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

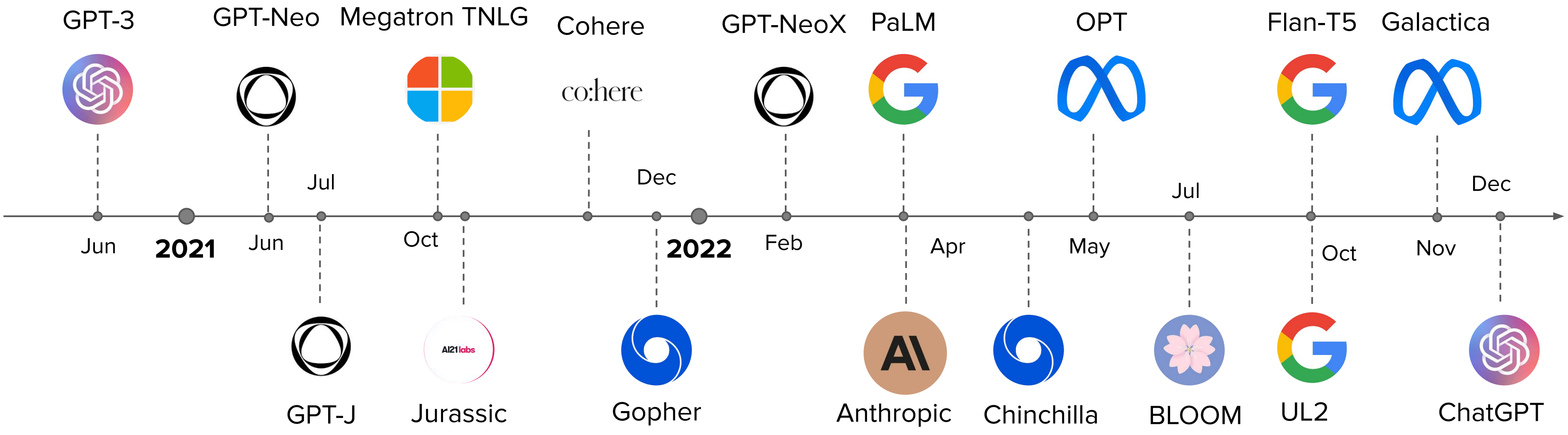
Systematic study of 75K models on HF

Part 2:

NLP Evaluation landscape

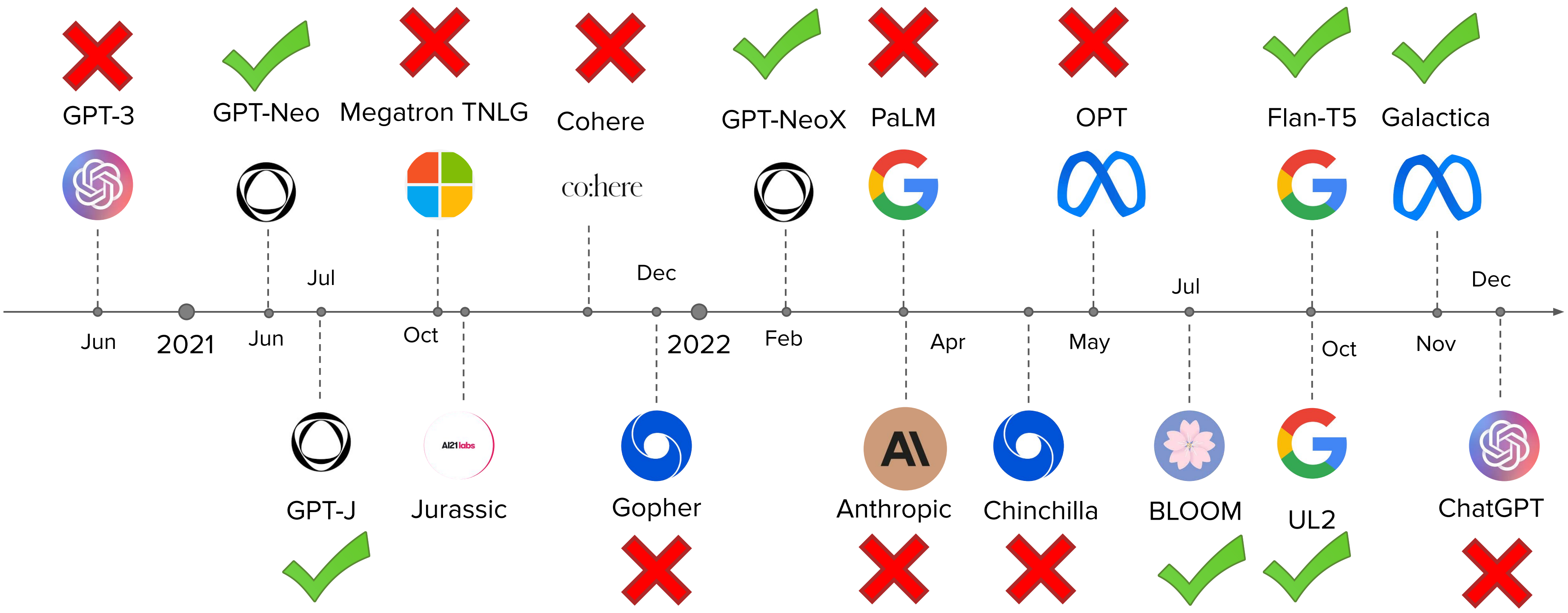
Challenges and opportunities in model evaluation and documentation

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

🔓 Open Access Models

Allows reproducing results and replicating parts of the model

Enable auditing and conducting risk analysis

Serves as a research artifact

Enables interpreting model output

🔒 Closed Access Models

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

🔒 Closed Access Models

Safety concerns

Competitive advantage

Expensive to setup guardrails for safe access

Model Access

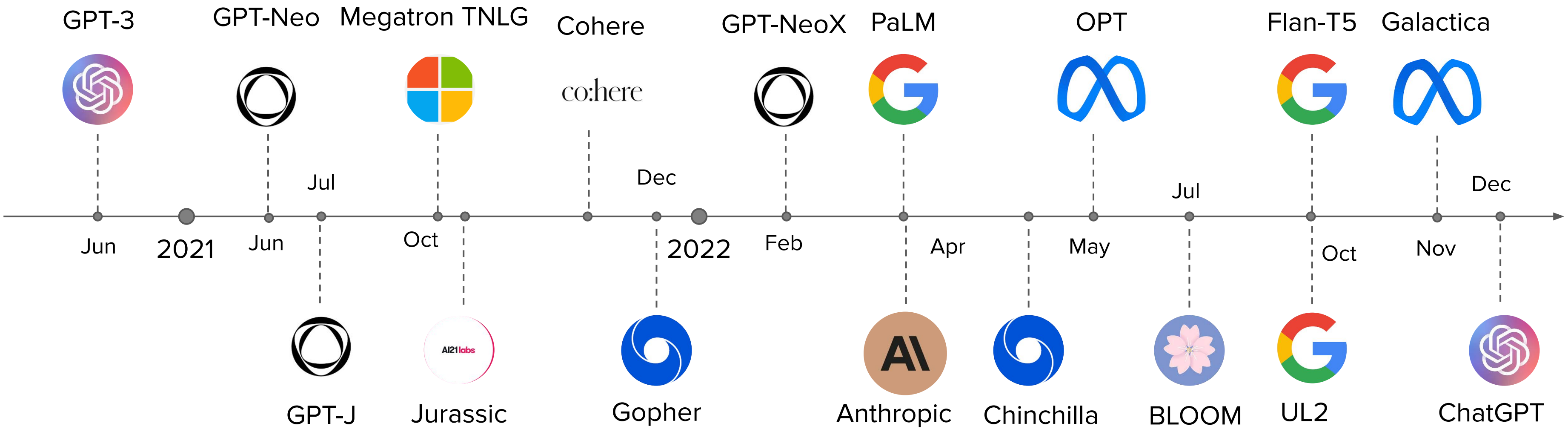
🔒

Open access

Limited access

Closed access

Large Language Models since GPT3



Open Access Large Language Models

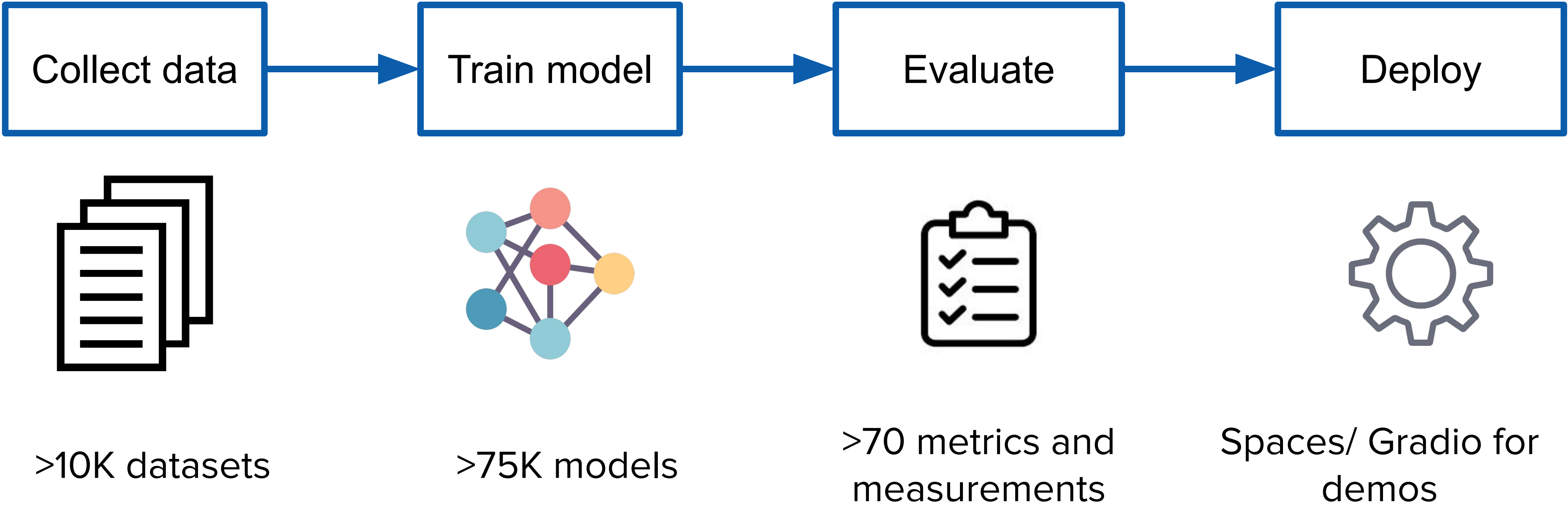
Research on policy, governance, AI safety and alignment

Community efforts like Eleuther, Big Science, LAION

Papers with several authors

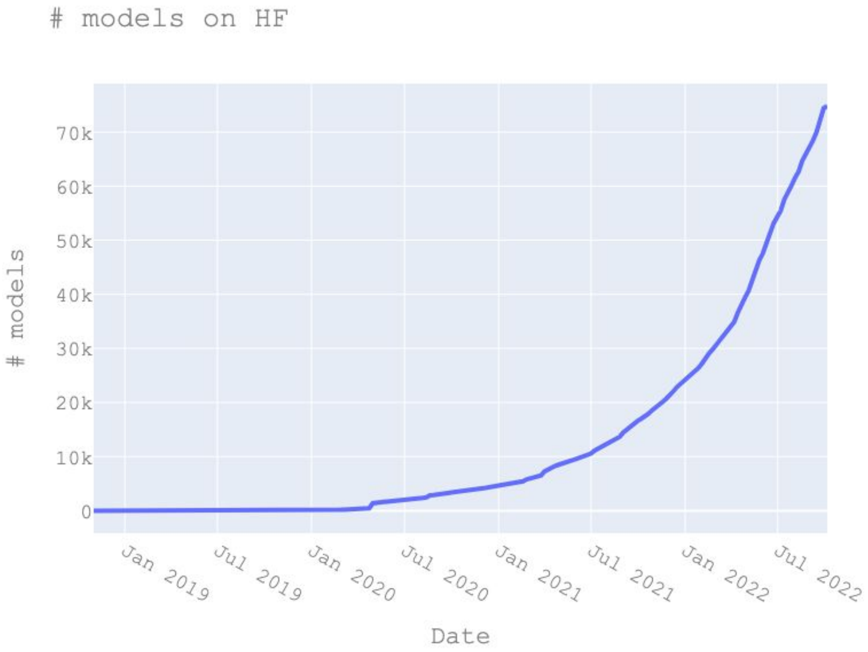
Open source ML has potential for huge impact

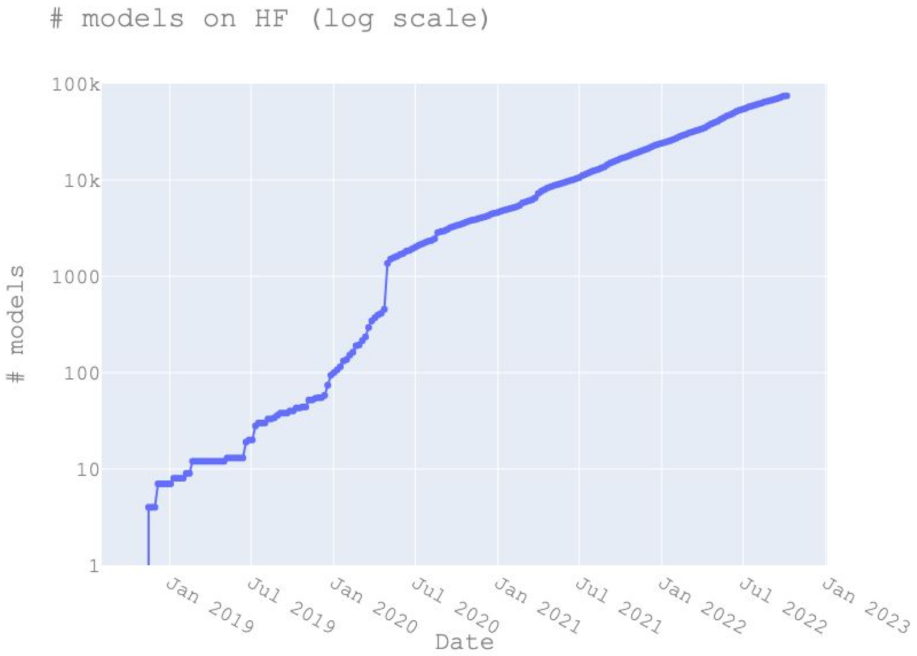
Ecosystem as part of the ML workflow



ML Modeling Landscape

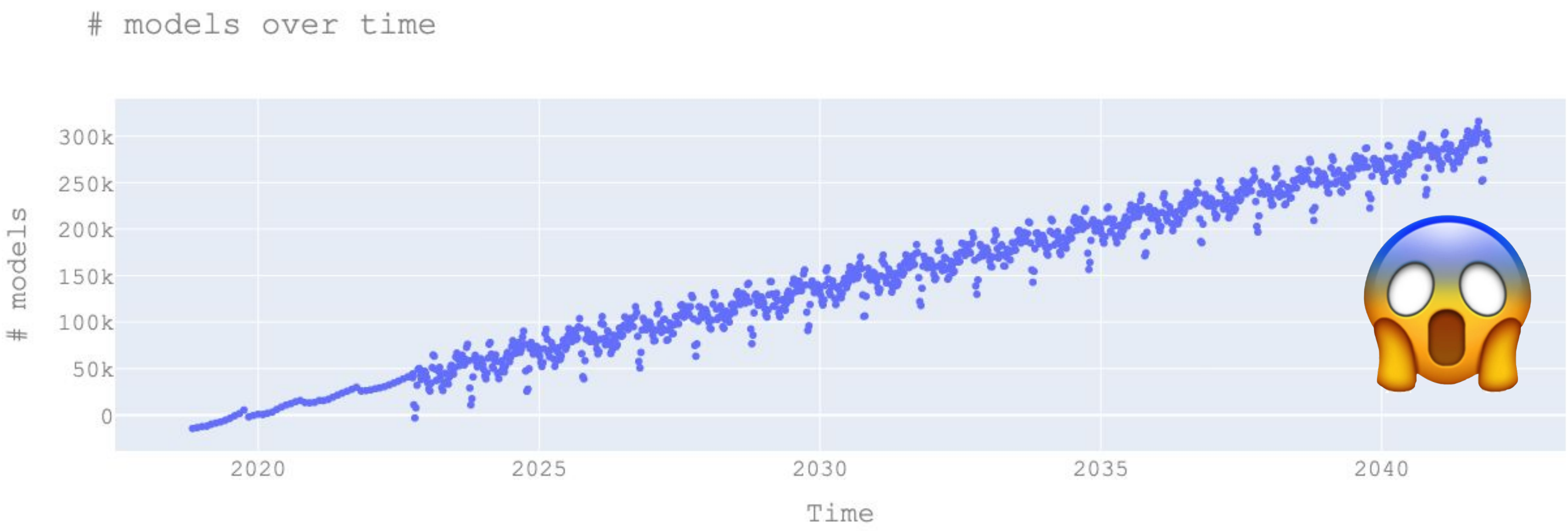
There is an exponential growth of ML models.



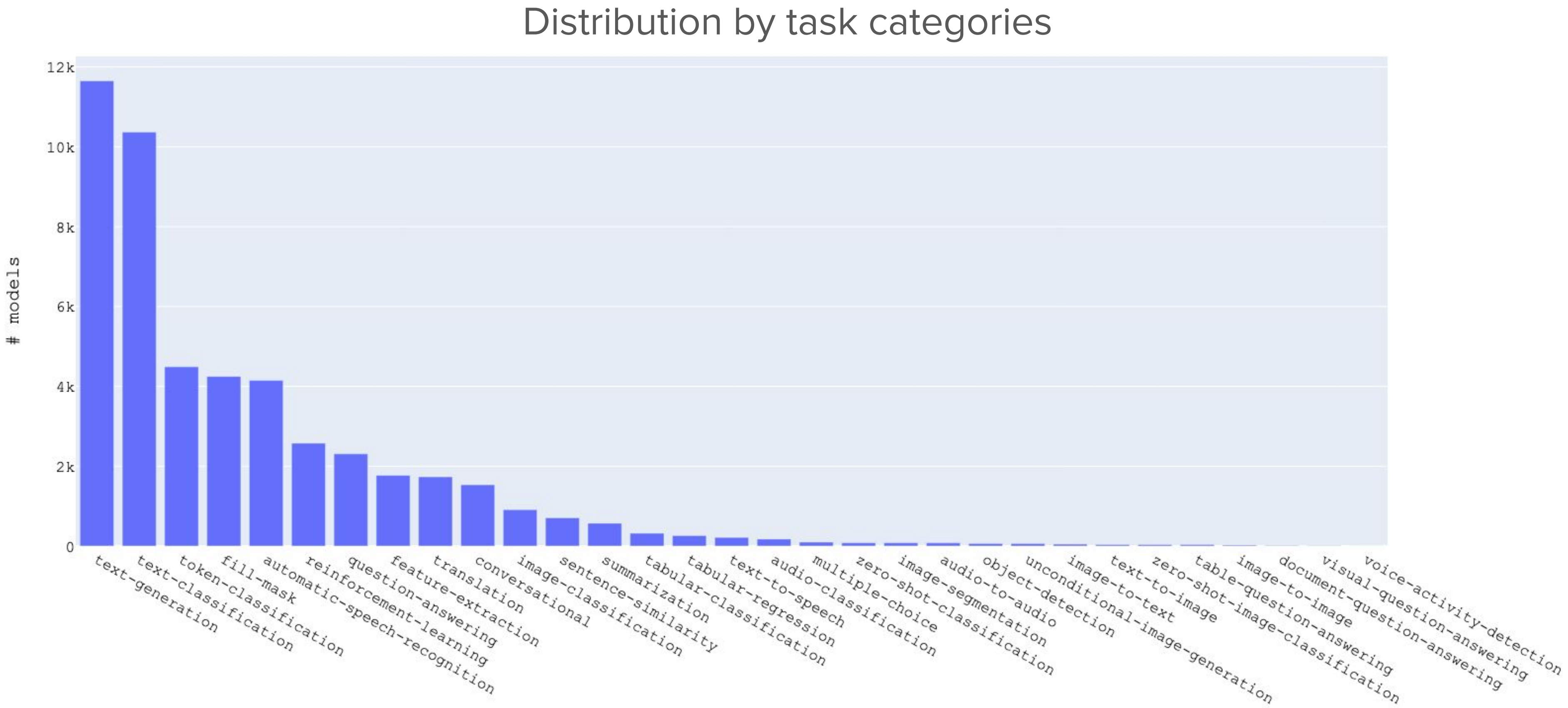


ML Modeling Landscape

There is an exponential growth of ML models.



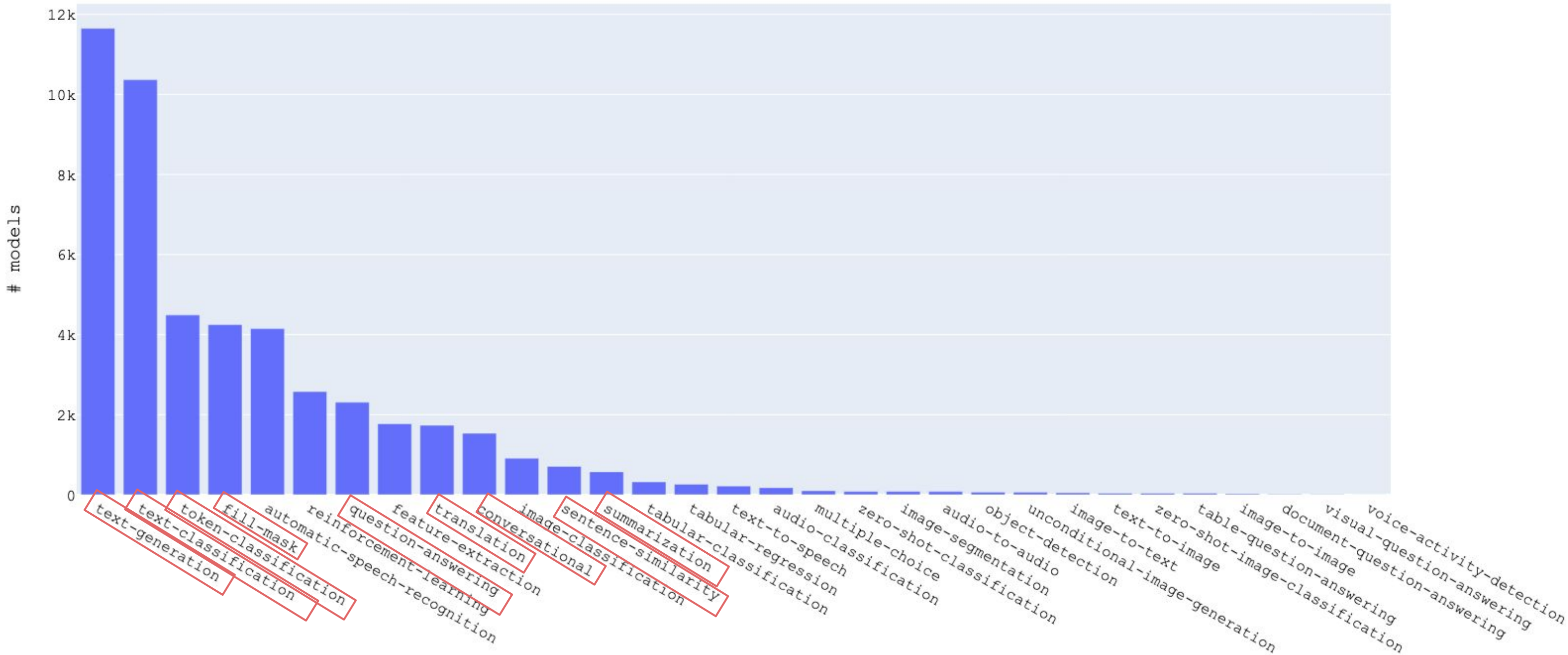
ML Modeling Landscape



NLP Modeling Landscape

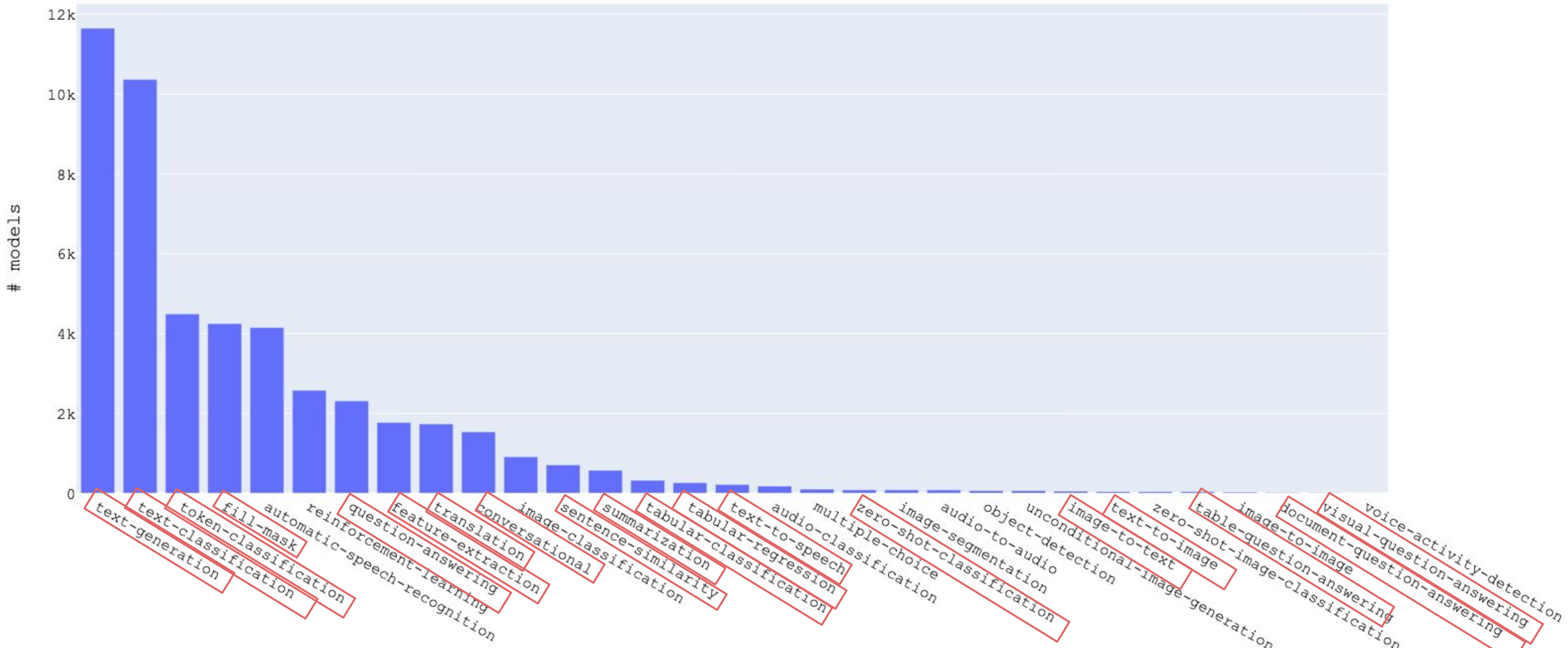
Approx 40% of the task categories are NLP

