

# The Wild West of NLP Modeling, Evaluation and Documentation

Nazneen Rajani | Robustness Research Lead @ Hugging Face | nazneen@hf.co | @nazneenrajani

# Outline

### Part 1:

NLP Modeling landscape

Systematic study of 75,000 models on HF

### Part 2:

NLP Evaluation landscape

Challenges and opportunities in model evaluation and documentation

# Outline

### Part 1:

NLP Modeling landscape

Systematic study of 75K models on HF

### Part 2:

NLP Evaluation landscape

Challenges and opportunities in model evaluation and documentation

### Large Language Models since GPT3



\*only LLMs with >1B parameters & EN as the main training language are shown. Comprehensive list: https://crfm.stanford.edu/helm/v1.0/?models=1

### Large Language Models since GPT3



\*only LLMs with >1B parameters & EN as the main training language are shown. Comprehensive list: https://crfm.stanford.edu/helm/v1.0/?models=1

### **Model Access**



Open access models

Closed access models

# Open Access Models

All model components are publicly available:

- Open source **code**
- Training **data** 
  - Sources and their distribution
  - Data preprocessing and curation steps
- Model weights
- Paper or blog summarizing
  - Architecture and training details
  - Evaluation results
  - Adaptation to the model
    - Safety filters
    - Training with human feedback



Allows reproducing results and replicating parts of the model

Enable auditing and conducting risk analysis

Serves as a research artifact

Enables interpreting model output



Only research paper or blog is available and may include overview of

- Training data
- Architecture and training details (including infrastructure)
- Evaluation results
- Adaptation to the model
  - Safety filters
  - Training with human feedback



Safety concerns

Competitive advantage

Expensive to setup guardrails for safe access

### **Model Access**



Open access

### Limited access

### **Closed** access

### Large Language Models since GPT3



### **Open Access Large Language Models**

Research on policy, governance, AI safety and alignment

Community efforts like Eleuther, Big Science, LAION

Papers with several authors

Open source ML has potential for huge impact





>10K datasets

>75K models

measurements

demos

There is an exponential growth of ML models.



There is an exponential growth of ML models.

# models over time



Distribution by task categories



models

Approx 40% of the task categories are NLP

#### Covering 78% of the models



Including multimodal – 55% task categories



Including multimodal – 55% task categories

Including speech – 72% task categories

#### Coverage – 90% of models



Distribution by language (based on 20% models reporting)



## Model Usage

Top 0.2% models (N=124) makeup >80% HF model usage



# Model Usage



98% of these models are trained on just text data



# Model Usage

Top 0.2% models (N=124) makeup >80% HF model usage

98% of these models are trained on just text data

Of these –

65% were created before 2021

33% were created in 2021

2% were created in 2022



Relation between model age and its usage

#### Relation between model age and its usage

Average model usage before 2021



Relation between model age and its usage

Average model usage before 2021



These models served as research artifacts for the later generation of models

#### Relation between model age and its usage

Average model usage before 2021



Average model usage in 2022



Factors:

- 1. Compute is becoming cheaper making model training more accessible
- 2. As more models are created, their usage is distributed
- Models are being replaced by their efficient counterparts (ex: BERT → DistilBERT)

## **Trend Width**

Step 1: Find all peaks in a signal

Step 2: Measure peak widths at base

Step 3: Take the max width



https://huggingface.co/spaces/nazneen/model-usage

Usage trend width for top models



bert-base-uncased

https://huggingface.co/spaces/nazneen/model-usage

Usage trend width for top models



bert-base-uncased

sentence-transformers/paraphrasexlm-r-multilingual-v1

https://huggingface.co/spaces/nazneen/model-usage

Usage trend width for top models



Trend width for models created before 2021



Trend width for models created before 2021





Trend width for models created in 2021
### Model Usage Trends

Trend width for models created in 2022



## Model Usage Trends

Average trend widths of models in 90th percentile of usage:

Created before 2021 → 60 weeks

- Created in 2021 → 45 weeks
- Created in 2022 → 24 weeks

# Model Usage

What other factors might affect model usage?

- What does the model do?
- How does it perform?
- What was it trained on?
- Is it easy to use?
- What are its limitations?

# Model Usage

What other factors might affect model usage?

- What does the model do?
- How good is the model?
- What was it trained on?
- Is it easy to use?
- What are its limitations?

Model documentation!



## Why document models?







#### Model Card - Toxicity in Text

Training Data

**Evaluation Data** 

Proprietary from Perspective API. Following details in [11]

labels of whether the comment is "toxic".

swapped into a variety of template sentences.

statements referencing a variety of groups.

Caveats and Recommendations

and [32], this includes comments from a online forums such

as Wikipedia and New York Times, with crowdsourced

"Toxic" is defined as "a rude, disrespectful, or unreasonable

comment that is likely to make you leave a discussion."

· A synthetic test set generated using a template-based ap-

proach, as suggested in [11], where identity terms are

· Synthetic data is valuable here because [11] shows that

real data often has disproportionate amounts of toxicity

directed at specific groups. Synthetic data ensures that we

evaluate on data that represents both toxic and non-toxic

· Synthetic test data covers only a small set of very specific

comments. While these are designed to be representative of

common use cases and concerns, it is not comprehensive

#### Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic Convolutional Neural Network.
- · Developed by Jigsaw in 2017.

#### Intended Use

- · Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- · Not intended for fully automated moderation. Not intended to make judgments about specific individuals.

#### Factors

- · Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race. Metrics
- · Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups

#### Ethical Considerations

· Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

#### Quantitative Analyses



### Model Card (Mitchell et al., 2019)

Basic Method	Information

- · Name, version, and application domain(s).
- Method purpose and appropriate uses. · Method definition, published literature, reference imple-
- mentation · Example input and output.

#### Safety and Troubleshooting

- · Inappropriate uses and common usage pitfalls.
- · Known weaknesses, biases, and privacy leakage.
- · How to detect biases in the model internals. · Common failure modes, potential root causes, and pos-
- · Fairness evaluation and subgroup comparison. sible mitigations via hyperparameter tuning or training Overfitting detection. Training and inference time efficiency.
- data expansion.

### Data Preparation

- · Input and output format, shape, and data type.
- · Data transformation and normalization. · Recommended sampling and balancing.
- · Recommended batching scheme and batch size. · Required data augmentation and shuffling.
- · Validation and train-test splitting schemes.

### Modelling and Optimization

- · Architecture family and components used. · A list of hyperparameters, along with applicable values and their known impact.
- · Training objective(s), loss(es), and optimizer(s).

### Method Card (Adkins et al., 2022)

@ ×			=
Conteractive Model Card	Quantitative Analysis	Model Performance Metrics	
Data is not permanently collected or stored from your interactions, but is temporarily cached during usage.	View the model's performance or visually explore the model's training and toping dataset Show:	Evaluation metrics include accuracy, precision, and recall.	
Show Marrings	Nodel Performance Metrics *	<ul> <li>Performance is shown for the training and testing set, as well as special groups within this dataset that have been protected groups</li> </ul>	automatically associated with US
fodel Details	Any groups you define via the prodysis actions will be automatically added to the steep	Flag (with a red border) subpopulations with fewer than the follow sentences:	
This model, distribut-base-second-firetaned-ast-2- ception is a sentiment analysis model. The model is trained to	the view	100	- •
analyze a piece of text and then to assess if it has an overall positive or negative sentiment.	Analysis Actions	All subpopulations with fewer than 100 sentences are reporting potentially unreliable results. These are identifies Click on the bars to see example sentences.	d with a red border around the bar.
This model is a fine-tune of a more general language model     called <u>UseIRERT</u> . +	Addify the quartitative analysis results by defining your own subpopulations in the data, including your own data by adding your own sentences or dataset. Explore new subpopulations in model data +	Investigation of the second se	rice
ntended Use Warning! Unintended uses cases are not reported! This model is primarily almed at classifying whether sentences have an overall positive or equative senteneet.	Explore with your own sentences - Write your own examples sentences, or clock Set Suggest Duamples' Hiller you, Hove you	productive: presentative: ettypibete: ettypibete: 02 04 08 00 04 08 02 64 08	
A positive section indicates the passage general conveys an happy, confident, or optimistic sentiment,	Get Suggested Example	Data Details	
A regative sections indicates the passage general conveys a	Model Prediction Summary	Customize Data Sample	+
sad, depressed, or pessimistic sentiment.	The sentiment model predicts that this sentence has an overall Positive Junitimere with dit Entremely High Probability (pr0.998)	The slice protected-age has a total size of skit sectores	
thical Considerations	Do you agree with the prediction?	Shown is a subsample of all the data to be sampled by kundus Sample     Sentences pertaining this US Protected Classes contain the following forms: next, young, old, mature, improve	
Warning Additional bias analysis was not conducted. wen if the training data used for this model could be characterized	Indicate your agreement below Agree Disagree	▲ Detecting US Protected classes by key word search is not perfect. Some sentences below may not be perfected to be a search is not perfect. Some sentences below may not be perfected to black convertences.	
s fairly neutral, this model can have biased predictions. It also sherits some of the bias of the <u>BOIT</u> base model and <u>DistiBEET</u>	Add to exisiting sentences		model label model probability
Addel Training & Evaluation	Explore with your own dotoset +	Abbough largely a heavy-handed indictment of parential failings and the indifference of Spanish social workers and legal system towards child abuse, the film retains ambiguities that make it well worth watching	Positive 1 0.9995 Sentiment
Warning: Dataset is more than five years old	Guidance +	97 the same tired old gags, modernized for the extreme sports generation .	Negative 0 0.9982

