\text{Large}}$ question generator (trained on $\text{SQuAD}_{10k} + \mathcal{D}_{\text{AQA}}$).

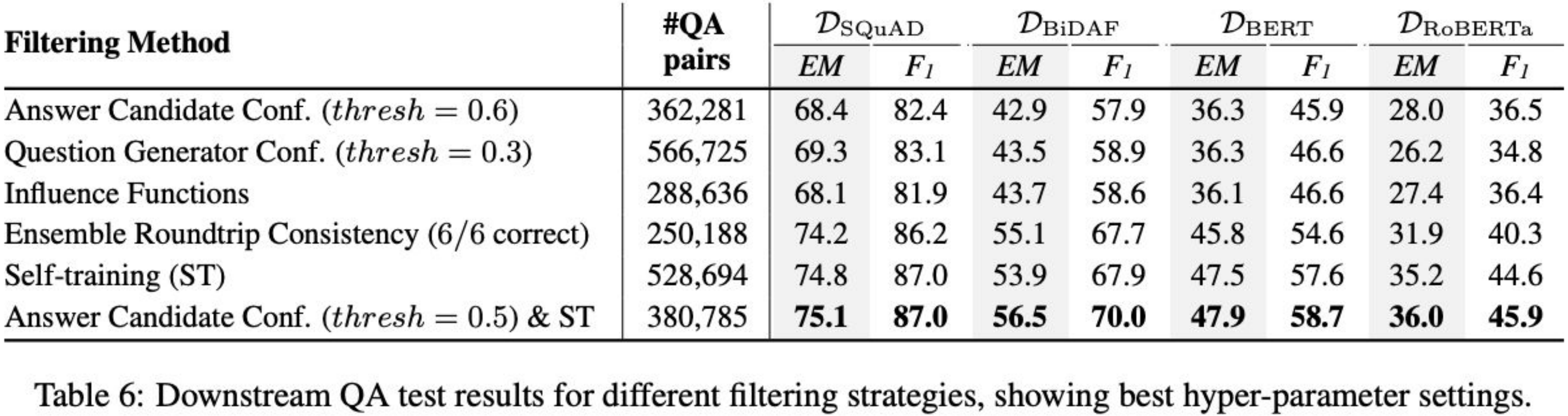
Step 3: Question generation



Method	#QA pairs	$\mathcal{D}_{\text{SQuAD}}$		$\mathcal{D}_{\text{BiDAF}}$		$\mathcal{D}_{\text{BERT}}$		$\mathcal{D}_{\text{RoBERTa}}$	
		EM	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1
R_{SQuAD}	87,599	73.2	86.3	48.9	64.3	31.3	43.5	16.1	26.7
$\text{R}_{\text{SQuAD+AQA}}$	117,599	74.2	86.9	57.4	72.2	53.9	65.3	43.4	54.2
SQuAD_{10k}	87,598	69.2	82.6	37.1	52.1	22.4	32.3	13.9	22.3
$\mathcal{D}_{\text{BiDAF}}$	87,598	67.1	80.4	41.4	56.5	**33.1**	43.8	22.0	32.5
$\mathcal{D}_{\text{BERT}}$	87,598	67.4	80.2	36.3	51.1	30.3	40.6	18.8	29.5
$\mathcal{D}_{\text{RoBERTa}}$	87,598	63.4	77.9	32.6	47.9	27.2	37.5	20.6	32.0
$\mathcal{D}_{\text{AQA}}$	87,598	65.5	80.1	37.0	53.0	31.1	40.9	**23.2**	**33.3**
$\text{SQuAD}_{10k} + \mathcal{D}_{\text{AQA}}$	87,598	**71.9**	**84.7**	**44.1**	**58.8**	32.9	**44.1**	19.1	28.8

Table 5: Downstream QA test results using generative models trained on different source data. We compare these results to baseline RoBERTa models trained on SQuAD, and on the combination of SQuAD and AdversarialQA.

Step 4: Filtering and self-training

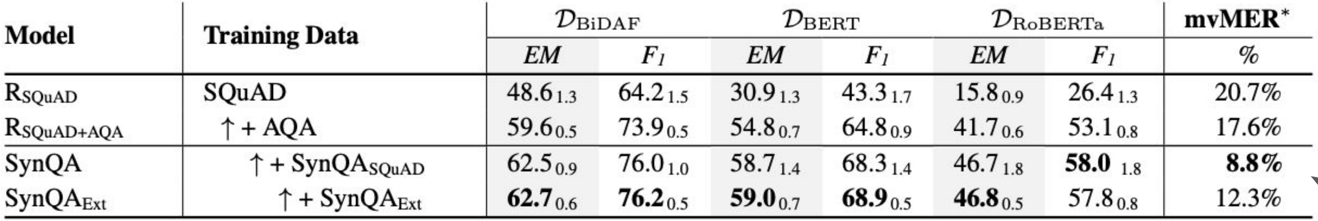


Filtering Method	**#QA pairs**	$\mathcal{D}_{\text{SQuAD}}$		$\mathcal{D}_{\text{BiDAF}}$		$\mathcal{D}_{\text{BERT}}$		$\mathcal{D}_{\text{RoBERTa}}$	
		EM	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1
Answer Candidate Conf. ($thresh = 0.6$)	362,281	68.4	82.4	42.9	57.9	36.3	45.9	28.0	36.5
Question Generator Conf. ($thresh = 0.3$)	566,725	69.3	83.1	43.5	58.9	36.3	46.6	26.2	34.8
Influence Functions	288,636	68.1	81.9	43.7	58.6	36.1	46.6	27.4	36.4
Ensemble Roundtrip Consistency ($6/6$ correct)	250,188	74.2	86.2	55.1	67.7	45.8	54.6	31.9	40.3
Self-training (ST)	528,694	74.8	87.0	53.9	67.9	47.5	57.6	35.2	44.6
Answer Candidate Conf. ($thresh = 0.5$) & ST	380,785	**75.1**	**87.0**	**56.5**	**70.0**	**47.9**	**58.7**	**36.0**	**45.9**

Table 6: Downstream QA test results for different filtering strategies, showing best hyper-parameter settings.

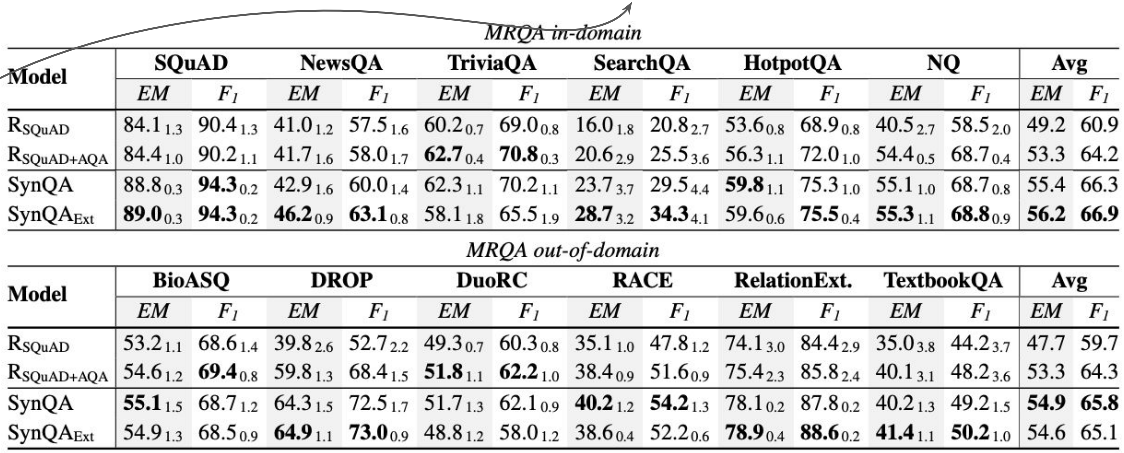
Findings

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



Model	**Training Data**	$\mathcal{D}_{\text{BiDAF}}$		$\mathcal{D}_{\text{BERT}}$		$\mathcal{D}_{\text{RoBERTa}}$		**mvMER***
		EM	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	%
R$_{\text{SQuAD}}$	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$_{\text{SQuAD+AQA}}$	↑ + 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$_{\text{SQuAD}}$	$62.5_{0.9}$	$76.0_{1.0}$	$58.7_{1.4}$	$68.3_{1.4}$	$46.7_{1.8}$	**58.0** $_{1.8}$	**8.8%**
SynQA$_{\text{Ext}}$	↑ + SynQA$_{\text{Ext}}$	**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%

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



MRQA in-domain

Model	**SQuAD**		**NewsQA**		**TriviaQA**		**SearchQA**		**HotpotQA**		**NQ**		**Avg**	
	EM	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*
R$_{\text{SQuAD}}$	$84.1_{1.3}$	$90.4_{1.3}$	$41.0_{1.2}$	$57.5_{1.6}$	$60.2_{0.7}$	$69.0_{0.8}$	$16.0_{1.8}$	$20.8_{2.7}$	$53.6_{0.8}$	$68.9_{0.8}$	$40.5_{2.7}$	$58.5_{2.0}$	49.2	60.9
R$_{\text{SQuAD+AQA}}$	$84.4_{1.0}$	$90.2_{1.1}$	$41.7_{1.6}$	$58.0_{1.7}$	**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$_{\text{Ext}}$	**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**

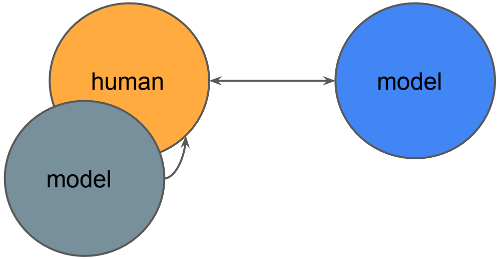
MRQA out-of-domain

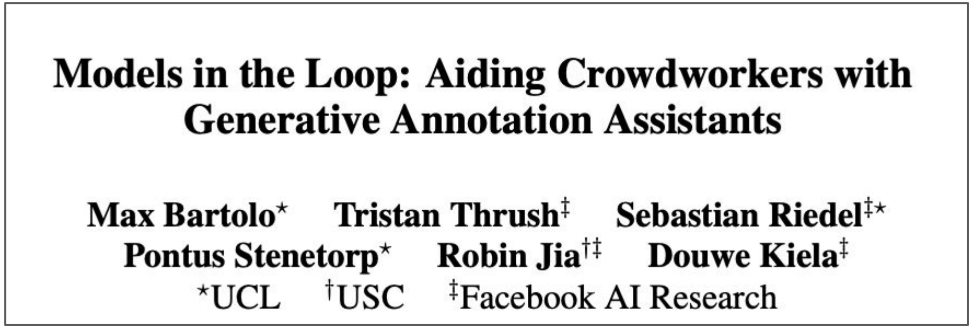
Model	**BioASQ**		**DROP**		**DuoRC**		**RACE**		**RelationExt.**		**TextbookQA**		**Avg**	
	EM	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*	*EM*	*F$_1$*
R$_{\text{SQuAD}}$	$53.2_{1.1}$	$68.6_{1.4}$	$39.8_{2.6}$	$52.7_{2.2}$	$49.3_{0.7}$	$60.3_{0.8}$	$35.1_{1.0}$	$47.8_{1.2}$	$74.1_{3.0}$	$84.4_{2.9}$	$35.0_{3.8}$	$44.2_{3.7}$	47.7	59.7
R$_{\text{SQuAD+AQA}}$	$54.6_{1.2}$	**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$_{\text{Ext}}$	$54.9_{1.3}$	$68.5_{0.9}$	**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

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?

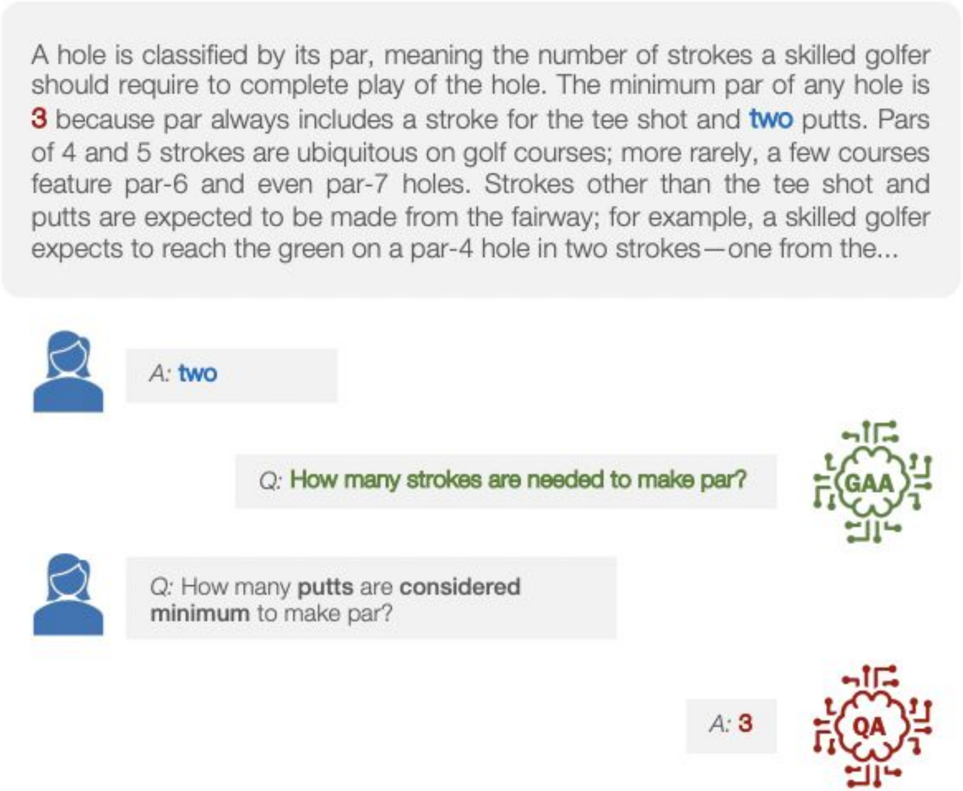




Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants

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**

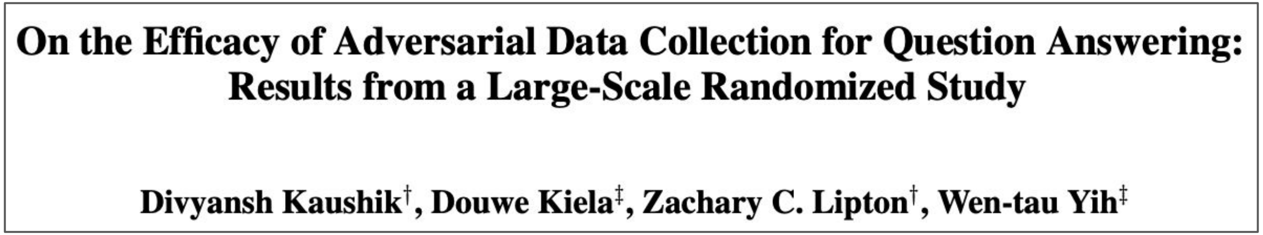
Q: **How many strokes are needed to make par?**