Covering 78% of the models



NLP Modeling Landscape

Including multimodal – 55% task categories

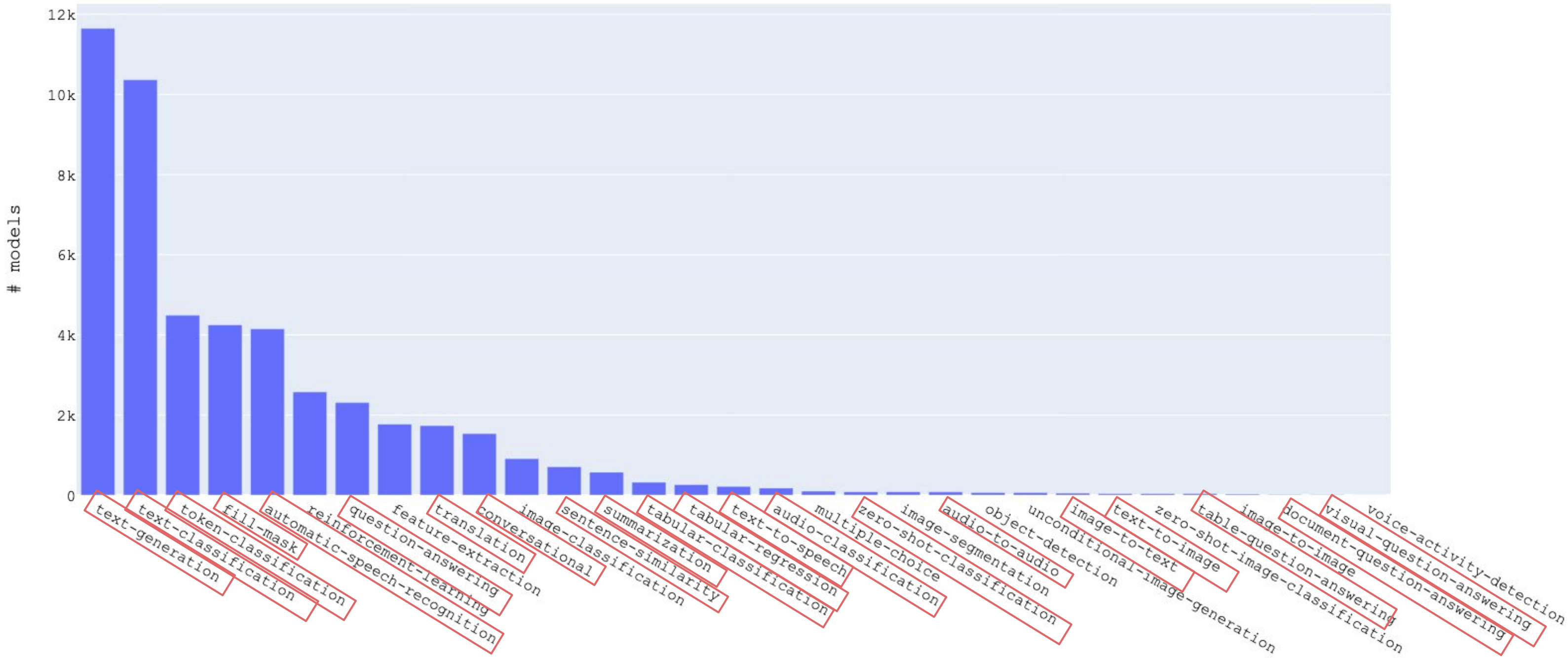


NLP Modeling Landscape

Including multimodal – 55% task categories

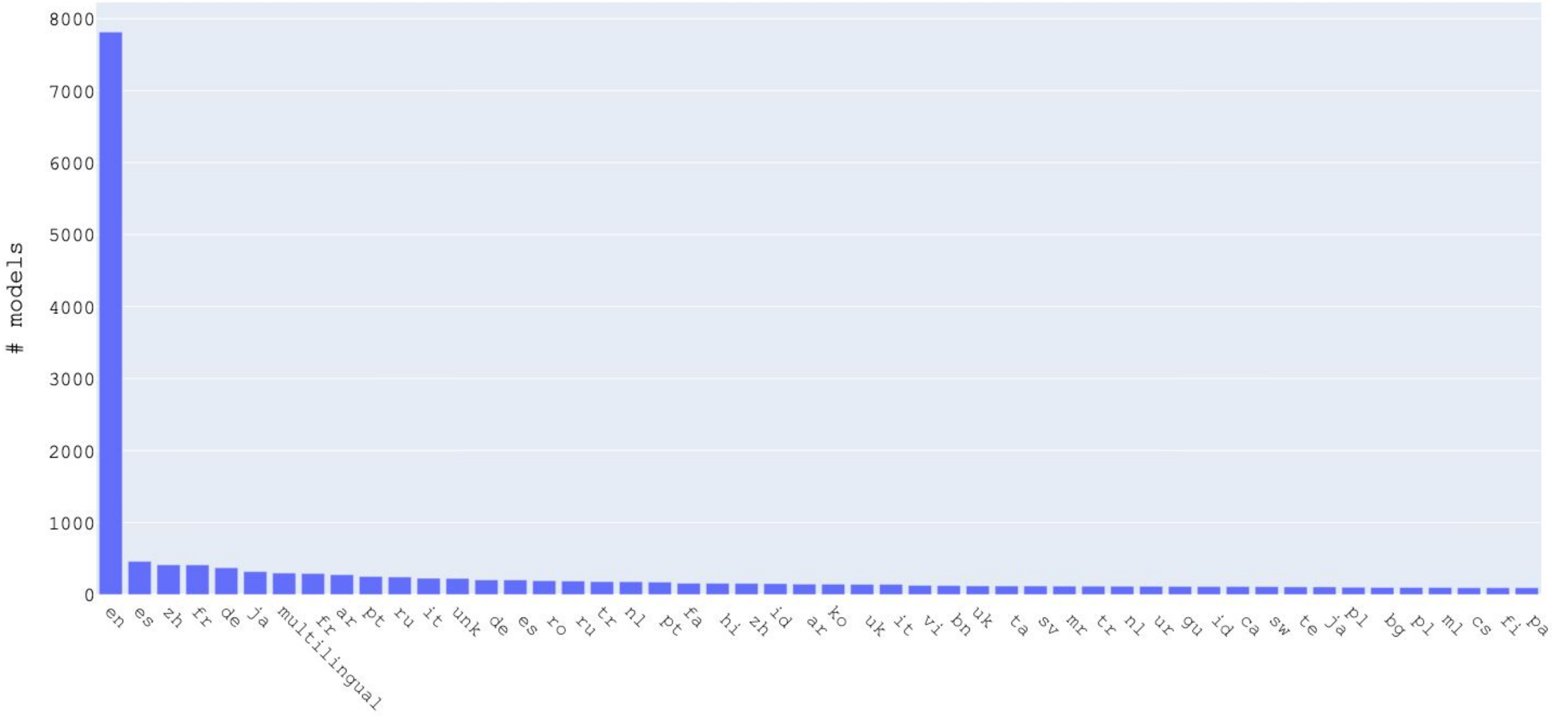
Including speech – 72% task categories

Coverage – 90% of models



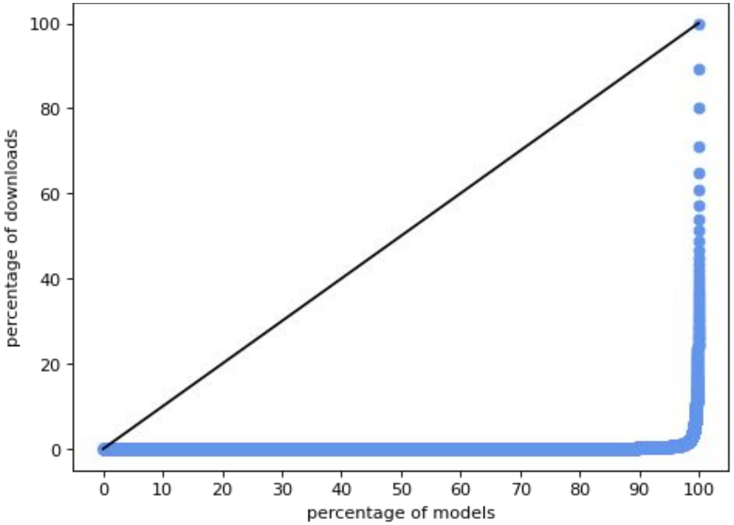
NLP Modeling Landscape

Distribution by language (based on 20% models reporting)



Model Usage

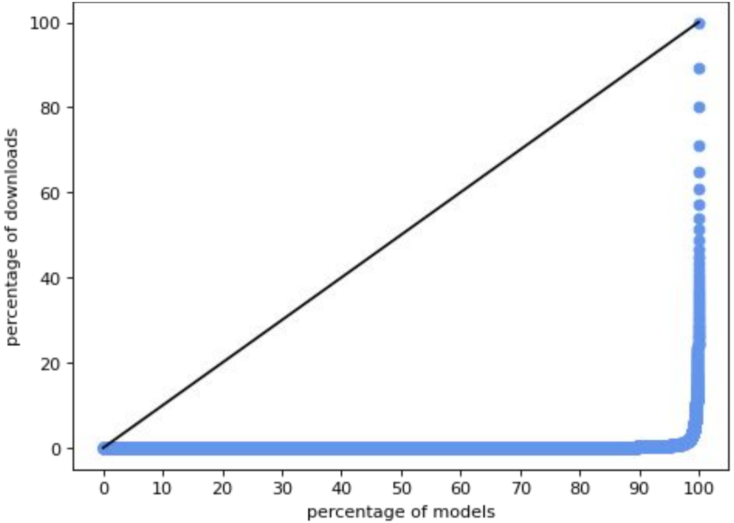
Top 0.2% models (N=124) makeup >80% HF model usage



Model Usage

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

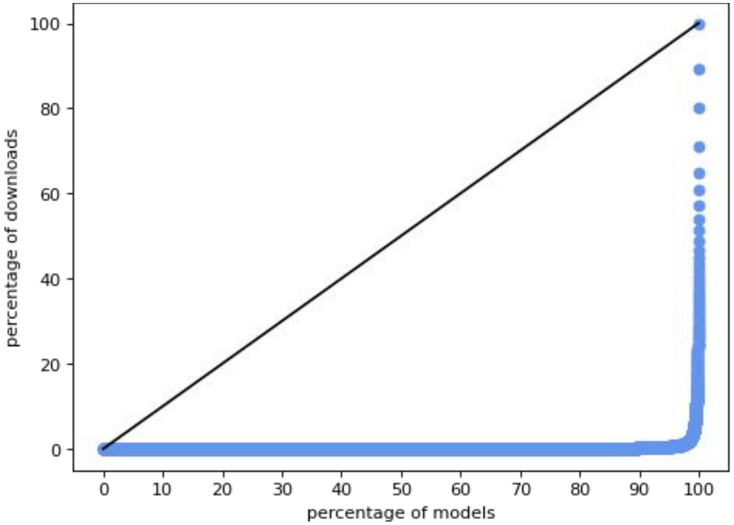
98% of these models are trained on just text data

Of these –

65% were created before 2021

33% were created in 2021

2% were created in 2022



Model Age vs. Usage

Relation between model age and its usage

Model Age vs. Usage

Relation between model age and its usage

Model Age vs. Usage

Relation between model age and its usage

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

Model Age vs. Usage

Relation between model age and its usage

Model Age vs. Usage

Factors:

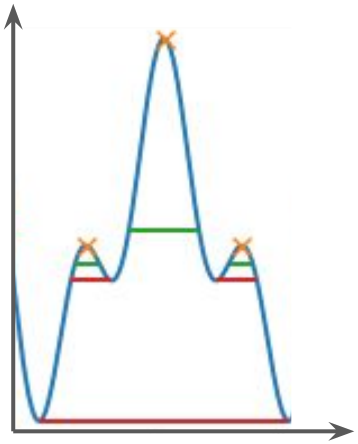
1. Compute is becoming cheaper making model training more accessible
2. As more models are created, their usage is distributed
3. Models are being replaced by their efficient counterparts (ex: BERT ➔ DistilBERT)

Trend Width

Step 1: Find all peaks in a signal

Step 2: Measure peak widths at base

Step 3: Take the max width



Model Usage Trends

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

Usage trend width for top models

bert-base-uncased

Model Usage Trends

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

Usage trend width for top models

bert-base-uncased

sentence-transformers/paraphrase-xlm-r-multilingual-v1

Model Usage Trends

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

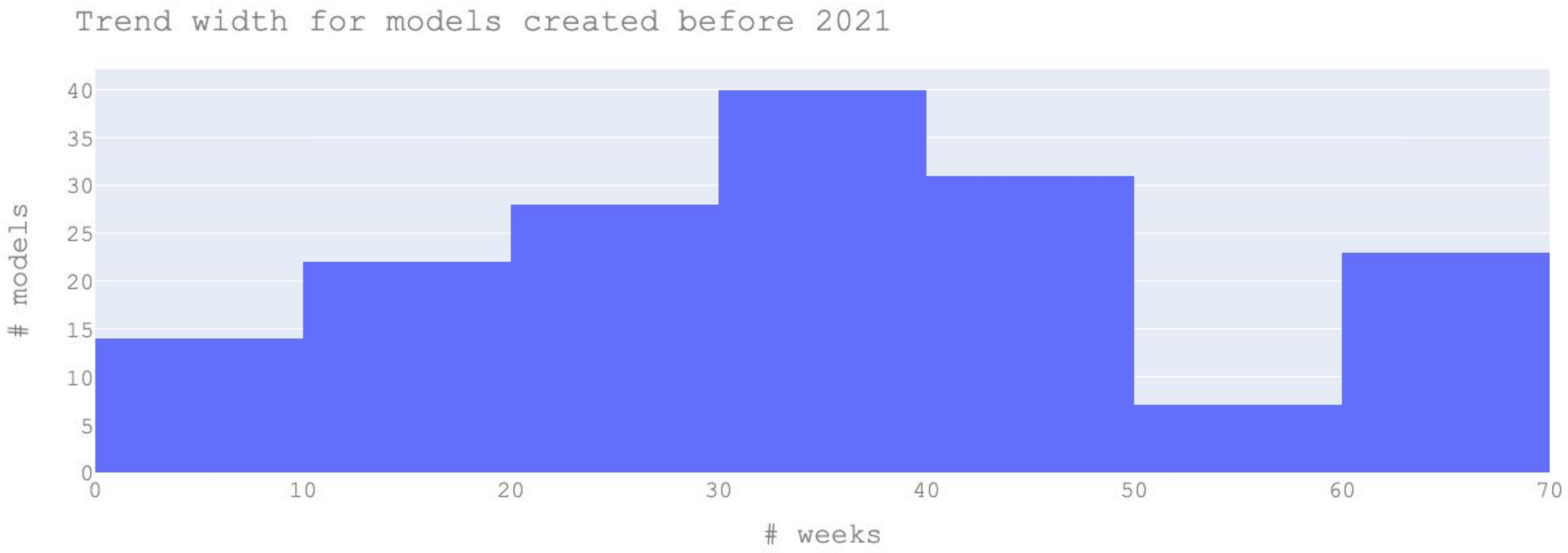
Usage trend width for top models

bert-base-uncased

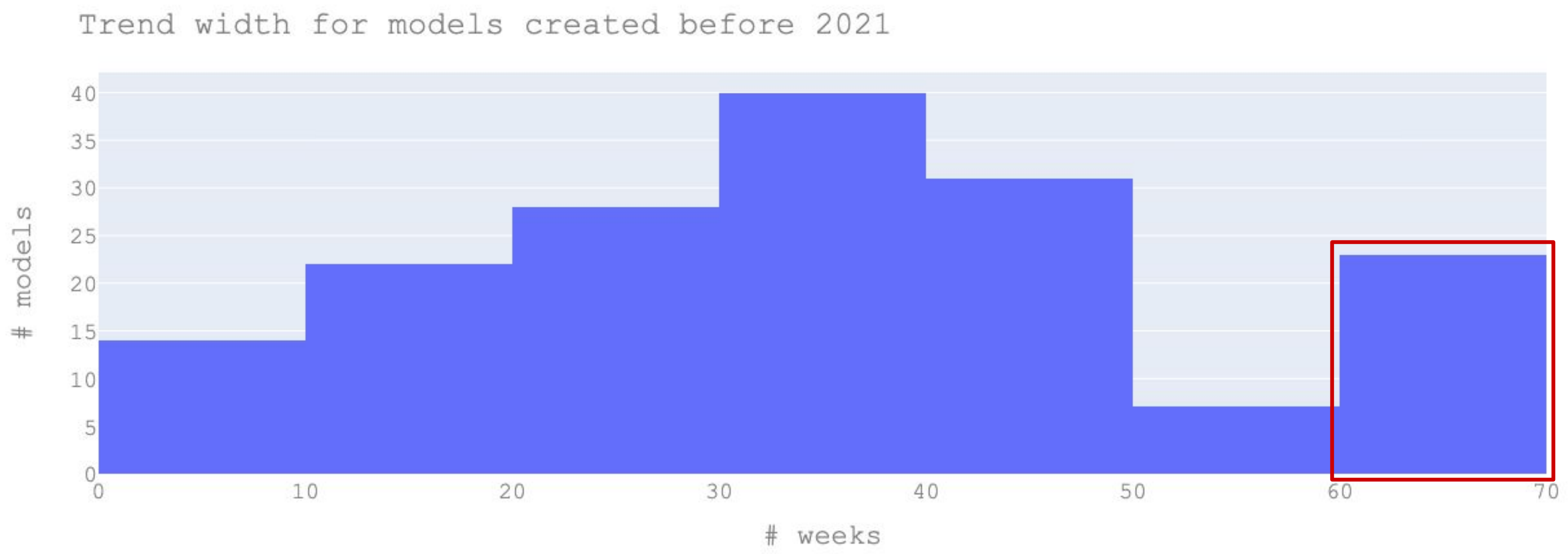
sentence-transformers/paraphrase-xlm-r-multilingual-v1

HateSpeech-CNERG/indic-abusive-allInOne-MuRIL

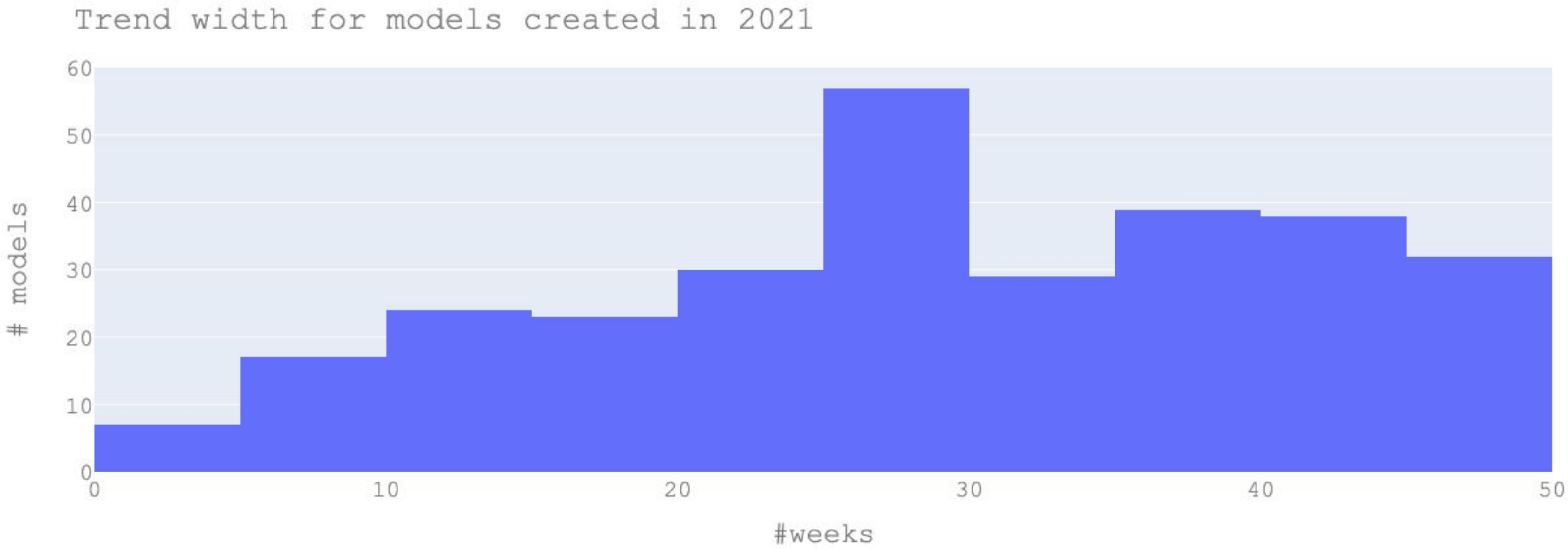
Model Usage Trends



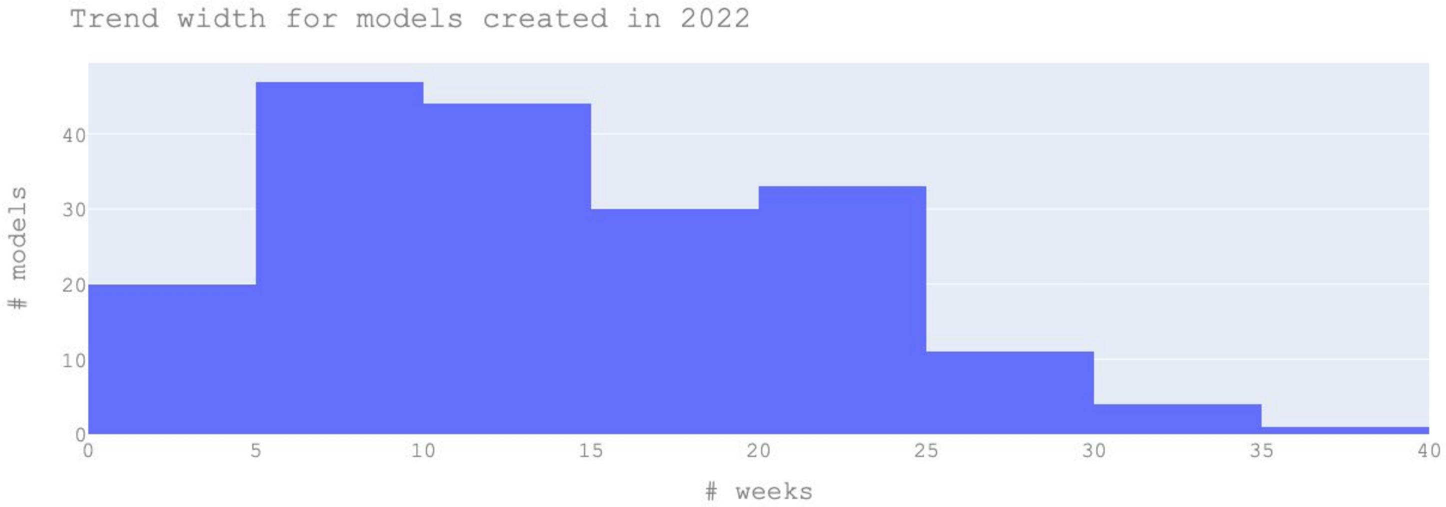
Model Usage Trends



Model Usage Trends



Model Usage Trends



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?