Accuracy

100 0

Robustness Report (Goel\*, Rajani\*, et al., NAACL 2021)

Low Constituency Tree Overlap (McCov, 2019)

High Constituency Tree Overlap (McCov, 2019)

Possessive Preposition @ hypothesis (Chen, 2020)

Temporal Preposition @ hypothesis (Chen, 2020)

Negation @ hypothesis (Naik, 2018)

Quantifier @ hypothesis (Chen, 2020)

Low Lexical Overlap (McCoy, 2019)

High Lexical Overlap (McCoy, 2019)

Easy Data Augmentation (Wei, 2019)

Keyboard Character Errors (Ma, 2019)

Synonym Substitution (Ma, 2019)

BAE (Garg, 2019)

SNLI (Bowman, 2015)

0

Negation @ premise (Naik, 2018)

F1

Clace Diet

53 24 23

22 17 61

31

34 27

34 28

13 61 25

52 29 19

13 58 29

33 33 28

33 33

33 33 24

33 33

100 E N C

20 33

20

31

Pred Dist

20 39 41

51 24 25

23 13 64

38 26 36

36 35 29

13 61 25

20 33 46

51 30 20

12 48 4

24 33 4

33

E N C

34 28

Size

2.1K

1.99K

109

39

585

170

106

2.04K

1.98K

2.92K

9.84K

9.14K

9.84K

9.84K

Interactive Model Cards (Crisan, Vig, Drouhard, and Rajani, FAccT2022)

### Method Card Template

· Parameter initialization / self pre-training / transfer from

a trained baseline (specify datasets).

Regularization scheme, capacity selection.

· If applicable, learning rate and schedulers.

· Possibilities to compile the model graph.

Method Benchmarking

Threshold selection.

· Available benchmarks.

Interpretability and Explainability

specific model prediction.

· Out-of-distribution behavior.

using the method.

mended mitigation.

Robustness

help explain model predictions.

· Weight quantization, recommended bit depth.

· Parallelization at training and inference time.

· Recommended model compression techniques.

· Performance metric(s) and applicable threshold(s).

· Applicable feature attribution methods, and how they can

· How to identify influential training instances behind a

· How to identify internal concepts and features learned

Known vulnerabilities to adversarial attacks, and recom-

· Detecting and mitigating data and model drifts.

#### Model Card - Toxicity in Text

Training Data

**Evaluation Data** 

#### Model Details

- · The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic. Convolutional Neural Network.
- · Developed by Jigsaw in 2017.

#### Intended Lise

- · Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- · Not intended for fully automated moderation. Not intended to make judgments about specific individuals.

#### Factors

- · Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race. Metrics
- · Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups

#### Ethical Considerations

· Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

#### Quantitative Analyses



### Model Card (Mitchell et al., 2019)

#### Method Card Template

mended mitigation.

· Out-of-distribution behavior.

Known vulnerabilities to adversarial attacks, and recom-

· Detecting and mitigating data and model drifts.

Robustness

Method Card (Adkins et al., 2022)

s fairly neutral, this model can have biased predictions. It also

#### **Basic Method Information**

- · Name, version, and application domain(s).
- Method purpose and appropriate uses.
- · Method definition, published literature, reference imple mentation

#### · Example input and output. Safety and Troubleshooting

- · Inappropriate uses and common usage pitfalls. · Known weaknesses, biases, and privacy leakage.
- · How to detect biases in the model internals.
- · Common failure modes, potential root causes, and pos-

· Data transformation and normalization.

· Recommended sampling and balancing.

· Recommended batching scheme and batch size.

· Required data augmentation and shuffling.

· Validation and train-test splitting schemes.

· Architecture family and components used.

· Training objective(s), loss(es), and optimizer(s).

· A list of hyperparameters, along with applicable values

Modelling and Optimization

and their known impact.

sible mitigations via hyperparameter tuning or training data expansion.

#### Data Preparation

- · Input and output format, shape, and data type.
- · Synthetic test data covers only a small set of very specific
- swapped into a variety of template sentences. · Synthetic data is valuable here because [11] shows that
- real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

#### Caveats and Recommendations

comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Proprietary from Perspective API. Following details in [11]

labels of whether the comment is "toxic".

and [32], this includes comments from a online forums such

as Wikipedia and New York Times, with crowdsourced

"Toxic" is defined as "a rude, disrespectful, or unreasonable

comment that is likely to make you leave a discussion."

· A synthetic test set generated using a template-based ap-

proach, as suggested in [11], where identity terms are

Interactive Model Card *	Quantitative Analysis	Model Performance Metrics							
Data is not permanently collected or stored from your interactions, but is temporarily cached during usage.	view the model's performance or visually explore the model's total ing and totaling dataset Shee:	Evaluation metrics include accuracy, anticision, and recall.     Performance is shown for the training and testing set, as well as special groups within this dataset that have been automatically as							
Show Marnings	Nodel Performance Metrics								
Model Details	Any groups you define via the prodysis actions will be automatically added to the standard	Flight have a closeff independences with feer that the follow entercore:           100         All subpopulations with Newer that 100 sentences are reporting potentially surreliable results. These are identified with a red be							
<ul> <li>This model, dist (bert-base-accessed-finetaned-sst-2- orgtion, is a sentiment analysis model. The model is trained to analyze a piece of text and then to assess if it has an overall</li> </ul>									
<ul> <li>positive or negative sentiment.</li> <li>This model is a fine ture of a more general language model</li> </ul>	Analysis Actions	Tech chick on the bars to see example sentences.							
called <u>Citable PT</u> .	Modify the quantitative analysis results by defining your own subpopulations in the data, including your own data by adding your own settences or dataset.	protected age Design of the Adaptive Design o							
Intended Use	Explore new subpopulations in model data + Explore with your own sentences -	protected-relativity protected							
Warning Unintended uses cases are not reported     This model is primarily almed at classifying whether sentences	Write your own example sentences, or click Vet Suggest Damples'	satippil-test, venice-10.0) satippil-tesk, venice-10.0) 00 04 08 00 04 08 00 04 08							
have an overall positive or equative sentiment. • A positive sentener indicates the passage general conveys an happy, confident, or optimizic sentiment.	Tilke you. Hove you Get Suggested Duample	Data Details							





Although largely a heavy-handed indictment of parental failings and the indifference of Spanish social Excluse with your own detected workers and legal system towards child abuse , the film retains ambiguities that make it well worth 0.9995 Indel Training & Evaluation Guidance

Interactive Model Cards (Crisan, Vig, Drouhard, and Rajani, FAccT2022)

1 Template		Accuracy	F1	Class Dist	Pred Dist	Size
				and the second second	the second se	
<ul> <li>Parameter initialization / self pre-training / transfer from</li> </ul>	Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20 39 41	20 39 41	2.1K
a trained baseline (specify datasets).	High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53 24 23	51 24 25	1.99K
<ul> <li>Regularization scheme, capacity selection.</li> </ul>	Negation @ hypothesis (Naik, 2018)	90.8	86.0	22 17 61	23 13 64	109
<ul> <li>If applicable, learning rate and schedulers.</li> <li>Weight quantization, recommended bit depth.</li> </ul>	Negation @ premise (Naik, 2018)	79.5	79.5	31 38 31	38 26 36	39 <sup>Sub</sup>
<ul> <li>Possibilities to compile the model graph.</li> </ul>	Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39 34 27	36 35 29	585 8
<ul> <li>Parallelization at training and inference time.</li> <li>Recommended model compression techniques.</li> </ul>	Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38 34 28	39 34 28	170 8
Method Benchmarking	Temporal Preposition @ hypothesis (Chen, 2020)	87.7	86.0	13 61 25	13 61 25	106
<ul> <li>Performance metric(s) and applicable threshold(s).</li> </ul>	Low Lexical Overlap (McCoy, 2019)	90.5	89.6	20 33 47	20 33 45	2.04K
<ul> <li>Threshold selection.</li> </ul>	High Lexical Overlap (McCoy, 2019)	92.7	91.9	52 29 19	51 30 20	1.98K
<ul> <li>Fairness evaluation and subgroup comparison.</li> </ul>						
<ul> <li>Overfitting detection.</li> <li>Training and inference time efficiency.</li> </ul>	BAE (Garg, 2019)	80.3	78.4	13 58 29	12 48 40	2.92K
Available benchmarks.						¥
Interpretability and Explainability	Easy Data Augmentation (Wei, 2019)	82.3	82.2	34 33 33	28 36 36	9.84K g
<ul> <li>Applicable feature attribution methods, and how they can</li> </ul>	Keyboard Character Errors (Ma, 2019)	65.8	65.4	34 33 33	24 33 44	9.14K
<ul> <li>help explain model predictions.</li> <li>How to identify influential training instances behind a</li> </ul>	Synonym Substitution (Ma, 2019)	75.4	75.1	34 33 33	<b>24</b> 36 40	9.84K B
specific model prediction.						e
<ul> <li>How to identify internal concepts and features learned</li> </ul>	SNLI (Bowman, 2015)	90.9	90.9	34 33 33	33 33 34	9.84K
using the method.		0 100	0 100	E N C	E N C	27

Robustness Report (Goel\*, Rajani\*, et al., NAACL 2021)

#### Model Card - Toxicity in Text

Training Data

#### Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic Convolutional Neural Network.
- · Developed by Jigsaw in 2017.