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

A: **3**

Standard (SDC) vs Adversarial (ADC) Data Collection

- Earlier finding: “Across a variety of [Question Answering] models and datasets, we find that models trained on adversarial data usually perform better on other adversarial datasets but worse on a diverse collection of out-of-domain evaluation sets.” (**Divyansh Kaushik** et al. ACL 2021)



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‡

Standard (SDC) vs Adversarial (ADC) Data Collection

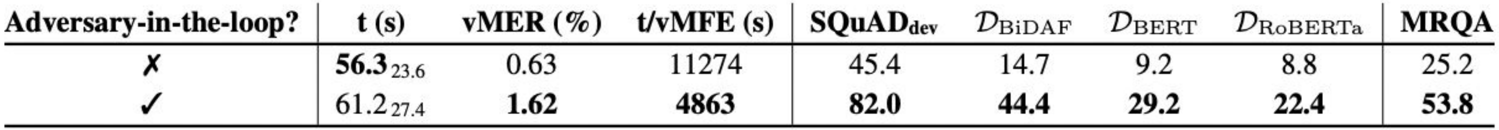
- New finding:
  (the preliminary take-away on smallish data is – be careful with setup)

Validated model error rate

Domain generalization

Median time per example

Time per model-fooling ex



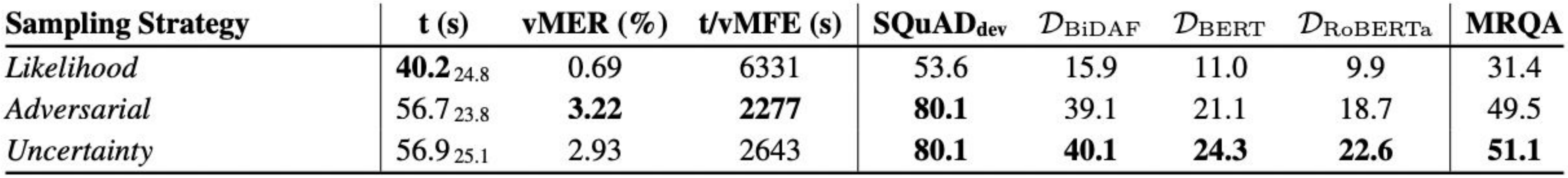
Adversary-in-the-loop?	t (s)	vMER (%)	t/vMFE (s)	SQuAD$_{\text{dev}}$	$\mathcal{D}_{\text{BiDAF}}$	$\mathcal{D}_{\text{BERT}}$	$\mathcal{D}_{\text{RoBERTa}}$	MRQA
✗	**56.3** $_{23.6}$	0.63	11274	45.4	14.7	9.2	8.8	25.2
✓	61.2 $_{27.4}$	**1.62**	**4863**	**82.0**	**44.4**	**29.2**	**22.4**	**53.8**

Standard

Adversarial QA

Improving SDC

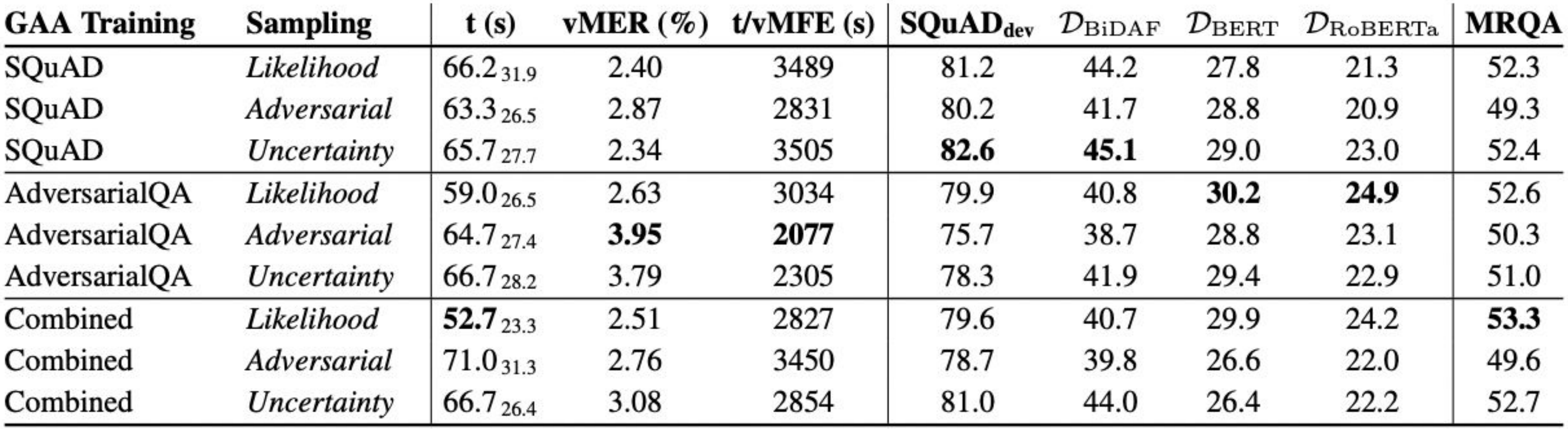
- Using a Generator-in-the-loop makes standard data collection more faster and much higher quality.
- Sampling strategies:
  a. Likelihood: sample candidates according to the generative model’s overall likelihood.
  b. Adversarial: sample questions according to the lowest F1 score queried against a QA model.
  c. Uncertainty: select generated questions in order of the least span selection confidence when queried against a QA model.



Sampling Strategy	**t (s)**	**vMER (%)**	**t/vMFE (s)**	**SQuAD**$_{\textbf{dev}}$	$\mathcal{D}_{\text{BiDAF}}$	$\mathcal{D}_{\text{BERT}}$	$\mathcal{D}_{\text{RoBERTa}}$	**MRQA**
Likelihood	**40.2**$_{24.8}$	0.69	6331	53.6	15.9	11.0	9.9	31.4
Adversarial	56.7$_{23.8}$	**3.22**	**2277**	**80.1**	39.1	21.1	18.7	49.5
Uncertainty	56.9$_{25.1}$	2.93	2643	**80.1**	**40.1**	**24.3**	**22.6**	**51.1**

Improving ADC

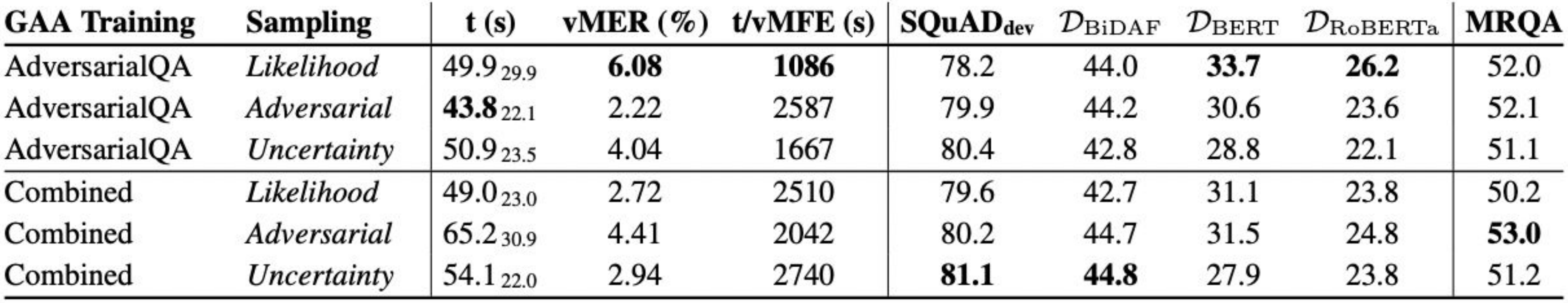
- Using a Generator-in-the-loop makes adversarial data collection as fast as standard data collection, with higher quality and better domain generalization.
- Generative Annotation Assistant (GAA) trained on SQuad, AQA or Combined.



GAA Training	Sampling	t (s)	vMER (%)	t/vMFE (s)	SQuAD$_{\text{dev}}$	$\mathcal{D}_{\text{BiDAF}}$	$\mathcal{D}_{\text{BERT}}$	$\mathcal{D}_{\text{RoBERTa}}$	MRQA
SQuAD	*Likelihood*	$66.2_{31.9}$	2.40	3489	81.2	44.2	27.8	21.3	52.3
SQuAD	*Adversarial*	$63.3_{26.5}$	2.87	2831	80.2	41.7	28.8	20.9	49.3
SQuAD	*Uncertainty*	$65.7_{27.7}$	2.34	3505	**82.6**	**45.1**	29.0	23.0	52.4
AdversarialQA	*Likelihood*	$59.0_{26.5}$	2.63	3034	79.9	40.8	**30.2**	**24.9**	52.6
AdversarialQA	*Adversarial*	$64.7_{27.4}$	**3.95**	**2077**	75.7	38.7	28.8	23.1	50.3
AdversarialQA	*Uncertainty*	$66.7_{28.2}$	3.79	2305	78.3	41.9	29.4	22.9	51.0
Combined	*Likelihood*	$\mathbf{52.7}_{23.3}$	2.51	2827	79.6	40.7	29.9	24.2	**53.3**
Combined	*Adversarial*	$71.0_{31.3}$	2.76	3450	78.7	39.8	26.6	22.0	49.6
Combined	*Uncertainty*	$66.7_{26.4}$	3.08	2854	81.0	44.0	26.4	22.2	52.7

Improving ADC further

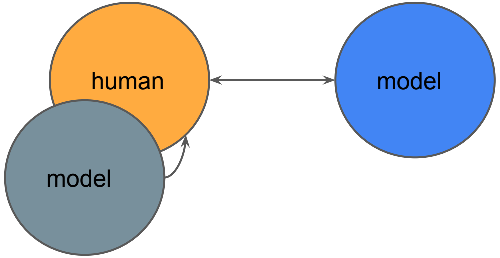
- If you do “answer prompting” where you don’t force annotators to pick the answer but suggest one, ADC gets even faster and much higher quality.
- Starting point, traditional data collection: vMER=0.63 with t=56.3
- End point, ADC with GAA: vMER=6.08 with t=43.8

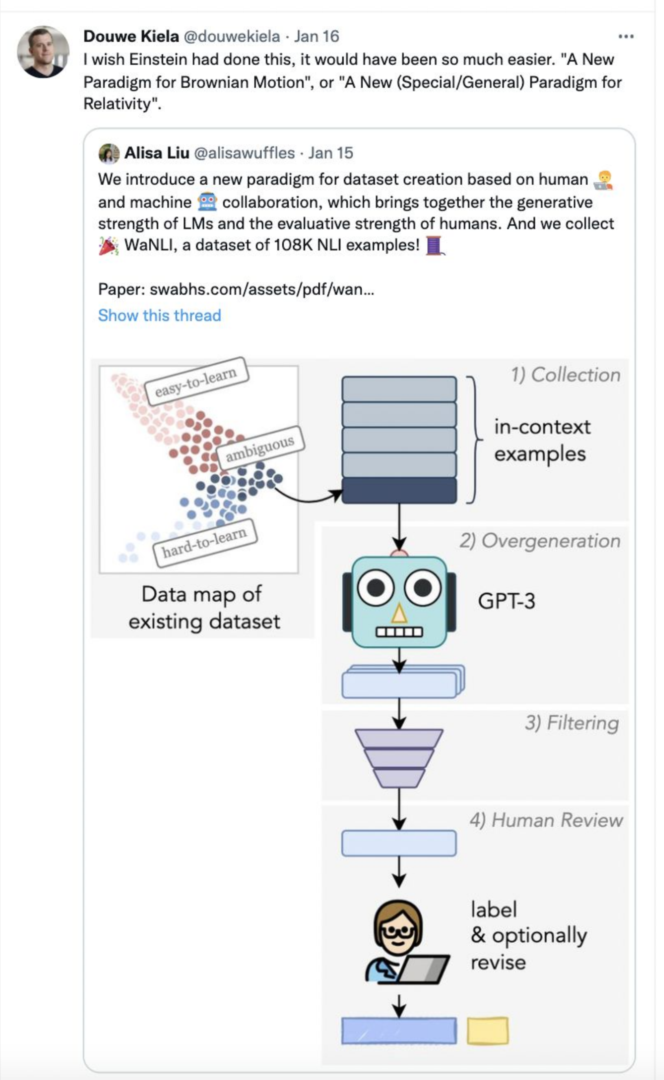


GAA Training	Sampling	t (s)	vMER (%)	t/vMFE (s)	SQuAD$_{\text{dev}}$	$\mathcal{D}_{\text{BiDAF}}$	$\mathcal{D}_{\text{BERT}}$	$\mathcal{D}_{\text{RoBERTa}}$	MRQA
AdversarialQA	*Likelihood*	$49.9_{29.9}$	**6.08**	**1086**	78.2	44.0	**33.7**	**26.2**	52.0
AdversarialQA	*Adversarial*	$\mathbf{43.8}_{22.1}$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

A “new paradigm”?

- ModelS in LoopS:
  a. **Yes**, we can collect much higher quality data than static data using this method.
  b. **Yes**, we can collect higher quality data than regular human-and-model-in-the-loop.
  c. **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...