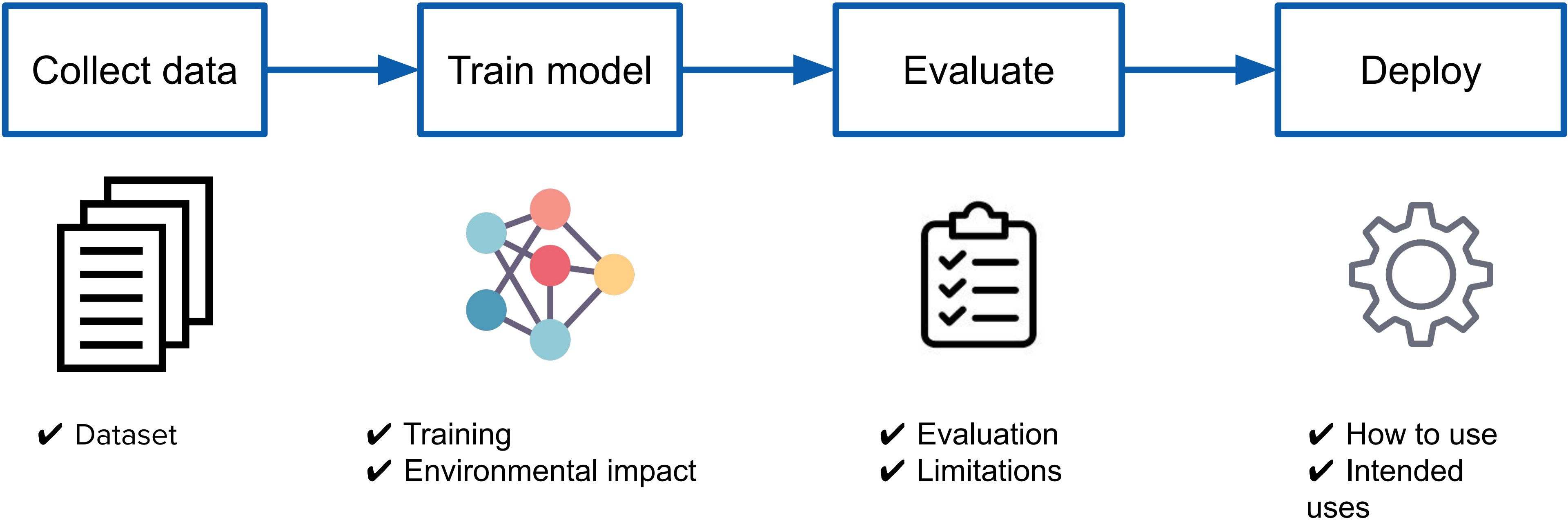
Model Usage

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!

Model Documentation



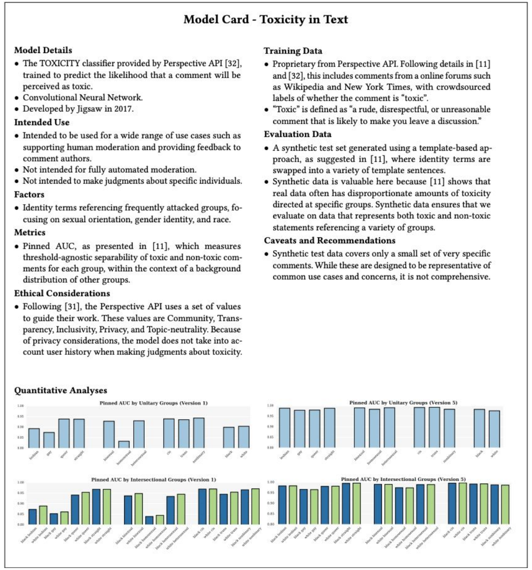
Why document models?

🔍Transparency

📈Reproducibility

📢Communication

Model Documentation Landscape



Model Card - Toxicity in Text

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.

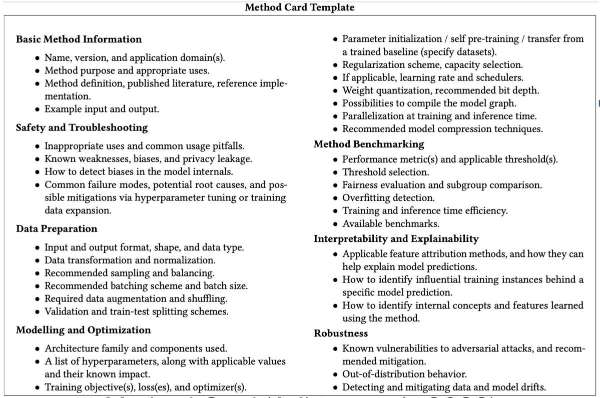
Training Data
- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data
- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are 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
- 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.

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 implementation.
- 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 possible mitigations via hyperparameter tuning or training 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.

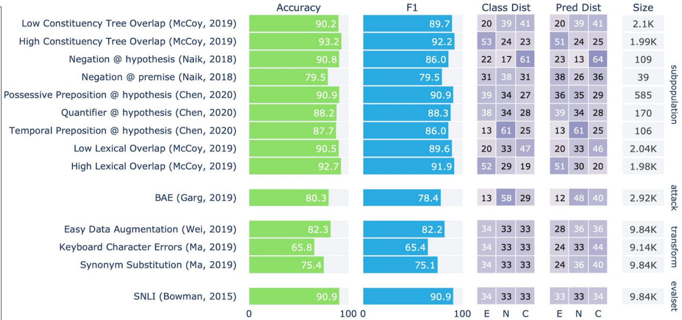
Modelling and Optimization
- Architecture family and components used.
- A list of hyperparameters, along with applicable values and their known impact.
- Training objective(s), loss(es), and optimizer(s).
- Parameter initialization / self pre-training / transfer from a trained baseline (specify datasets).
- Regularization scheme, capacity selection.
- If applicable, learning rate and schedulers.
- Weight quantization, recommended bit depth.
- Possibilities to compile the model graph.
- Parallelization at training and inference time.
- Recommended model compression techniques.

Method Benchmarking
- Performance metric(s) and applicable threshold(s).
- Threshold selection.
- Fairness evaluation and subgroup comparison.
- Overfitting detection.
- Training and inference time efficiency.
- Available benchmarks.

Interpretability and Explainability
- Applicable feature attribution methods, and how they can help explain model predictions.
- How to identify influential training instances behind a specific model prediction.
- How to identify internal concepts and features learned using the method.

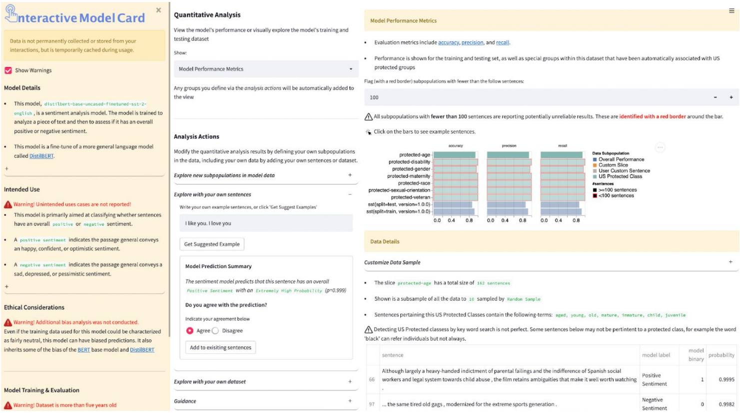
Robustness
- Known vulnerabilities to adversarial attacks, and recommended mitigation.
- Out-of-distribution behavior.
- Detecting and mitigating data and model drifts.

Method Card (Adkins et al., 2022)



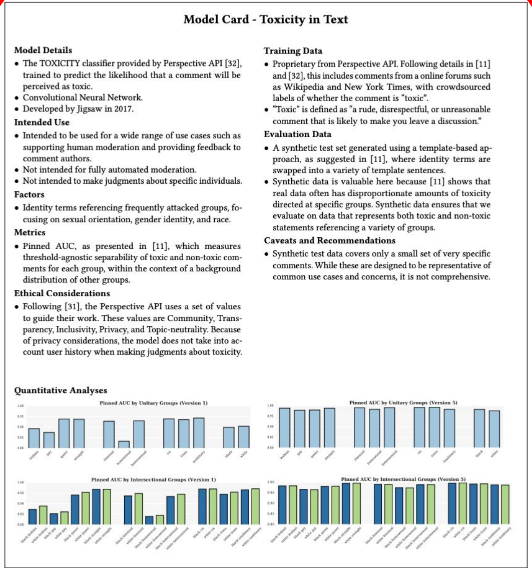
		Accuracy	F1	Class Dist (E N C)	Pred Dist (E N C)	Size
subpopulation	Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20 39 41	20 39 41	2.1K
	High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53 24 23	51 24 25	1.99K
	Negation @ hypothesis (Naik, 2018)	90.8	86.0	22 17 61	23 13 64	109
	Negation @ premise (Naik, 2018)	79.5	79.5	31 38 31	38 26 36	39
	Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39 34 27	36 35 29	585
	Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38 34 28	39 34 28	170
	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
attack	BAE (Garg, 2019)	80.3	78.4	13 58 29	12 48 40	2.92K
transform	Easy Data Augmentation (Wei, 2019)	82.3	82.2	34 33 33	28 36 36	9.84K
	Keyboard Character Errors (Ma, 2019)	65.8	65.4	34 33 33	24 33 44	9.14K
	Synonym Substitution (Ma, 2019)	75.4	75.1	34 33 33	24 36 40	9.84K
evalset	SNLI (Bowman, 2015)	90.9	90.9	34 33 33	33 33 34	9.84K

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

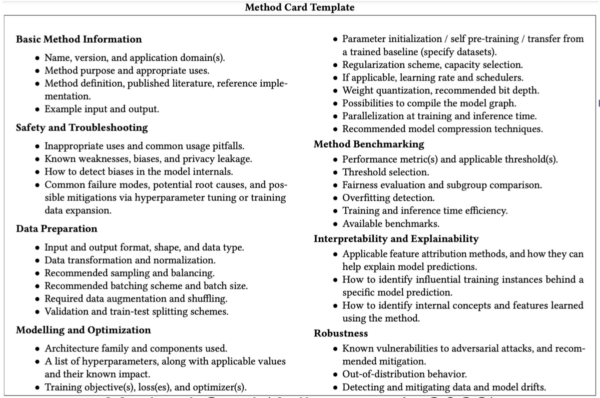


Interactive Model Cards (Crisan, Vig,Drouhard, and **Rajani**, FAccT2022)

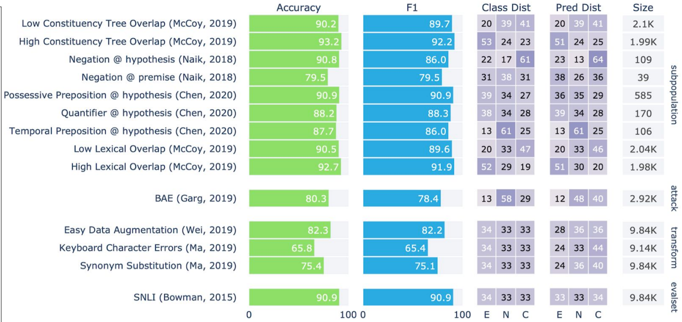
Model Documentation Landscape



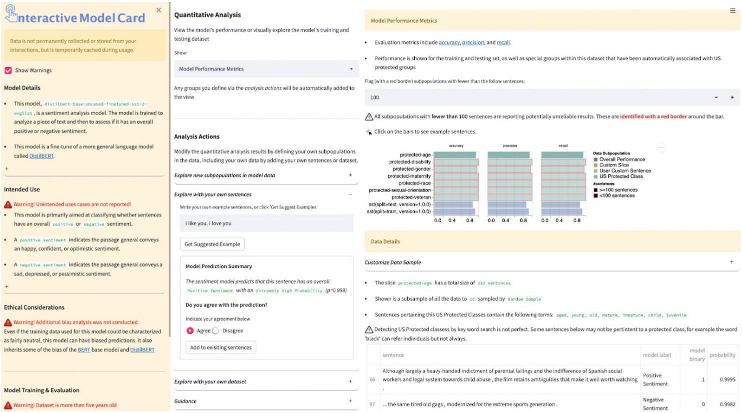
Model Card (Mitchell et al., 2019)



Method Card (Adkins et al., 2022)

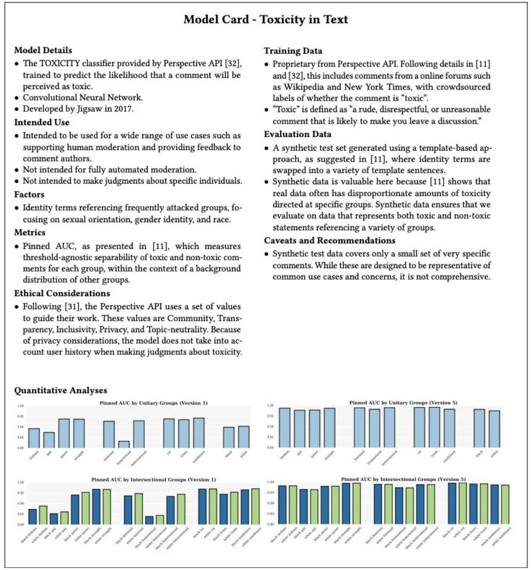


Robustness Report (Goel*, **Rajani***, et al., NAACL 2021)



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

Model Documentation Landscape



Model Card - Toxicity in Text

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.

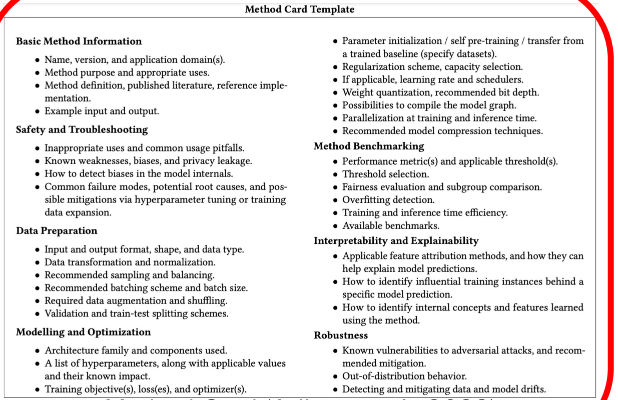
Training Data
- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data
- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are 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
- 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.

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 implementation.
- 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 possible mitigations via hyperparameter tuning or training 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.

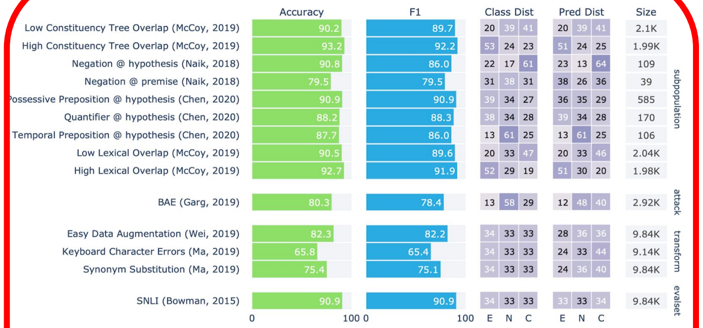
Modelling and Optimization
- Architecture family and components used.
- A list of hyperparameters, along with applicable values and their known impact.
- Training objective(s), loss(es), and optimizer(s).
- Parameter initialization / self pre-training / transfer from a trained baseline (specify datasets).
- Regularization scheme, capacity selection.
- If applicable, learning rate and schedulers.
- Weight quantization, recommended bit depth.
- Possibilities to compile the model graph.
- Parallelization at training and inference time.
- Recommended model compression techniques.

Method Benchmarking
- Performance metric(s) and applicable threshold(s).
- Threshold selection.
- Fairness evaluation and subgroup comparison.
- Overfitting detection.
- Training and inference time efficiency.
- Available benchmarks.

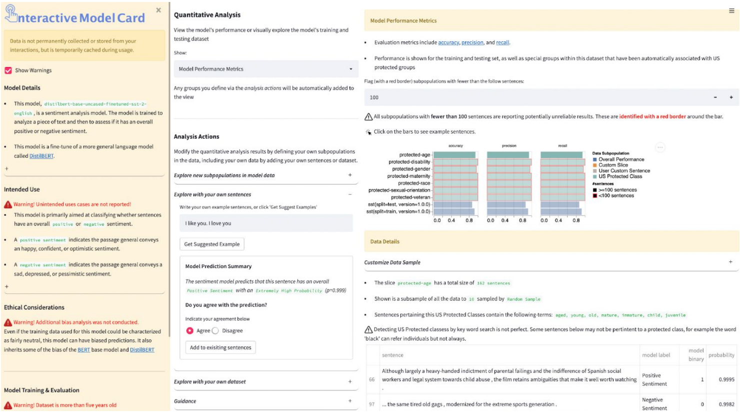
Interpretability and Explainability
- Applicable feature attribution methods, and how they can help explain model predictions.
- How to identify influential training instances behind a specific model prediction.
- How to identify internal concepts and features learned using the method.

Robustness
- Known vulnerabilities to adversarial attacks, and recommended mitigation.
- Out-of-distribution behavior.
- Detecting and mitigating data and model drifts.

Method Card (Adkins et al., 2022)

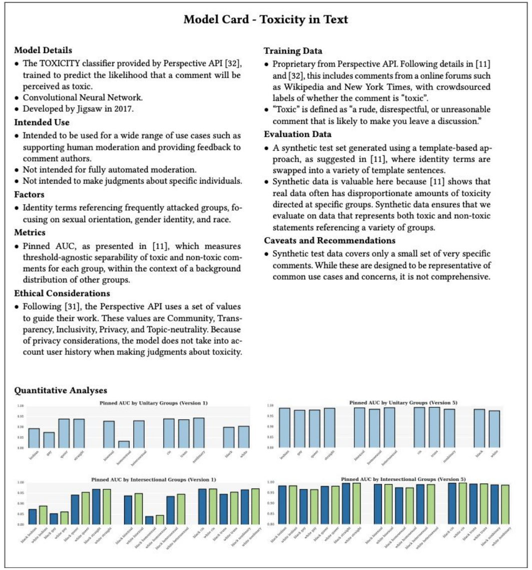


Robustness Report (Goel*, **Rajani***, et al., NAACL 2021)



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

Model Documentation Landscape


Model Card - Toxicity in Text

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.

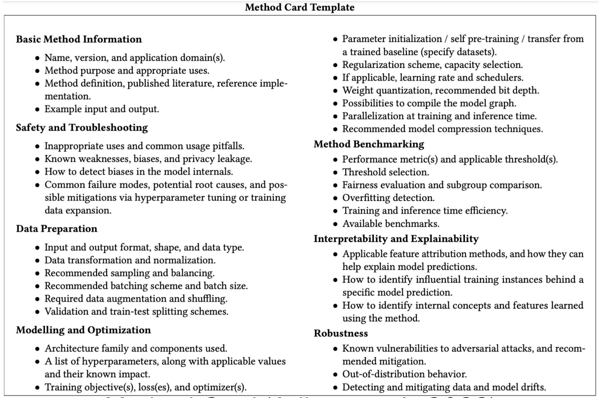
Training Data
- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data
- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are 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
- 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.

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 implementation.
- 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 possible mitigations via hyperparameter tuning or training 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.

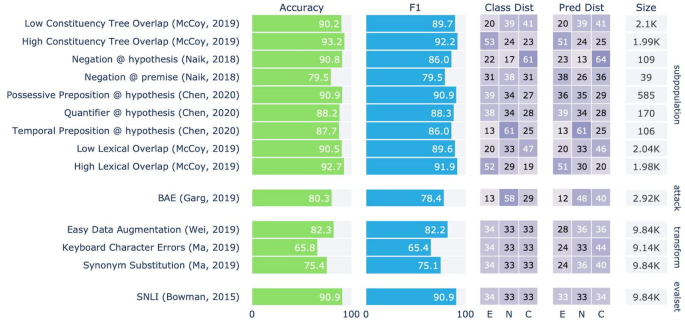
Modelling and Optimization
- Architecture family and components used.
- A list of hyperparameters, along with applicable values and their known impact.
- Training objective(s), loss(es), and optimizer(s).
- Parameter initialization / self pre-training / transfer from a trained baseline (specify datasets).
- Regularization scheme, capacity selection.
- If applicable, learning rate and schedulers.
- Weight quantization, recommended bit depth.
- Possibilities to compile the model graph.
- Parallelization at training and inference time.
- Recommended model compression techniques.

Method Benchmarking
- Performance metric(s) and applicable threshold(s).
- Threshold selection.
- Fairness evaluation and subgroup comparison.
- Overfitting detection.
- Training and inference time efficiency.
- Available benchmarks.

Interpretability and Explainability
- Applicable feature attribution methods, and how they can help explain model predictions.
- How to identify influential training instances behind a specific model prediction.
- How to identify internal concepts and features learned using the method.

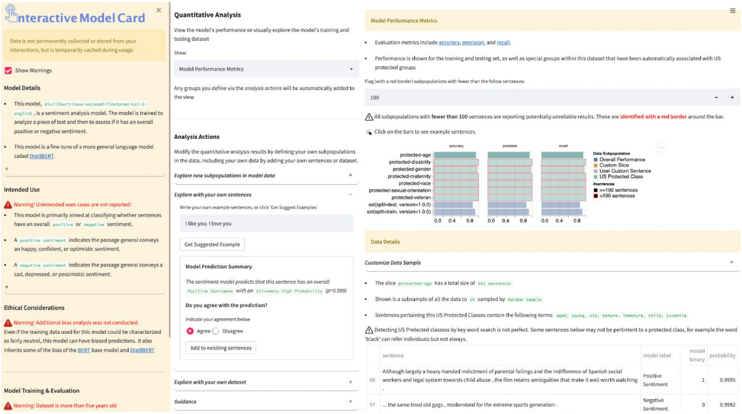
Robustness
- Known vulnerabilities to adversarial attacks, and recommended mitigation.
- Out-of-distribution behavior.
- Detecting and mitigating data and model drifts.

Method Card (Adkins et al., 2022)



	Accuracy	F1	Class Dist E	N	C	Pred Dist E	N	C	Size	
Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20	39	41	20	39	41	2.1K	subpopulation
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	
Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39	34	27	36	35	29	585	
Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38	34	28	39	34	28	170	
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	12	48	40	2.92K	attack
Easy Data Augmentation (Wei, 2019)	82.3	82.2	34	33	33	28	36	36	9.84K	transform
Keyboard Character Errors (Ma, 2019)	65.8	65.4	34	33	33	24	33	44	9.14K	
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	33	33	34	9.84K	evalset

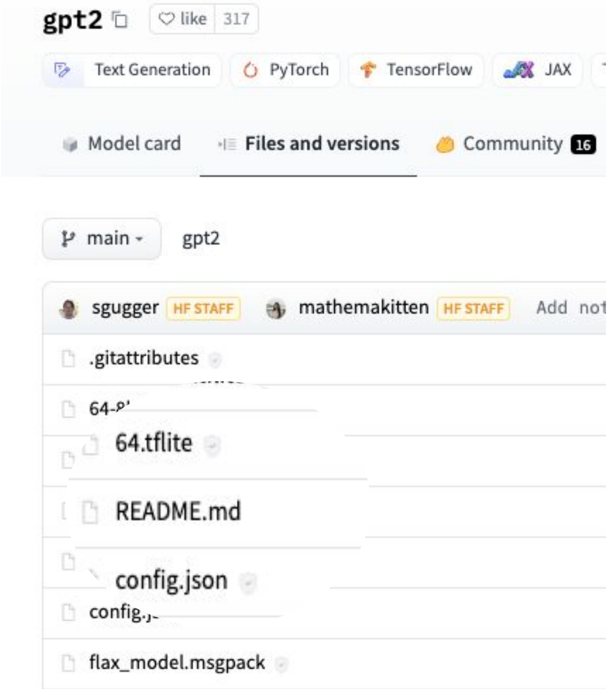
Robustness Report (Goel*, **Rajani***, et al., NAACL 2021)



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

Model Documentation in 🤗

Model documentation is part of the repo's README



Model Documentation for GPT2

Model description

GPT-2 is a transformers model pretrained on a very large corpus of English data in a self-supervised 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.

Model Documentation for GPT2

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 here.

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.