#### Intended Lise

- · Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- · Not intended for fully automated moderation. Not intended to make judgments about specific individuals.

#### Factors

- · Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race. Metrics
- · Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups

#### Ethical Considerations

· Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

#### Quantitative Analyses



### Model Card (Mitchell et al., 2019)

### **Basic Method Information**

- · Name, version, and application domain(s).
- · Method purpose and appropriate uses.
- · Method definition, published literature, reference imple mentation

#### · Example input and output. Safety and Troubleshooting

data expansion.

Modelling and Optimization

Data Preparation

- · Inappropriate uses and common usage pitfalls.
- · Known weaknesses, biases, and privacy leakage.
- · How to detect biases in the model internals. · Common failure modes, potential root causes, and pos-

· Input and output format, shape, and data type.

· Recommended batching scheme and batch size.

· Required data augmentation and shuffling.

· Validation and train-test splitting schemes.

Data transformation and normalization.

· Recommended sampling and balancing.

- sible mitigations via hyperparameter tuning or training
  - · Overfitting detection. · Training and inference time efficiency.

Method Card Template

#### Interpretability and Explainability

mended mitigation.

Out-of-distribution behavior

Method Benchmarking

- · Applicable feature attribution methods, and how they can help explain model predictions.
- · How to identify influential training instances behind specific model prediction.

· Detecting and mitigating data and model drifts.

· How to identify internal concepts and features learne using the method. · Known vulnerabilities to adversarial attacks, and recom-

#### Robustness

Intera

This model, analyze a pie This model is

ntended Use Warning Uni an happy, con sad, depresse

- · Architecture family and components used. · A list of hyperparameters, along with applicable values and their known impact.
- · Training objective(s), loss(es), and optimizer(s).

### Method Card (Adkins et al., 2022)

ıplate		Accuracy	F1	Class Dist	Pred Dist	Size
	Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20 39 41	20 39 41	2.1
<ul> <li>Parameter initialization / self pre-training / transfer from a trained baseline (specify datasets).</li> </ul>	High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53 24 23	51 24 25	1.99
Regularization scheme, capacity selection.	Negation @ hypothesis (Naik, 2018)	93.2	86.0	22 17 61	23 13 64	1.99
<ul> <li>If applicable, learning rate and schedulers.</li> </ul>						
<ul> <li>Weight quantization, recommended bit depth.</li> </ul>	Negation @ premise (Naik, 2018)	79.5	79.5	31 38 31	38 26 36	39
<ul> <li>Possibilities to compile the model graph.</li> </ul>	Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39 34 27	36 35 29	585
Parallelization at training and inference time.     Recommended model compression techniques.	Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38 34 28	39 34 28	170
thod Benchmarking	Temporal Preposition @ hypothesis (Chen, 2020)	87.7	86.0	13 61 25	13 61 25	106
<ul> <li>Performance metric(s) and applicable threshold(s).</li> </ul>	Low Lexical Overlap (McCoy, 2019)	90.5	89.6	20 33 47	20 33 46	2.04
<ul> <li>Threshold selection.</li> </ul>	High Lexical Overlap (McCoy, 2019)	92.7	91.9	52 29 19	51 30 20	1.98
<ul> <li>Fairness evaluation and subgroup comparison.</li> </ul>						
Overfitting detection.	BAE (Garg, 2019)	80.3	78,4	13 58 29	12 48 40	2.92
Training and inference time efficiency.     Available benchmarks.	Une (Guig, 2013)	-00.5	70.4	20 00 29	10 40	2.92
erpretability and Explainability	Easy Data Augmentation (Wei, 2019)	82.3	82.2	34 33 33	<b>28</b> 36 36	9.84
Applicable feature attribution methods, and how they can	Keyboard Character Errors (Ma, 2019)	65.8	65.4	34 33 33	24 33 44	9.14
help explain model predictions.	Synonym Substitution (Ma, 2019)	75.4	75.1	34 33 33	24 36 40	9.84
<ul> <li>How to identify influential training instances behind a</li> </ul>	Synonym Substitution (Ma, 2019)	75.4	75.1			9.04
specific model prediction.	01011 (0.00000 00010)					
<ul> <li>How to identify internal concepts and features learned using the method.</li> </ul>	SNLI (Bowman, 2015)	90.9	90.9	34 33 33	33 33 34	9.84
using the method.		0 100 0	) 100	E N C	E N C	

### Robustness Report (Goel\*, Rajani\*, et al., NAACL 2021)

active Model Card	Quantitative Analysis	Performance Metrics					=	
manently collected or stored from your ut is temporarily cached during usage.	View the model's performance or visually explore the model's training and toping detaset Shee:	uation metrics include accurs		ial groups within this dataset that have been		uncisted at	95.05	
ngs	Nodel Performance Metrics *	icted groups						
distillent-box-uncased-firetared-sst-2- sentiment analysis model. The model is trained to	Any groups you define via the environ ections will be automatically added to the view		th fewer than the follow sentences:				- +	
ree of text and then to assess if it has an overall egative sentiment. s a fine turns of a more general language model SERI-	Analysis Actions Modify the quarkitative analysis results by defining your own subpopulations in the data, indusing your own data by adding your own sentences or dataset. <i>Explore new subpopulations in model</i> data	on the bars to see example s		ntial y unreliable results. These are identified mail Data Subpopulation III During Performance Using Control Data Using Control Data	co torico	ber around t	De Sal	
intended uses cases are not reported! a primarily almed at classifying whither sentences all positive or equitive sentiment.	Explare with year own Sentences	protected-race ct-sexual-orientation protected-veteran- elect, vension=1.0.0) train, vension=1.0.0) 0.0 0.	4 0.8 0.0 0.4 0.8	Asathemes >=100 sentences = 100 sentences				
unctionet indicates the passage general conveys infident, or optimistic sentiment.	Get Suggested Example	etails						
ant inert indicates the passage general conveys a ed, or pessimistic sentiment.	Model Prediction Summary The sentiment model predicts that this sentence has an oversil	e Doto Sample ilice protected-age has a to	talsize of siz servesces				+	
erations Efforts flux analysis was not conducted. Agato used for this model could be characterized this model can have biased prodictions. It also the bias of the gott base model and Disability.	Practice: Intercomer With a prediction? Do you agree with the prediction? Holdmay or upper Tables      Agree () Integree      Af the resisting sonstress	Show is subanyled of the data to a subject by user, sport     Book subanyled of the data to a subject by the data to a sport of the						
& Evaluation	Explore with your own dotoset +			and the indifference of Spanish social mbiguities that make it well worth watching	Positive Sentiment	1	0.9995	
	Guidance +	the same blood and some or	described des bles esternes seconds		Negative		0.0000	

Interactive Model Cards (Crisan, Vig, Drouhard, and Rajani, FAccT2022)

#### Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic". "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." **Evaluation Data** · A synthetic test set generated using a template-based ap-

swapped into a variety of template sentences.

statements referencing a variety of groups.

**Caveats and Recommendations** 

proach, as suggested in [11], where identity terms are

· Synthetic data is valuable here because [11] shows that

real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we

evaluate on data that represents both toxic and non-toxic

· Synthetic test data covers only a small set of very specific

comments. While these are designed to be representative of

common use cases and concerns, it is not comprehensive.

#### Model Card - Toxicity in Text

Training Data

**Evaluation Data** 

Proprietary from Perspective API. Following details in [11]

labels of whether the comment is "toxic"

swapped into a variety of template sentences.

statements referencing a variety of groups.

Caveats and Recommendations

and [32], this includes comments from a online forums such

as Wikipedia and New York Times, with crowdsourced

"Toxic" is defined as "a rude, disrespectful, or unreasonable

comment that is likely to make you leave a discussion."