Show this thread

Recent work out of the Dynabench team

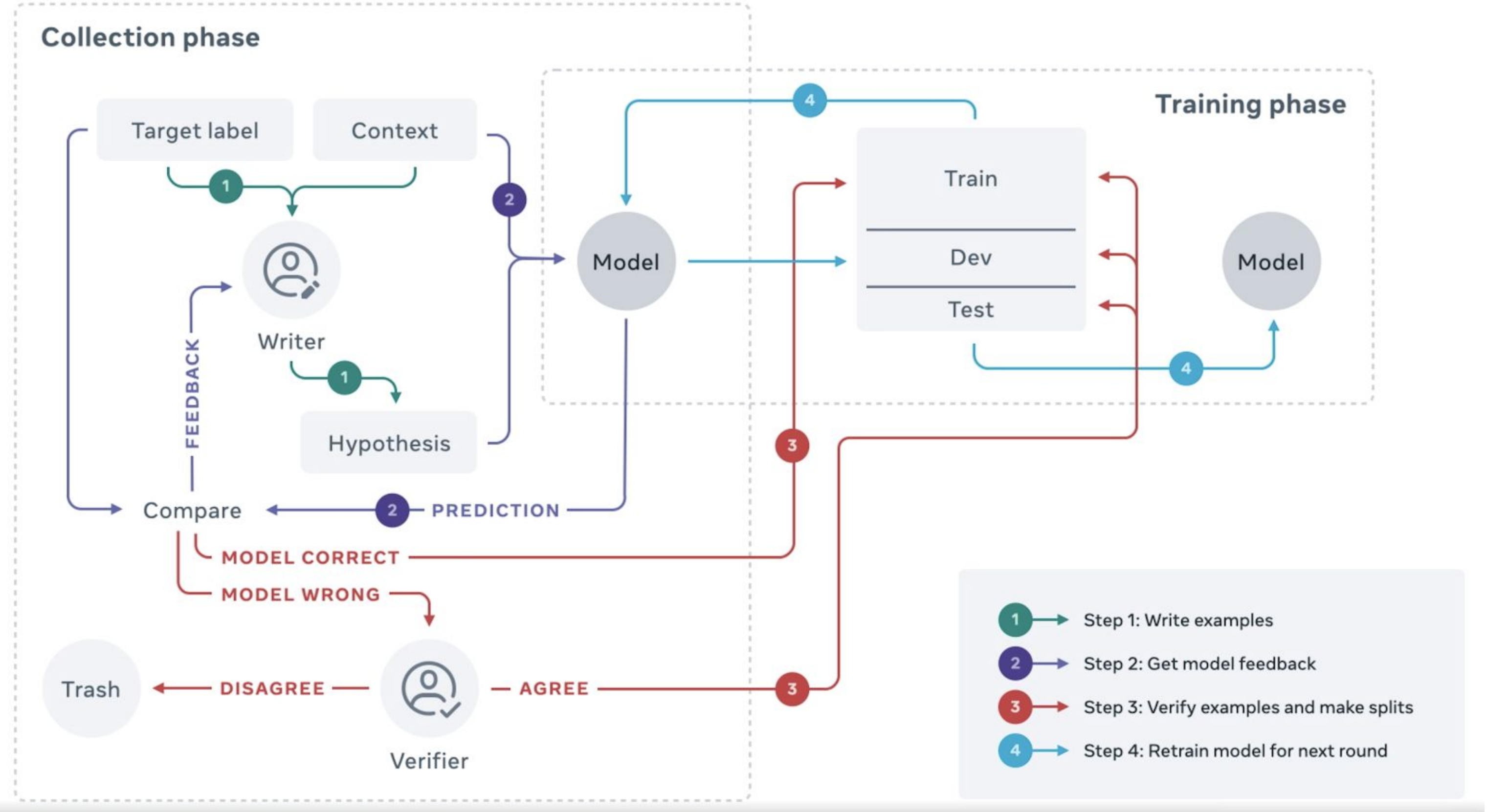
- Kiela et al. (NAACL21). **Dynabench: Rethinking Benchmarking in NLP**

- 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. (2021). **Analyzing Dynamic Adversarial Training Data in the Limit**

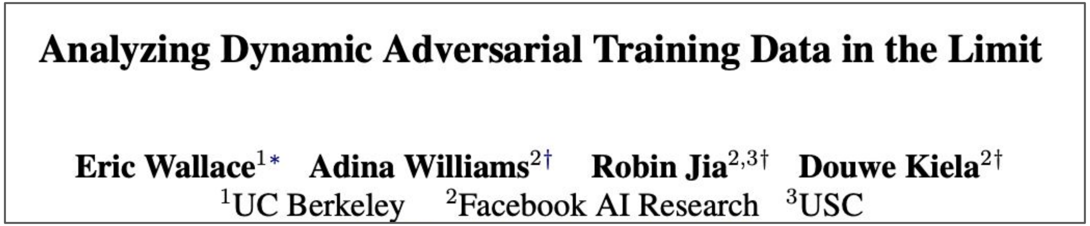
Dynamic adversarial data collection in the limit



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

Work by **Eric Wallace** et al.

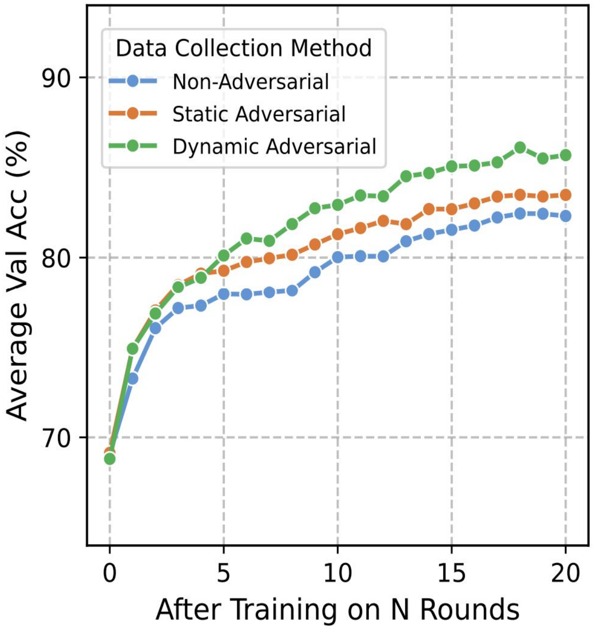


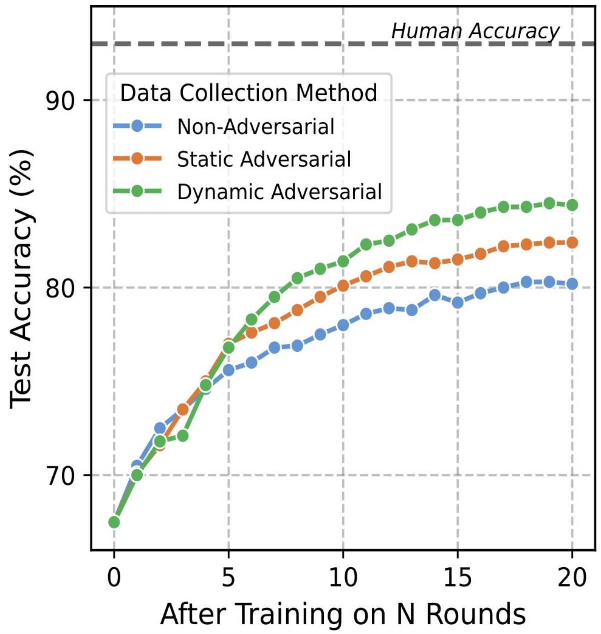
Analyzing Dynamic Adversarial Training Data in the Limit

Eric Wallace[1*] **Adina Williams**[2†] **Robin Jia**[2,3†] **Douwe Kiela**[2†]
[1]UC Berkeley [2]Facebook AI Research [3]USC

Findings: A virtuous cycle

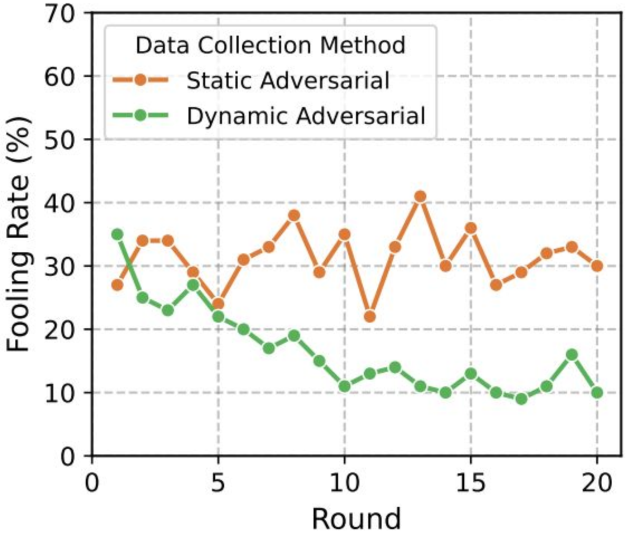
Promising results when exploring Dynamic Adversarial Data Collection in the limit:

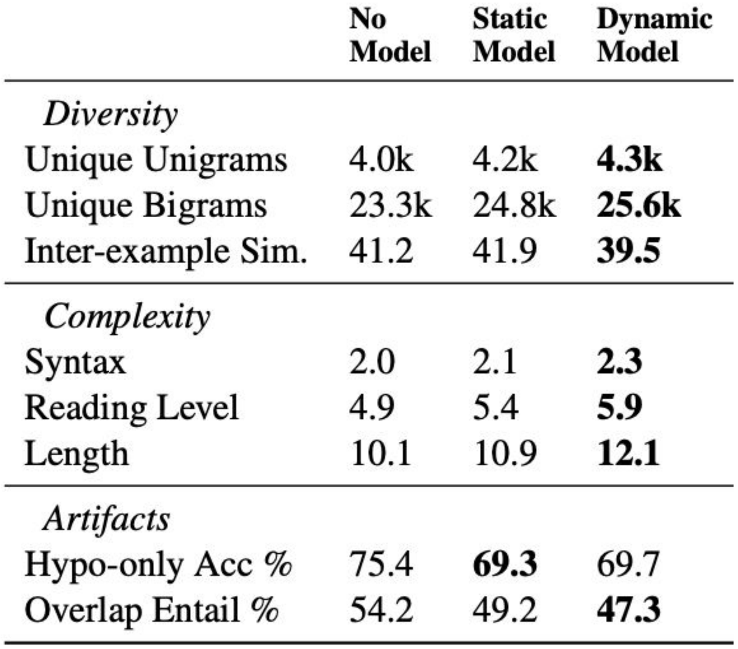




Findings: Diversity is key

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





	No Model	Static Model	Dynamic Model
Diversity			
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**
Complexity			
Syntax	2.0	2.1	**2.3**
Reading Level	4.9	5.4	**5.9**
Length	10.1	10.9	**12.1**
Artifacts			
Hypo-only Acc %	75.4	**69.3**	69.7
Overlap Entail %	54.2	49.2	**47.3**

Method take-aways

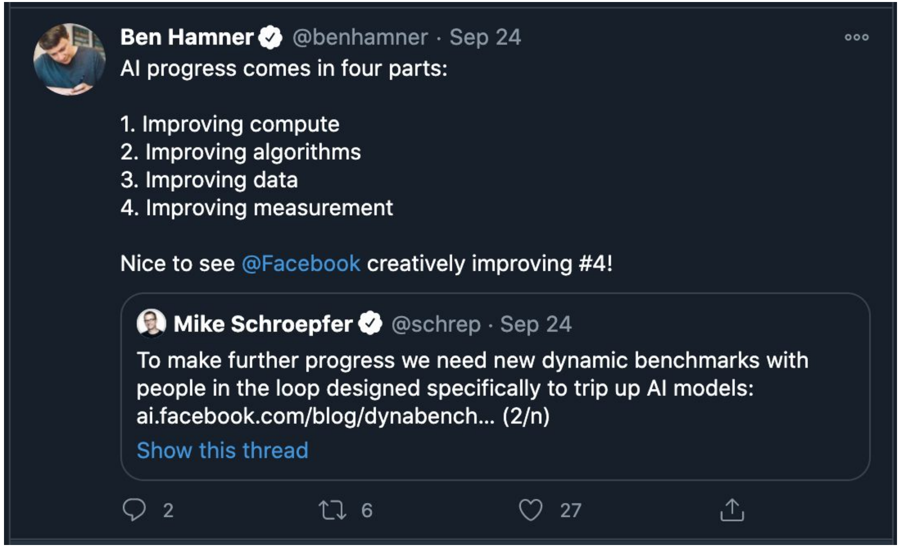
- ADC looks like a good alternative for traditional crowdworker data collection.
- This is a nice side benefit, considering that the original goal was evaluation.
- Human-and-model-in-the-loop / human feedback holds a lot of promise (see OpenAI’s recent papers on this as well, or “red teaming”)
- Further 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 comes next?

We’ve opened everything up:

- Fully open source (MIT licensed)
- Dynatask: Anyone can add tasks
- Keep growing the community
- Keep pushing the boundaries

- Exploring synergies with 🤗 ?



Ben Hamner @benhamner · Sep 24

AI progress comes in four parts:

1. Improving compute
2. Improving algorithms
3. Improving data
4. Improving measurement

Nice to see @Facebook creatively improving #4!

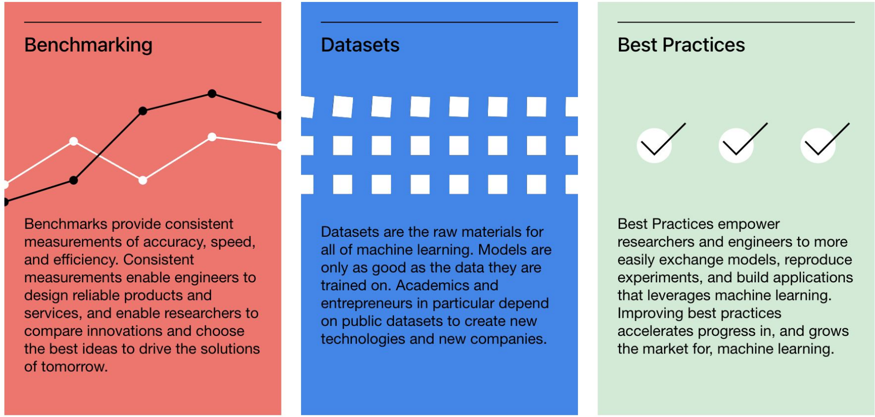
Mike Schroepfer @schrep · Sep 24

To make further progress we need new dynamic benchmarks with people in the loop designed specifically to trip up AI models: ai.facebook.com/blog/dynabench... (2/n)

Show this thread

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:



Benchmarking

Benchmarks provide consistent measurements of accuracy, speed, and efficiency. Consistent measurements enable engineers to design reliable products and services, and enable researchers to compare innovations and choose the best ideas to drive the solutions of tomorrow.

Datasets

Datasets are the raw materials for all of machine learning. Models are only as good as the data they are trained on. Academics and entrepreneurs in particular depend on public datasets to create new technologies and new companies.

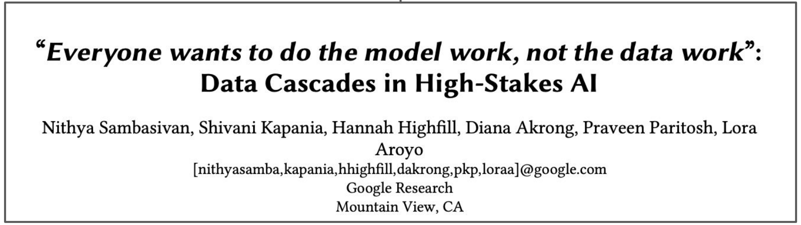
Best Practices

Best Practices empower researchers and engineers to more easily exchange models, reproduce experiments, and build applications that leverages machine learning. Improving best practices accelerates progress in, and grows the market for, machine learning.