Model Documentation for GPT2

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 model card:

> *"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 model hub 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 moc
```

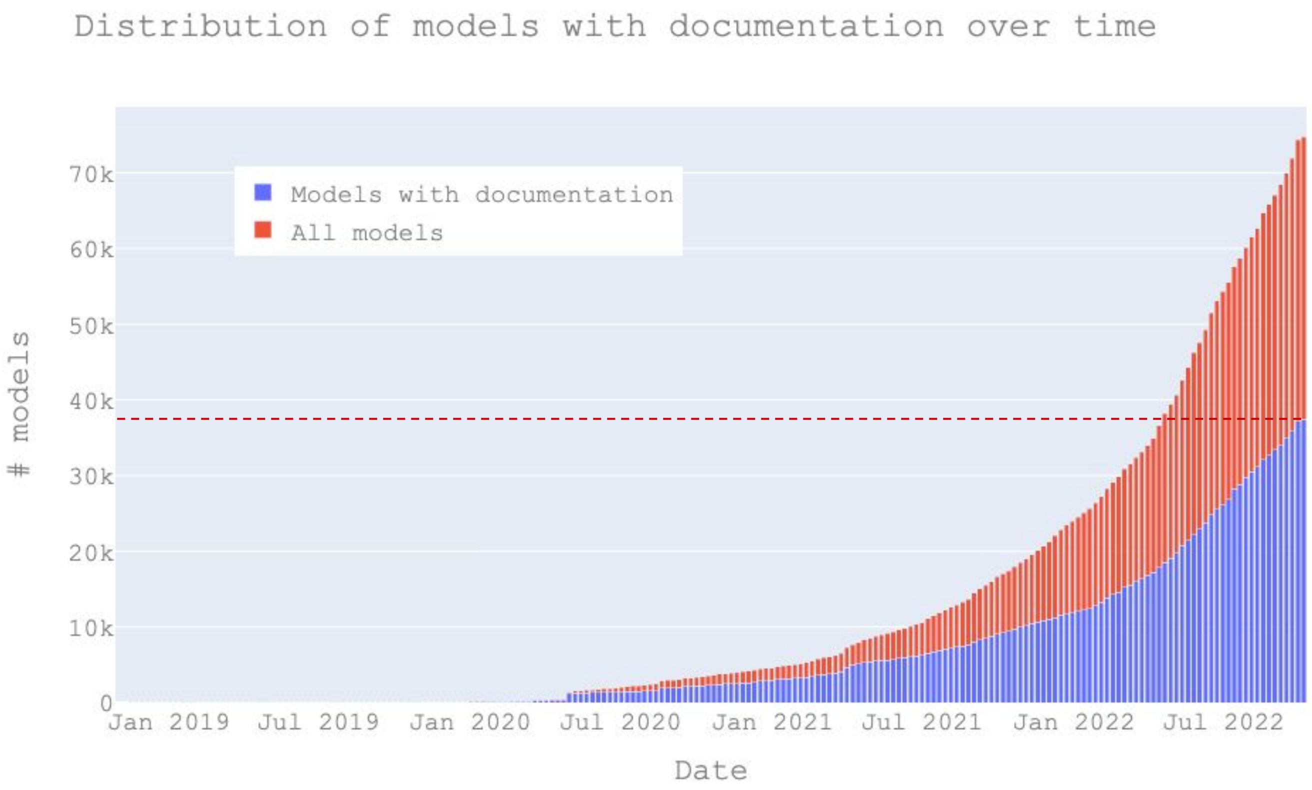
Model Documentation for GPT2

Evaluation results

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

Dataset	LAMBADA	LAMBADA	CBT-CN	CBT-NE	WikiText2	PTB	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



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

Model Documentation RCT

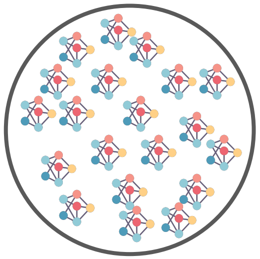
Observation: Only 50% models have model cards but contribute 98% of total usage
Goal: Study the relation between model usage and documentation
Hypothesis: model documentation drives model usage

Randomized Control Trial (RCT) for models:

Model Documentation RCT

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:

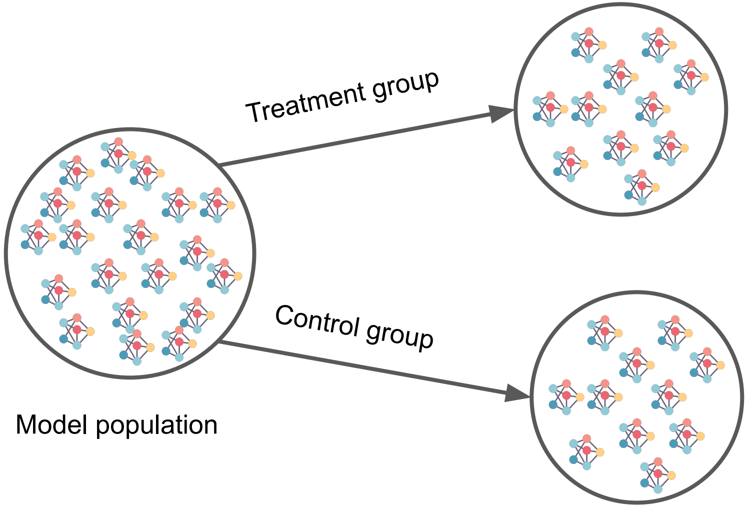


Model population

Model Documentation RCT

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:



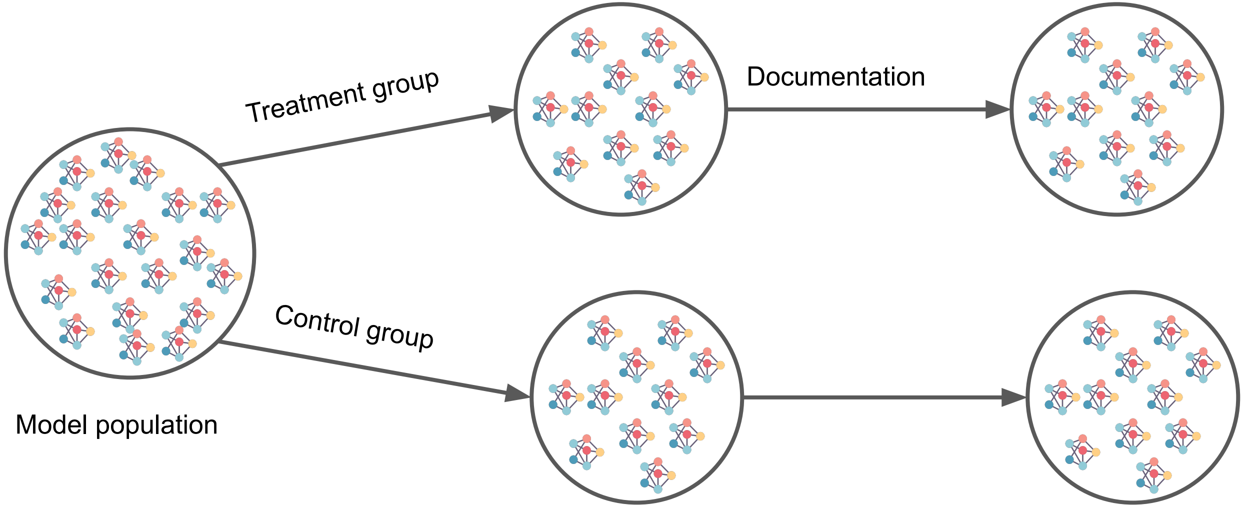
Model Documentation RCT

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

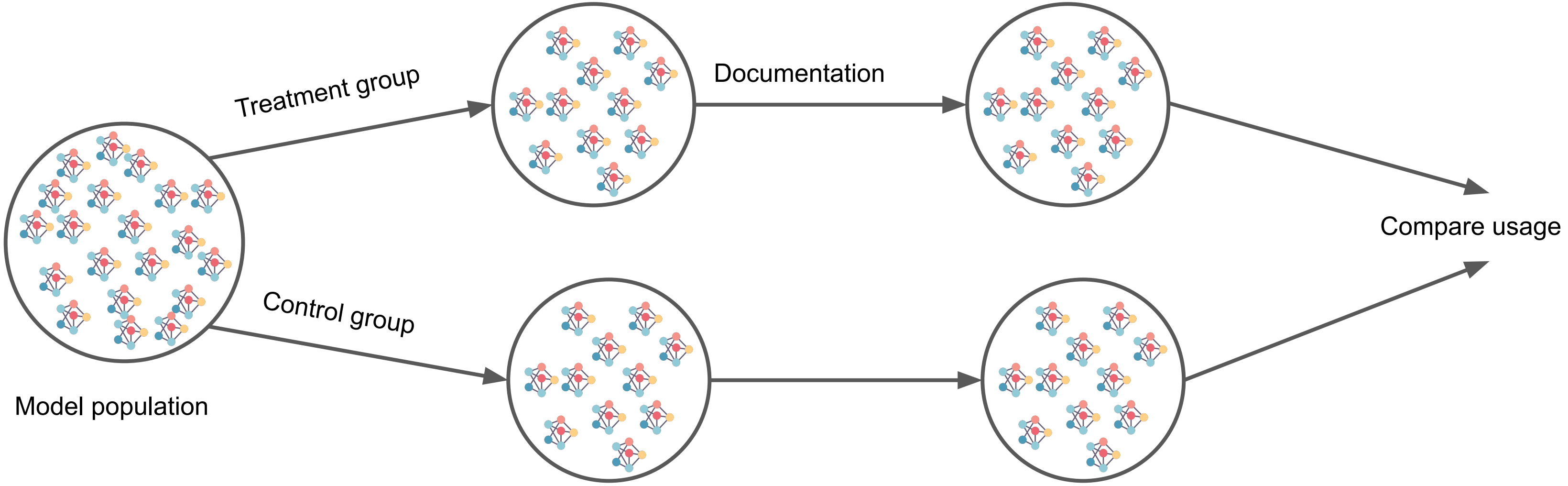
Randomized Control Trial (RCT) for models:



Model Documentation RCT

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

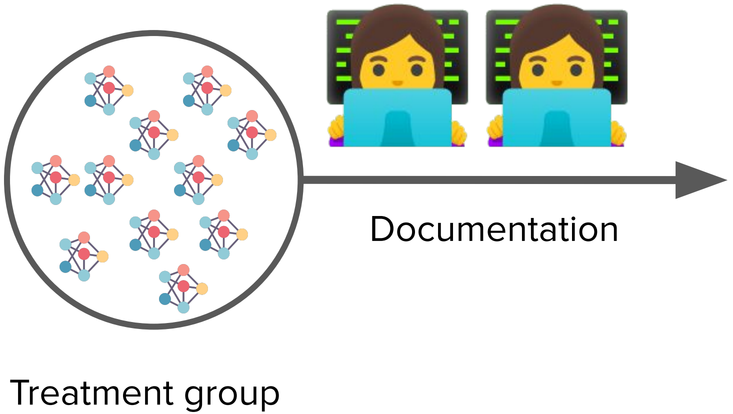
Randomized Control Trial (RCT) for models:



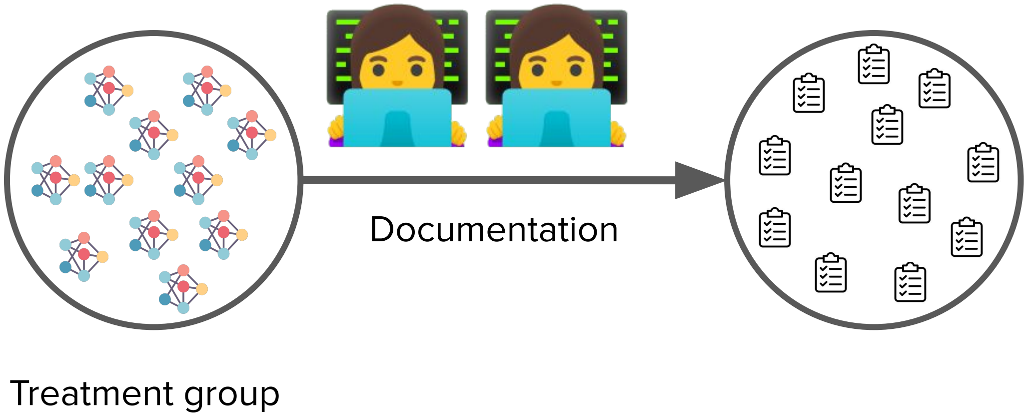
Randomized Control Trial Process

Treatment group

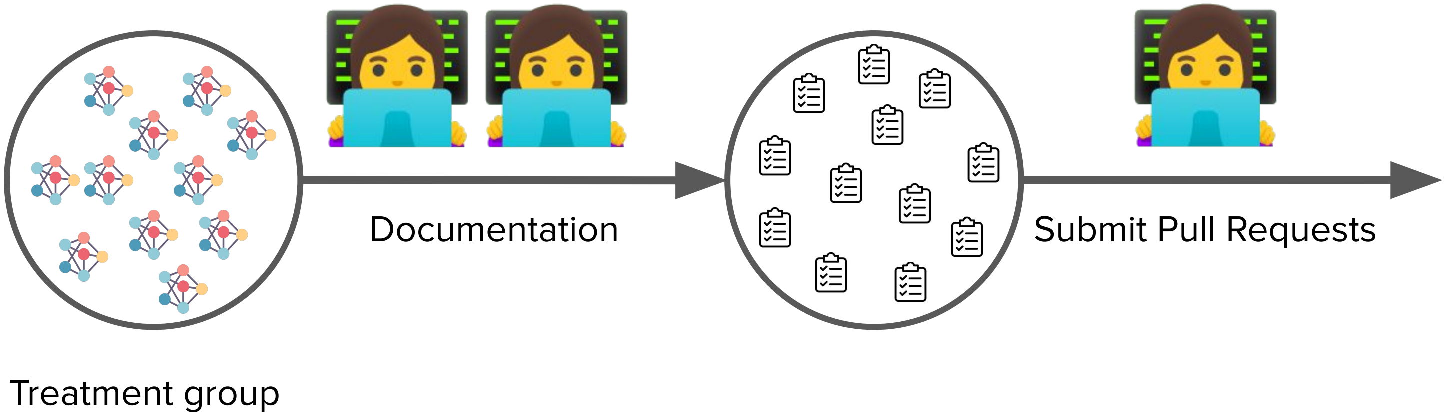
Randomized Control Trial Process



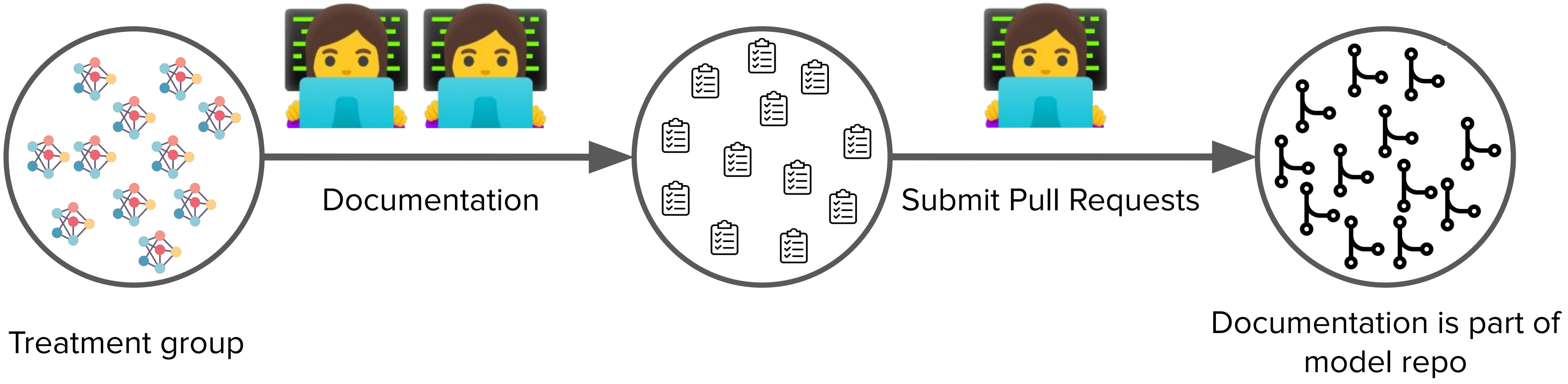
Randomized Control Trial Process



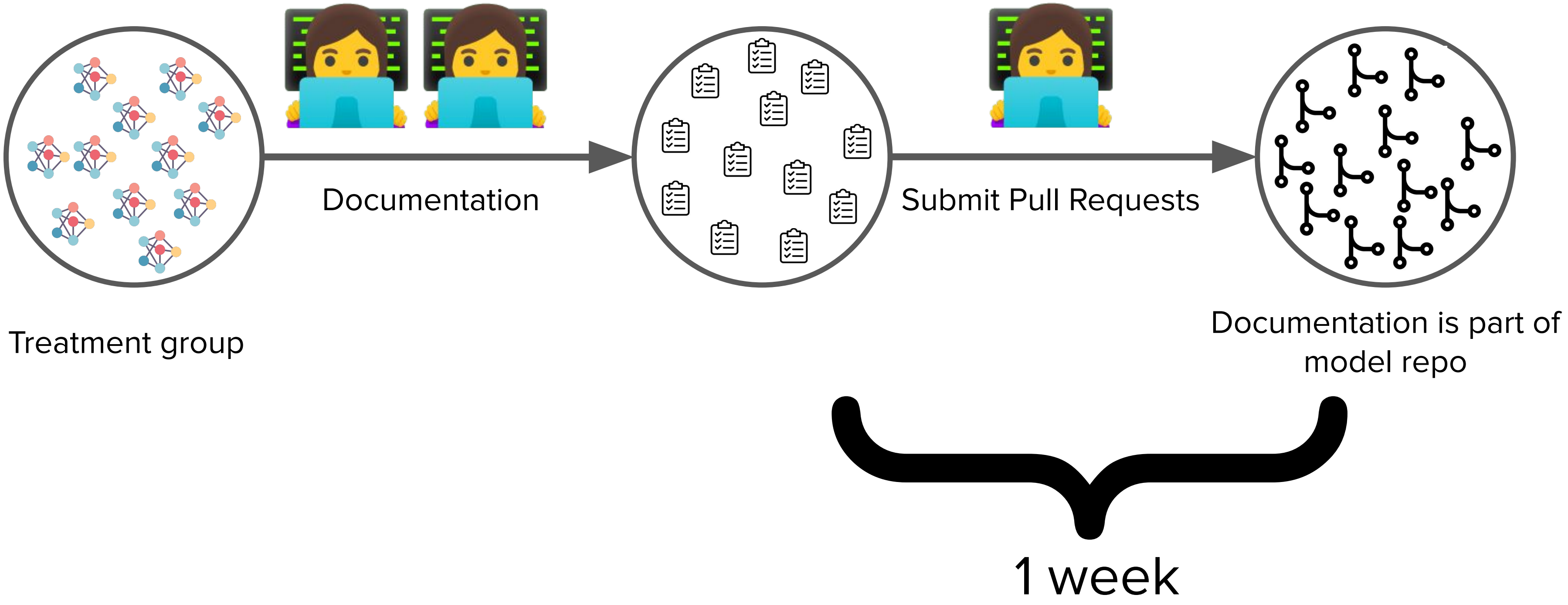
Randomized Control Trial Process



Randomized Control Trial Process

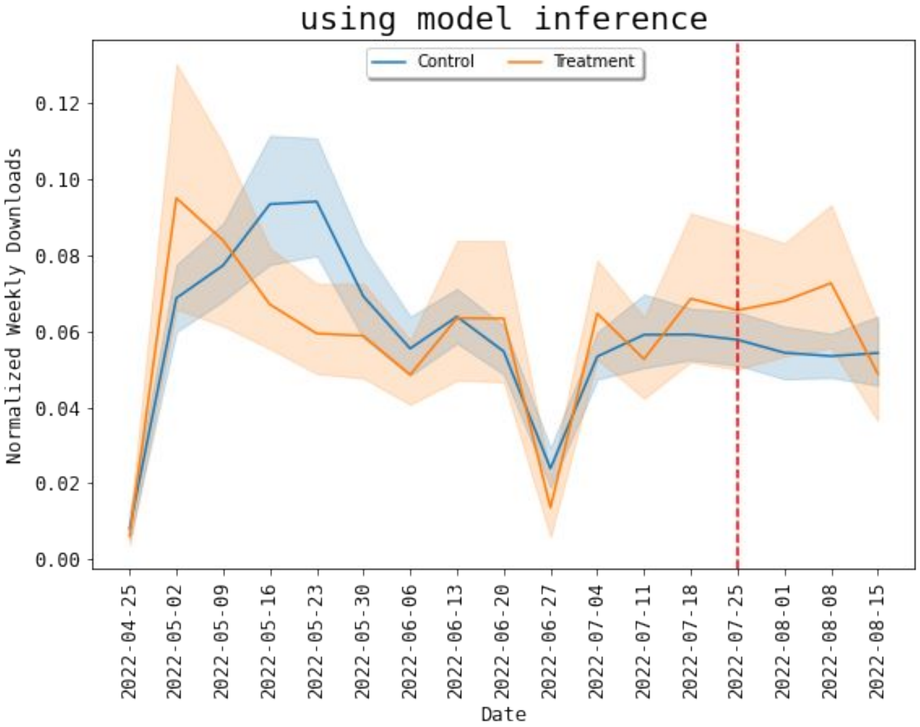


Randomized Control Trial Process



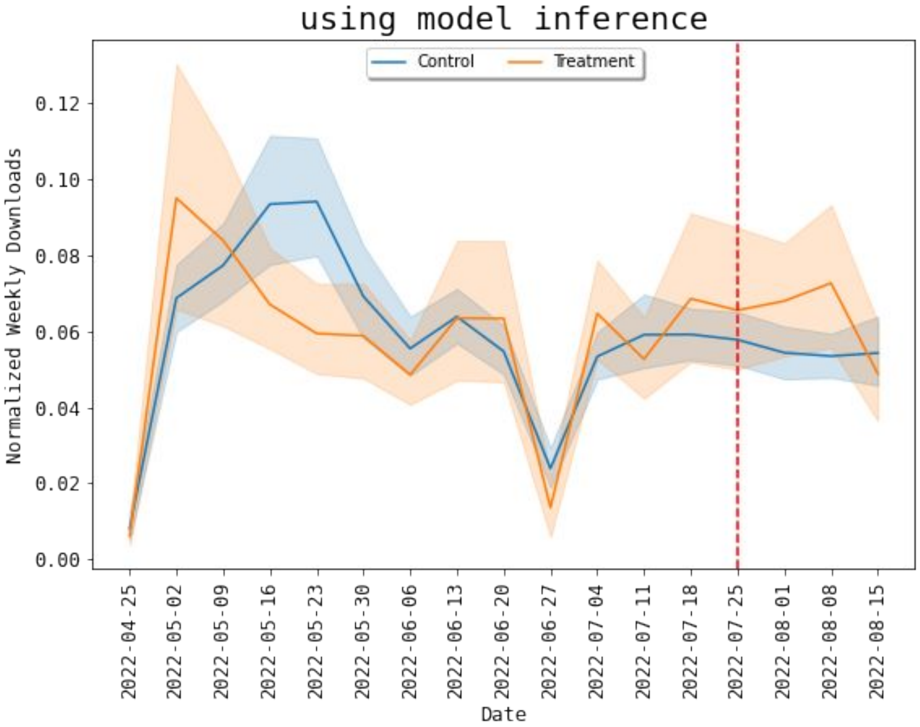
RCT Results

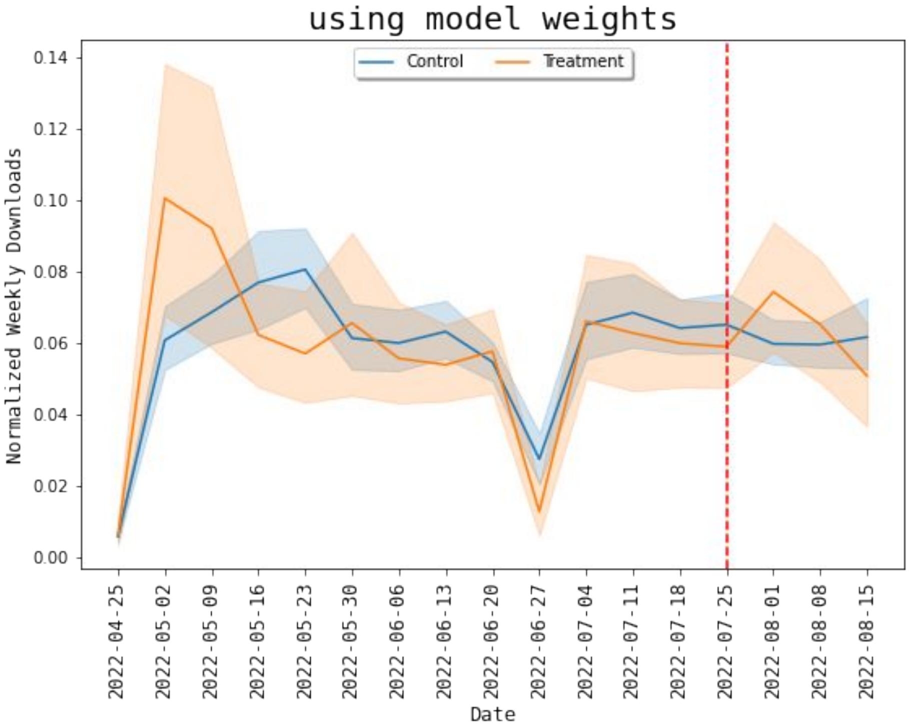
Red line indicates week when treatment was administered



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

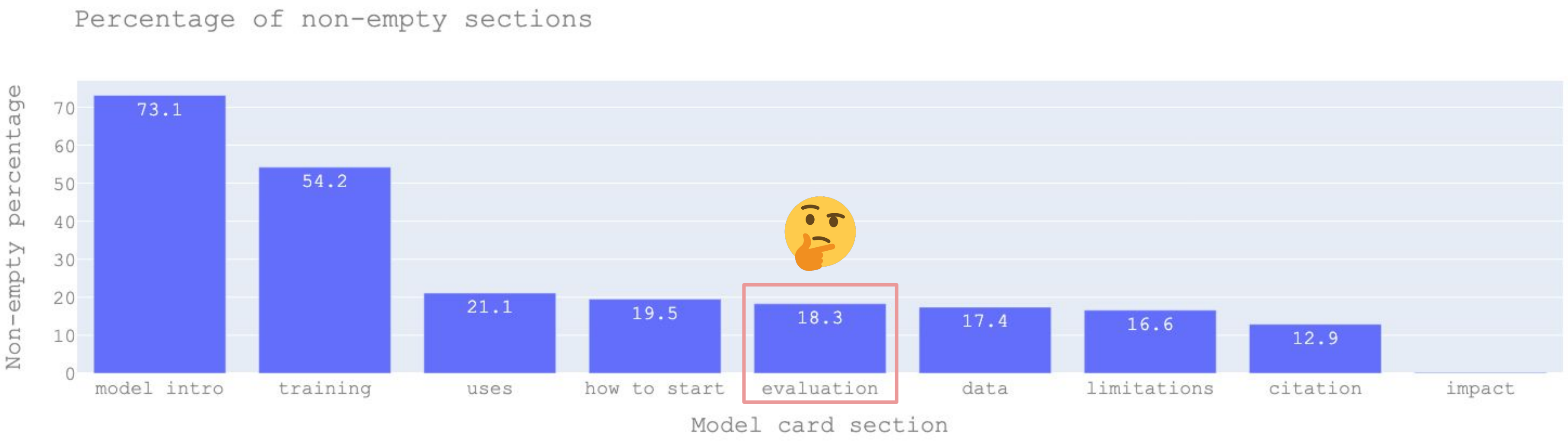
Model documentation drives model usage

What do developers document about models?

Distribution of sections in model cards

What do developers document about models?

Distribution of sections in model cards



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

NLP Evaluation Landscape

Slew of work on evaluation in NLP

NLP Evaluation Landscape

Slew of work on evaluation in NLP

Tools

snorkel

OpenAttack

Dyna Bench

HELM

Evaluate

Errudite: Scalable, Reproducible, and Testable Error Analysis

Tongshuang Wu[1], Marco Tulio Ribeiro[2], Jeffrey Heer[1], and Daniel S. Weld[1]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]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, Univ. of Washington, guestrin@cs.uw.edu
Sameer Singh, Univ. of California, Irvine, sameer@uci.edu

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

John X. Morris[1], Eli Lifland[1], Jin Yong Yoo[1], Jake Grigsby[1], Di Jin[2], Yanjun Qi[1]

[1] Department of Computer Science, University of Virginia
[2] Computer Science and Artificial Intelligence Laboratory, MIT
{jm8wx, yq2h}@virginia.edu

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

Alex Wang*, New York University
Yada Pruksachatkun*, New York University
Nikita Nangia*, 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

NLP Evaluation Landscape

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

Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan[★◇] **Swabha Swayamdipta**[★♡]
Omer Levy[♣] **Roy Schwartz**[♠♣] **Samuel R. Bowman**[†] **Noah A. Smith**[♣]
[◇] Department of Linguistics, University of Washington, Seattle, WA, USA
[♡] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[♣] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[♠] Allen Institute for Artificial Intelligence, Seattle, WA, USA
[†] Center for Data Science and Department of Linguistics, New York University, New York, NY, USA
{sg01,swabha,omerlevy,roysch,nasmith}@cs.washington.edu bowman@nyu.edu

Universal Adversarial Triggers for Attacking and Analyzing NLP

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

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

Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie[*], **Adina Williams**[†], **Emily Dinan**[†], **Mohit Bansal**[*], **Jason Weston**[†], **Douwe Kiela**[†]
[*]UNC Chapel Hill
[†]Facebook AI Research

How well do NLI models capture verb veridicality?

Alexis Ross
Harvard University
alexis_ross@college.harvard.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Stress Test Evaluation for Natural Language Inference

Aakanksha Naik[1*], **Abhilasha Ravichander**[1*],
Norman Sadeh[2], **Carolyn Rose**[1], **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]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

NLP Evaluation Idioms

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

NLP Evaluation Idioms

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)

NLP Evaluation Idioms

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

2. Transformations – natural perturbations to original evaluation instances

NLP Evaluation Idioms

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)

NLP Evaluation Idioms

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

NLP Evaluation Idioms

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)

NLP Evaluation Idioms

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

NLP Evaluation Idioms

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

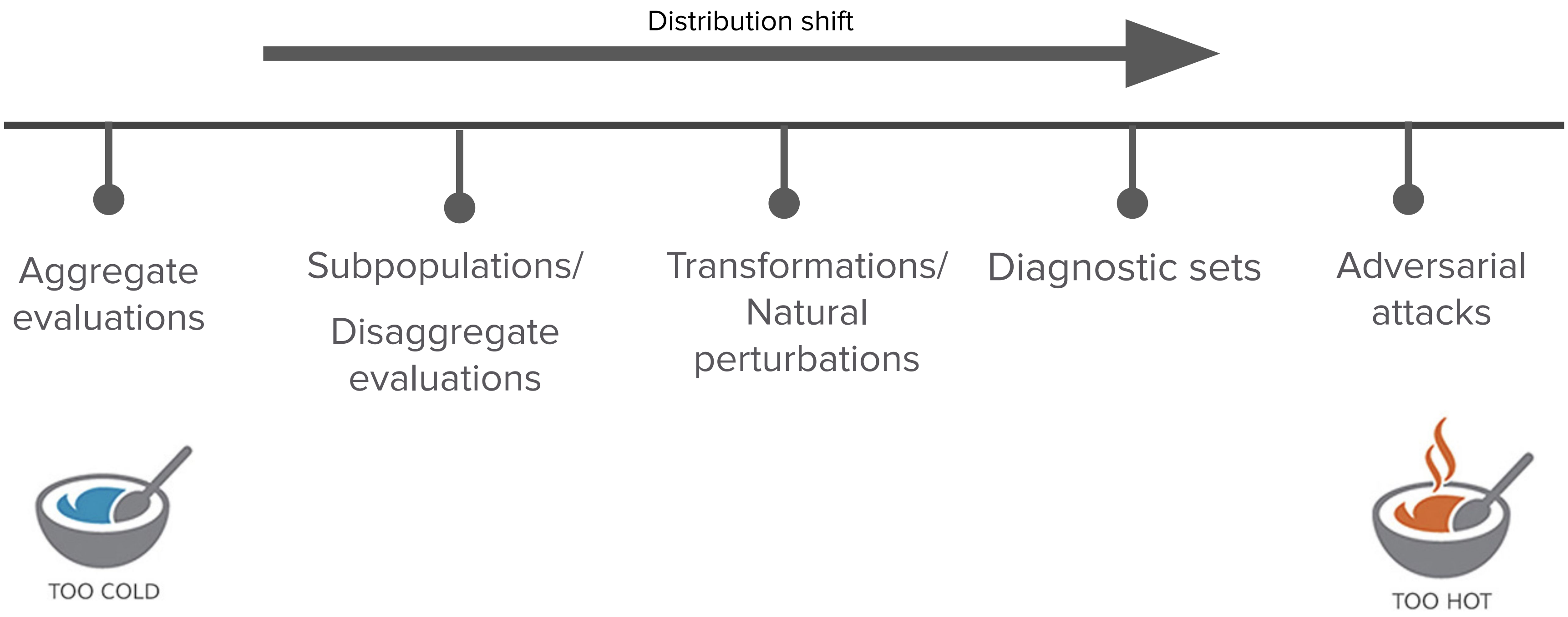
Tools: TextAttack (Morris et al., 2020), OpenAttack (Zeng et al., 2020)

NLP Evaluation Landscape

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

Evaluation Idiom	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

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

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

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 AI recognition in other racial categories. ... remained blocked on Google Photos after Alciné's tweet

Jan 12, 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

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

VB VentureBeat

AI 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 Today

Idiomatic Lock-In

	Tool A	Tool B
Subpopulations	✓	✗
Transformations	✗	✓
Attacks	✗	✓
Evaluation sets	✗	✗

Workflow Fragmentation

Scattered evaluation

Difficulty reporting

Challenges with evaluation

Challenges Today

Idiomatic Lock-In

	Tool A	Tool B
Subpopulations	✓	✗
Transformations	✗	✓
Attacks	✗	✓
Evaluation sets	✗	✗

Workflow Fragmentation

Scattered evaluation

Difficulty reporting

Challenges with evaluation

Challenges Today

Idiomatic Lock-In

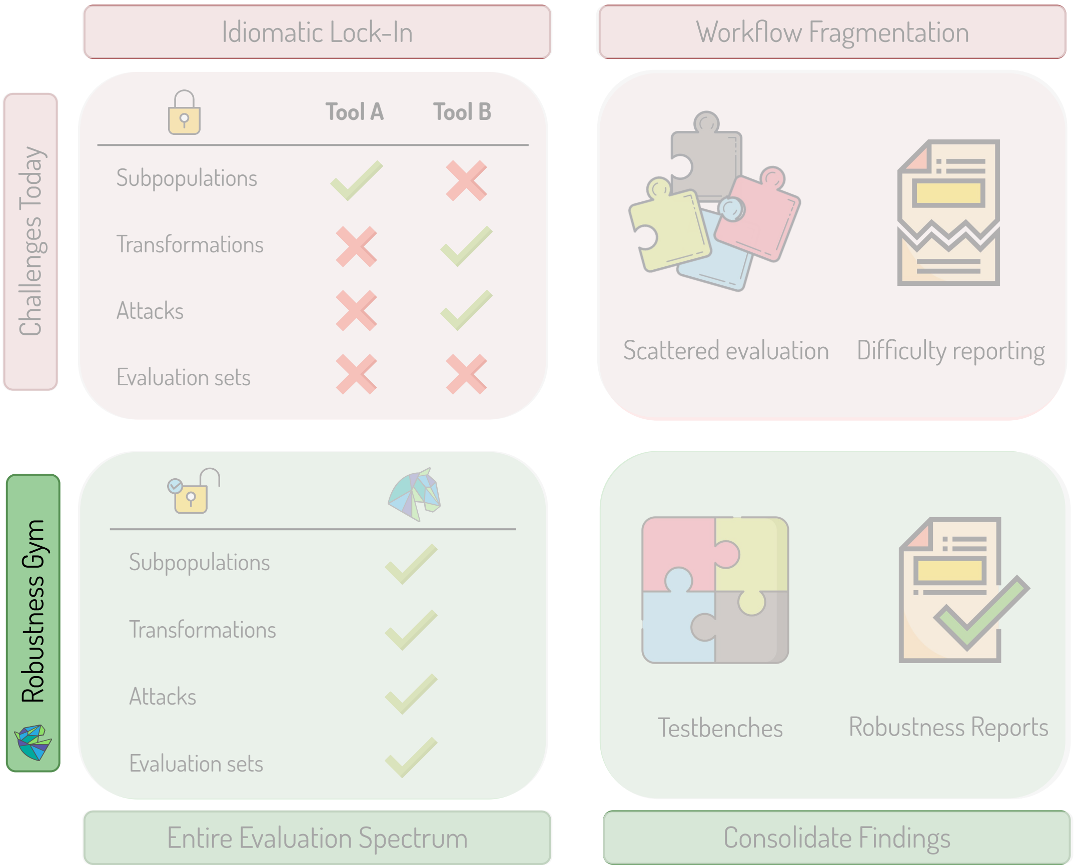
	Tool A	Tool B
Subpopulations	✓	✗
Transformations	✗	✓
Attacks	✗	✓
Evaluation sets	✗	✗

Workflow Fragmentation

Scattered evaluation

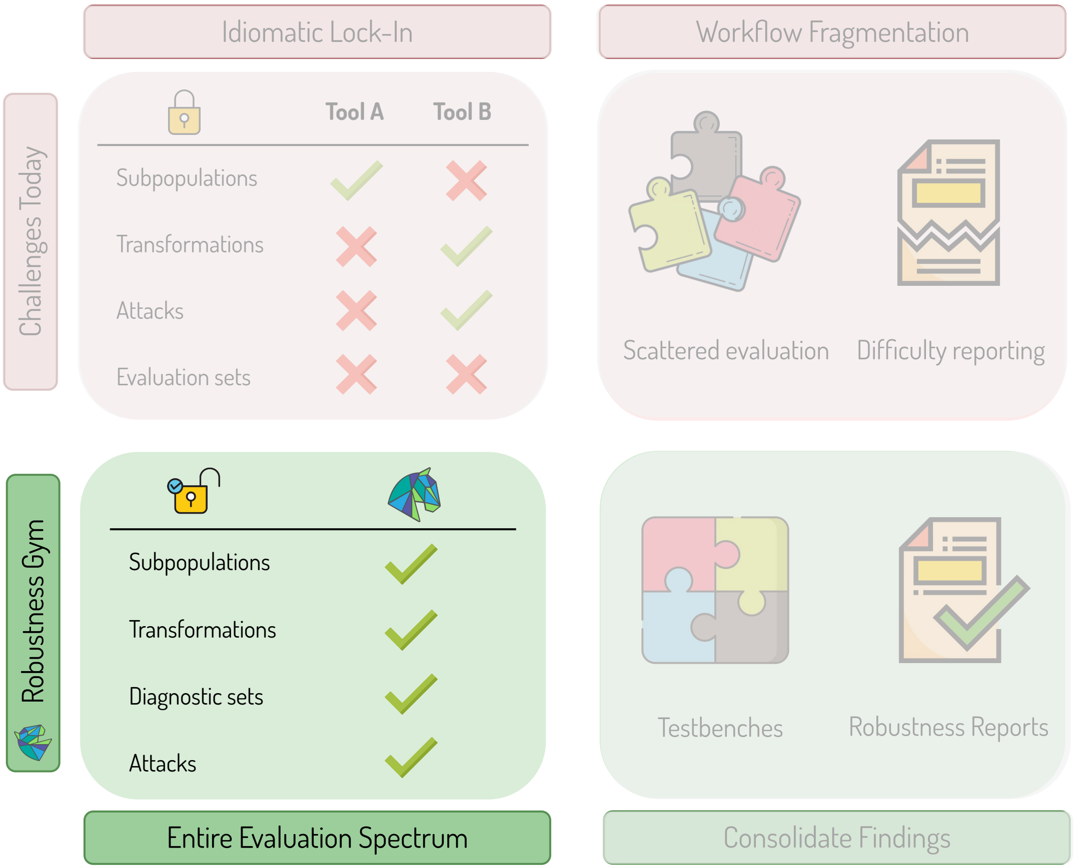
Difficulty reporting

Robustness Gym