· A synthetic test set generated using a template-based ap-

proach, as suggested in [11], where identity terms are

· Synthetic data is valuable here because [11] shows that

real data often has disproportionate amounts of toxicity

directed at specific groups. Synthetic data ensures that we

evaluate on data that represents both toxic and non-toxic

· Synthetic test data covers only a small set of very specific

comments. While these are designed to be representative of

common use cases and concerns, it is not comprehensive

#### Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic
- Convolutional Neural Network. · Developed by Jigsaw in 2017.

#### Intended Lie

- · Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- · Not intended for fully automated moderation. Not intended to make judgments about specific individuals.

#### Factors

- · Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race. Metrics
- · Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups

#### Ethical Considerations

· Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

#### Quantitative Analyses



### Model Card (Mitchell et al., 2019)

#### Method Card Template

### **Basic Method Information**

- · Name, version, and application domain(s).
- · Method purpose and appropriate uses.
- · Method definition, published literature, reference imple mentation
- · Example input and output.

#### Safety and Troubleshooting

data expansion.

Modelling and Optimization

and their known impact.

Data Preparation

- · Inappropriate uses and common usage pitfalls.
- · Known weaknesses, biases, and privacy leakage. · How to detect biases in the model internals.
- · Common failure modes, potential root causes, and possible mitigations via hyperparameter tuning or training

· Input and output format, shape, and data type.

· Recommended batching scheme and batch size.

· Required data augmentation and shuffling.

· Validation and train-test splitting schemes.

· Architecture family and components used.

· A list of hyperparameters, along with applicable values

Data transformation and normalization.

· Recommended sampling and balancing.

- · Overfitting detection.
  - · Training and inference time efficiency. · Available benchmarks.

#### Interpretability and Explainability

Method Benchmarking

Threshold selection

· Applicable feature attribution methods, and how they can help explain model predictions.

· Parameter initialization / self pre-training / transfer from

a trained baseline (specify datasets).

Regularization scheme, capacity selection.

· If applicable, learning rate and schedulers.

- · How to identify influential training instances behind a specific model prediction.
- · How to identify internal concepts and features learned using the method.
- Robustness
- · Known vulnerabilities to adversarial attacks, and recom-
- mended mitigation.
- · Out-of-distribution behavior. · Detecting and mitigating data and model drifts.

### · Training objective(s), loss(es), and optimizer(s). Method Card (Adkins et al., 2022)



Low Constituency Tree Overlap (McCov, 2019)

High Constituency Tree Overlap (McCov, 2019)

Negation @ hypothesis (Naik, 2018)

Accuracy

F1

Clace Diet

53 24 23

22 17 61

20

Pred Dist

20 39 41

51 24 25

23 13 64

Size

2.1K

1.99K

109

39

585

170

106

2.04K

1.98K

2.92K

9.84K

9.14K

9.84K

9.84K

Model Card *	Quantitative Analysis	Model Performance Metrics	
Data is not permanently collected or stored from your	View the model's performance or visually explore the model's training and toping dataset Shew:	Evaluation metrics include accuracy, practicito, and recall.     Performance is above for the training and testing up, as well as special groups within this dataset that have been autowatically associated w	ah US
Show Warnings	Model Performance Metrica •	<ul> <li>Perturbative is stored or reacting and resong ter, as well as special groups within this causes, that have seen automaticary associated with protected groups.</li> </ul>	0103
Model Details This model, distillant-base-uncased-firetaned-ast-2-	Any groups you define via the environ ections will be automatically added to the view	Flig left a ref border subpapalation with four the tables serverses:	
This model is a fine turne of a more general language model     called <u>triatilitERT</u> . +	Analysis Actions Modify the quantizative analysis results by defining your own subpopulations in the data, including your own data by adding your own settences or distance. Explore new subpopulations in model datas +	All Hougedons with lever the LB or correction requiring straticity services multiple multi- 1 G d on the fact to ser countil services an equiring straticity multiple multinter multinter multiple multiple multiple multiple multiple multi	Debai.
Intended Use  Intended Use  Iterating Unknowled uses cases are not reported  This model is privarily aimed at classifying which ar someone have an overall assifture of regaries settiment.  A prestructure sectore indicates the passage general conveys an abayous confidence regaring assistance.	Epiptine men angespectation at master assa  Epiptine with your even sentences  Nither your own reample antencos, or clos Vart Suggest Examples'  Hiller you. I love you	Description (United View) Paradia Statut description (United View) Par	
	Get Suggested Exemple	Dans Details	
A regative section is to document structure.     A regative section is indicated the passage general conveys a sad, dopressed, or pessimistic serviment.     +	Model Prediction Summary The sensitivest model predicts that this sentence has an overall Prediction Exercises: with de Extremely stigh Prediction (by GP0.599)	Constantian Botto Georgia • There Silver processed upper Natl and Botto Silver of 1932 surfacements	
Ethical Considerations  Kerning: Additional bias analysis was not conducted. Even if the mining data used for this model could be characterised at fairly wards (1) is model can have biased predictions. It also	Do you agree with the prediction? Indicate your approment below Agree D Disagree	Shows & a subcarder of all the data to as sampled by some single     some same paralleling this to Potential Classes sential the following terms aper, young, sitz, minure, minure, citiz, journils     Detecting 50 Potential classes by word march is not perfect. Serie sentences believe may not be perfect sent as presented class, the event     word can refer indexidual location and word refers. Series sentences believe may not be perfect sent as presented class, the event	ple the wo
inherits some of the bias of the BERT base model and DistREET	Add to existing sentences	sentence model label model binary	probabili
Model Training & Evaluation	Explore with your own dotoset +	Although largely a heavy-handed indictment of parental failings and the indifference of Spanish social 6% workers and legal system towards child abuse, the film retains ambiguities that make it well worth watching 5entiment.	0.99
Warning: Dataset is more than five years old	Guidance +	97 the same tired old gags , modernized for the extreme sports generation . Negative 0 Sectioners	0.99

### Robustness Report (Goel\*, Rajani\*, et al., NAACL 2021)



Model documentation is part of the repo's README

B	Text Generation 🕜 PyTorch 🕈 TensorFlow 🔐 JAX
9	Model card HE Files and versions 🖉 Community 16
Y	main - gpt2
	sgugger HF STAFF 🕘 mathemakitten HF STAFF Add not
C	.gitattributes
B	64- <sup>p</sup> '
D	64.tflite 🕘
	README.md
B	config.json 💿
B	config. <sub>J-</sub>

### **Model description**

GPT-2 is a transformers model pretrained on a very large corpus of English data in a selfsupervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was trained to guess the next word in sentences.

More precisely, inputs are sequences of continuous text of a certain length and the targets are the same sequence, shifted one token (word or piece of word) to the right. The model uses internally a mask-mechanism to make sure the predictions for the token i only uses the inputs from 1 to i but not the future tokens.

This way, the model learns an inner representation of the English language that can then be used to extract features useful for downstream tasks. The model is best at what it was pretrained for however, which is generating texts from a prompt.

This is the smallest version of GPT-2, with 124M parameters.

### Training data

The OpenAI team wanted to train this model on a corpus as large as possible. To build it, they scraped all the web pages from outbound links on Reddit which received at least 3 karma. Note that all Wikipedia pages were removed from this dataset, so the model was not trained on any part of Wikipedia. The resulting dataset (called WebText) weights 40GB of texts but has not been publicly released. You can find a list of the top 1,000 domains present in WebText <u>here</u>.

### Preprocessing

The texts are tokenized using a byte-level version of Byte Pair Encoding (BPE) (for unicode characters) and a vocabulary size of 50,257. The inputs are sequences of 1024 consecutive tokens.

The larger model was trained on 256 cloud TPU v3 cores. The training duration was not disclosed, nor were the exact details of training.

### **Limitations and bias**

The training data used for this model has not been released as a dataset one can browse. We know it contains a lot of unfiltered content from the internet, which is far from neutral. As the openAI team themselves point out in their <u>model card</u>:

"Because large-scale language models like GPT-2 do not distinguish fact from fiction, we don't support use-cases that require the generated text to be true.

Additionally, language models like GPT-2 reflect the biases inherent to the systems they were trained on, so we do not recommend that they be deployed into systems that interact with humans > unless the deployers first carry out a study of biases relevant to the intended use-case. We found no statistically significant difference in gender, race, and religious bias probes between 774M and 1.5B, implying all versions of GPT-2 should be approached with similar levels of caution around use cases that are sensitive to biases around human attributes."

### **Intended uses & limitations**

You can use the raw model for text generation or fine-tune it to a downstream task. See the <u>model hub</u> to look for fine-tuned versions on a task that interests you.