Announcement and Call for Participation

December 14, 2021

Whitepaper - Working Group - Email List



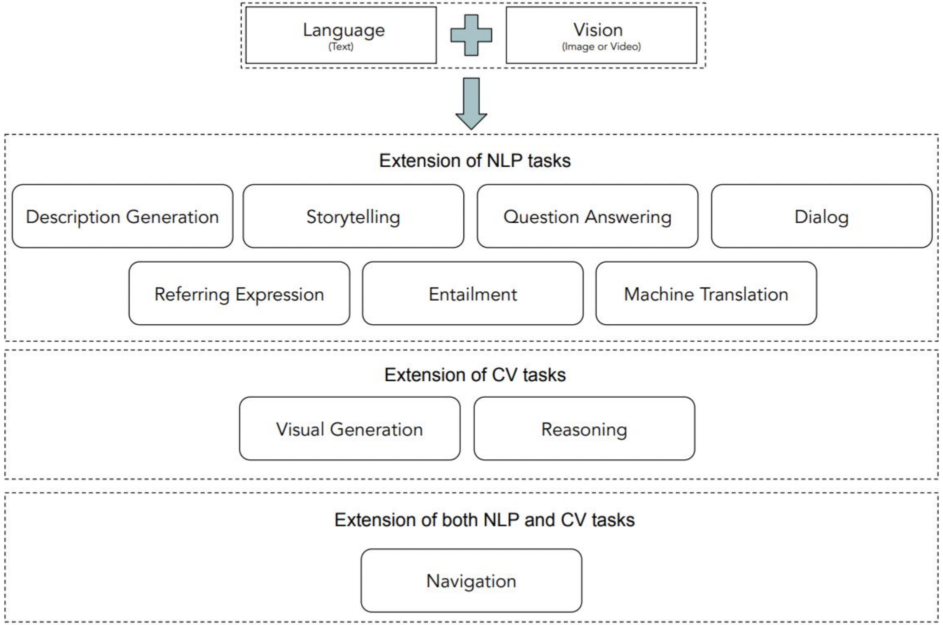
"Everyone wants to do the model work, not the data work":
Data Cascades in High-Stakes AI

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

Outline

1. Dynabench
    a. Overview
    b. Common Objections & Misconceptions
2. Progress in Dynamic Adversarial Data Collection
    a. Humans and Models in Loops
    b. Dynamic Adversarial Training Data
3. Adventures in Multimodal ML
    a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
    b. Foundation Models: FLAVA

Vision and Language Tasks & Datasets



Source: Mogadala et al 2021

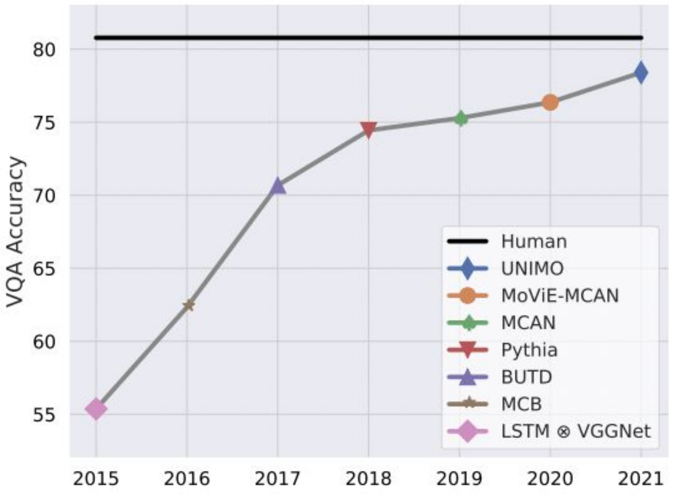
Citations as of 4/4/22:

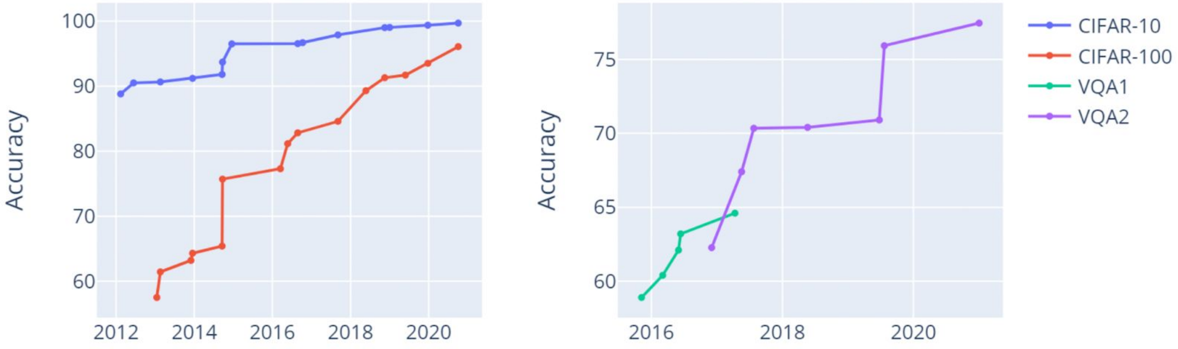
- VQA/VQA2 - 3409/1227
- Visual Genome - 2776
- COCO - 1240 (22724)
- Flickr30k - 908
- VisDial - 715
- NLVR2 - 189

Power law distribution with VQA as the dominant task.

Visual Question Answering

VQA is plateauing and arguably/almost saturated





a) **Consecutive saturation**
Image classification on
CIFAR-10 and CIFAR-100

b) **Consecutive versions**
Visual question answering on
VQA1 and VQA2

Source: Barbosa-Silva et al 2022

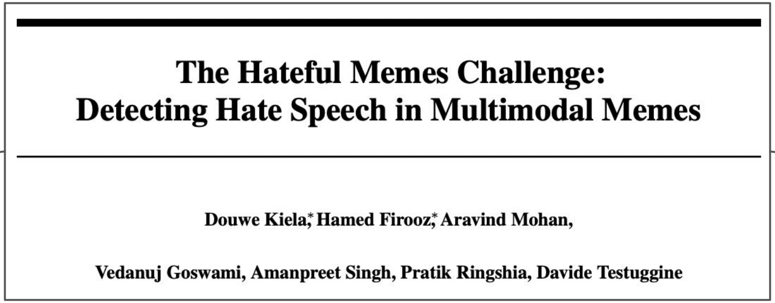
What do we want?

- Ideally, evaluation sets are:
  - High-quality and without error
  - Not too expensive
  - Not too easy
  - Discriminative between models
  - Realistic and representative of practical use-cases
  - Straightforwardly measured
- Multimodal evaluation sets, in addition, ideally are:
  - Not dominated by a specific modality
  - Actually measuring multimodal rather than unimodal performance

    (cf “making the V in VQA matter”)

Multimodal Evaluation

NeurIPS 2020

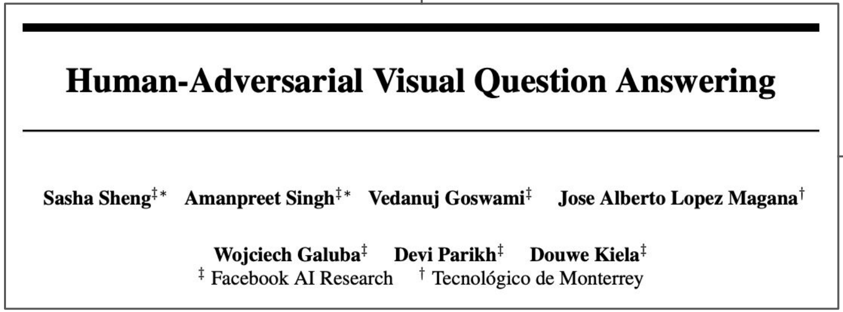


The Hateful Memes Challenge:
Detecting Hate Speech in Multimodal Memes

Douwe Kiela[*], Hamed Firooz[*], Aravind Mohan,

Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine

NeurIPS 2021

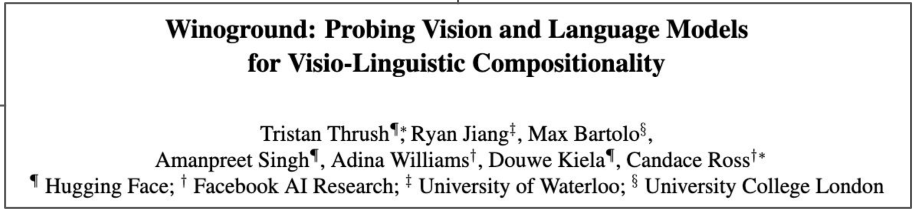


Human-Adversarial Visual Question Answering

Sasha Sheng[‡*] Amanpreet Singh[‡*] Vedanuj Goswami[‡] Jose Alberto Lopez Magana[†]

Wojciech Galuba[‡] Devi Parikh[‡] Douwe Kiela[‡]
[‡] Facebook AI Research [†] Tecnológico de Monterrey

CVPR 2022

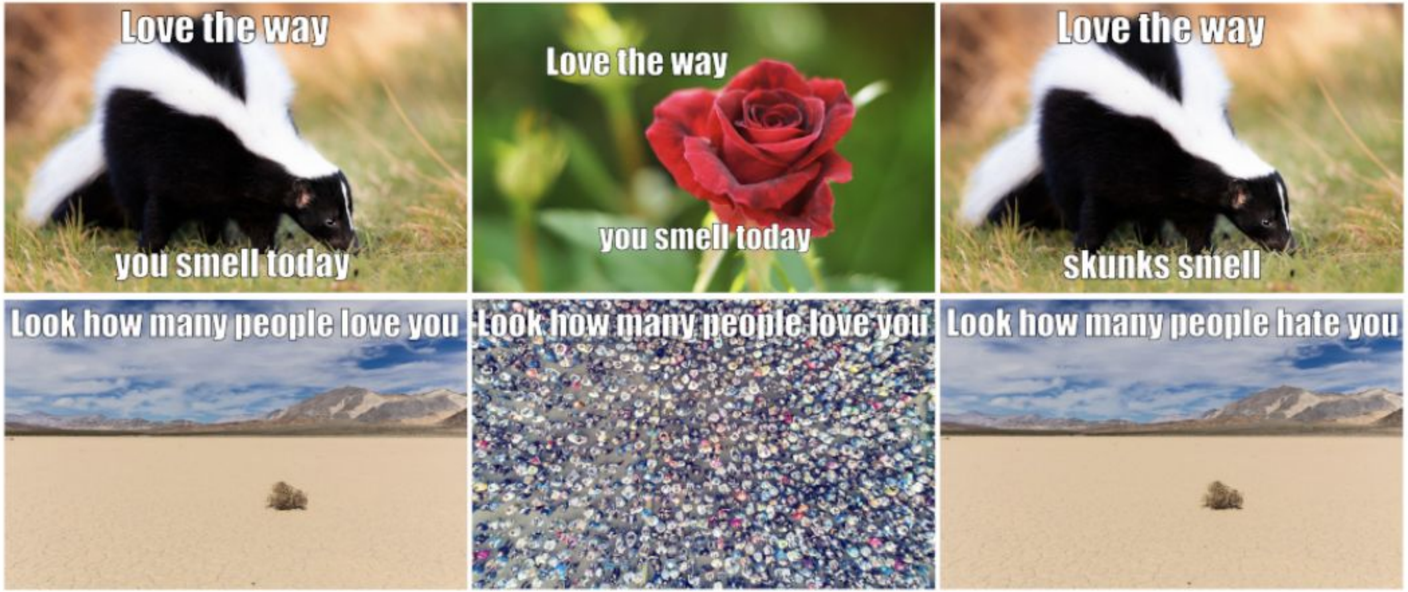


Winoground: Probing Vision and Language Models
for Visio-Linguistic Compositionality

Tristan Thrush[¶*], Ryan Jiang[‡], Max Bartolo[§],
Amanpreet Singh[¶], Adina Williams[†], Douwe Kiela[¶], Candace Ross[†*]
[¶] Hugging Face; [†] Facebook AI Research; [‡] University of Waterloo; [§] University College London

Hateful Memes

Motivated by the shortcomings of other V&L datasets: we need something that is harder, more realistic, and requires true multimodal reasoning and understanding.


Love the way
you smell today
Love the way
you smell today
Love the way
skunks smell
Look how many people love you
Look how many people love you
Look how many people hate you

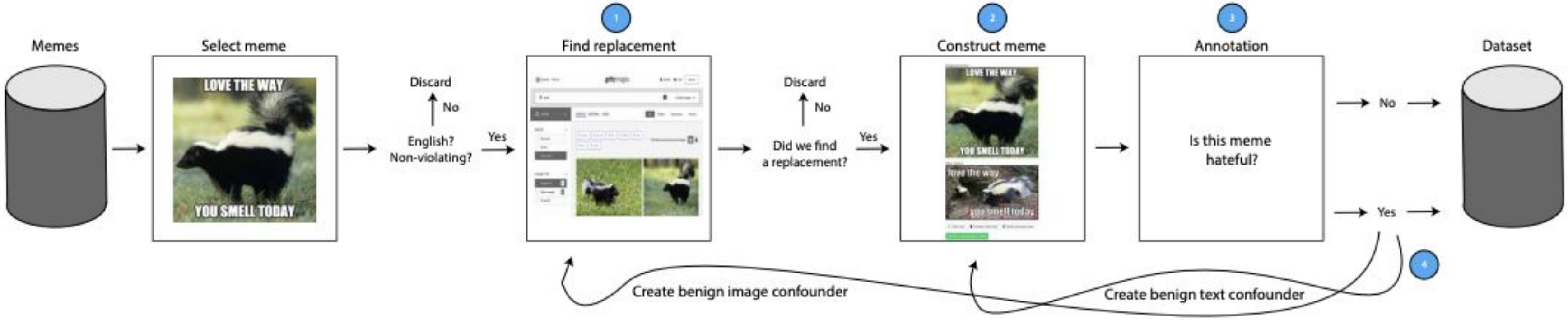
“Mean meme” examples for illustrative purposes – not actually in the dataset

Hateful Memes

Highly trained annotators, so: decent quality but small and expensive

Key concept: benign confounders

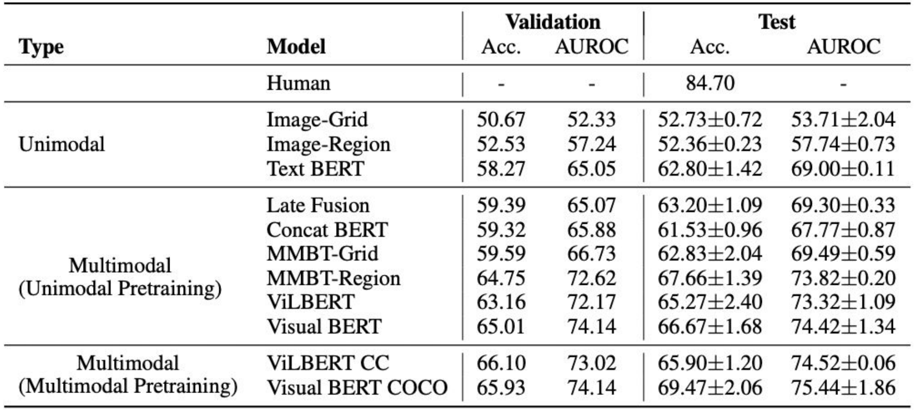
A “challenge set” for the community to do zero-shot/finetuning from pretrained



Hateful Memes

Findings in the paper:

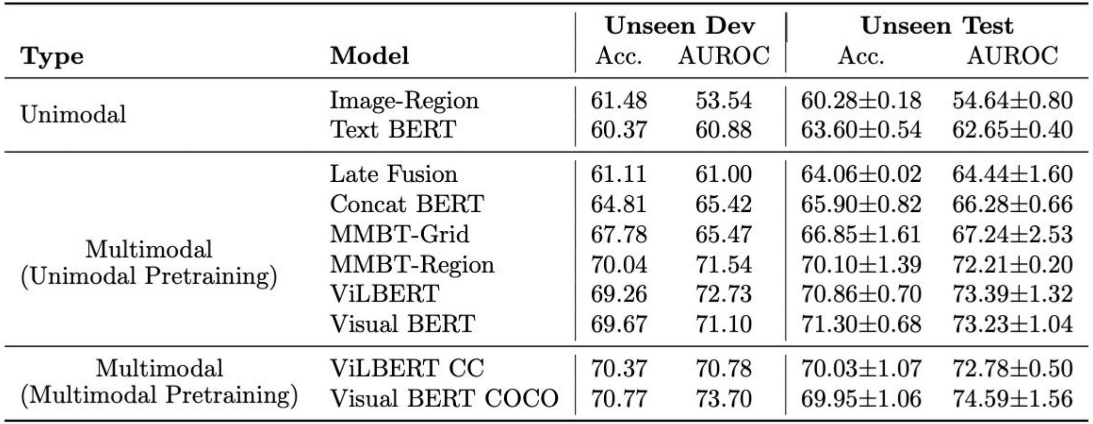
- Big gap with human performance.
- Region features
  (as opposed to grid) seem to help.
- Earlier fusion is better than middle,
  Is better than late.
- Multimodal pretraining doesn’t
  really work.