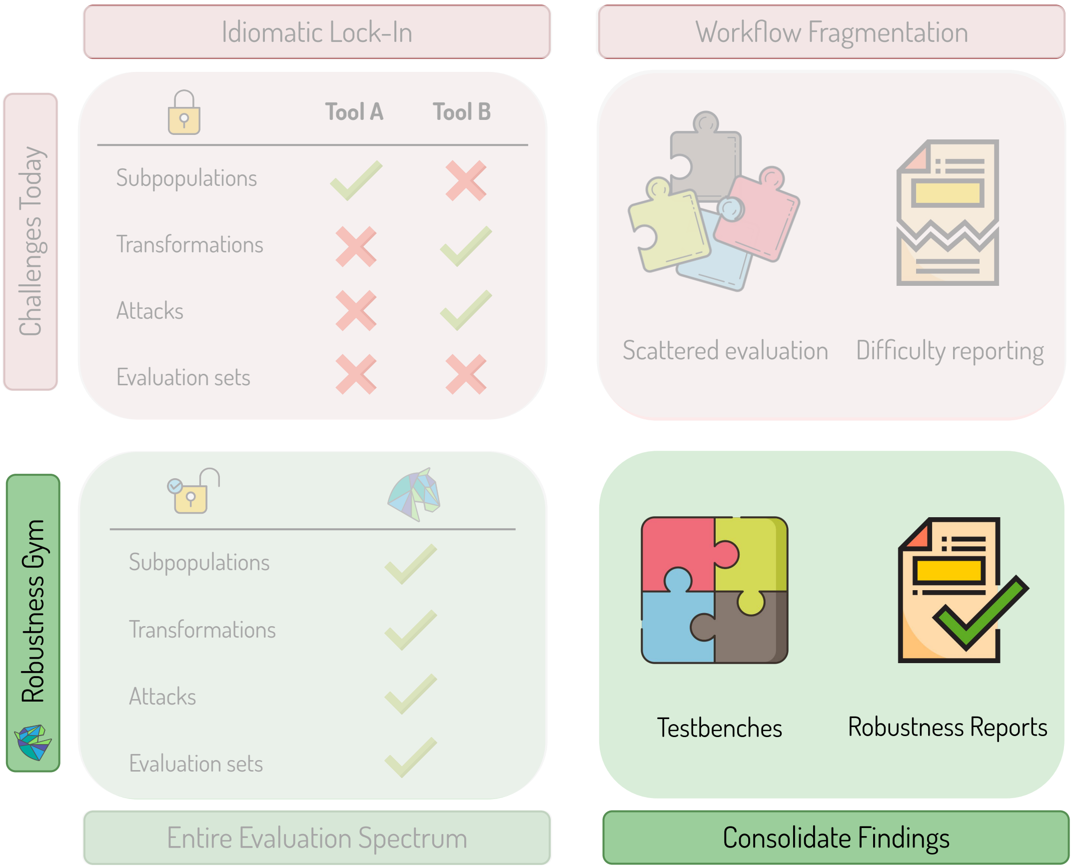
(Goel*, **Rajani***, et al., NAACL 2021)

Robustness Gym



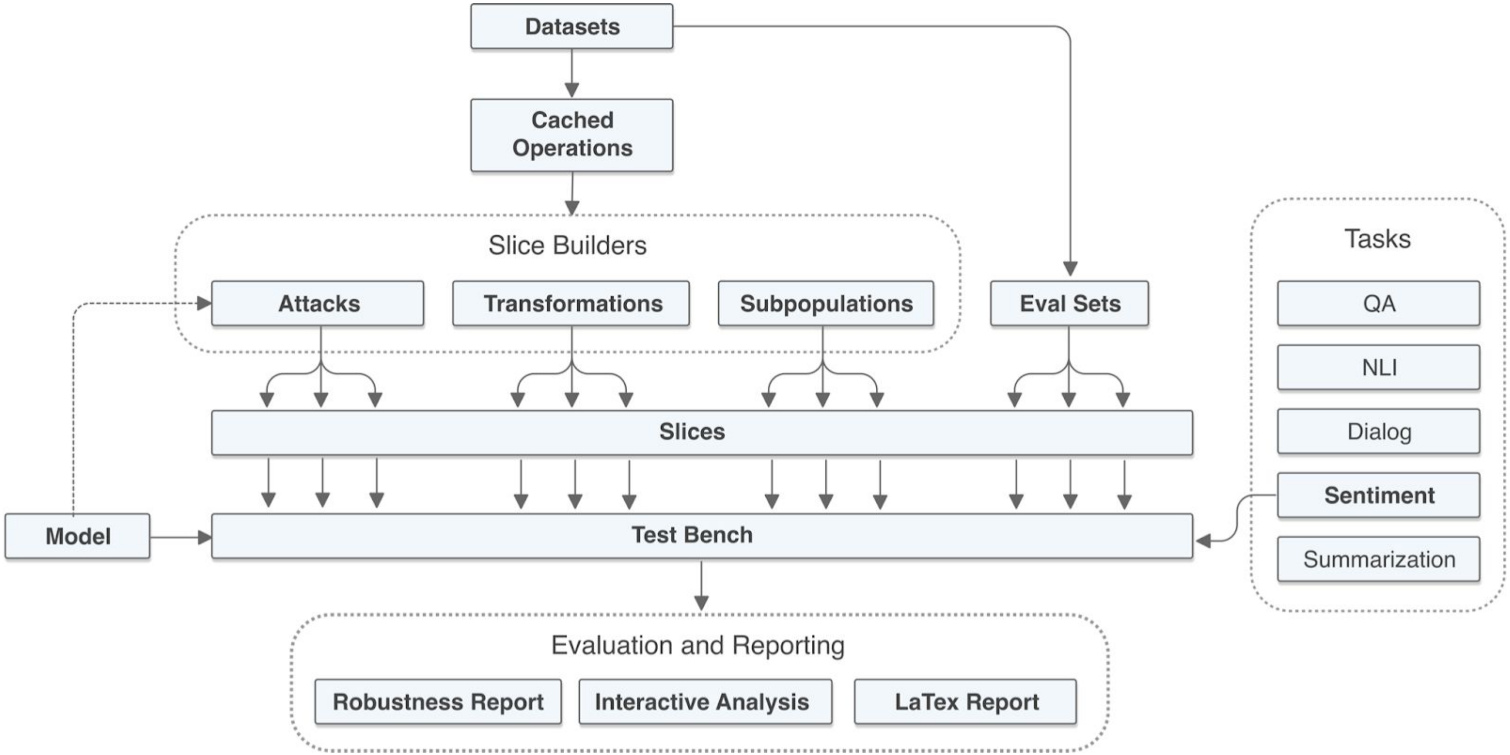
(Goel*, **Rajani***, et al., NAACL 2021)

Robustness Gym

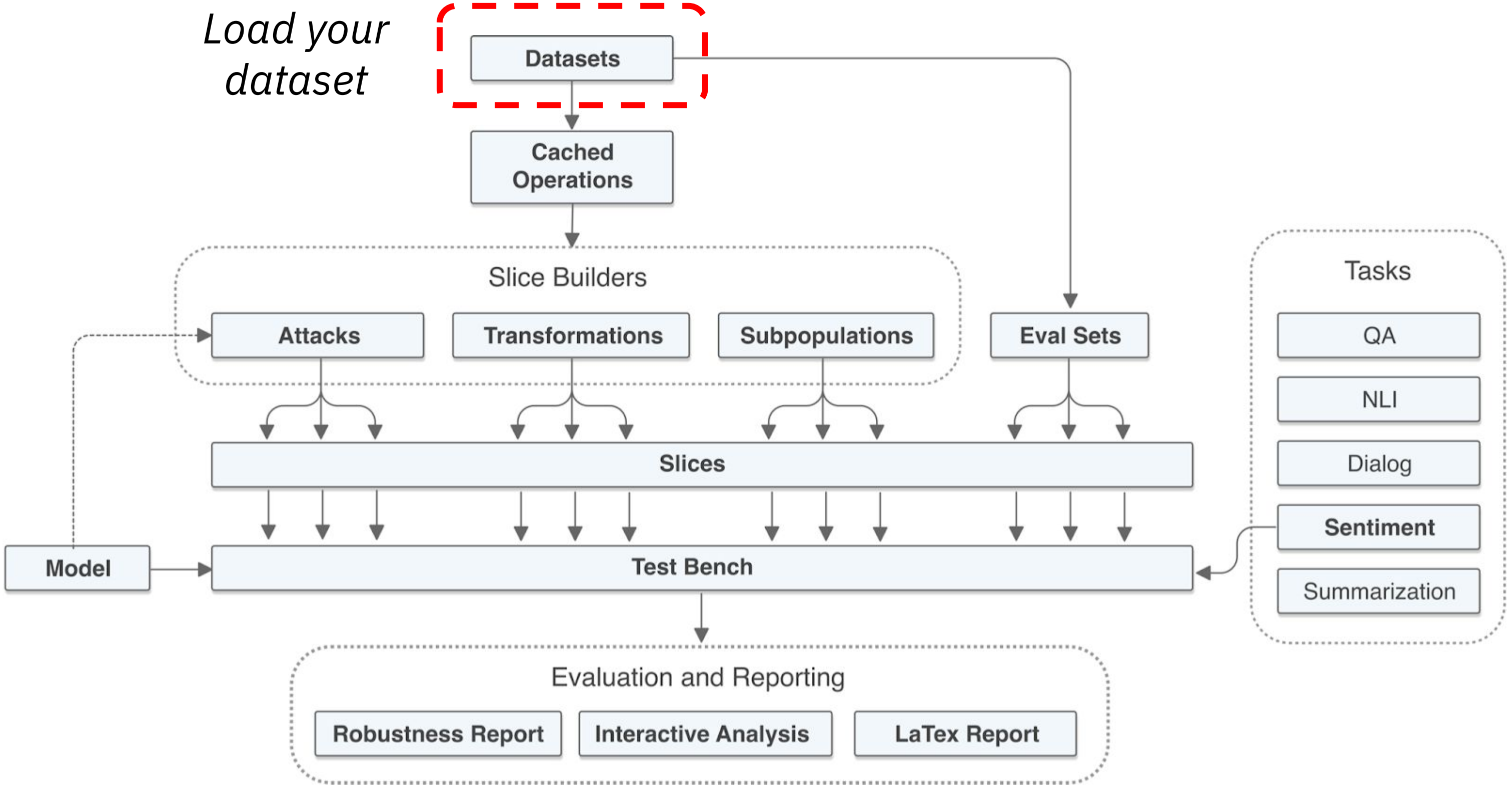


(Goel*, **Rajani***, et al., NAACL 2021)

Robustness Gym Workflow

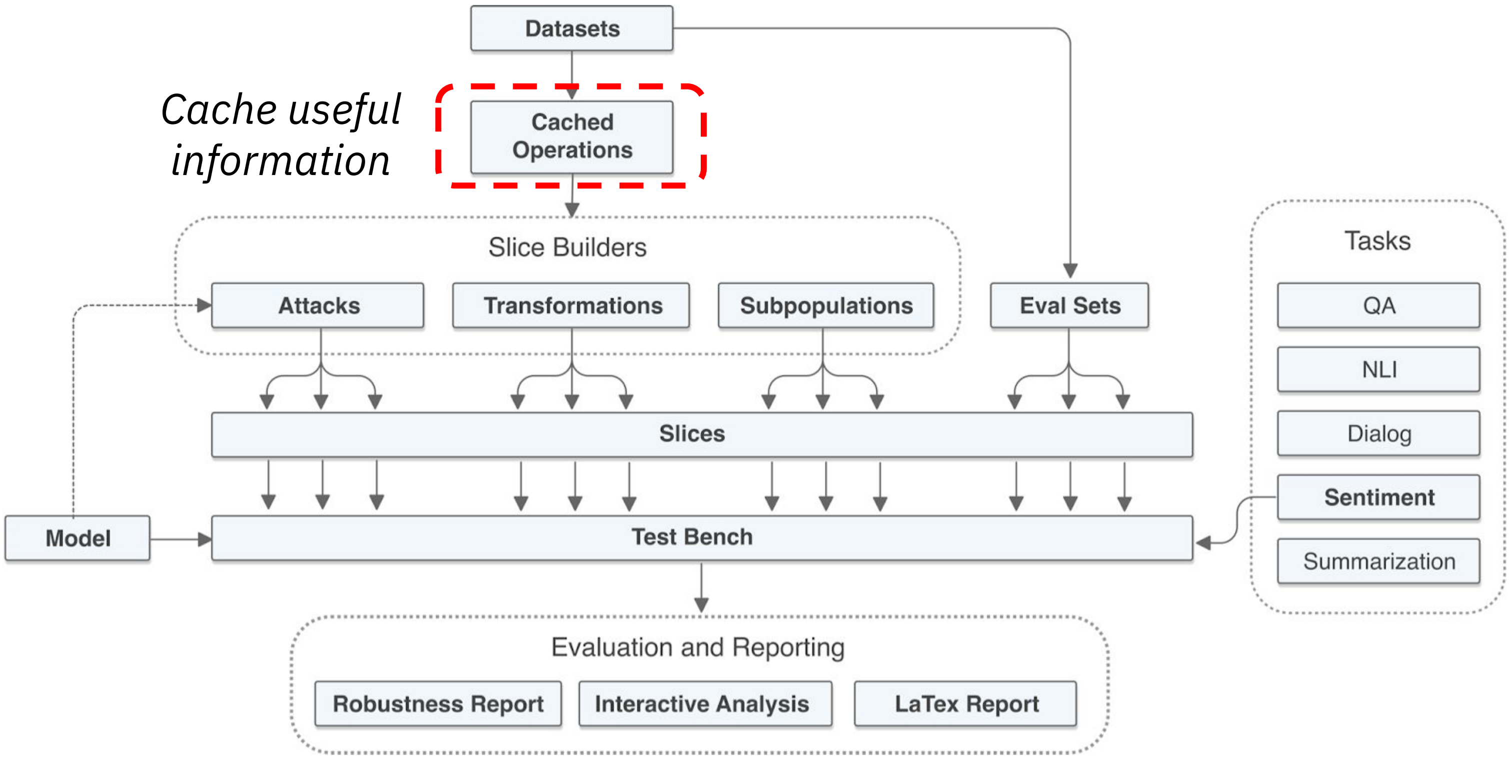


Robustness Gym Workflow

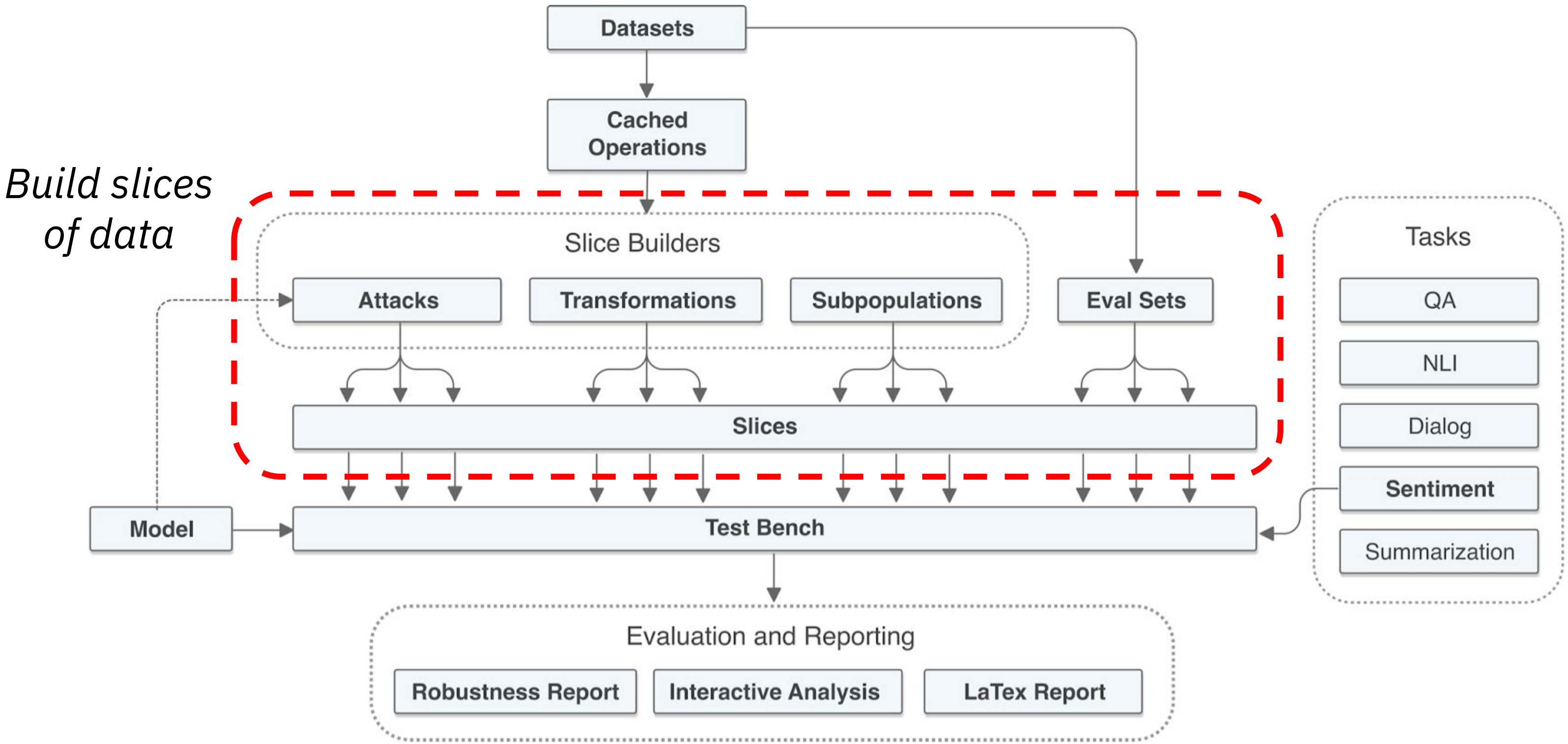


Robustness Gym Workflow

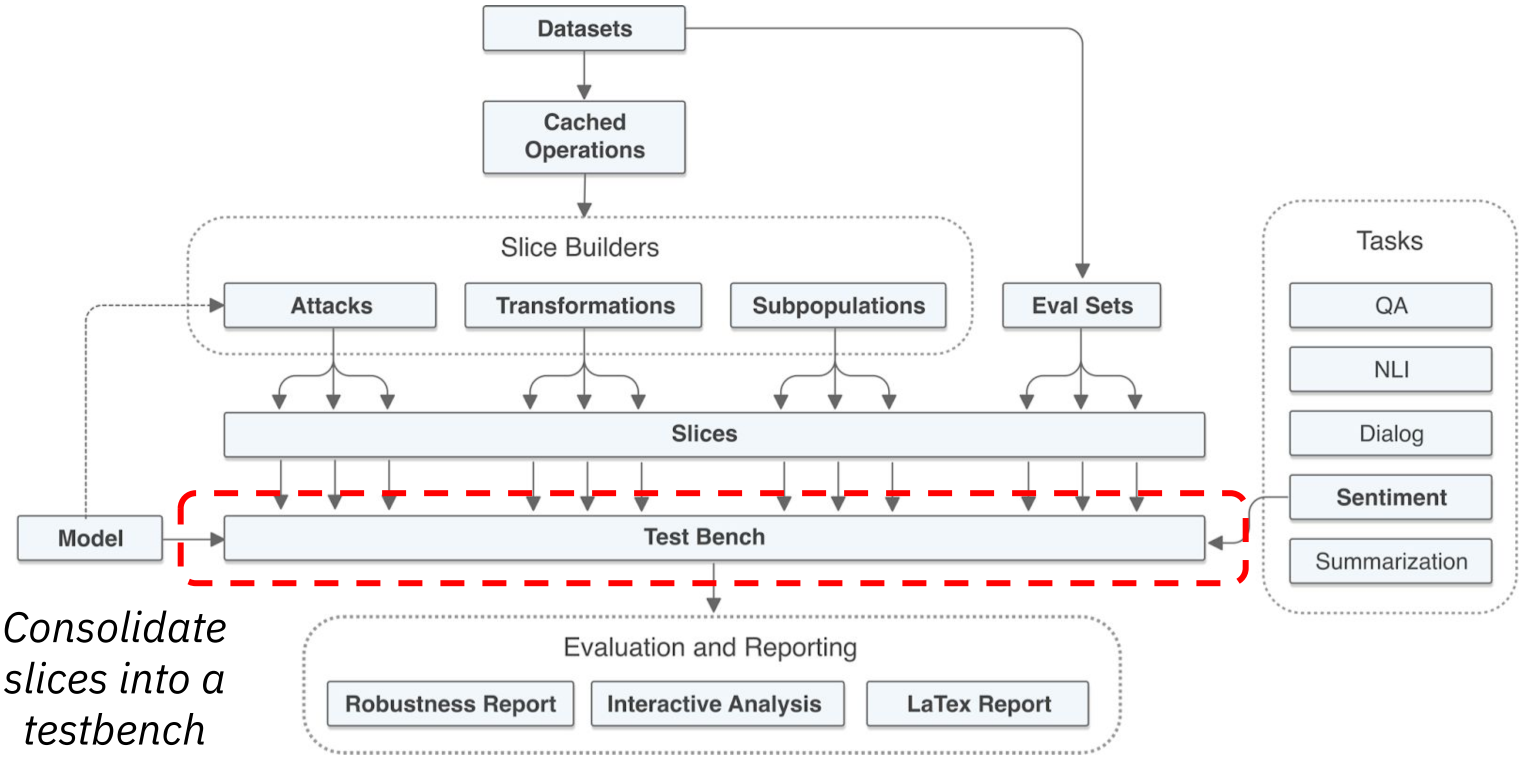
Cache useful information



Robustness Gym Workflow

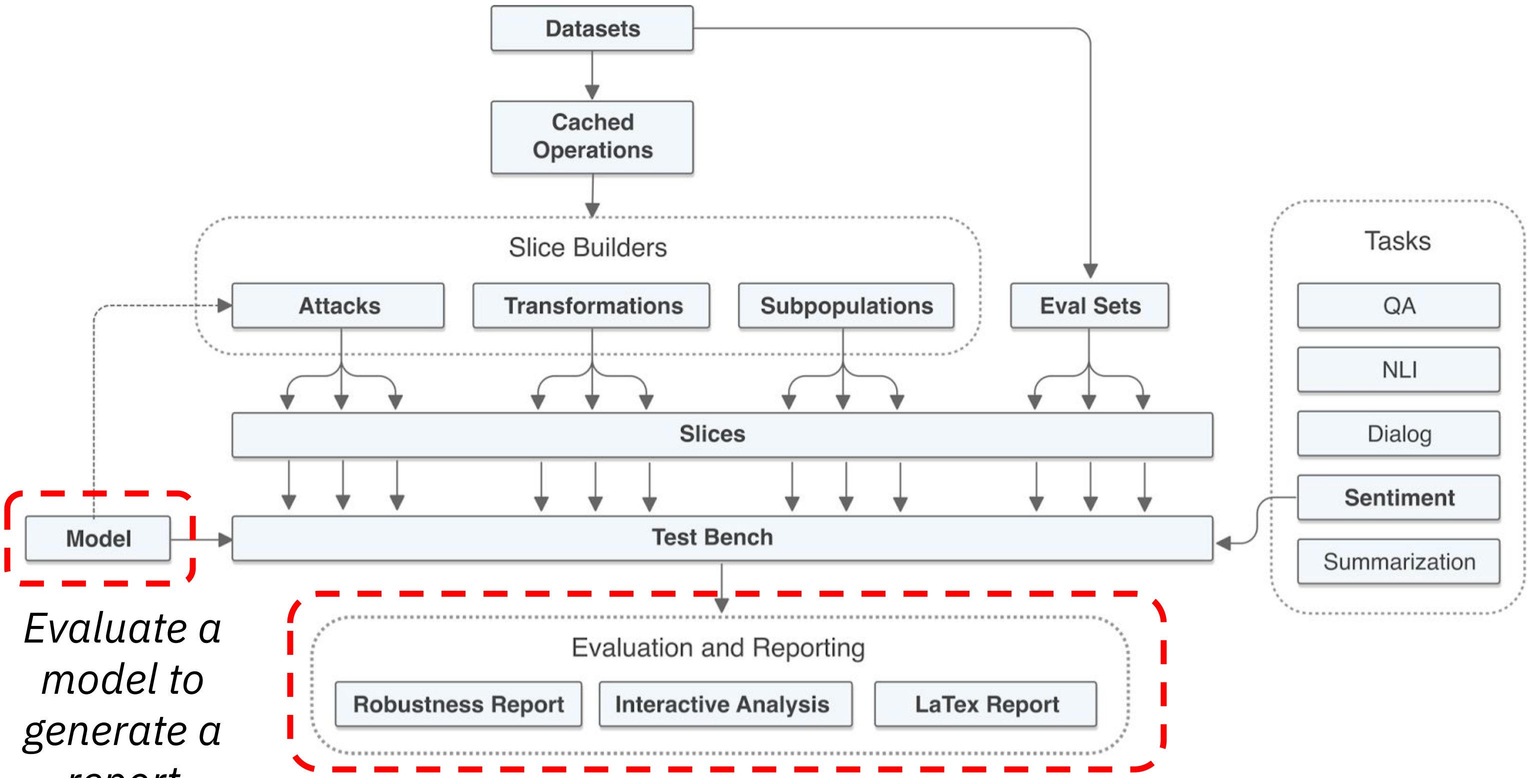


Robustness Gym Workflow

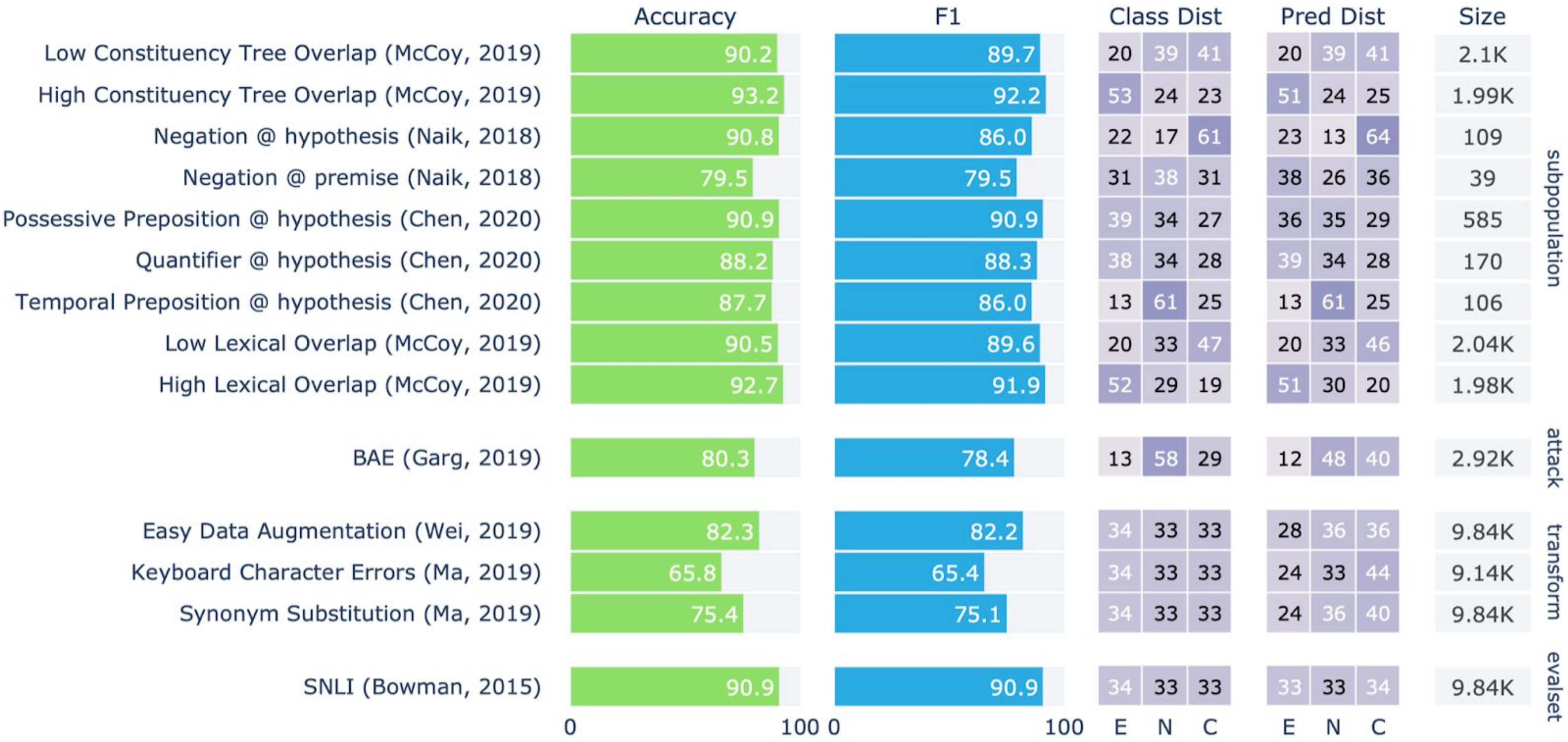


Consolidate slices into a testbench

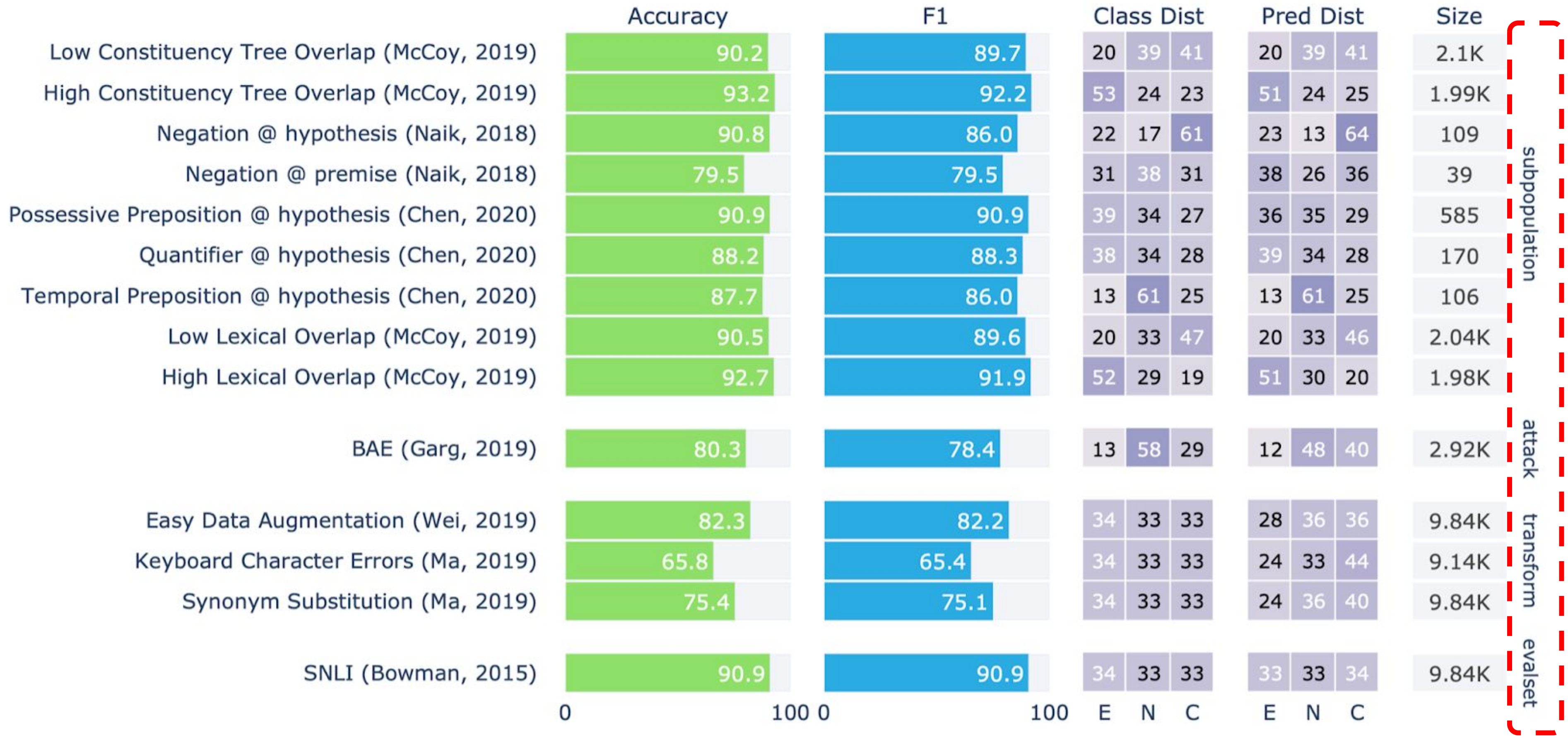
Robustness Gym Workflow



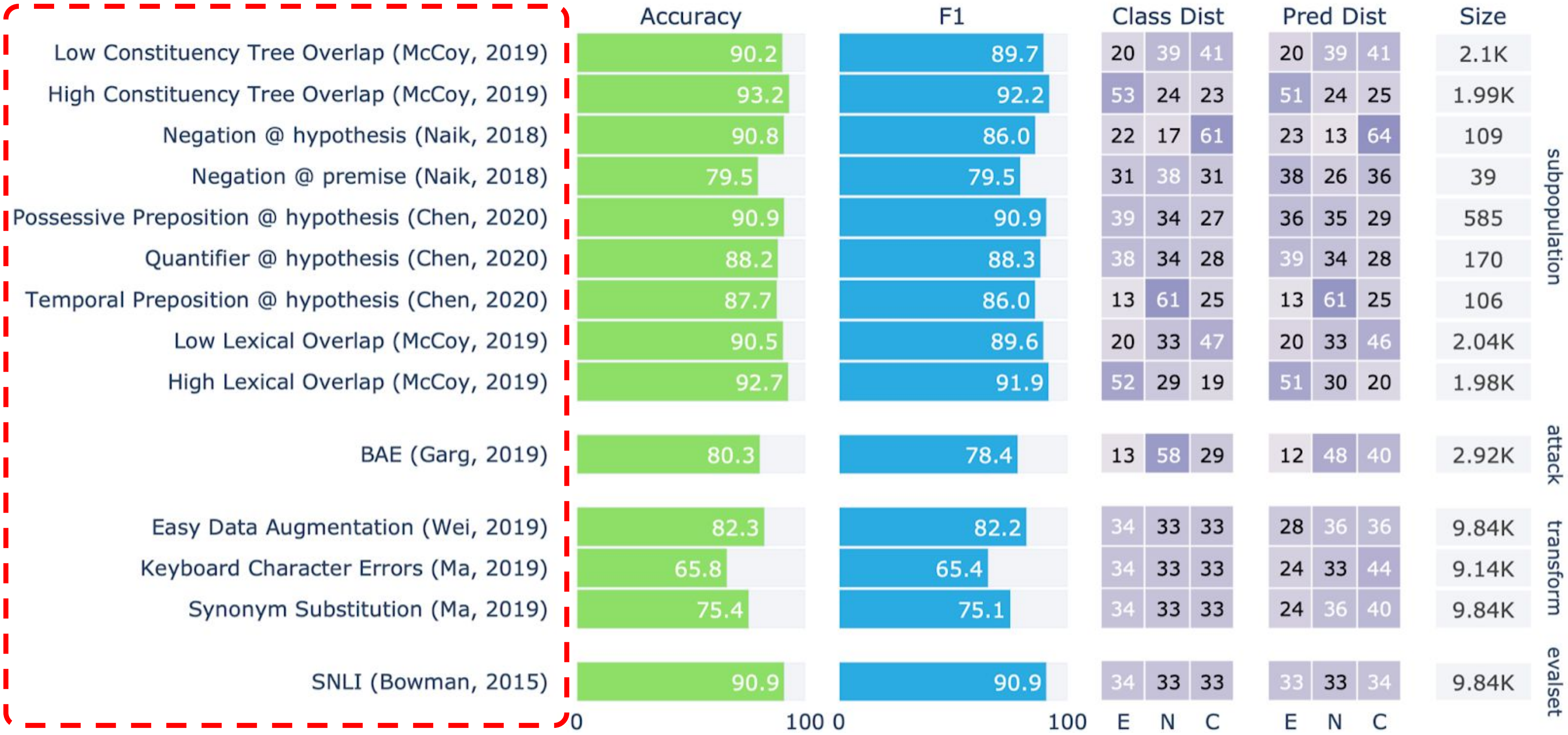
Evaluate a model to generate a report



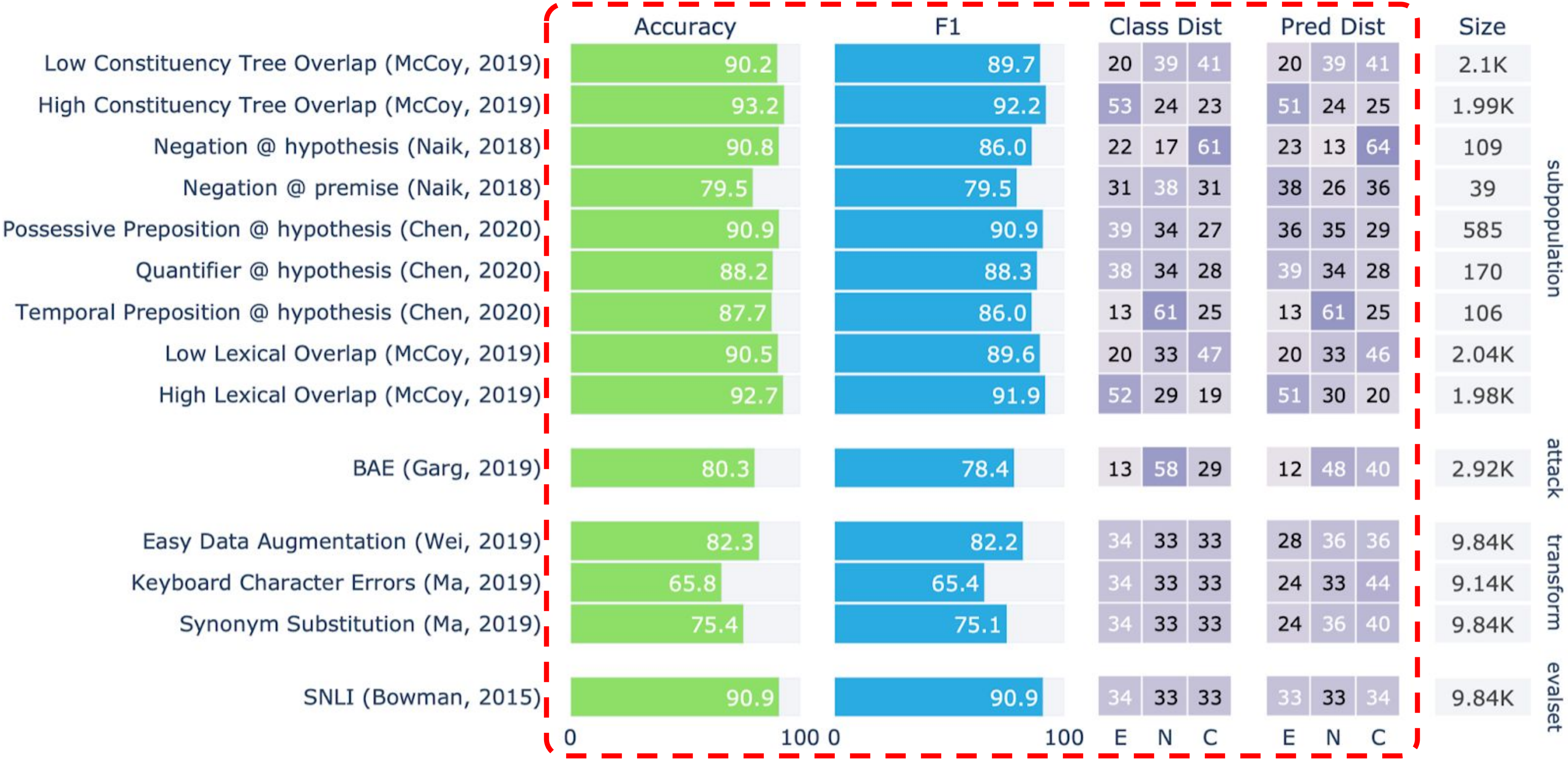
Robustness Report for Natural Language Inference using *bert-base-uncased* on SNLI



Robustness Report for Natural Language Inference using *bert-base-uncased* on SNLI

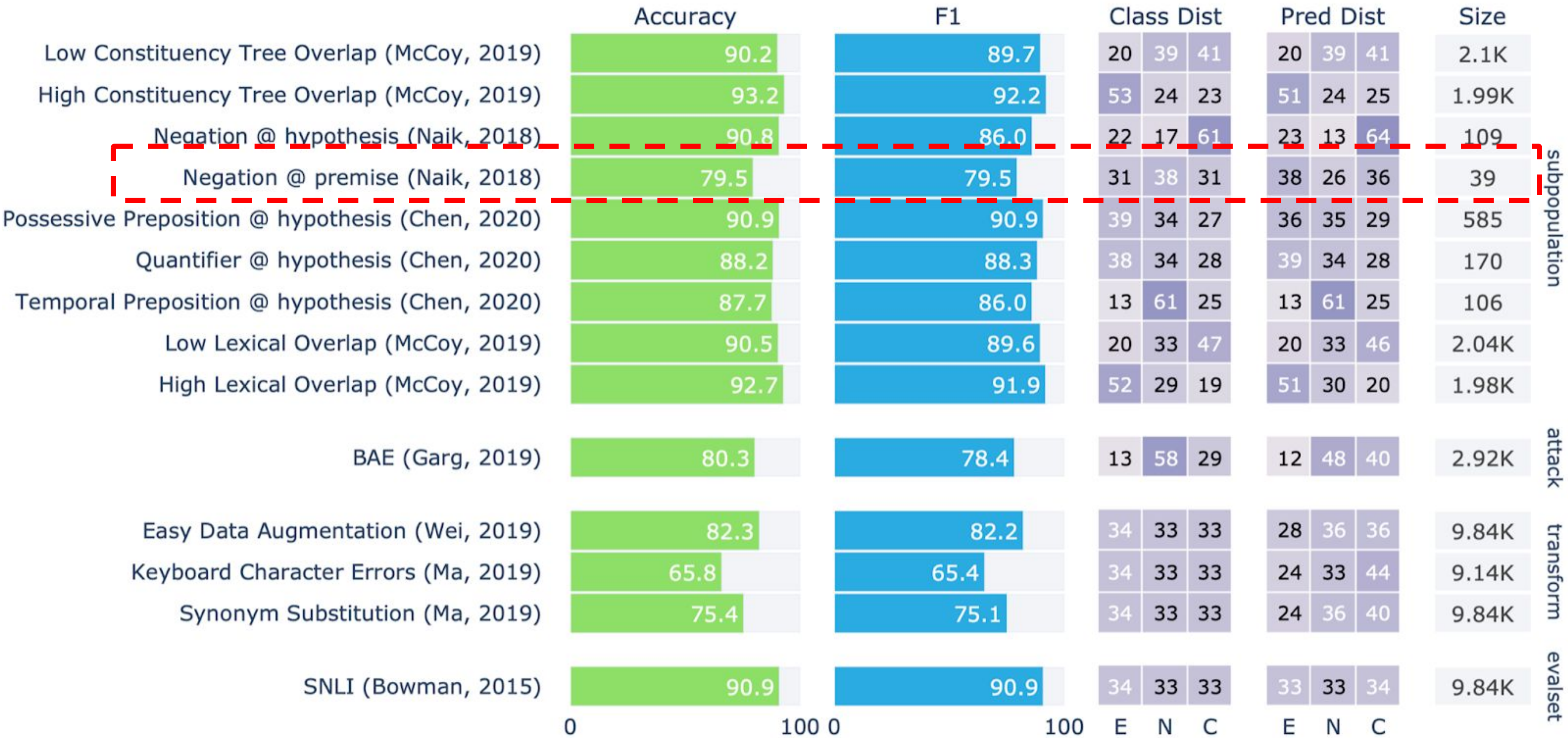


Robustness Report for Natural Language Inference using *bert-base-uncased* on SNLI

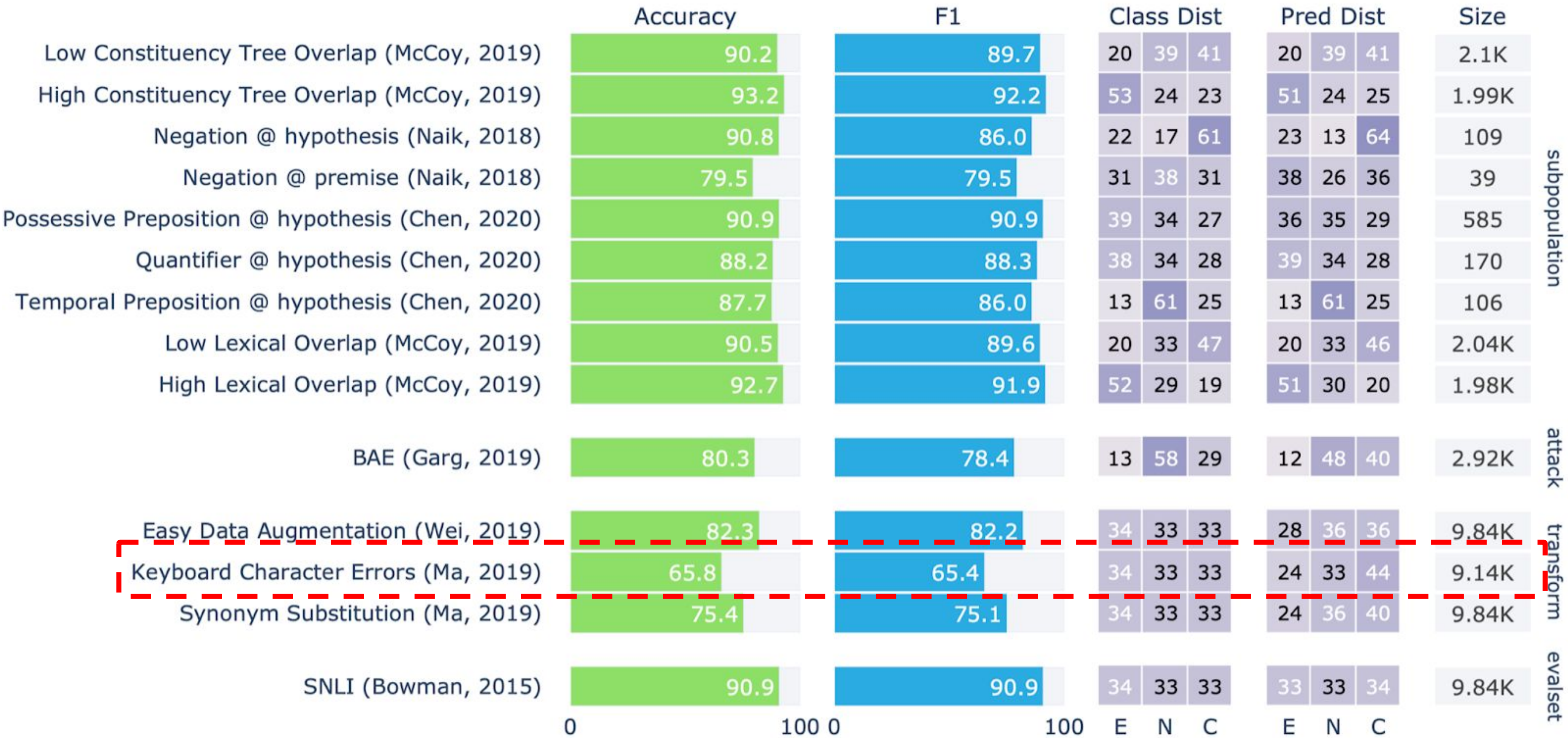


	Accuracy	F1	Class Dist E	Class Dist N	Class Dist C	Pred Dist E	Pred Dist N	Pred Dist C	Size
Low Constituency Tree Overlap (McCoy, 2019)	90.2	89.7	20	39	41	20	39	41	2.1K
High Constituency Tree Overlap (McCoy, 2019)	93.2	92.2	53	24	23	51	24	25	1.99K
Negation @ hypothesis (Naik, 2018)	90.8	86.0	22	17	61	23	13	64	109
Negation @ premise (Naik, 2018)	79.5	79.5	31	38	31	38	26	36	39
Possessive Preposition @ hypothesis (Chen, 2020)	90.9	90.9	39	34	27	36	35	29	585
Quantifier @ hypothesis (Chen, 2020)	88.2	88.3	38	34	28	39	34	28	170
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	12	48	40	2.92K
Easy Data Augmentation (Wei, 2019)	82.3	82.2	34	33	33	28	36	36	9.84K
Keyboard Character Errors (Ma, 2019)	65.8	65.4	34	33	33	24	33	44	9.14K
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	33	33	34	9.84K

Robustness Report for Natural Language Inference using *bert-base-uncased* on SNLI



Robustness Report for Natural Language Inference using *bert-base-uncased* on SNLI



Robustness Report for Natural Language Inference using *bert-base-uncased* on SNLI

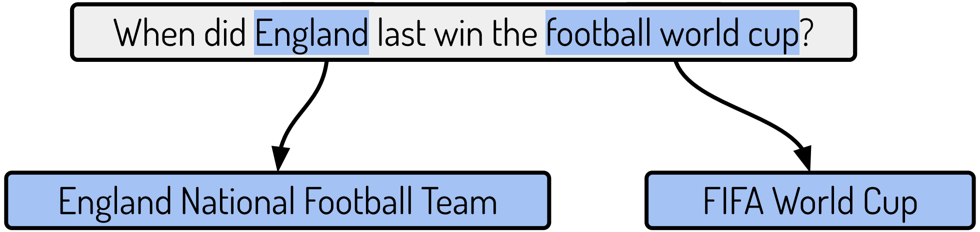
Experiments with Commercial APIs for Named Entity Linking

Named Entity Linking
map "strings" to "things" in a knowledge base like Wikipedia

When did England last win the football world cup?

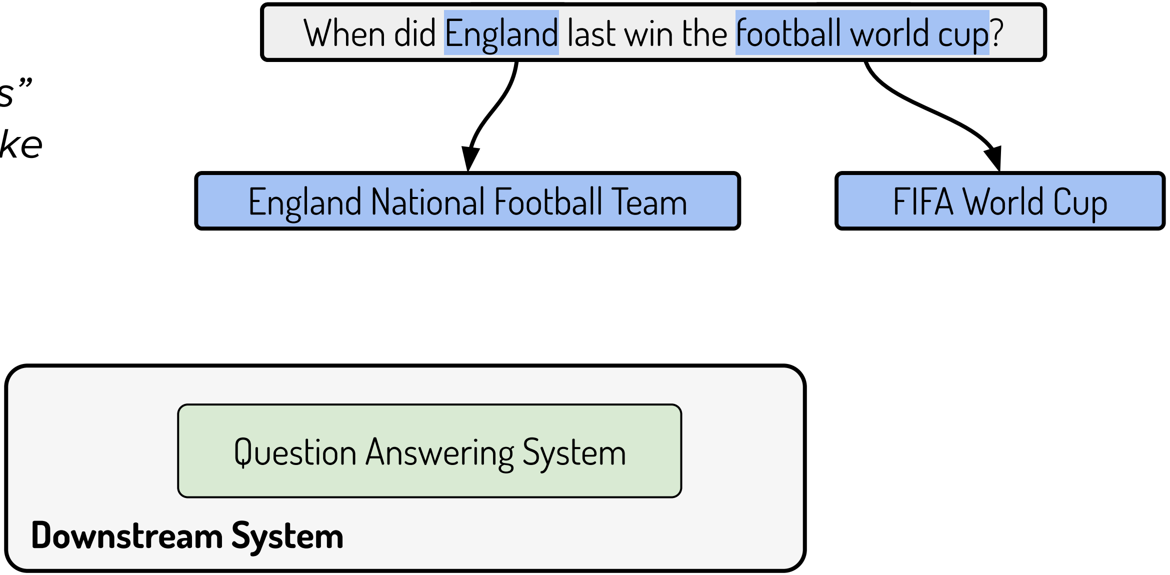
Experiments with Commercial APIs for Named Entity Linking

Named Entity Linking
map “strings” to “things” in a knowledge base like Wikipedia



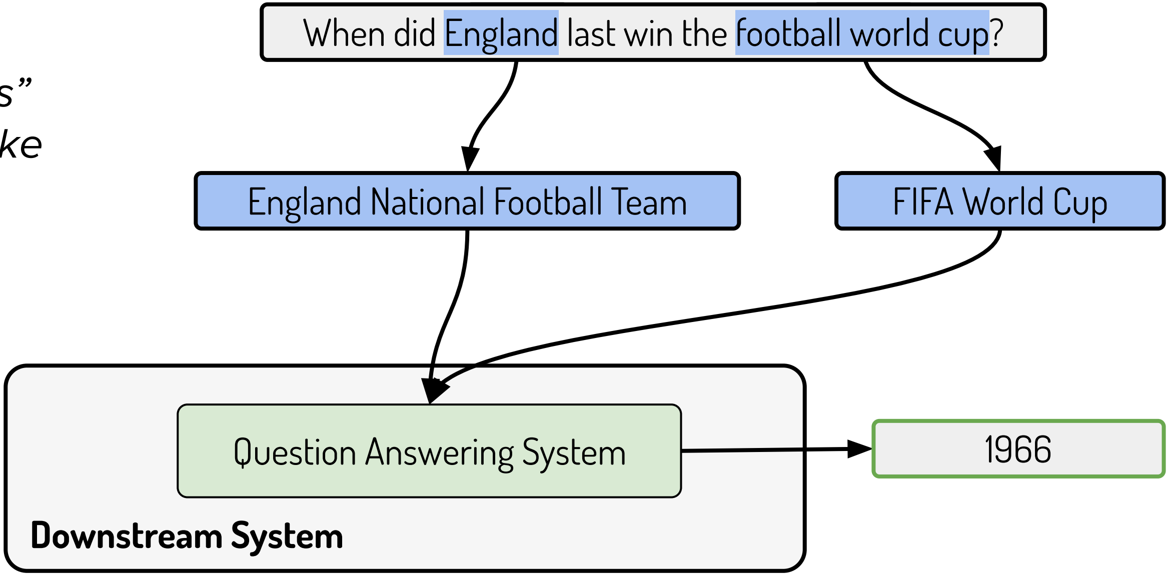
Experiments with Commercial APIs for Named Entity Linking

Named Entity Linking
map "strings" to "things" in a knowledge base like Wikipedia



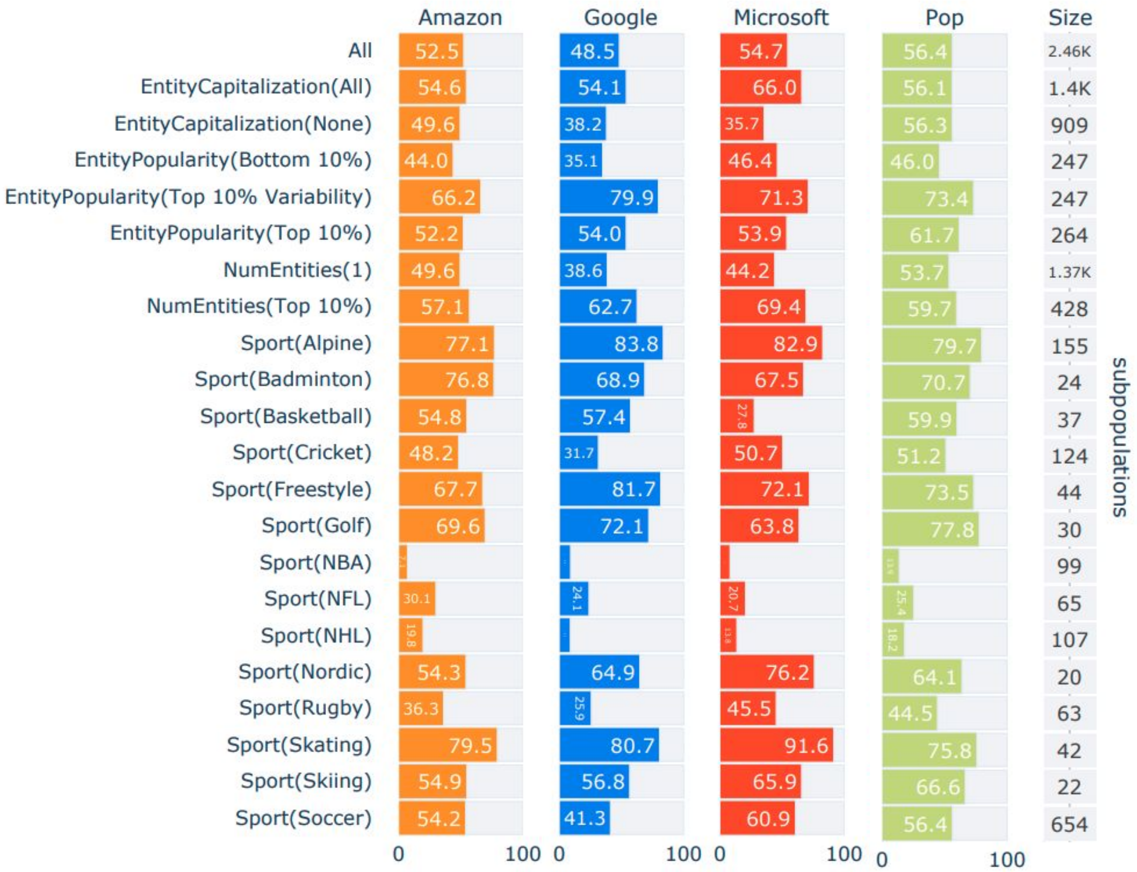
Experiments with Commercial APIs for Named Entity Linking

Named Entity Linking
map "strings" to "things" in a knowledge base like Wikipedia



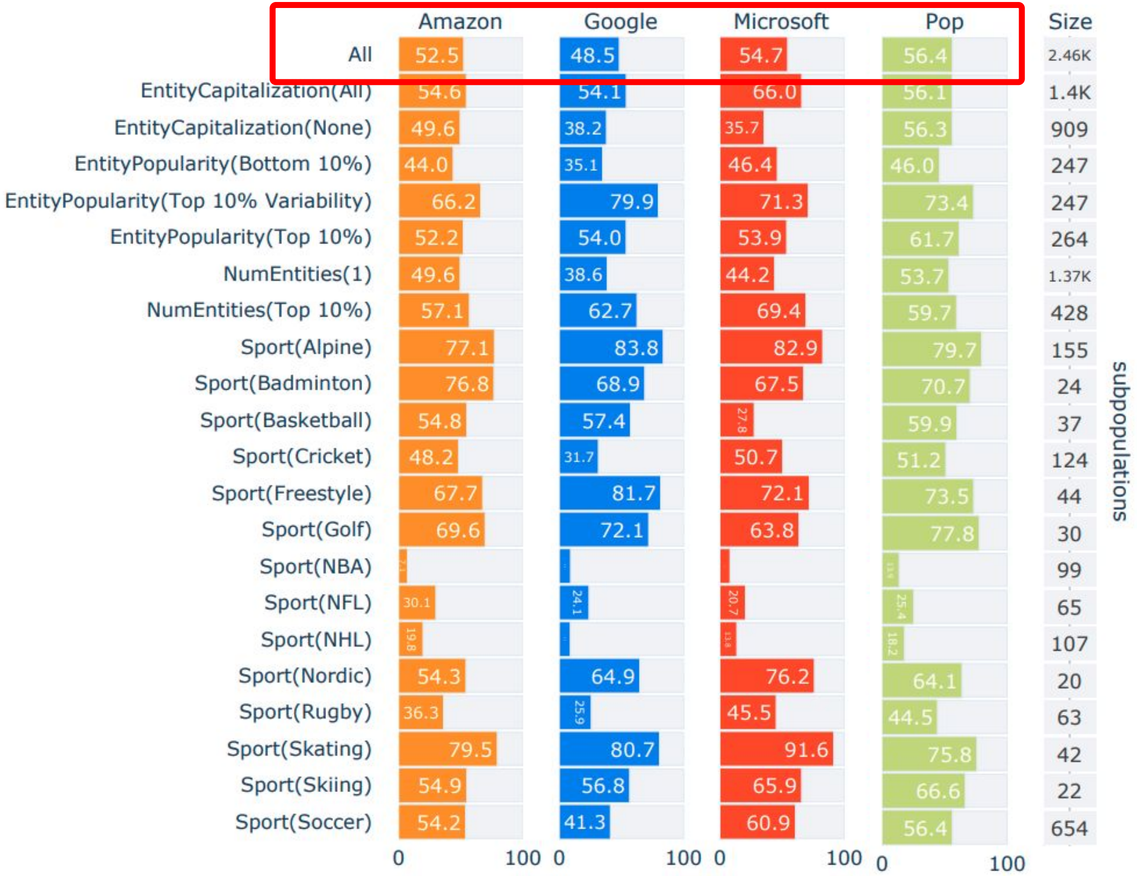
A correct NEL is required for the downstream system!

Experiments with Commercial APIs for Named Entity Linking



Robustness Report for NEL on AIDA-b dataset

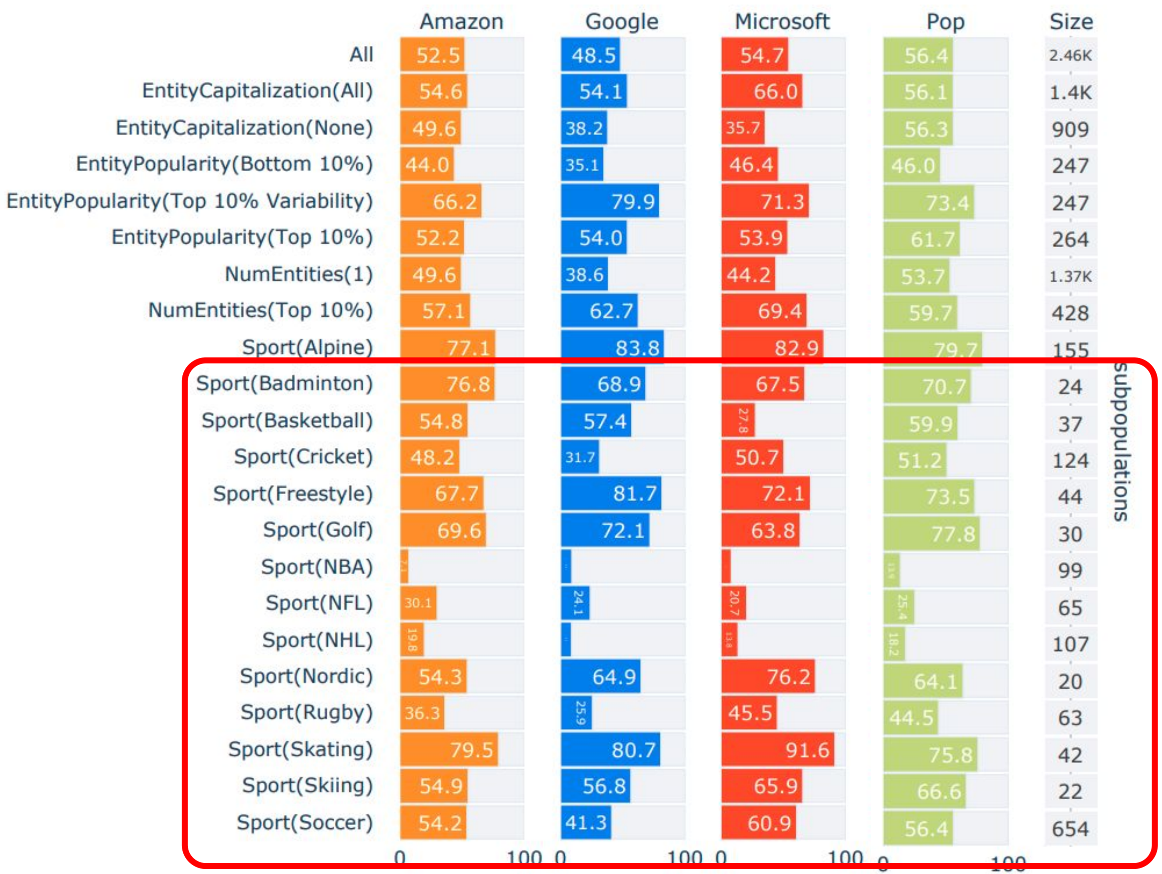
Experiments with Commercial APIs for Named Entity Linking



Popularity heuristic outperforms all commercial systems

Robustness Report for NEL on AIDA-b dataset

Experiments with Commercial APIs for Named Entity Linking