### How to use

You can use this model directly with a pipeline for text generation. Since the generation relies on some randomness, we set a seed for reproducibility:

>>> from transformers import pipeline, set\_seed
>>> generator = pipeline('text-generation', model='gpt2')
>>> set\_seed(42)

>>> generator("Hello, I'm a language model,", max\_length=30, num\_retu

[{'generated\_text': "Hello, I'm a language model, a language for thir {'generated\_text': "Hello, I'm a language model, a compiler, a compi {'generated\_text': "Hello, I'm a language model, and also have more {'generated\_text': "Hello, I'm a language model, a system model. I w {'generated\_text': 'Hello, I\'m a language model, not a language mocel

### **Evaluation results**

The model achieves the following results without any fine-tuning (zero-shot):

			CBT-	CBT-					
Dataset	LAMBADA	LAMBADA	CN	NE	WikiText2	ΡΤΒ	enwiki8	text8	WikiText1
(metric)	(PPL)	(ACC)	(ACC)	(ACC)	(PPL)	(PPL)	(BPB)	(BPC)	(PPL)
	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1,17	37.50

### **Model documentation statistics**

Distribution of models with documentation over time



Newer models are less likely to have model cards



## **Model Documentation vs. Usage**

**Observation:** Only 50% models have model cards but contribute 98% of total usage

## **Model Documentation vs. Usage**

**Observation:** Only 50% models have model cards but contribute 98% of total usage

Goal: Study the relation between model usage and documentation

## **Model Documentation vs. Usage**

**Observation:** Only 50% models have model cards but contribute 98% of total usage

**Goal:** Study the relation between model usage and documentation **Hypothesis:** model documentation drives model usage

**Observation:** Only 50% models have model cards but contribute 98% of total usage

**Goal:** Study the relation between model usage and documentation **Hypothesis:** model documentation drives model usage

**Observation:** Only 50% models have model cards but contribute 98% of total usage

**Goal:** Study the relation between model usage and documentation **Hypothesis:** model documentation drives model usage

Randomized Control Trial (RCT) for models:



Model population

**Observation:** Only 50% models have model cards but contribute 98% of total usage

**Goal:** Study the relation between model usage and documentation **Hypothesis:** model documentation drives model usage



**Observation:** Only 50% models have model cards but contribute 98% of total usage

**Goal:** Study the relation between model usage and documentation **Hypothesis:** model documentation drives model usage



**Observation:** Only 50% models have model cards but contribute 98% of total usage

**Goal:** Study the relation between model usage and documentation **Hypothesis:** model documentation drives model usage













Treatment group

Documentation is part of model repo



### **RCT Results**

Red line indicates week when treatment was administered



## **RCT Results**

Red line indicates week when treatment was administered



## **Model Documentation RCT Findings**

Increased usage of models in treatment group compared to control group

More prominent for model weights downloads

Model documentation drives model usage

### What do developers document about models?

Distribution of sections in model cards

Percentage of non-empty sections



## What do developers document about models?

Distribution of sections in model cards

Percentage of non-empty sections



# Outline

### Part 1:

NLP Modeling landscape

Systematic study of 75K models on HF

### Part 2:

NLP Evaluation landscape

Challenges and opportunities in model evaluation and documentation
Slew of work on evaluation in NLP

Slew of work on evaluation in NLP







### Tools



Errudite: Scalable, Reproducible, and Testable Error Analysis

Tongshuang Wu<sup>1</sup>, Marco Tulio Ribeiro<sup>2</sup>, Jeffrey Heer<sup>1</sup>, and Daniel S. Weld<sup>1</sup>

<sup>1</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington <sup>2</sup>Microsoft Research {wtshuang, jheer, weld}@cs.washington.edu marcotcr@microsoft.com

#### Beyond Accuracy: Behavioral Testing of NLP Models with CHECKLIST

Marco Tulio Ribeiro Microsoft Research marcotcr@microsoft.com Tongshuang Wu Univ. of Washington wtshuang@cs.uw.edu Carlos Guestrin Sameer Singh Univ. of Washington Univ. of California, Irvine guestrin@cs.uw.edu sameer@uci.edu



TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP

John X. Morris<sup>1</sup>, Eli Lifland<sup>1</sup>, Jin Yong Yoo<sup>1</sup>, Jake Grigsby<sup>1</sup>, Di Jin<sup>2</sup>, Yanjun Qi<sup>1</sup> <sup>1</sup> Department of Computer Science, University of Virginia <sup>2</sup> Computer Science and Artificial Intelligence Laboratory, MIT {jm8wx, yq2h}@virginia.edu

#### SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

Alex Wang*	Yada Pruksachati	Nikita Nangia*	
New York University	New York Univer	sity	New York University
Amanpreet Singh* Facebook AI Research	Julian Michael University of Washington	Felix Hill DeepMind	Omer Levy Facebook AI Research

Samuel R. Bowman New York University

Slew of work on evaluation in NLP

### **Papers**

Behavior Analysis of NLI Models: Uncovering the Influence of Three Factors on Robustness

V. Ivan Sanchez Carmona and Jeff Mitchell and Sebastian Riedel University College London Department of Computer Science {i.sanchezcarmona, j.mitchell, s.riedel}@cs.ucl.ac.uk

#### Universal Adversarial Triggers for Attacking and Analyzing NLP

WARNING: This paper contains model outputs which are offensive in nature.

Eric Wallace<sup>1</sup>, Shi Feng<sup>2</sup>, Nikhil Kandpal<sup>3</sup>, Matt Gardner<sup>1</sup>, Sameer Singh<sup>4</sup> <sup>1</sup>Allen Institute for Artificial Intelligence, <sup>2</sup>University of Maryland <sup>3</sup>Independent Researcher, <sup>4</sup>University of California, Irvine ericw@allenai.org, sameer@uci.edu

#### How well do NLI models capture verb veridicality?

Alexis Ross Ellie Pavlick Harvard University Brown University alexis\_ross@college.harvard.edu ellie\_pavlick@brown.edu

#### Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan<sup>★</sup> Swabha Swayamdipta<sup>★</sup> Omer Levy<sup>♣</sup> Roy Schwartz<sup>♣</sup> Samuel R. Bowman<sup>↑</sup> Noah A. Smith<sup>♣</sup>

#### Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie\*, Adina Williams<sup>†</sup>, Emily Dinan<sup>†</sup>, Mohit Bansal\*, Jason Weston<sup>†</sup>, Douwe Kiela<sup>†</sup> \*UNC Chapel Hill <sup>†</sup>Facebook AI Research

#### Stress Test Evaluation for Natural Language Inference

Aakanksha Naik<sup>1</sup>', Abhilasha Ravichander<sup>1</sup>', Norman Sadeh<sup>2</sup>, Carolyn Rose<sup>1</sup>, Graham Neubig<sup>1</sup> <sup>1</sup>Language Technologies Institute, Carnegie Mellon University <sup>2</sup>Institute of Software Research, Carnegie Mellon University {anaik, aravicha, sadeh, cprose, gneubig}@cs.cmu.edu

### LEARNING THE DIFFERENCE THAT MAKES A DIFFERENCE WITH COUNTERFACTUALLY-AUGMENTED DATA

Divyansh Kaushik, Eduard Hovy, Zachary C. Lipton Carnegie Mellon University Pittsburgh PA, USA {dkaushik, hovy, zlipton}@cmu.edu

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

**Example:** short reviews (< 50 words) in the IMDB sentiment dataset

Tools: Snorkel (Ratner et al., 2017), Errudite (Wu et al., 2019)

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

2. Transformations – natural perturbations to original evaluation instances

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

2. **Transformations** – natural perturbations to original evaluation instances

### **Example:** substitute words with their synonyms in the IMDB dataset

Tools: NLPAug (Ma, 2019)

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

2. Transformations – natural perturbations to original evaluation instances

3. Evaluation sets – evaluation on diagnostic sets

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

2. Transformations – natural perturbations to original evaluation instances

3. **Evaluation sets** – evaluation on diagnostic sets

**Example:** write new movie reviews in the style of a newspaper columnist

**Tools:** CheckList (Ribeiro et al., 2020)

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

2. Transformations – natural perturbations to original evaluation instances

3. Evaluation sets – evaluation on diagnostic sets

4. Attacks – adversarial evaluation

1. Subpopulations – disaggregate evaluation on slice or subpopulation of data

2. Transformations – natural perturbations to original evaluation instances

3. Evaluation sets – evaluation on diagnostic sets

4. Attacks – adversarial evaluation

**Example:** add "aabbccaa" to reviews because it makes the model predict positive sentiment

**Tools:** TextAttack (Morris et al., 2020), OpenAttack (Zeng et al., 2020)

Slew of work on evaluation in NLP -- tools and research papers

<b>Evaluation Idiom</b>	Tools Available	Research Literature (focusing on NLI)
Subpopulations	Snorkel [Ratner et al., 2017], Errudite [Wu et al., 2019]	Hard/easy sets [Gururangan et al., 2018] Compositional-sensitivity [Nie et al., 2019]
Transformations	NLPAug [Ma, 2019]	Counterfactuals [Kaushik et al., 2019], Stress test [Naik et al., 2018], Bias factors [Sanchez et al., 2018], Verb veridicality [Ross and Pavlick, 2019]
Attacks	TextAttack [Morris et al., 2020], OpenAttack [Zeng et al., 2020] Dynabench [Kiela et al., 2020]	Universal Adversarial Triggers [Wallace et al., 2019], Adversarial perturbations [Glockner et al., 2018], ANLI [Nie et al., 2020]
Evaluation Sets	SuperGLUE diagnostic sets [Wang et al., 2019] Checklist [Ribeiro et al., 2020]	FraCaS [Cooper et al., 1994], RTE [Dagan et al., 2005], SICK [Marelli et al., 2014], SNLI [Bowman et al., 2015], MNLI [Williams et al., 2018], HANS [McCoy et al., 2019], Quantified NLI [Geiger et al., 2018], MPE [Lai et al., 2017], EQUATE [Ravichander et al., 2019], DNC [Poliak et al., 2018], ImpPres [Jeretic et al., 2020], Systematicity [Yanaka et al., 2020] ConjNLI [Saha et al., 2020], SherLIiC [Schmitt and Schütze, 2019]