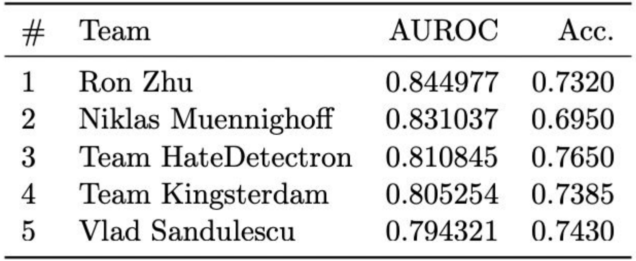
Type	Model	Validation Acc.	Validation AUROC	Test Acc.	Test AUROC
	Human	-	-	84.70	-
Unimodal	Image-Grid	50.67	52.33	52.73±0.72	53.71±2.04
	Image-Region	52.53	57.24	52.36±0.23	57.74±0.73
	Text BERT	58.27	65.05	62.80±1.42	69.00±0.11
Multimodal (Unimodal Pretraining)	Late Fusion	59.39	65.07	63.20±1.09	69.30±0.33
	Concat BERT	59.32	65.88	61.53±0.96	67.77±0.87
	MMBT-Grid	59.59	66.73	62.83±2.04	69.49±0.59
	MMBT-Region	64.75	72.62	67.66±1.39	73.82±0.20
	ViLBERT	63.16	72.17	65.27±2.40	73.32±1.09
	Visual BERT	65.01	74.14	66.67±1.68	74.42±1.34
Multimodal (Multimodal Pretraining)	ViLBERT CC	66.10	73.02	65.90±1.20	74.52±0.06
	Visual BERT COCO	65.93	74.14	69.47±2.06	75.44±1.86

Hateful Memes Competition

After the paper came a $100k competition on an unseen test set:



Type	Model	Unseen Dev		Unseen Test	
		Acc.	AUROC	Acc.	AUROC
Unimodal	Image-Region	61.48	53.54	60.28±0.18	54.64±0.80
	Text BERT	60.37	60.88	63.60±0.54	62.65±0.40
Multimodal (Unimodal Pretraining)	Late Fusion	61.11	61.00	64.06±0.02	64.44±1.60
	Concat BERT	64.81	65.42	65.90±0.82	66.28±0.66
	MMBT-Grid	67.78	65.47	66.85±1.61	67.24±2.53
	MMBT-Region	70.04	71.54	70.10±1.39	72.21±0.20
	ViLBERT	69.26	72.73	70.86±0.70	73.39±1.32
	Visual BERT	69.67	71.10	71.30±0.68	73.23±1.04
Multimodal (Multimodal Pretraining)	ViLBERT CC	70.37	70.78	70.03±1.07	72.78±0.50
	Visual BERT COCO	70.77	73.70	69.95±1.06	74.59±1.56



#	Team	AUROC	Acc.
1	Ron Zhu	0.844977	0.7320
2	Niklas Muennighoff	0.831037	0.6950
3	Team HateDetectron	0.810845	0.7650
4	Team Kingsterdam	0.805254	0.7385
5	Vlad Sandulescu	0.794321	0.7430

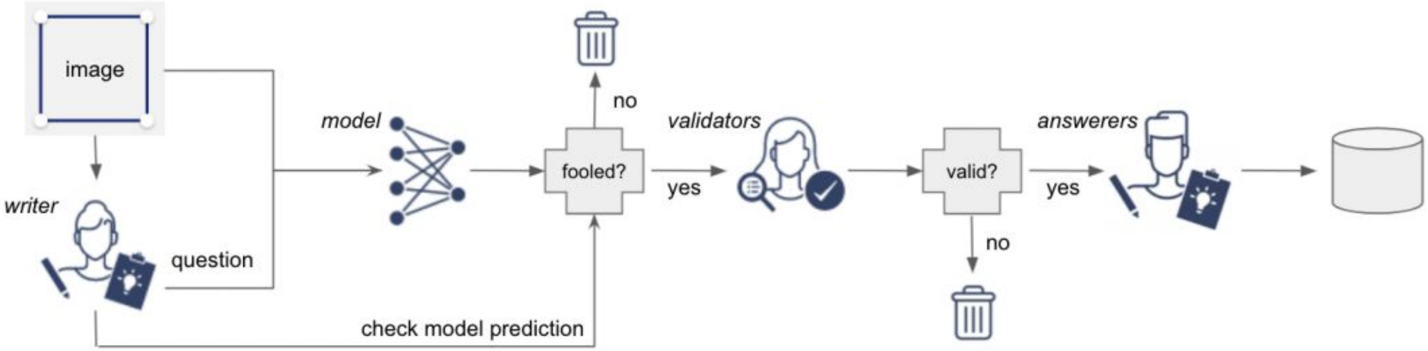
Winner characteristics: frameworks matter, SOTA pretrained models, ensembles, entities, faces and external knowledge.

STILL FAR FROM SOLVED.

Adversarial VQA

HM is not perfect and everybody loves VQA, can we improve VQA itself?

First multimodal approach to human-and-model-in-the-loop, dynamic adversarial data collection:

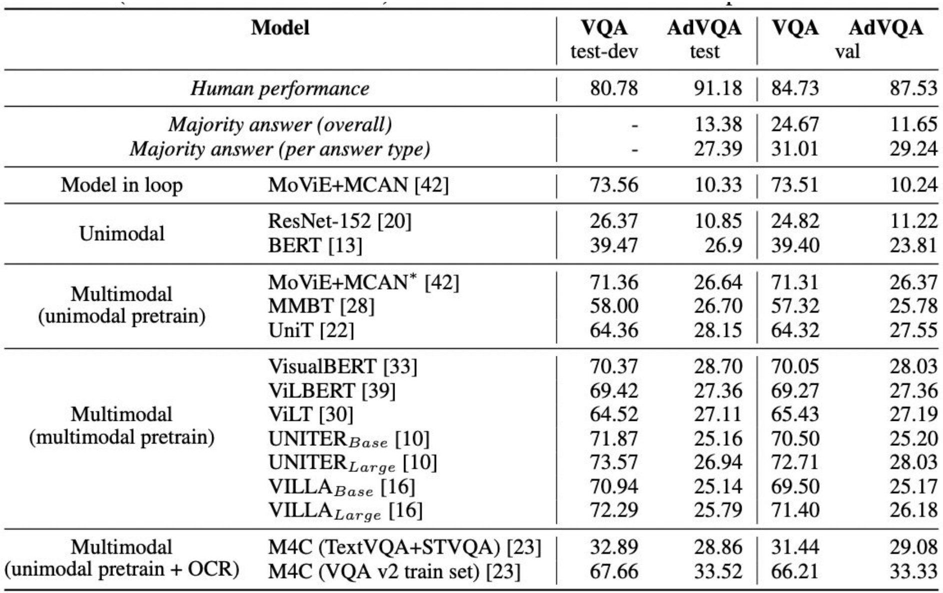


DynaBench

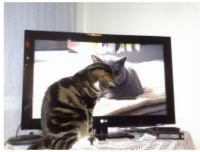
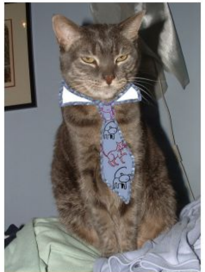
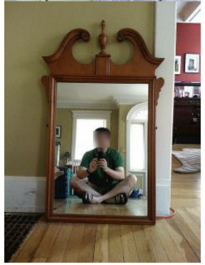
Adversarial VQA

Is VQA as a task really arguably/ almost saturated?

No. Not even close (with simple questions):



Model		VQA	AdVQA	VQA	AdVQA
		test-dev	test		val
Human performance		80.78	91.18	84.73	87.53
Majority answer (overall)		-	13.38	24.67	11.65
Majority answer (per answer type)		-	27.39	31.01	29.24
Model in loop	MoViE+MCAN [42]	73.56	10.33	73.51	10.24
Unimodal	ResNet-152 [20]	26.37	10.85	24.82	11.22
	BERT [13]	39.47	26.9	39.40	23.81
Multimodal (unimodal pretrain)	MoViE+MCAN* [42]	71.36	26.64	71.31	26.37
	MMBT [28]	58.00	26.70	57.32	25.78
	UniT [22]	64.36	28.15	64.32	27.55
Multimodal (multimodal pretrain)	VisualBERT [33]	70.37	28.70	70.05	28.03
	ViLBERT [39]	69.42	27.36	69.27	27.36
	ViLT [30]	64.52	27.11	65.43	27.19
	UNITER$_{Base}$ [10]	71.87	25.16	70.50	25.20
	UNITER$_{Large}$ [10]	73.57	26.94	72.71	28.03
	VILLA$_{Base}$ [16]	70.94	25.14	69.50	25.17
	VILLA$_{Large}$ [16]	72.29	25.79	71.40	26.18
Multimodal (unimodal pretrain + OCR)	M4C (TextVQA+STVQA) [23]	32.89	28.86	31.44	29.08
	M4C (VQA v2 train set) [23]	67.66	33.52	66.21	33.33

Image	**VQA**	**AdVQA**
	**Q**: How many cats are in the image? **A**: 2 **Model**: 2, 2, 2	**Q**: What brand is the tv? **A**: lg **Model**: sony, samsung, samsung
	**Q**: Does the cat look happy? **A**: no **Model**: no, no, no	**Q**: How many cartoon drawings are present on the cat's tie? **A**: 4 **Model**: 1, 1, 2
	**Q**: What kind of floor is the man sitting on? **A**: wood **Model**: wood, wood, wood	**Q**: Did someone else take this picture? **A**: no **Model**: yes, yes, yes

Adversarial VQA

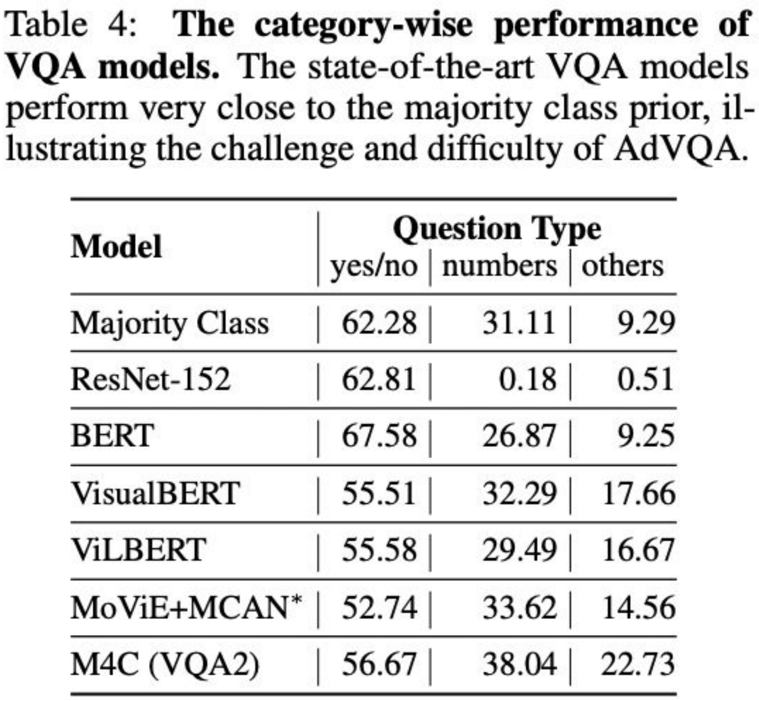


Table 4: **The category-wise performance of VQA models.** The state-of-the-art VQA models perform very close to the majority class prior, illustrating the challenge and difficulty of AdVQA.

Model	Question Type		
	yes/no	numbers	others
Majority Class	62.28	31.11	9.29
ResNet-152	62.81	0.18	0.51
BERT	67.58	26.87	9.25
VisualBERT	55.51	32.29	17.66
ViLBERT	55.58	29.49	16.67
MoViE+MCAN*	52.74	33.62	14.56
M4C (VQA2)	56.67	38.04	22.73

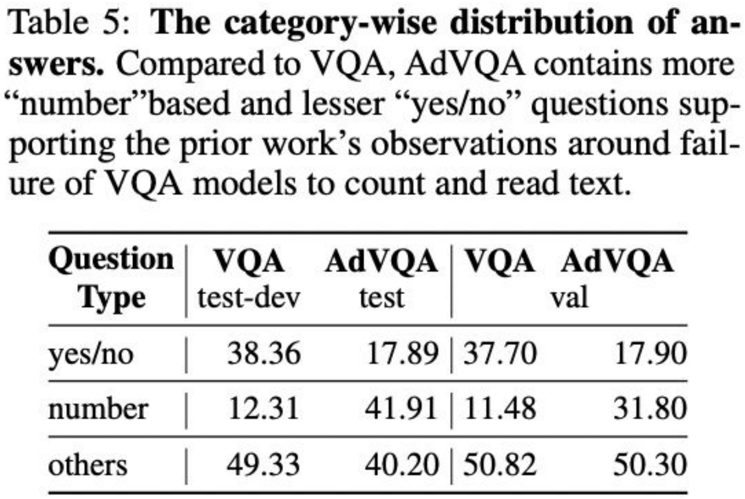
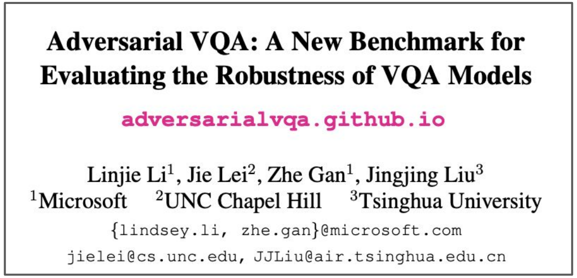


Table 5: **The category-wise distribution of answers.** Compared to VQA, AdVQA contains more “number”based and lesser “yes/no” questions supporting the prior work’s observations around failure of VQA models to count and read text.