Robustness Report for NEL on AIDA-b dataset

Commercial APIs are not any more robust than popularity heuristic

Experiments with Commercial APIs for Named Entity Linking

Commercial systems are capitalization sensitive

Robustness Report for NEL on AIDA-b dataset

Experiments with Commercial APIs for Named Entity Linking

Type of Systematic Error!

Robustness Report for NEL on AIDA-b dataset

Systematic Error Analysis and Labeling (SEAL)

Evaluation is a creative process

Systematic errors are difficult to detect:

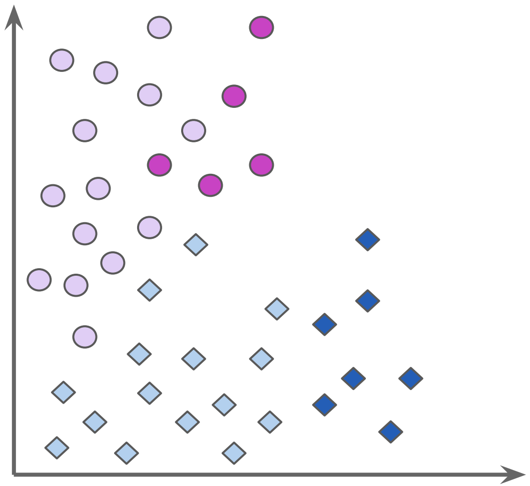
- High dimension of the learned representations
- Extracting and labeling semantics in the error group requires human-in-the-loop

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

(Rajani et al, EMNLP ‘22 demo)

Systematic Error Analysis and Labeling (SEAL)

Identify candidate groups with high systematic errors

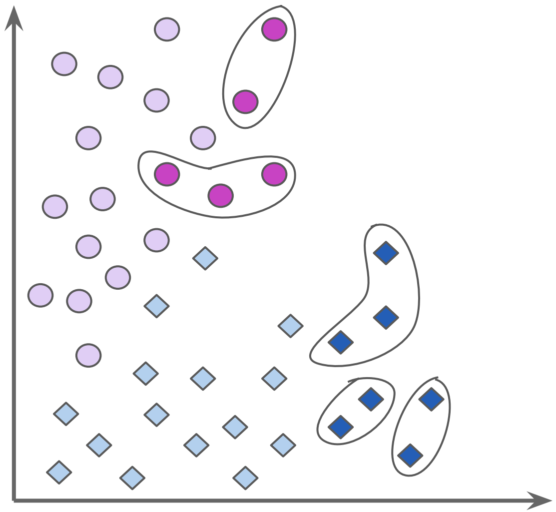


1. Embed

(Rajani et al, EMNLP ‘22 demo)

Systematic Error Analysis and Labeling (SEAL)

Identify candidate groups with high systematic errors

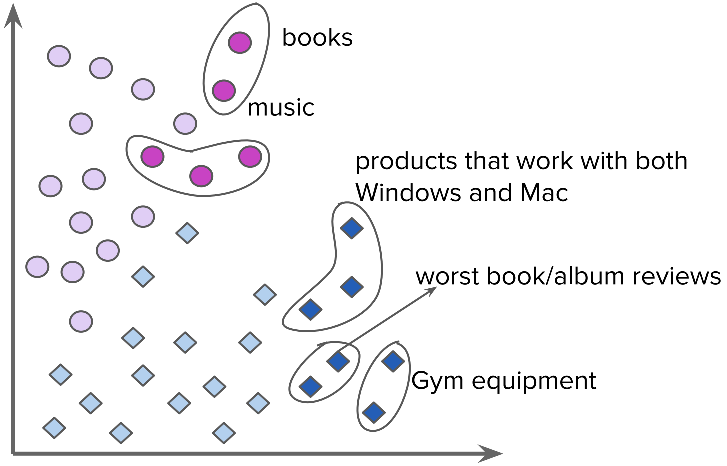


2. Cluster

(Rajani et al, EMNLP '22 demo)

Systematic Error Analysis and Labeling (SEAL)

Generate semantic labels using LLMs



3. Semantic Labeling

(Rajani et al, EMNLP '22 demo)

Systematic Error Analysis and Labeling (SEAL)

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

Dataset

yelp_polarity

Model

distilbert-base-uncas...

Loss Quantile

0.99

0.90 1.00

Cluster error group?

True False

clusters

11

1 60

data points to visualize

1000

1000 5000

Cluster #:

1 − +

Build prompt from data

Error Groups

How to read this table: +

	content	label	pred	loss	clust
19102	Food is always good.	0	1	8.99	4
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
14812	Edible is the best I can muster.	0	1	8.77	4
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
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
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
9622	Average Japanese food at amazing Japanese food prices.	0	1	8.65	4
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
12566	Its just ok	0	1	8.53	4
12336	Wild menu..huge portions .. Just ok.	0	1	8.52	4

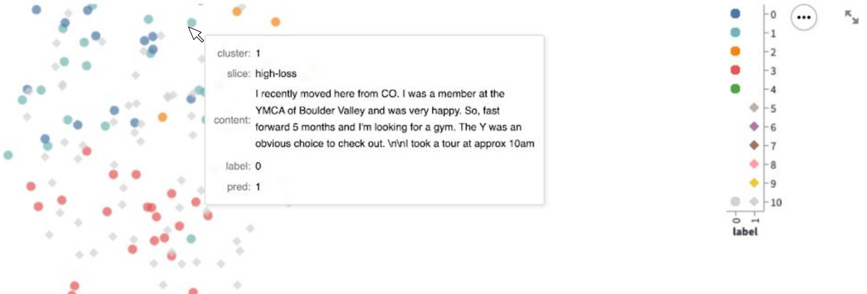
Word Distribution in Error Groups

How to read this table: +

	Token	Freq	Freq err	lrs
0	##now	0.0118	0.0332	2.28
1	delight	0.0027	0.0117	2.16
2	tips	0.0051	0.0166	2.14
3	stepping	0.0009	0.0068	2.01
4	points	0.0068	0.0186	2.00
5	combined	0.0029	0.0107	2.00
6	colored	0.0024	0.0098	2.00
7	gas	0.0068	0.0186	1.99
8	unlike	0.0056	0.0156	1.95
9	level	0.0136	0.0312	1.95

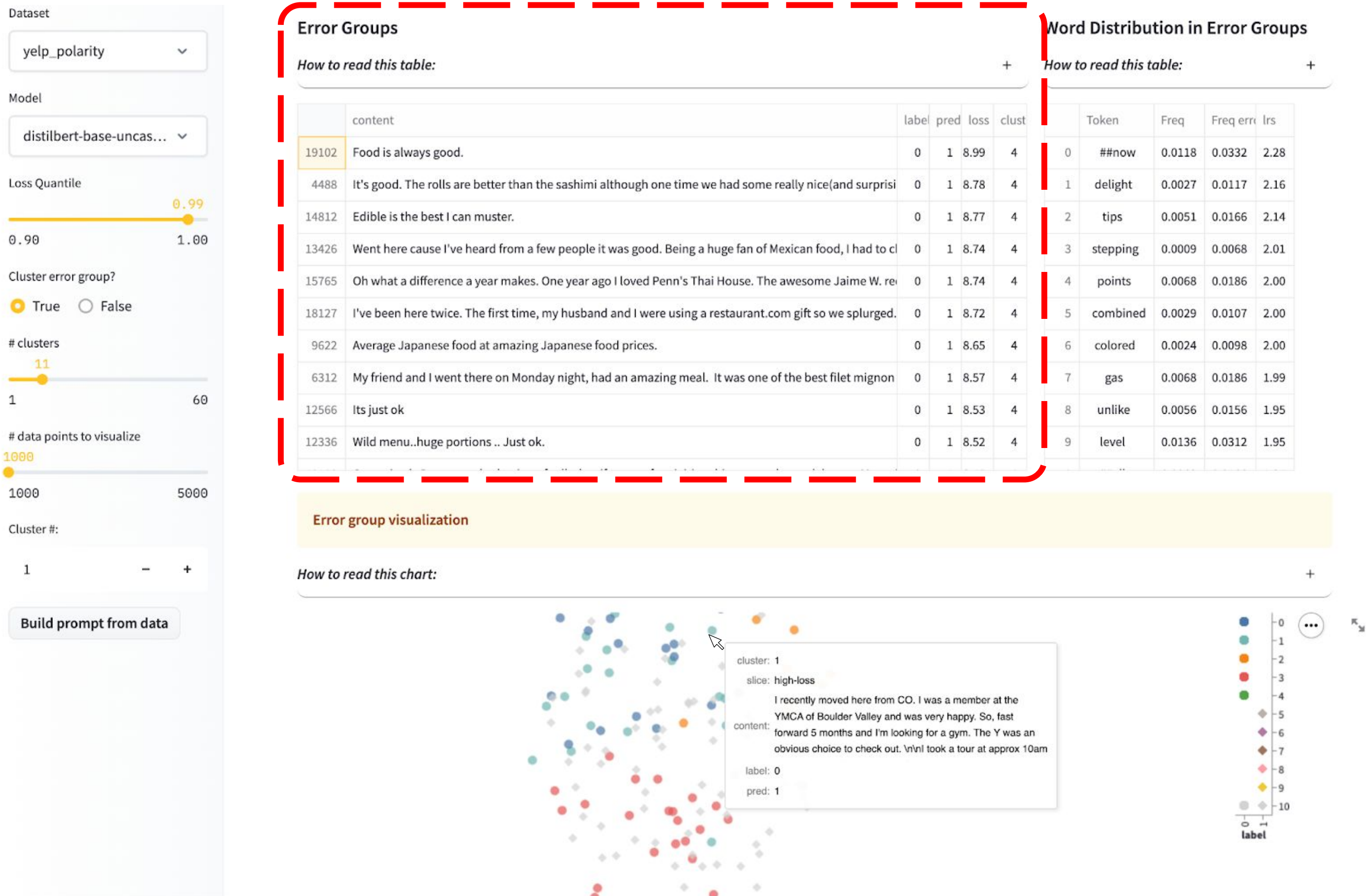
Error group visualization

How to read this chart: +



Systematic Error Analysis and Labeling (SEAL)

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

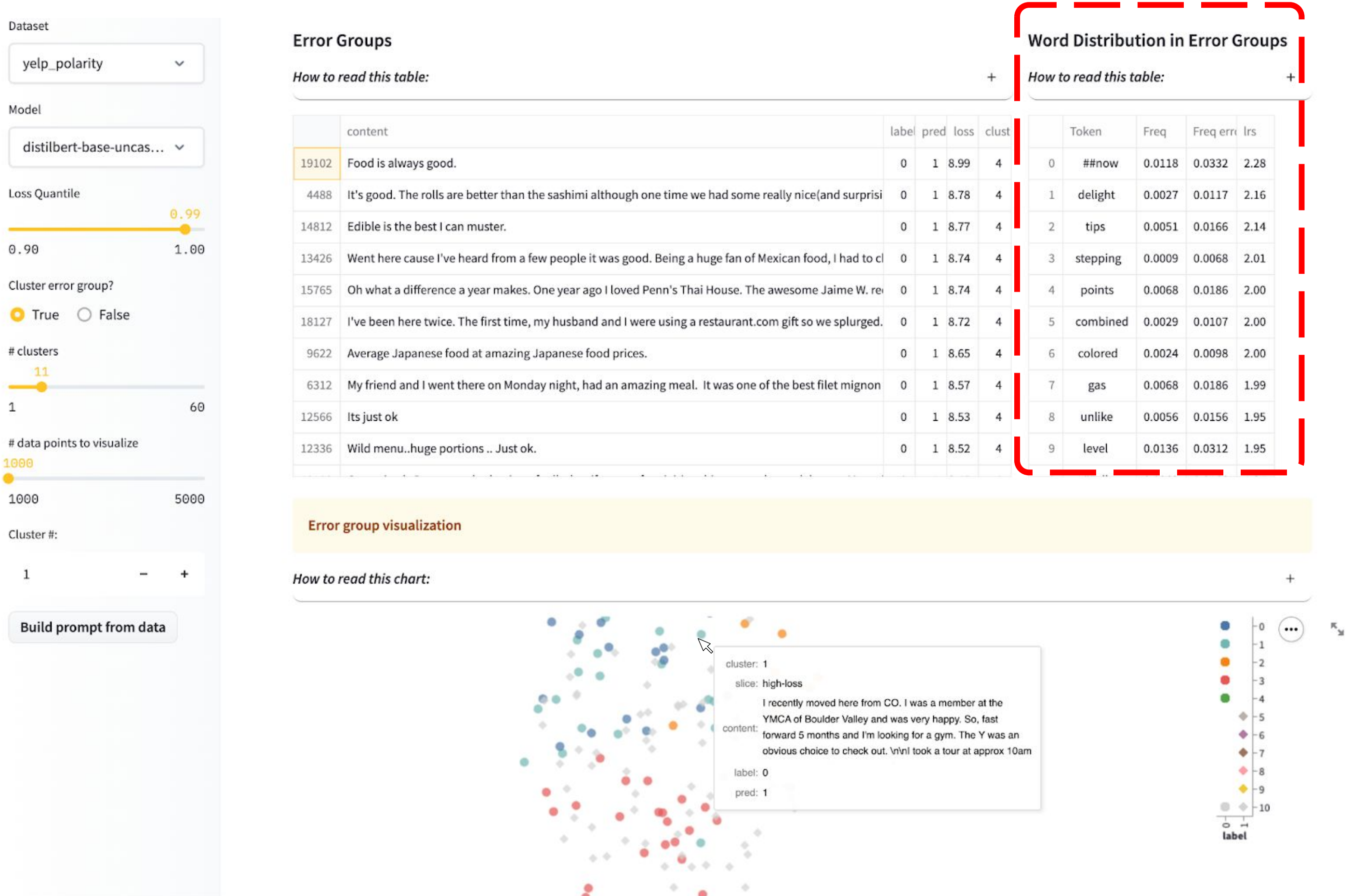


	content	label	pred	loss	clust
19102	Food is always good.	0	1	8.99	4
4488	It's good. The rolls are better than the sashimi although one time we had some really nice(and surprisi	0	1	8.78	4
14812	Edible is the best I can muster.	0	1	8.77	4
13426	Went here cause I've heard from a few people it was good. Being a huge fan of Mexican food, I had to cl	0	1	8.74	4
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
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
9622	Average Japanese food at amazing Japanese food prices.	0	1	8.65	4
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
12566	Its just ok	0	1	8.53	4