# **Goldilocks spectrum for Model Evaluation**



# **Challenges with Evaluation**

#### A Nerdist

### Twitter's Cropping Algorithm Shows Evidence of Racial Bias

(Note: you need to view the tweets on Twitter, and open the images, in order to see the algorithm's selections.) I wonder if Twitter does this to ... 1 month ago

#### V The Verge

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Google said it was "appalled" at the mistake, apologized to Alciné, ... The publication also found that Google had restricted its Al recognition in other racial categories. ... remained blocked on Google Photos after Alciné's tweet Jan 12, 2018

#### WIRED



### The Apple Card Didn't 'See' Gender—and That's the Problem

WIRED. The Apple Card Didn't 'See' Gender—and That's the Problem ... Even Apple's amiable cofounder, Steve Wosniak, wondered, more politely, ... bank for the Apple Card, insisted right away that there isn't any gender Nov 19, 2019

#### Reuters

### Amazon scraps secret AI recruiting tool that showed bias against women

Amazon scraps secret AI recruiting tool that showed bias against women ... uncovered a big problem: their new recruiting engine did not like women. ... has more than tripled to 575,700 workers, regulatory filings show. Oct 10, 2018



#### QZ Quartz

Microsoft's Zo chatbot is a politically correct version of her sister Tay-except she's much, much worse

Microsoft's politically correct chatbot is even worse than its racist one. zo screenshot chatbot. Screenshot/Microsoft. There's nothing loljk about ... Jul 31, 2018



#### VB VentureBeat

### Al Weekly: Facebook's discriminatory ad targeting illustrates the dangers of biased algorithms

This summer has been littered with stories about algorithms gone awry. For one example, a recent study found evidence Facebook's ad ... 1 month ago



### **Challenges with Evaluation**

**Clever Hans effect** 



\* Translation: What is ten plus ten?

## **Challenges with evaluation**



# **Challenges with evaluation**



## **Challenges with evaluation**





### **Robustness Gym**



### (Goel\*, Rajani\*, et al., NAACL 2021)

### **Robustness Gym**



(Goel\*, Rajani\*, et al., NAACL 2021)

### **Robustness Gym**



(Goel\*, Rajani\*, et al., NAACL 2021)













	Accuracy	F1	Class I	Dist	Pre	ed D	ist	Size	
Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20 39	41	20	39	41	2.1K	
High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53 24	23	51	24	25	1.99K	
Negation @ hypothesis (Naik, 2018)	90.8	86.0	22 17	61	23	13	64	109	(0
Negation @ premise (Naik, 2018)	79.5	79.5	31 38	31	38	26	36	39	subpopulation
Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39 34	27	36	35	29	585	opul
Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38 34	28	39	34	28	170	atior
Temporal Preposition @ hypothesis (Chen, 2020)	87.7	86.0	13 61	25	13	61	25	106	2
Low Lexical Overlap (McCoy, 2019)	90.5	89.6	20 33	47	20	33	46	2.04K	
High Lexical Overlap (McCoy, 2019)	92.7	91.9	52 29	19	51	30	20	1.98K	
				_			_		at
BAE (Garg, 2019)	80.3	78.4	13 58	29	12	48	40	2.92K	attack
Easy Data Augmentation (Wei, 2019)	82.3	82.2	34 33	33	28	36	36	9.84K	
							30		transform
Keyboard Character Errors (Ma, 2019)	65.8	65.4	34 33	33	24	33	44	9.14K	sfor
Synonym Substitution (Ma, 2019)	75.4	75.1	34 33	33	24	36	40	9.84K	'n
									ev
SNLI (Bowman, 2015)	90.9	90.9	34 33	33	33	33	34	9.84K	evalset
	0 100	0 100	E N	С	Е	Ν	С		t

	Accuracy	F1	Class Dist		Pred Dist		ist	Size	<b>~</b> -
Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20 3	9 41	20	39	41	2.1K	1
High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53 <mark>2</mark>	4 23	51	24	25	1.99K	1
Negation @ hypothesis (Naik, 2018)	90.8	86.0	22 1	7 61	23	13	64	109	S
Negation @ premise (Naik, 2018)	79.5	79.5	31 3	3 31	38	26	36	39	ubp
Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39 <b>3</b>	4 27	36	35	29	585	subpopulation
Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38 3	4 28	39	34	28	170	atior
Temporal Preposition @ hypothesis (Chen, 2020)	87.7	86.0	13 6	1 25	13	61	25	106	2
Low Lexical Overlap (McCoy, 2019)	90.5	89.6	20 3	3 47	20	33	46	2.04K	1
High Lexical Overlap (McCoy, 2019)	92.7	91.9	52 <mark>2</mark>	9 19	51	30	20	1.98K	с ¦
						_			at
BAE (Garg, 2019)	80.3	78.4	13 5	8 29	12	48	40	2.92K	attack
Easy Data Augmentation (Wei, 2019)	82.3	82.2	34 3	3 33	28	36	36	0.841	<b>_</b>
Keyboard Character Errors (Ma, 2019)	65.8	65.4	34 3		24		10	9.84K 9.14K	ransfo
Synonym Substitution (Ma, 2019)	75.4	75.1	34 <b>3</b>		24		44	9.84K	form
Synonym Substitution (Ma, 2019)	/5.4	75.1	34 3	5 55	24	30	40	9.04K	
SNLI (Bowman, 2015)	90.9	90.9	34 3	3 33	33	33	34	9.84K	evals
	0 100 (	0 100	ΕN	С	Е	N	С		set

<pre></pre>	Accuracy	F1	Class Dist	Pred Dist	Size
Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20 39 41	20 39 41	2.1K
High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53 24 23	51 24 25	1.99K
Negation @ hypothesis (Naik, 2018)	90.8	86.0	22 17 61	23 13 64	109
Negation @ premise (Naik, 2018)	79.5	79.5	31 38 31	38 26 36	39 subp
Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39 34 27	36 35 29	39 subpopulation
Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38 34 28	39 34 28	170 atio
Temporal Preposition @ hypothesis (Chen, 2020)	87.7	86.0	13 61 25	13 61 25	106
Low Lexical Overlap (McCoy, 2019)	90.5	89.6	20 33 47	20 33 46	2.04K
High Lexical Overlap (McCoy, 2019)	92.7	91.9	52 29 19	51 30 20	1.98K
BAE (Garg, 2019)	80.3	78.4	13 58 29	<b>12</b> 48 40	2.92K attack
Easy Data Augmentation (Wei, 2019)	82.3	82.2	34 33 33	<b>28</b> 36 36	9.84K 5
Keyboard Character Errors (Ma, 2019)	65.8	65.4	34 33 33	24 33 44	9.84K transform 9.14K 9.84K
Synonym Substitution (Ma, 2019)	75.4	75.1	34 33 33	24 36 40	9.84K
SNLI (Bowman, 2015)	90.9	90.9	34 33 33 E N C	33 <b>33</b> 34	9.84K evalset
	100 0	100			







### **Experiments with Commercial APIs for Named Entity Linking**

### Named Entity Linking

When did England last win the football world cup?