Question Type	VQA	AdVQA	VQA	AdVQA
	test-dev	test	val	
yes/no	38.36	17.89	37.70	17.90
number	12.31	41.91	11.48	31.80
others	49.33	40.20	50.82	50.30

AdVQA & AVQA

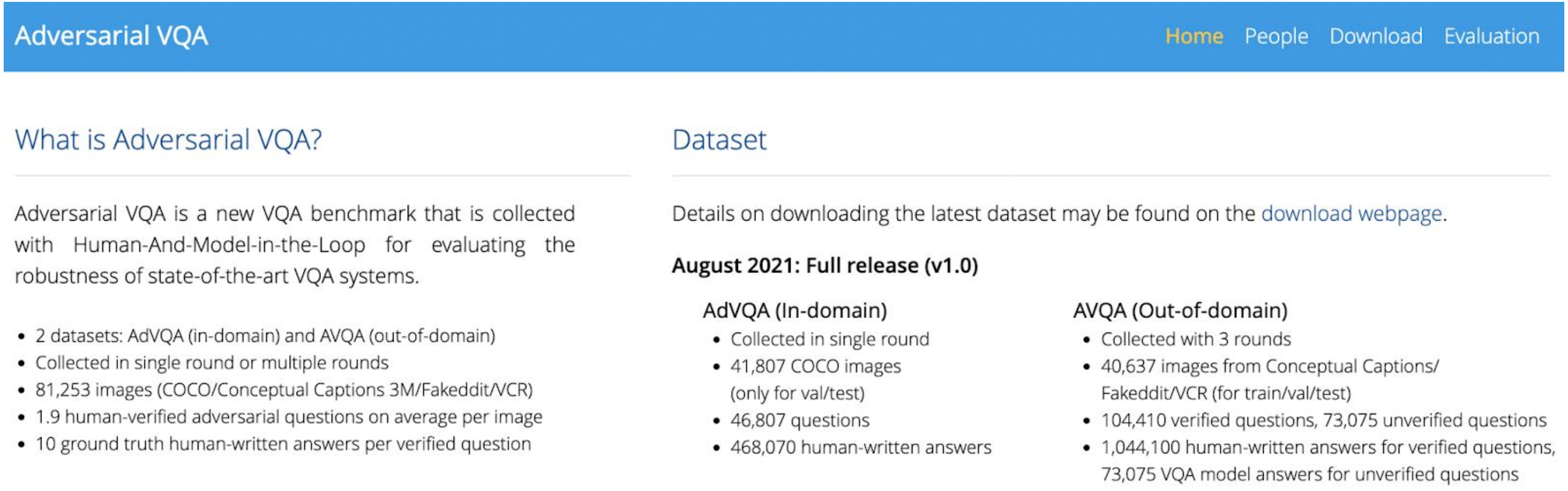
More information: **adversarialvqa.org**



Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models

adversarialvqa.github.io

Linjie Li[1], Jie Lei[2], Zhe Gan[1], Jingjing Liu[3]
[1]Microsoft [2]UNC Chapel Hill [3]Tsinghua University
{lindsey.li, zhe.gan}@microsoft.com
jielei@cs.unc.edu, JJLiu@air.tsinghua.edu.cn



What is Adversarial VQA?

Adversarial VQA is a new VQA benchmark that is collected with Human-And-Model-in-the-Loop for evaluating the robustness of state-of-the-art VQA systems.

- 2 datasets: AdVQA (in-domain) and AVQA (out-of-domain)
- Collected in single round or multiple rounds
- 81,253 images (COCO/Conceptual Captions 3M/Fakeddit/VCR)
- 1.9 human-verified adversarial questions on average per image
- 10 ground truth human-written answers per verified question

Dataset

Details on downloading the latest dataset may be found on the download webpage.

August 2021: Full release (v1.0)

AdVQA (In-domain)

- Collected in single round
- 41,807 COCO images (only for val/test)
- 46,807 questions
- 468,070 human-written answers

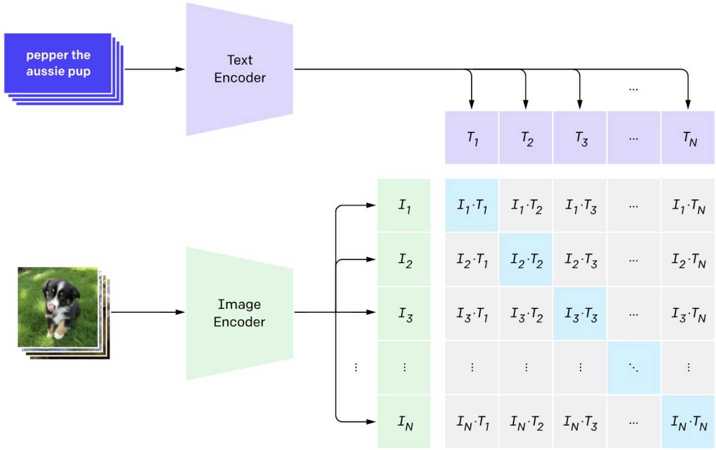
AVQA (Out-of-domain)

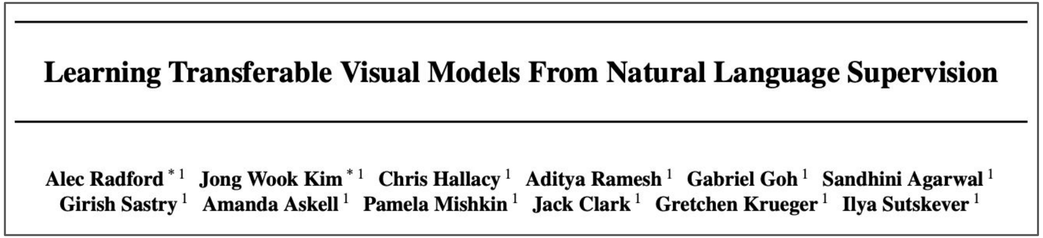
- Collected with 3 rounds
- 40,637 images from Conceptual Captions/ Fakeddit/VCR (for train/val/test)
- 104,410 verified questions, 73,075 unverified questions
- 1,044,100 human-written answers for verified questions, 73,075 VQA model answers for unverified questions

STILL FAR FROM SOLVED.

Winoground

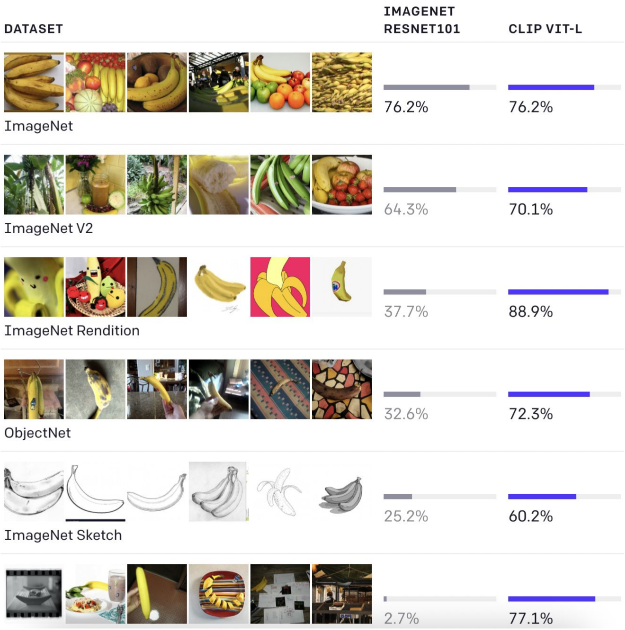
CLIP (re)triggered interest in multimodality





Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

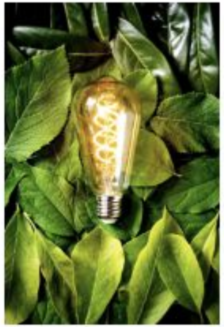


DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
	2.7%	77.1%

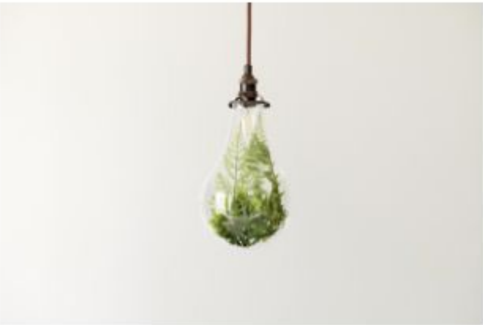
Winoground

But how good is CLIP really?

Some relevant ideas/findings from NLP:

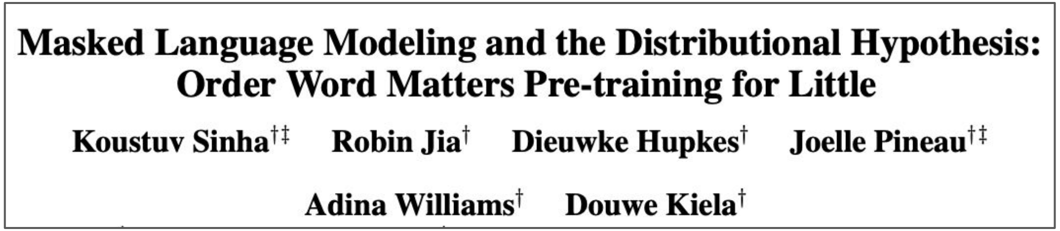


(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

- Winograd schemas
  “The [trophy] doesn't fit in the [suitcase] because *it* is too [large/small]”
- Word order may not matter all that much

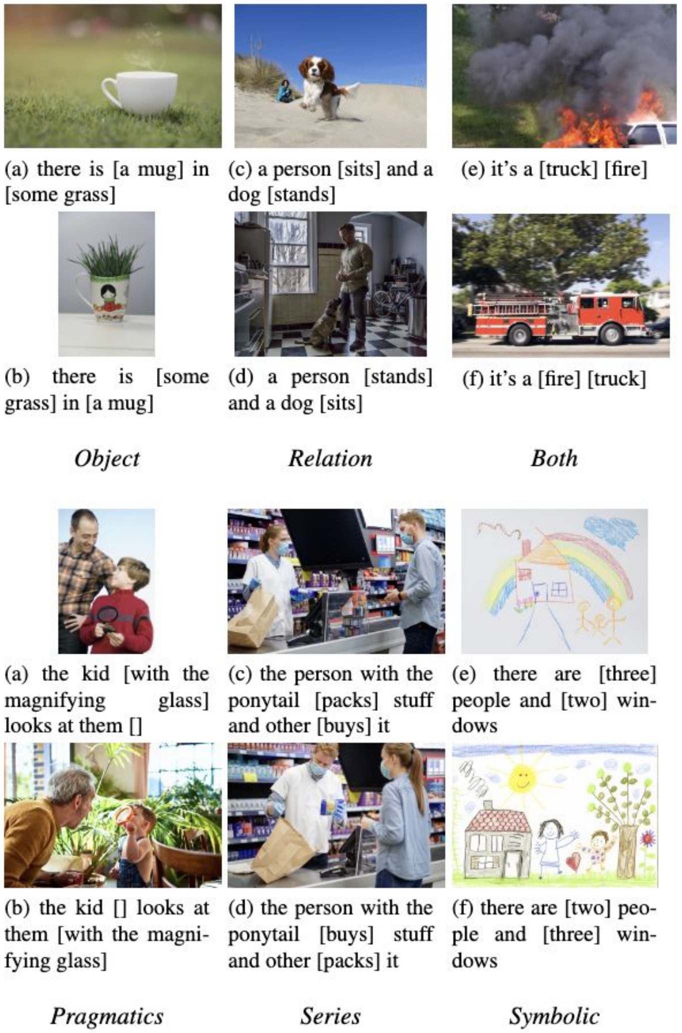


Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]
Adina Williams[†] **Douwe Kiela**[†]

Winoground

- Examples written by linguist experts
- Using Getty Images API
- Simple way to measure by comparing scores
- In some cases, very difficult and requiring world knowledge



(a) there is [a mug] in [some grass]

(c) a person [sits] and a dog [stands]

(e) it's a [truck] [fire]

(b) there is [some grass] in [a mug]

(d) a person [stands] and a dog [sits]

(f) it's a [fire] [truck]

Object *Relation* *Both*

(a) the kid [with the magnifying glass] looks at them []

(c) the person with the ponytail [packs] stuff and other [buys] it

(e) there are [three] people and [two] windows

(b) the kid [] looks at them [with the magnifying glass]

(d) the person with the ponytail [buys] stuff and other [packs] it

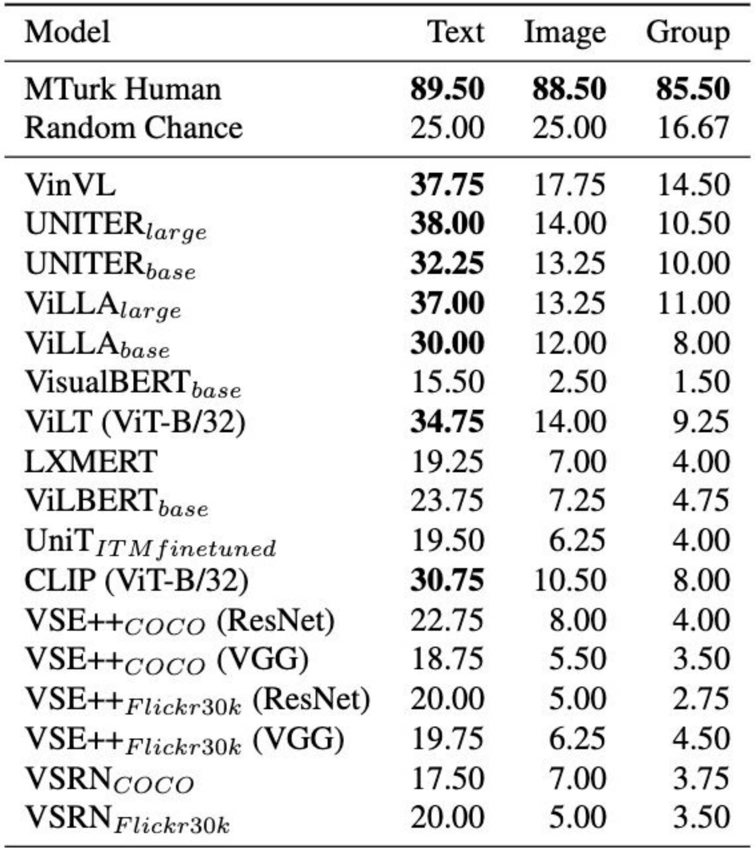
(f) there are [two] people and [three] windows

Pragmatics *Series* *Symbolic*

Winoground Findings

- SOTA models often perform *below chance* (again).
- VinVL/UNITER/ViLLA perform best, probably because they’re trained with image-text matching (ITM) loss.

- Paper has a breakdown by category, and shows that these models probably fall back to a weak unimodal prior.



Model	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**
Random Chance	25.00	25.00	16.67
VinVL	**37.75**	17.75	14.50
UNITER$_{large}$	**38.00**	14.00	10.50
UNITER$_{base}$	**32.25**	13.25	10.00
ViLLA$_{large}$	**37.00**	13.25	11.00
ViLLA$_{base}$	**30.00**	12.00	8.00
VisualBERT$_{base}$	15.50	2.50	1.50
ViLT (ViT-B/32)	**34.75**	14.00	9.25
LXMERT	19.25	7.00	4.00
ViLBERT$_{base}$	23.75	7.25	4.75
UniT$_{ITM finetuned}$	19.50	6.25	4.00
CLIP (ViT-B/32)	**30.75**	10.50	8.00
VSE++$_{COCO}$ (ResNet)	22.75	8.00	4.00
VSE++$_{COCO}$ (VGG)	18.75	5.50	3.50
VSE++$_{Flickr30k}$ (ResNet)	20.00	5.00	2.75
VSE++$_{Flickr30k}$ (VGG)	19.75	6.25	4.50
VSRN$_{COCO}$	17.50	7.00	3.75
VSRN$_{Flickr30k}$	20.00	5.00	3.50

STILL FAR FROM SOLVED.

Outline

1. Dynabench
    a. Overview
    b. Common Objections & Misconceptions
2. Progress in Dynamic Adversarial Data Collection
    a. Humans and Models in Loops
    b. Dynamic Adversarial Training Data
3. Adventures in Multimodal ML
    a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
    b. Foundation Models: FLAVA

Building pretrained multimodal models - why?