12336	Wild menu..huge portions .. Just ok.	0	1	8.52	4

	Token	Freq	Freq err	lrs
0	##now	0.0118	0.0332	2.28
1	delight	0.0027	0.0117	2.16
2	tips	0.0051	0.0166	2.14
3	stepping	0.0009	0.0068	2.01
4	points	0.0068	0.0186	2.00
5	combined	0.0029	0.0107	2.00
6	colored	0.0024	0.0098	2.00
7	gas	0.0068	0.0186	1.99
8	unlike	0.0056	0.0156	1.95
9	level	0.0136	0.0312	1.95

Systematic Error Analysis and Labeling (SEAL)

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



Systematic Error Analysis and Labeling (SEAL)

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

Dataset

yelp_polarity

Model

distilbert-base-uncas...

Loss Quantile

0.99

0.90 1.00

Cluster error group?

True False

clusters

11

1 60

data points to visualize

1000

1000 5000

Cluster #:

1 − +

Build prompt from data

Error Groups

How to read this table: +

	content	label	pred	loss	clust
19102	Food is always good.	0	1	8.99	4
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
14812	Edible is the best I can muster.	0	1	8.77	4
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
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
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
9622	Average Japanese food at amazing Japanese food prices.	0	1	8.65	4
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
12566	Its just ok	0	1	8.53	4
12336	Wild menu..huge portions .. Just ok.	0	1	8.52	4

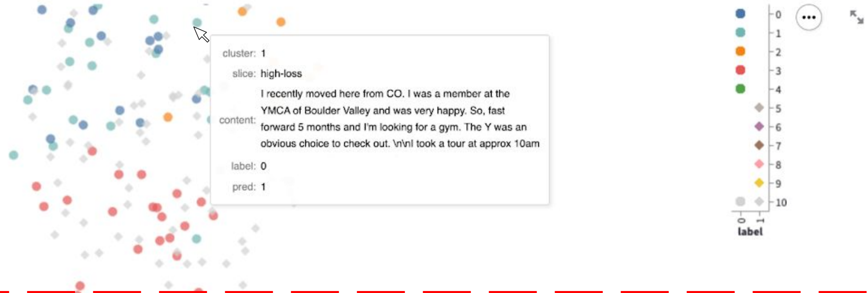
Word Distribution in Error Groups

How to read this table: +

	Token	Freq	Freq err	lrs
0	##now	0.0118	0.0332	2.28
1	delight	0.0027	0.0117	2.16
2	tips	0.0051	0.0166	2.14
3	stepping	0.0009	0.0068	2.01
4	points	0.0068	0.0186	2.00
5	combined	0.0029	0.0107	2.00
6	colored	0.0024	0.0098	2.00
7	gas	0.0068	0.0186	1.99
8	unlike	0.0056	0.0156	1.95
9	level	0.0136	0.0312	1.95

Error group visualization

How to read this chart: +



Systematic Error Analysis and Labeling (SEAL)

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

Dataset

yelp_polarity

Model

distilbert-base-uncas...

Loss Quantile

0.99

0.90 1.00

Cluster error group?

True False

clusters

11

1 60

data points to visualize

1000

1000 5000

Cluster #:

1 − +

Build prompt from data

Error Groups

How to read this table: +

	content	label	pred	loss	clust
19102	Food is always good.	0	1	8.99	4
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
14812	Edible is the best I can muster.	0	1	8.77	4
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
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
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
9622	Average Japanese food at amazing Japanese food prices.	0	1	8.65	4
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
12566	Its just ok	0	1	8.53	4
12336	Wild menu..huge portions .. Just ok.	0	1	8.52	4

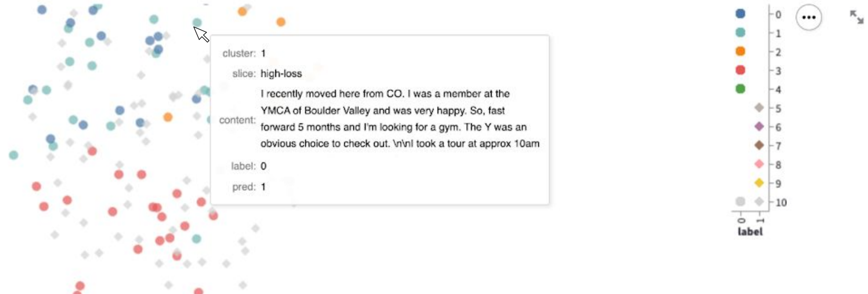
Word Distribution in Error Groups

How to read this table: +

	Token	Freq	Freq err	lrs
0	##now	0.0118	0.0332	2.28
1	delight	0.0027	0.0117	2.16
2	tips	0.0051	0.0166	2.14
3	stepping	0.0009	0.0068	2.01
4	points	0.0068	0.0186	2.00
5	combined	0.0029	0.0107	2.00
6	colored	0.0024	0.0098	2.00
7	gas	0.0068	0.0186	1.99
8	unlike	0.0056	0.0156	1.95
9	level	0.0136	0.0312	1.95

Error group visualization

How to read this chart: +



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%

Takeaways

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

Takeaways

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

Takeaways

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

Takeaways

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