map "strings" to "things" in a knowledge base like Wikipedia

### **Experiments with Commercial APIs for Named Entity Linking**

### **Named Entity Linking**

map "strings" to "things" in a knowledge base like Wikipedia



### **Experiments with Commercial APIs for Named Entity Linking**

### **Named Entity Linking**

map "strings" to "things" in a knowledge base like Wikipedia



Question Answering System

**Downstream System**


### A correct NEL is required for the downstream system!

	Amazon	Google	Microsoft	Рор	Size	
All	52.5	48.5	54.7	56.4	2.46K	
EntityCapitalization(All)	54.6	54.1	66.0	56.1	1.4K	
EntityCapitalization(None)	49.6	38.2	35.7	56.3	909	
EntityPopularity(Bottom 10%)	44.0	35.1	46.4	46.0	247	
EntityPopularity(Top 10% Variability)	66.2	79.9	71.3	73.4	247	
EntityPopularity(Top 10%)	52.2	54.0	53.9	61.7	264	
NumEntities(1)	49.6	38.6	44.2	53.7	1.37K	
NumEntities(Top 10%)	57.1	62.7	69.4	59.7	428	
Sport(Alpine)	77.1	83.8	82.9	79.7	155	10
Sport(Badminton)	76.8	68.9	67.5	70.7	24	subj
Sport(Basketball)	54.8	57.4	27.8	59.9	37	pop
Sport(Cricket)	48.2	31.7	50.7	51.2	124	subpopulations
Sport(Freestyle)	67.7	81.7	72.1	73.5	44	ion
Sport(Golf)	69.6	72.1	63.8	77.8	30	S
Sport(NBA)		1		8	99	
Sport(NFL)	30.1	24.1	20.7	25.4	65	
Sport(NHL)	8.61	2	<b>2</b>	18,2	107	
Sport(Nordic)	54.3	64.9	76.2	64.1	20	
Sport(Rugby)	36.3	25.9	45.5	44.5	63	
Sport(Skating)	79.5	80.7	91.6	75.8	42	
Sport(Skiing)	54.9	56.8	65.9	66.6	22	
Sport(Soccer)	54.2	41.3	60.9	56.4	654	
	0 100		0 0 100			
Robustness Re	port fo	r NEL o	n AIDA-k	o datase	t	
	<b>—</b>					

Popularity heuristic outperforms all commercial systems



	Amazon	Google	Microsoft	Рор	Size	
All	52.5	48.5	54.7	56.4	2.46K	
EntityCapitalization(All)	54.6	54.1	66.0	56.1	1.4K	
EntityCapitalization(None)	49.6	38.2	35.7	56.3	909	
EntityPopularity(Bottom 10%)	44.0	35.1	46.4	46.0	247	
EntityPopularity(Top 10% Variability)	66.2	79.9	71.3	73.4	247	
EntityPopularity(Top 10%)	52.2	54.0	53.9	61.7	264	
NumEntities(1)	49.6	38.6	44.2	53.7	1.37K	
NumEntities(Top 10%)	57.1	62.7	69.4	59.7	428	
Sport(Alpine)	77.1	83.8	82.9	79.7	155	
Sport(Badminton)	76.8	68.9	67.5	70.7	24	subpopulations
Sport(Basketball)	54.8	57.4	27.8	59.9	37	pop
Sport(Cricket)	48.2	31.7	50.7	51.2	124	ulat
Sport(Freestyle)	67.7	81.7	72.1	73.5	44	ion
Sport(Golf)	69.6	72.1	63.8	77.8	30	S
Sport(NBA)				2	99	
Sport(NFL)	30.1	24.1	20.7	25,4	65	
Sport(NHL)	19.8		<b>a</b>	18,2	107	
Sport(Nordic)	54.3	64.9	76.2	64.1	20	
Sport(Rugby)	36.3	25.9	45.5	44.5	63	
Sport(Skating)	79.5	80.7	91.6	75.8	42	
Sport(Skiing)	54.9	56.8	65.9	66.6	22	
Sport(Soccer)	54.2	41.3	60.9	56.4	654	
	0 10	0 0 10	0 0 100	0 10	0	

Commercial APIs are not any more robust than popularity heuristic

Robustness Report for NEL on AIDA-b dataset

Commercial systems are capitalization sensitive

ſ		Amazon	Google	Microsoft	Рор	Size	
	All	52.5	48.5	54.7	56.4	2.46K	
	EntityCapitalization(All)	54.6	54.1	66.0	56.1	1.4K	
E	EntityCapitalization(None)	49.6	38.2	35.7	56.3	909	
Entity	yPopularity(Bottom 10%)	44.0	35.1	46.4	46.0	247	
EntityPopula	arity(Top 10% Variability)	66.2	79.9	71.3	73.4	247	
E	ntityPopularity(Top 10%)	52.2	54.0	53.9	61.7	264	
	NumEntities(1)	49.6	38.6	44.2	53.7	1.37K	
	NumEntities(Top 10%)	57.1	62.7	69.4	59.7	428	
	Sport(Alpine)	77.1	83.8	82.9	79.7	155	10
	Sport(Badminton)	76.8	68.9	67.5	70.7	24	lqns
	Sport(Basketball)	54.8	57.4	27.8	59.9	37	doc
	Sport(Cricket)	48.2	31.7	50.7	51.2	124	subpopulations
	Sport(Freestyle)	67.7	81.7	72.1	73.5	44	ion
	Sport(Golf)	69.6	72.1	63.8	77.8	30	S
	Sport(NBA)				8	99	
	Sport(NFL)	30.1	24.1	20,7	25,4	65	
	Sport(NHL)	19.8		E Contra	18.2	107	
	Sport(Nordic)	54.3	64.9	76.2	64.1	20	
	Sport(Rugby)	36.3	25.9	45.5	44.5	63	
	Sport(Skating)	79.5	80.7	91.6	75.8	42	
	Sport(Skiing)	54.9	56.8	65.9	66.6	22	
	Sport(Soccer)	54.2	41.3	60.9	56.4	654	
			0 0 100			-	
Ro	obustness Re	<b>port</b> fo	r NEL or	n AIDA-k	o datase	t	

	Amazon	Google	Microsoft	Рор	Size
All	52.5	48.5	54.7	56.4	2.46K
EntityCapitalization(All)	54.6	54.1	66.0	56.1	1.4K
EntityCapitalization(None)	49.6	38.2	35.7	56.3	909
EntityPopularity(Bottom 10%)	44.0	35.1	46.4	46.0	247
EntityPopularity(Top 10% Variability)	66.2	79.9	71.3	73.4	247
EntityPopularity(Top 10%)	52.2	54.0	53.9	61.7	264
NumEntities(1)	49.6	38.6	44.2	53.7	1.37K
NumEntities(Top 10%)	57.1	62.7	69.4	59.7	428
Sport(Alpine)	77.1	83.8	82.9	79.7	155
Sport(Badminton)	76.8	68.9	67.5	70.7	24
Sport(Basketball)	54.8	57.4	27.8	59.9	24 Suppopulations 37 124 44
Sport(Cricket)	48.2	31.7	50.7	51.2	124
Sport(Freestyle)	67.7	81.7	72.1	73.5	44
Sport(Golf)	69.6	72.1	63.8	77.8	30
Sport(NBA)				2	99
Sport(NFL)	30.1	24.1	20,7	25,4	65
Sport(NHL)	19.8		12	18.2	107
Sport(Nordic)	54.3	64.9	76.2	64.1	20
Sport(Rugby)	36.3	25.9	45.5	44.5	63
Sport(Skating)	79.5	80.7	91.6	75.8	42
Sport(Skiing)	54.9	56.8	65.9	66.6	22
Sport(Soccer)	54.2	41.3	60.9	56.4	654
			0 0 100		
Robustness Re	<b>port</b> fo	r NEL o	n AIDA-b	o datase	et

Type of Systematic Error!



Evaluation is a creative process

Systematic errors are difficult to detect:

- High dimension of the learned representations
- Extracting and labeling semantics in the error group requires human-in-the-loop

Interactive tool to identify and label candidate data slices with high systematic errors



Identify candidate groups with high systematic errors



(Rajani et al, EMNLP '22 demo)



Identify candidate groups with high systematic errors



(Rajani et al, EMNLP '22 demo)

Generate semantic labels using LLMs



3. Semantic Labeling

(Rajani et al, EMNLP '22 demo)

#### https://huggingface.co/spaces/nazneen/seal

		Error	Groups
yelp_polarity	~	How to	read this t
Model			
distilbert-base-unc	as v		content
		19102	Food is al
Loss Quantile	0.99	4488	It's good.
	0.99	14812	Edible is t
0.90	1.00	13426	Went here
Cluster error group?		15765	Oh what a
🔾 True 🔿 False		18127	I've been
# clusters		9622	Average J
11		6312	My friend
1	60	12566	lts just ok
# data points to visualize	2	12336	Wild men
000			
1000	5000		
Cluster #:		Error	group vis
1			read this c