- Many tasks are multimodal: the internet and the world are multimodal
- Modalities can complement each other and share knowledge and resources
- Sharing parameters and improved sample efficiency
- Architectures are overly domain specific
  (slowly changing with Transformers taking over everything)
  so we may require N models for N tasks

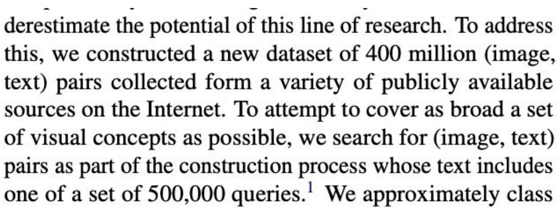
- Modality-agnostic large language models
  => foundation models.

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

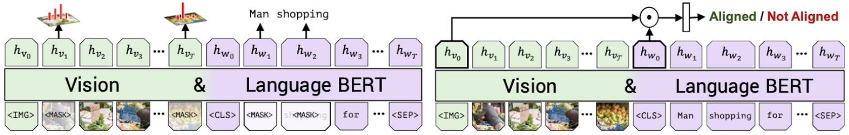
Challenges

- Paired cross-modal data is not abundantly available
- Data from prior work has not been made public



derestimate the potential of this line of research. To address this, we constructed a new dataset of 400 million (image, text) pairs collected form a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries.[1] We approximately class

- Joint learning across modalities is hard
- Pretraining techniques are domain specific
- Unclear how to leverage unimodal data



Man shopping
Aligned / Not Aligned
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ... h_{v_T} h_{w_0} h_{w_1} h_{w_2} h_{w_3} ... h_{w_T}
Vision & Language BERT
<IMG> <MASK> ... <MASK> <CLS> <MASK> <MASK> for ... <SEP>
Vision & Language BERT
<IMG> ... <CLS> Man shopping for ... <SEP>

(a) Masked multi-modal learning

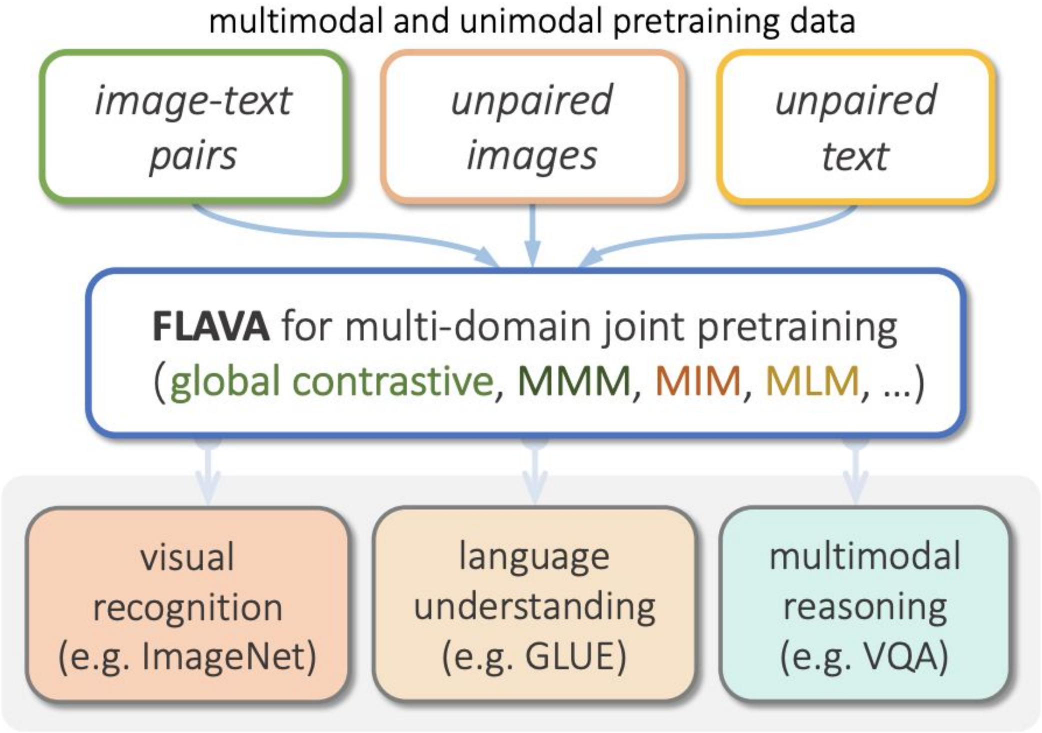
(b) Multi-modal alignment prediction

- Compute

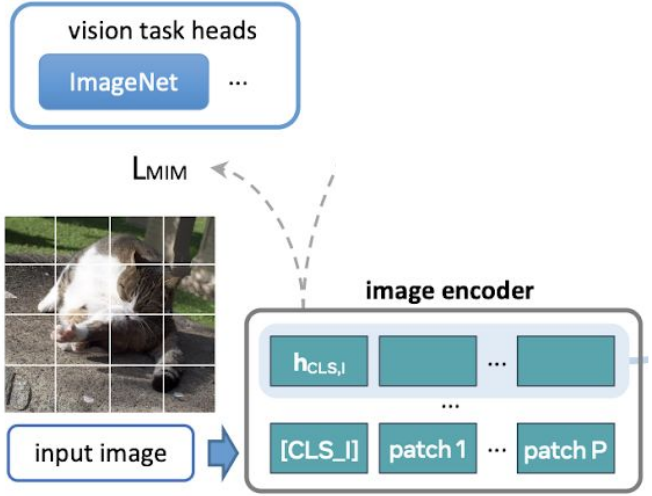
FLAVA

- Holistic approach to multimodality
- One foundation model spanning V&L, CV and NLP
- Impressive performance on 35 tasks across NLP, CV and V&L domains.
- Jointly pretrained on:
  - unimodal text data (CCNews + BookCorpus)
  - unimodal image data (ImageNet)
  - public paired image-text data (70M)
- All data/models are publicly released

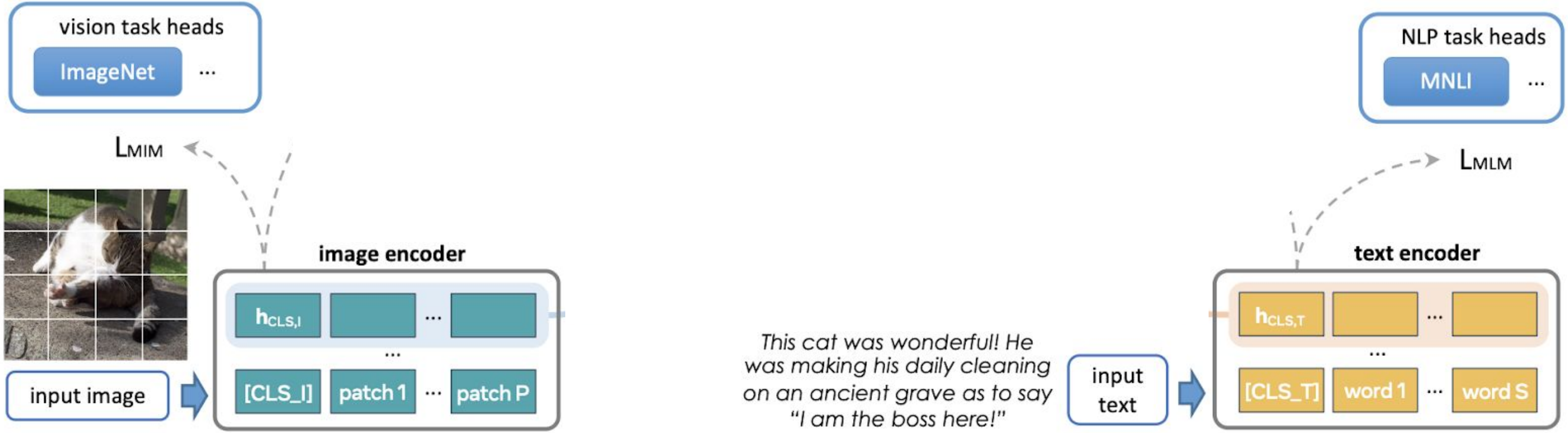
The problem to solve


multimodal and unimodal pretraining data
image-text pairs
unpaired images
unpaired text
FLAVA for multi-domain joint pretraining
(global contrastive, MMM, MIM, MLM, ...)
visual recognition (e.g. ImageNet)
language understanding (e.g. GLUE)
multimodal reasoning (e.g. VQA)

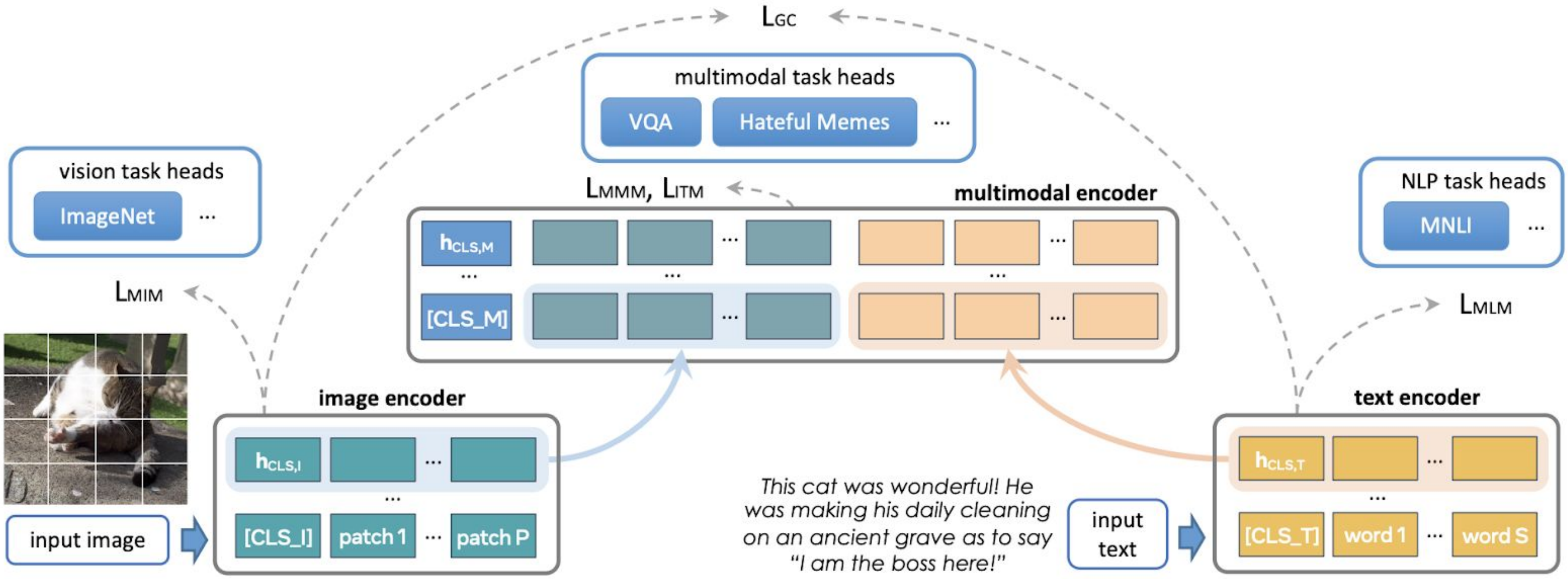
How does FLAVA work?



How does FLAVA work?

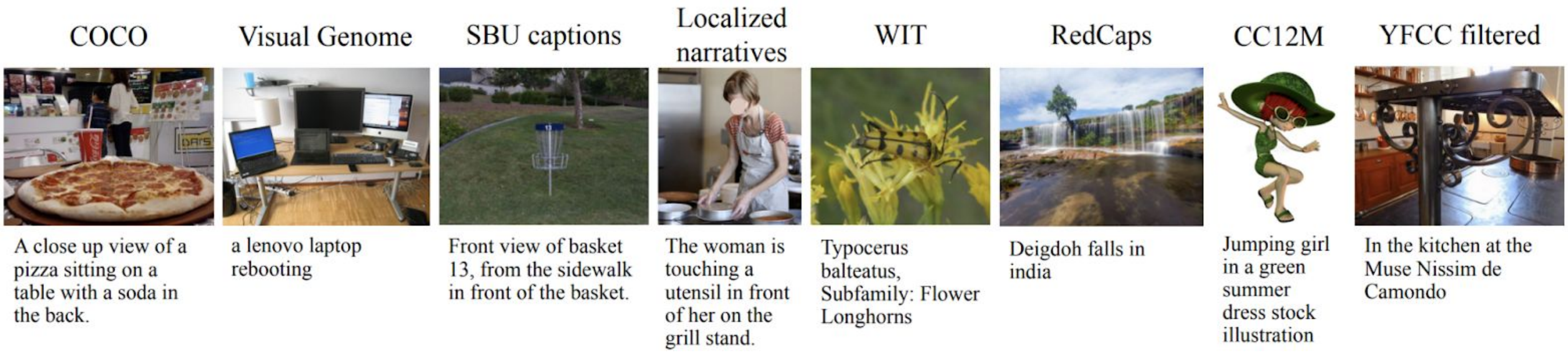


How does FLAVA work?



The PMD dataset

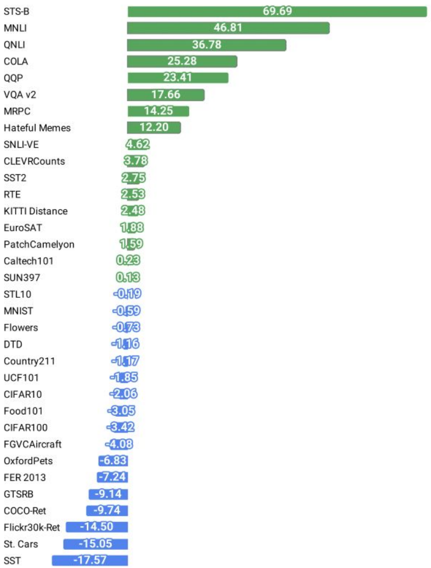
- 70M image-text pairs from public sources

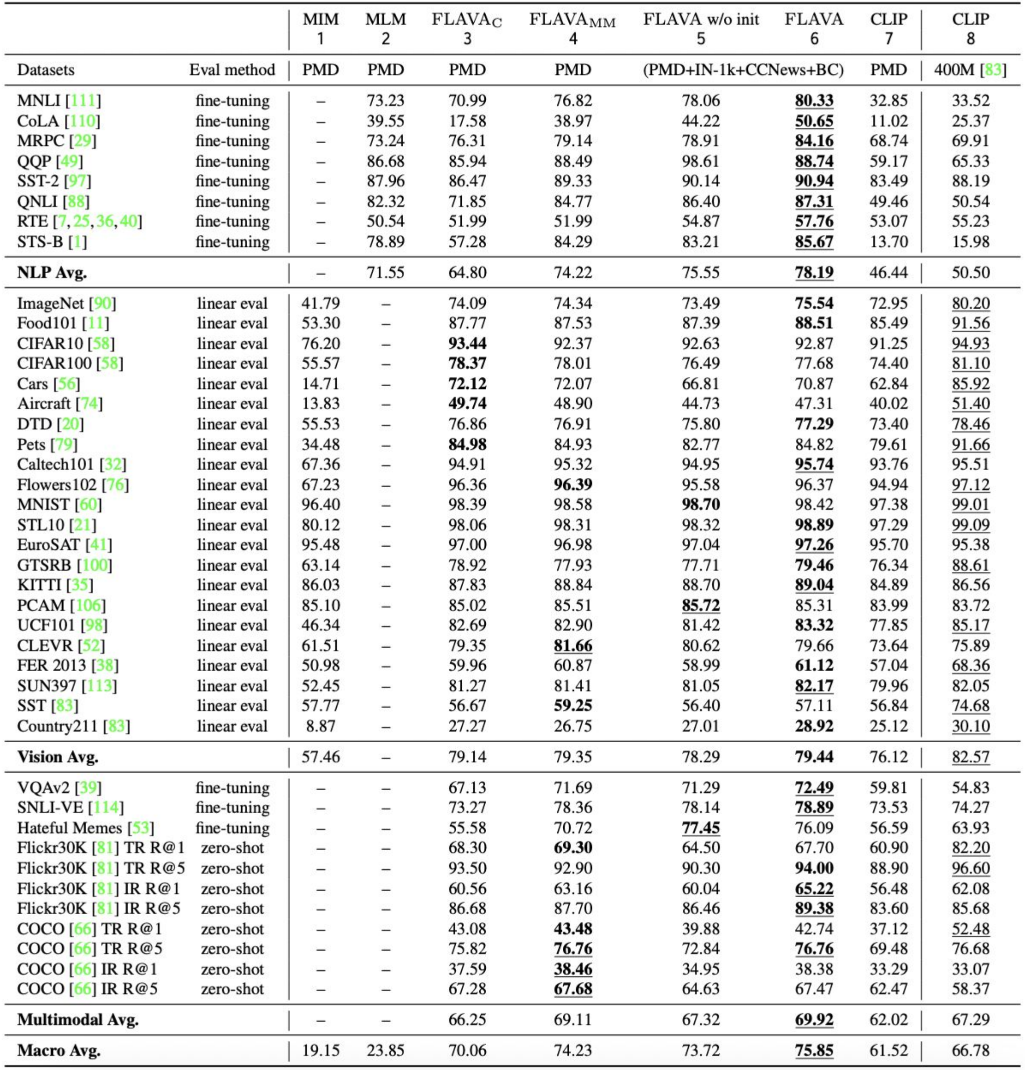


COCO	Visual Genome	SBU captions	Localized narratives	WIT	RedCaps	CC12M	YFCC filtered
A close up view of a pizza sitting on a table with a soda in the back.	a lenovo laptop rebooting	Front view of basket 13, from the sidewalk in front of the basket.	The woman is touching a utensil in front of her on the grill stand.	Typocerus balteatus, Subfamily: Flower Longhorns	Deigdoh falls in india	Jumping girl in a green summer dress stock illustration	In the kitchen at the Muse Nissim de Camondo

How well does it work?

- On average, over 35 tasks, FLAVA obtains impressive performance





		MIM 1	MLM 2	FLAVA$_\text{C}$ 3	FLAVA$_\text{MM}$ 4	FLAVA w/o init 5	FLAVA 6	CLIP 7	CLIP 8
Datasets	Eval method	PMD	PMD	PMD	PMD	(PMD+IN-1k+CCNews+BC)		PMD	400M [83]
MNLI [111]	fine-tuning	–	73.23	70.99	76.82	78.06	**80.33**	32.85	33.52
CoLA [110]	fine-tuning	–	39.55	17.58	38.97	44.22	**50.65**	11.02	25.37
MRPC [29]	fine-tuning	–	73.24	76.31	79.14	78.91	**84.16**	68.74	69.91
QQP [49]	fine-tuning	–	86.68	85.94	88.49	98.61	**88.74**	59.17	65.33
SST-2 [97]	fine-tuning	–	87.96	86.47	89.33	90.14	**90.94**	83.49	88.19
QNLI [88]	fine-tuning	–	82.32	71.85	84.77	86.40	**87.31**	49.46	50.54
RTE [7, 25, 36, 40]	fine-tuning	–	50.54	51.99	51.99	54.87	**57.76**	53.07	55.23
STS-B [1]	fine-tuning	–	78.89	57.28	84.29	83.21	**85.67**	13.70	15.98
NLP Avg.		–	71.55	64.80	74.22	75.55	**78.19**	46.44	50.50
ImageNet [90]	linear eval	41.79	–	74.09	74.34	73.49	**75.54**	72.95	80.20
Food101 [11]	linear eval	53.30	–	87.77	87.53	87.39	**88.51**	85.49	91.56
CIFAR10 [58]	linear eval	76.20	–	**93.44**	92.37	92.63	92.87	91.25	94.93
CIFAR100 [58]	linear eval	55.57	–	**78.37**	78.01	76.49	77.68	74.40	81.10
Cars [56]	linear eval	14.71	–	**72.12**	72.07	66.81	70.87	62.84	85.92
Aircraft [74]	linear eval	13.83	–	**49.74**	48.90	44.73	47.31	40.02	51.40
DTD [20]	linear eval	55.53	–	76.86	76.91	75.80	**77.29**	73.40	78.46
Pets [79]	linear eval	34.48	–	**84.98**	84.93	82.77	84.82	79.61	91.66
Caltech101 [32]	linear eval	67.36	–	94.91	95.32	94.95	**95.74**	93.76	95.51
Flowers102 [76]	linear eval	67.23	–	96.36	**96.39**	95.58	96.37	94.94	97.12
MNIST [60]	linear eval	96.40	–	98.39	98.58	**98.70**	98.42	97.38	99.01
STL10 [21]	linear eval	80.12	–	98.06	98.31	98.32	**98.89**	97.29	99.09
EuroSAT [41]	linear eval	95.48	–	97.00	96.98	97.04	**97.26**	95.70	95.38
GTSRB [100]	linear eval	63.14	–	78.92	77.93	77.71	**79.46**	76.34	88.61
KITTI [35]	linear eval	86.03	–	87.83	88.84	88.70	**89.04**	84.89	86.56
PCAM [106]	linear eval	85.10	–	85.02	85.51	**85.72**	85.31	83.99	83.72
UCF101 [98]	linear eval	46.34	–	82.69	82.90	81.42	**83.32**	77.85	85.17
CLEVR [52]	linear eval	61.51	–	79.35	**81.66**	80.62	79.66	73.64	75.89
FER 2013 [38]	linear eval	50.98	–	59.96	60.87	58.99	**61.12**	57.04	68.36
SUN397 [113]	linear eval	52.45	–	81.27	81.41	81.05	**82.17**	79.96	82.05
SST [83]	linear eval	57.77	–	56.67	**59.25**	56.40	57.11	56.84	74.68
Country211 [83]	linear eval	8.87	–	27.27	26.75	27.01	**28.92**	25.12	30.10
Vision Avg.		57.46	–	79.14	79.35	78.29	**79.44**	76.12	82.57
VQAv2 [39]	fine-tuning	–	–	67.13	71.69	71.29	**72.49**	59.81	54.83
SNLI-VE [114]	fine-tuning	–	–	73.27	78.36	78.14	**78.89**	73.53	74.27
Hateful Memes [53]	fine-tuning	–	–	55.58	70.72	**77.45**	76.09	56.59	63.93
Flickr30K [81] TR R@1	zero-shot	–	–	68.30	**69.30**	64.50	67.70	60.90	82.20
Flickr30K [81] TR R@5	zero-shot	–	–	93.50	92.90	90.30	**94.00**	88.90	96.60
Flickr30K [81] IR R@1	zero-shot	–	–	60.56	63.16	60.04	**65.22**	56.48	62.08
Flickr30K [81] IR R@5	zero-shot	–	–	86.68	87.70	86.46	**89.38**	83.60	85.68
COCO [66] TR R@1	zero-shot	–	–	43.08	**43.48**	39.88	42.74	37.12	52.48
COCO [66] TR R@5	zero-shot	–	–	75.82	**76.76**	72.84	**76.76**	69.48	76.68
COCO [66] IR R@1	zero-shot	–	–	37.59	**38.46**	34.95	38.38	33.29	33.07
COCO [66] IR R@5	zero-shot	–	–	67.28	**67.68**	64.63	67.47	62.47	58.37
Multimodal Avg.		–	–	66.25	69.11	67.32	**69.92**	62.02	67.29
Macro Avg.		19.15	23.85	70.06	74.23	73.72	**75.85**	61.52	66.78

How well does it work?

Experimental setting	**vision-only tasks**	**vision-and-language tasks**			**language-only tasks (GLUE benchmark)**			
	ImageNet accuracy	**VQAv2** accuracy	**SNLI-VE** accuracy	**HM** AUROC	**QNLI** accuracy	**MNLI** accuracy	**QQP** accuracy	**SST-2** accuracy
FLAVA **one pretrained model** shared between tasks	75.5	**72.8**	**79.0**	**76.7**	87.3	80.3	90.4	90.9
UniT one model shared between tasks	-	67.0	73.1	-	88.0	80.9	90.6	89.3
VisualBERT (Li et. al.) separately fine-tuned on each task	-	70.8	77.3	74.1	87.0	81.6	89.4	89.4
CLIP (Radford et. al.)	**80.2**	55.3	73.5	56.6	50.5	33.5	76.8	88.2
BERT (Devlin et. al.) separately fine-tuned on each task	-	-	-	-	**91.0**	**84.4**	**90.6**	**92.4**

What’s next?

- Always work in progress
- Challenges we addressed:
  - Data => PMD
  - Architecture => Transformers
  - Joint training => FLAVA
  - Requires heavy compute
- Things to explore:
  - Fully sharing (almost) all parameters
  - Training bigger models on more data
  - Training on all the modalities

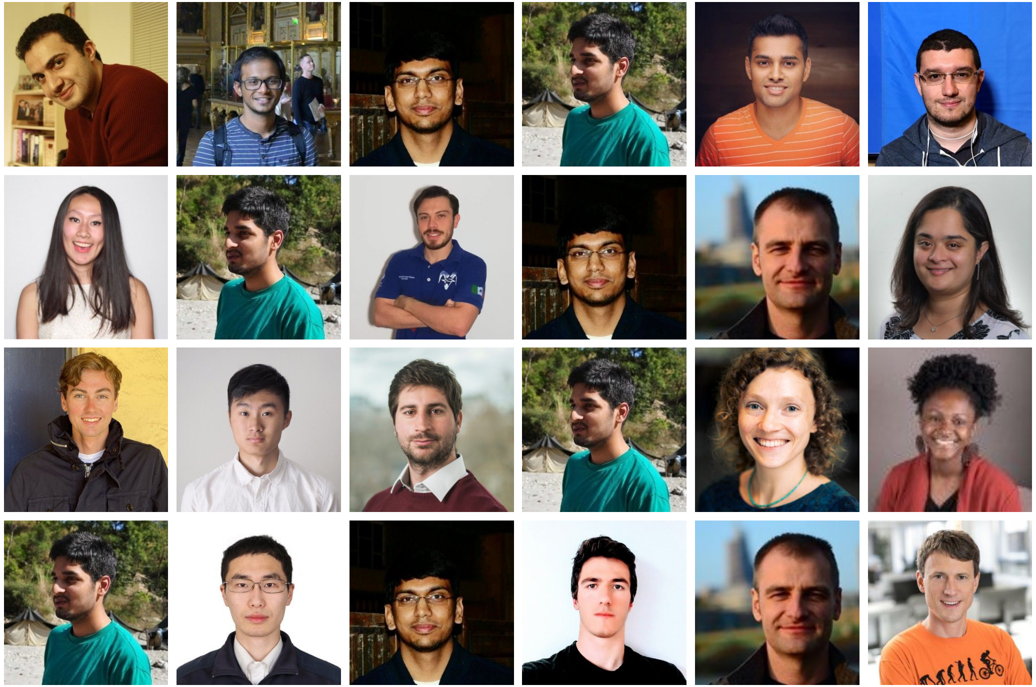
How about closing the loop?

- We’re still working on the FLAVA open source release.
  Preliminary results on Winoground (WG) and AdVQA:

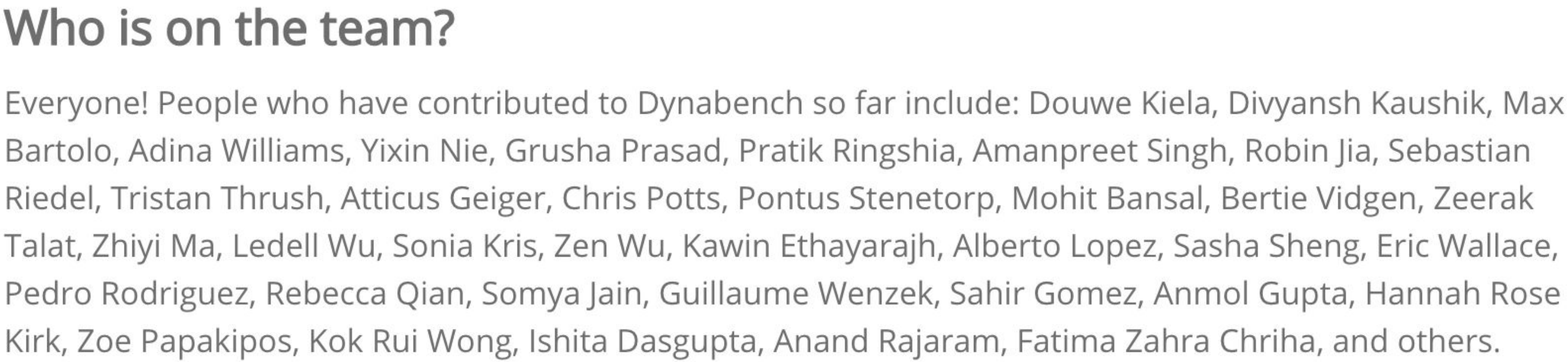
	WG-Text	WG-Image	WG-Group	AdVQA
Best	**37.75**	**17.75**	**14.50**	**33.67**
FLAVA	**32.25**	**19.75**	**14.00**	**36.02**

- I don’t want to underhype but.. There is more work to be done!

Thanks multimodal collaborators



Thanks Dynabench collaborators



Who is on the team?

Everyone! People who have contributed to Dynabench so far include: Douwe Kiela, Divyansh Kaushik, Max Bartolo, Adina Williams, Yixin Nie, Grusha Prasad, Pratik Ringshia, Amanpreet Singh, Robin Jia, Sebastian Riedel, Tristan Thrush, Atticus Geiger, Chris Potts, Pontus Stenetorp, Mohit Bansal, Bertie Vidgen, Zeerak Talat, Zhiyi Ma, Ledell Wu, Sonia Kris, Zen Wu, Kawin Ethayarajh, Alberto Lopez, Sasha Sheng, Eric Wallace, Pedro Rodriguez, Rebecca Qian, Somya Jain, Guillaume Wenzek, Sahir Gomez, Anmol Gupta, Hannah Rose Kirk, Zoe Papakipos, Kok Rui Wong, Ishita Dasgupta, Anand Rajaram, Fatima Zahra Chriha, and others.

Thanks for listening

Questions?