How to	read this table:				+	How	to read this t	able:		. 4
	content	labe	pred	loss	clust		Token	Freq	Freq erre	Irs
19102	Food is always good.	0	1	8.99	4	0	##now	0.0118	0.0332	2.28
4488	It's good. The rolls are better than the sashimi although one time we had some really nice(and surprise	0	1	8.78	4	1	delight	0.0027	0.0117	2.16
14812	Edible is the best I can muster.	0	1	8.77	4	2	tips	0.0051	0.0166	2.14
13426	Went here cause I've heard from a few people it was good. Being a huge fan of Mexican food, I had to c	0	1	8.74	4	3	stepping	0.0009	0.0068	2.01
15765	Oh what a difference a year makes. One year ago I loved Penn's Thai House. The awesome Jaime W. re	0	1	8.74	4	4	points	0.0068	0.0186	2.00
18127	I've been here twice. The first time, my husband and I were using a restaurant.com gift so we splurged.	0	1	8.72	4	5	combined	0.0029	0.0107	2.00
9622	Average Japanese food at amazing Japanese food prices.	0	1	8.65	4	6	colored	0.0024	0.0098	2.00
6312	My friend and I went there on Monday night, had an amazing meal. It was one of the best filet mignon	0	1	8.57	4	7	gas	0.0068	0.0186	1.99
12566	lts just ok	0	1	8.53	4	8	unlike	0.0056	0.0156	1.95
12336	Wild menuhuge portions Just ok.	0	1	8.52	4	9	level	0.0136	0.0312	1.95

Word Distribution in Error Groups

#### Error group visualization



### https://huggingface.co/spaces/nazneen/seal

_polarity ~											
	How to	read this table:				+	How	to read this i	able:		
			l. k.					<b>T</b> .1	F	P	1 m
lbert-base-uncas 🗸	19102	content			d loss	4		Token	Freq	Freq err	
antile		Food is always good. It's good. The rolls are better than the sashimi although one time we had some really nice(and surprisi	0		8.99		0	##now	0.0118	0.0332	
0.99	4488				8.78	4		delight	0.0027	0.0117	
1.00	14812	Edible is the best I can muster.	0		8.77	-	2	tips	0.0051	0.0166	2.14
error group?	13426	Went here cause I've heard from a few people it was good. Being a huge fan of Mexican food, I had to cl			8.74	4	3	stepping	0.0009	0.0068	2.01
ue 🔘 False	15765	Oh what a difference a year makes. One year ago I loved Penn's Thai House. The awesome Jaime W. re			8.74	4	4	points	0.0068	0.0186	
	18127	I've been here twice. The first time, my husband and I were using a restaurant.com gift so we splurged.			8.72	4	5	combined	0.0029	0.0107	2.00
rs	9622	Average Japanese food at amazing Japanese food prices.	0		8.65	4	6	colored	0.0024		2.00
60	6312	My friend and I went there on Monday night, had an amazing meal. It was one of the best filet mignon	0			4	7	gas	0.0068	0.0186	1.99
points to visualize		Its just ok	0		8.53	4	8	unlike	0.0056	0.0156	
Joints to visualize	12336	Wild menuhuge portions Just ok.	0	1	8.52	4	9	level	0.0136	0.0312	1.95
5000			-	-	_	-					
#:	Error	group visualization									
- +	How to	read this chart:									1
d prompt from data											+0 (.
a prompt from data											
		cluster: 1									-2
		slice: high-loss									- 3
		P I recently moved here from	CO. I wa	is a n	nember a	at the					-4
		YMCA of Boulder Valley and	d was ve	ry ha	ppy. So,	fast					-5
		content: forward 5 months and I'm lo	oking for	r a gy	m. The '	r was an					-6
		obvious choice to check out	. \n\nl to	ok a t	tour at a	oprox 10a	am				• -7
		e label: 0									-8
		pred: 1									
										-	

.

#### https://huggingface.co/spaces/nazneen/seal

yelp_polarity	~
Model	
distilbert-base-unc	as v
.oss Quantile	0.99
0.90	1.00
Cluster error group? True False clusters	
L	60
# data points to visualize	2
1000	5000
Cluster #:	
1	- +

ow to	read this table:				+	How t	to read this t	able:		
	content	labe	prec	loss	clust		Token	Freq	Freq erre	lrs
9102	Food is always good.	0	1	8.99	4	0	##now	0.0118	0.0332	2.28
4488	It's good. The rolls are better than the sashimi although one time we had some really nice(and surprisi	0	1	8.78	4	1	delight	0.0027	0.0117	2.16
4812	Edible is the best I can muster.	0	1	8.77	4	2	tips	0.0051	0.0166	2.14
3426	Went here cause I've heard from a few people it was good. Being a huge fan of Mexican food, I had to cl	0	1	8.74	4	3	stepping	0.0009	0.0068	2.01
5765	Oh what a difference a year makes. One year ago I loved Penn's Thai House. The awesome Jaime W. re	0	1	8.74	4	4	points	0.0068	0.0186	2.00
8127	I've been here twice. The first time, my husband and I were using a restaurant.com gift so we splurged.	0	1	8.72	4	5	combined	0.0029	0.0107	2.00
9622	Average Japanese food at amazing Japanese food prices.	0	1	8.65	4	6	colored	0.0024	0.0098	2.00
6312	My friend and I went there on Monday night, had an amazing meal. It was one of the best filet mignon	0	1	8.57	4	7	gas	0.0068	0.0186	1.99
2566	Its just ok	0	1	8.53	4	8	unlike	0.0056	0.0156	1.95
2336	Wild menuhuge portions Just ok.	0	1	8.52	4	9	level	0.0136	0.0312	1.95

#### Error group visualization



### https://huggingface.co/spaces/nazneen/seal



к,

### https://huggingface.co/spaces/nazneen/seal

yelp_polarity	~
Model	
distilbert-base-u	ncas 🗸
Loss Quantile	0.99
0.90	1.00
Cluster error group?	
🔾 True 🔿 Fals	e
# clusters	
1	60
# data points to visual	ize
1000	5000
Cluster #:	
1	- +
Build prompt fro	m data

#### Word Distribution in Error Groups Error Groups How to read this table: How to read this table: content label pred loss clust Token Freq Freq erri Irs 19102 Food is always good. 0 1 8.99 4 0 ##now 0.0118 0.0332 2.28 4488 It's good. The rolls are better than the sashimi although one time we had some really nice(and surprisi 0 1 8.78 4 1 delight 0.0027 0.0117 2.16 14812 Edible is the best I can muster. 0 1 8.77 4 2 tips 0.0051 0.0166 2.14 13426 Went here cause I've heard from a few people it was good. Being a huge fan of Mexican food, I had to cl 0 1 8.74 4 3 stepping 0.0009 0.0068 2.01 15765 Oh what a difference a year makes. One year ago I loved Penn's Thai House. The awesome Jaime W. rec 0 1 8.74 4 4 points 0.0068 0.0186 2.00 18127 I've been here twice. The first time, my husband and I were using a restaurant.com gift so we splurged. 0 1 8.72 4 0.0029 0.0107 2.00 5 combined 9622 Average Japanese food at amazing Japanese food prices. 0 1 8.65 4 6 colored 0.0024 0.0098 2.00 6312 My friend and I went there on Monday night, had an amazing meal. It was one of the best filet mignon 0 1 8.57 4 0.0068 0.0186 1.99 7 gas 12566 Its just ok 0 1 8.53 4 8 unlike 0.0056 0.0156 1.95

#### Error group visualization

12336 Wild menu. huge portions .. Just ok.



0 1 8.52 4

9 level

0.0136 0.0312 1.95

### **SEAL Experimental Results**

Group label	Size	Group acc.
Albert Base v2 on Yelp (	overall acc:	0.95)
Club reviews	574	0.90 (-5%)
Movie theater reviews	231	0.85 (-10%)
Dentist reviews	69	0.88 (-7%)
Chain restaurant reviews	61	0.88 (-7%)
Frozen custard reviews	37	0.83 (-12%)
Waterfront business reviews	11	0.72 (-23%)

## **SEAL Experimental Results**

Group label	Size	Group acc.
Albert Base v2 on Yelp (	overall acc:	0.95)
Club reviews	574	0.90 (-5%)
Movie theater reviews	231	0.85 (-10%)
Dentist reviews	69	0.88 (-7%)
Chain restaurant reviews	61	0.88 (-7%)
Frozen custard reviews	37	0.83 (-12%)
Waterfront business reviews	11	0.72 (-23%)

SEAL identified data groups where the model performance drops between 5% to 28%

1. Open-sourcing ML research artifacts is becoming the norm

- 1. Open-sourcing ML research artifacts is now the default
- 2. The most popular Hugging Face models are those that are older and well-documented

- 1. Open-sourcing ML research artifacts is becoming the norm
- 2. The most popular Hugging Face models are those that are older and well-documented
- 3. Model evaluation can be actionable RG toolkit supports this goal via fine-grained evaluation

- 1. Open-sourcing ML research artifacts is becoming the norm
- 2. The most popular Hugging Face models are those that are older and well-documented
- 3. Model evaluation can be actionable RG toolkit supports this goal via fine-grained evaluation
- 4. LLMs can help label systematic errors in models in a human interpretable way

### **Collaborators**

Systematic study of HF models and SEAL



Weixin Liang (Stanford)



Xinyu Yang (ZJU)



Meg Mitchell (Hugging Face)



James Zou (Stanford)

### Robustness Gym



Karan Goel (Stanford)



Jesse Vig (Salesforce)



Chris Re (Stanford)



Mohit Bansal (UNC)

# **Thanks for listening**

