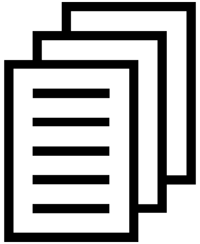
Takeaways from a Systematic Study of 75,000 ML Models

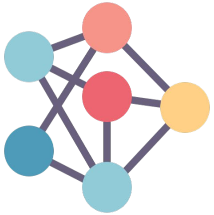
Nazneen Rajani | Robustness Research Lead @ Hugging Face | nazneen@hf.co | @nazneenrajani

Ecosystem as part of the ML workflow

Collect data → Train model → Evaluate → Deploy

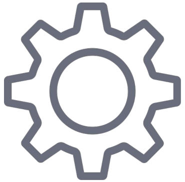


>10K datasets



>75K models

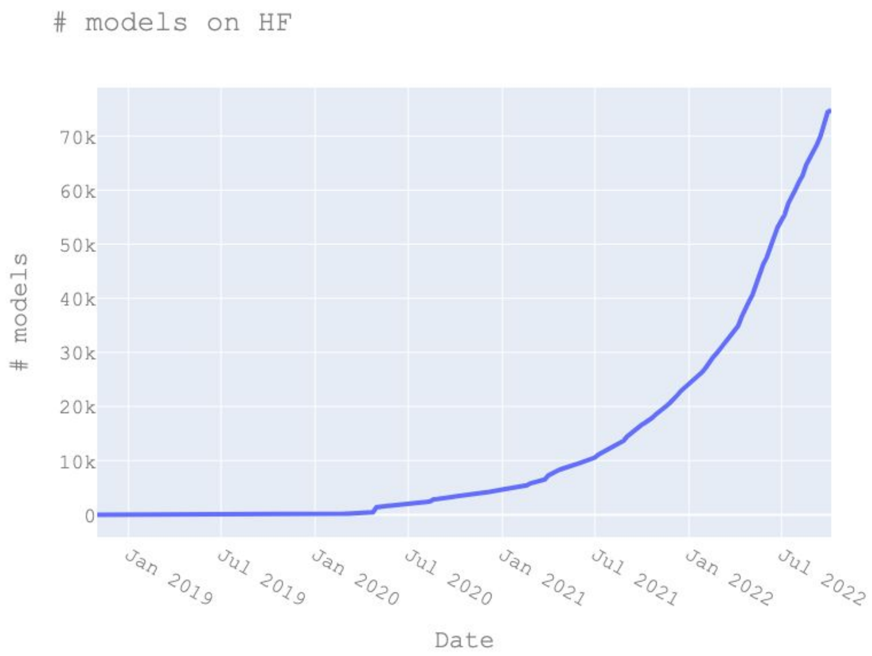
>70 metrics and measurements



Spaces/ Gradio for demos

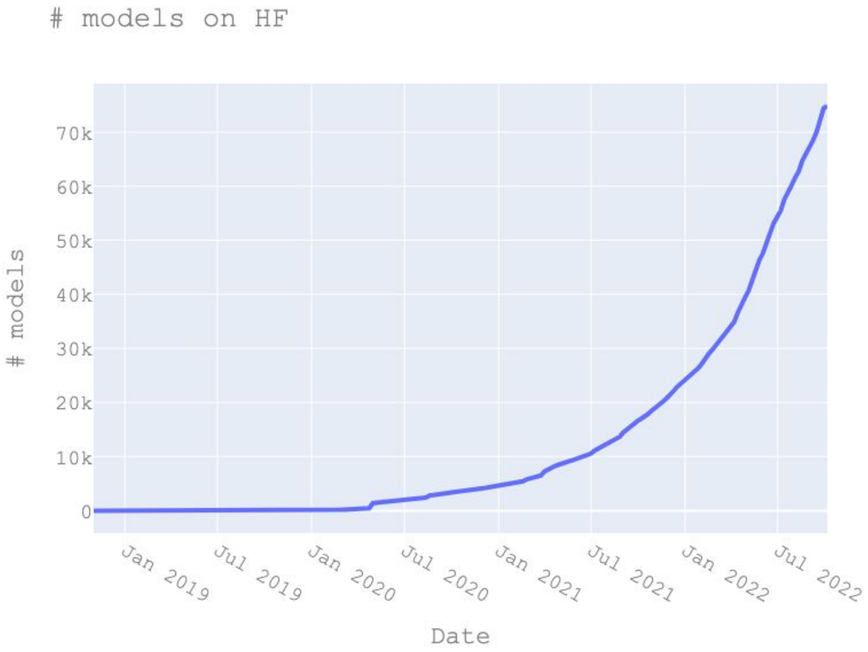
ML Modeling Landscape

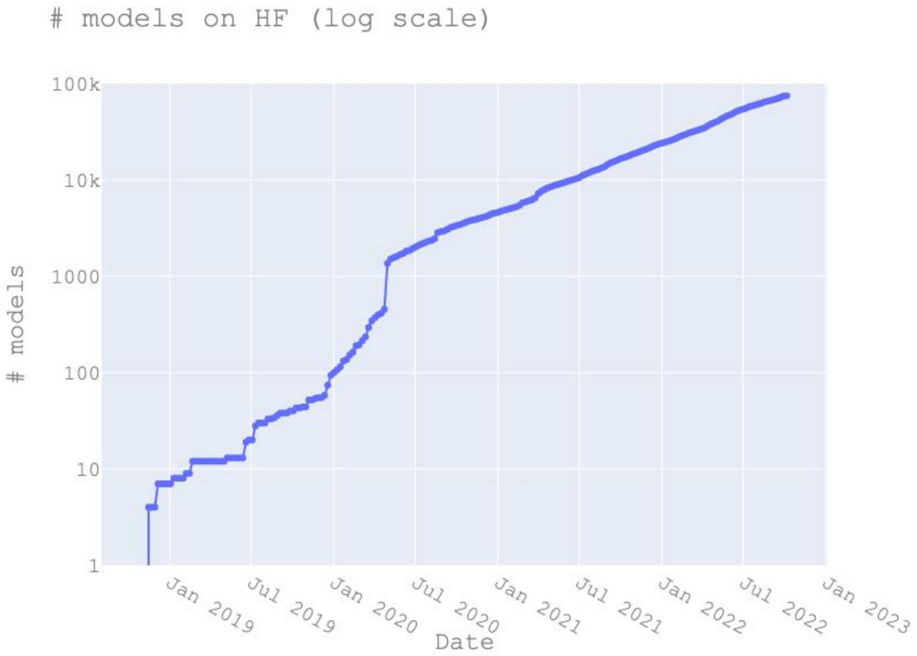
There is an exponential growth of ML models.



ML Modeling Landscape

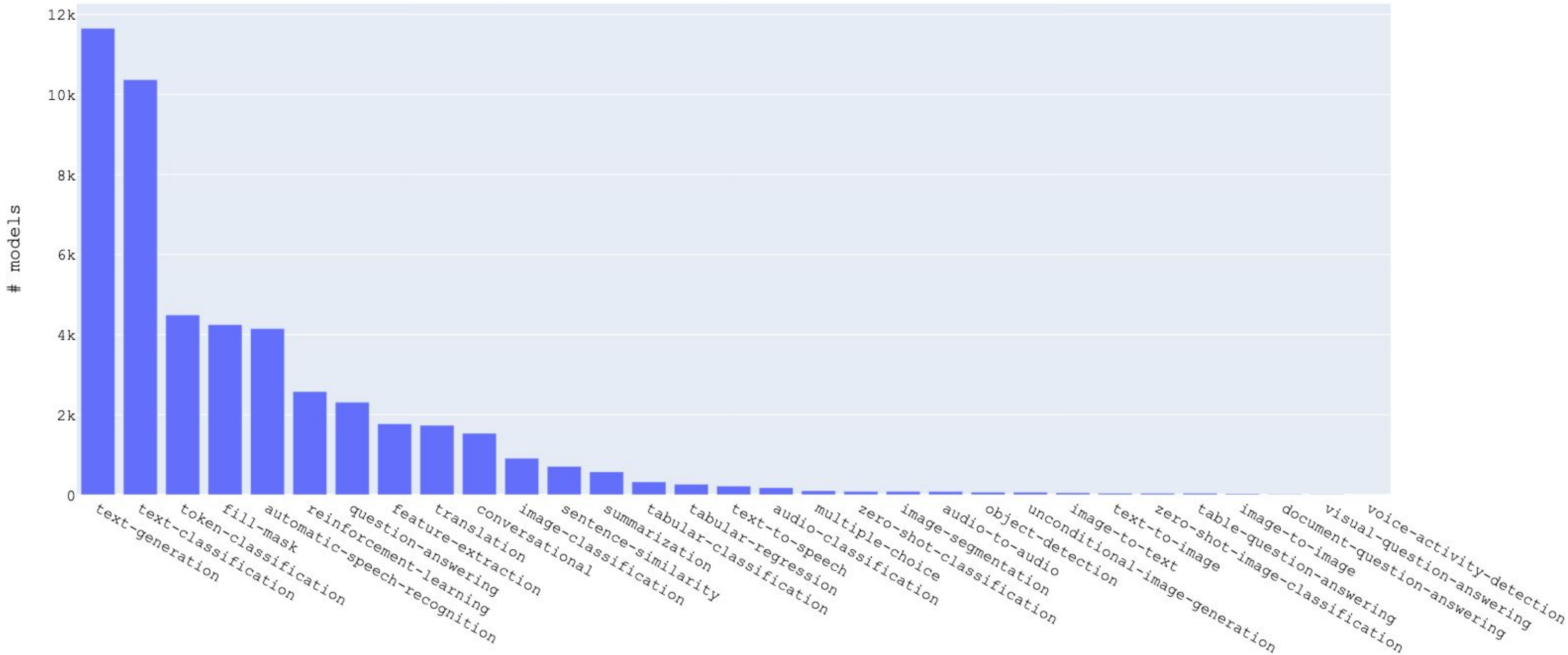
There is an exponential growth of ML models.





ML Modeling Landscape

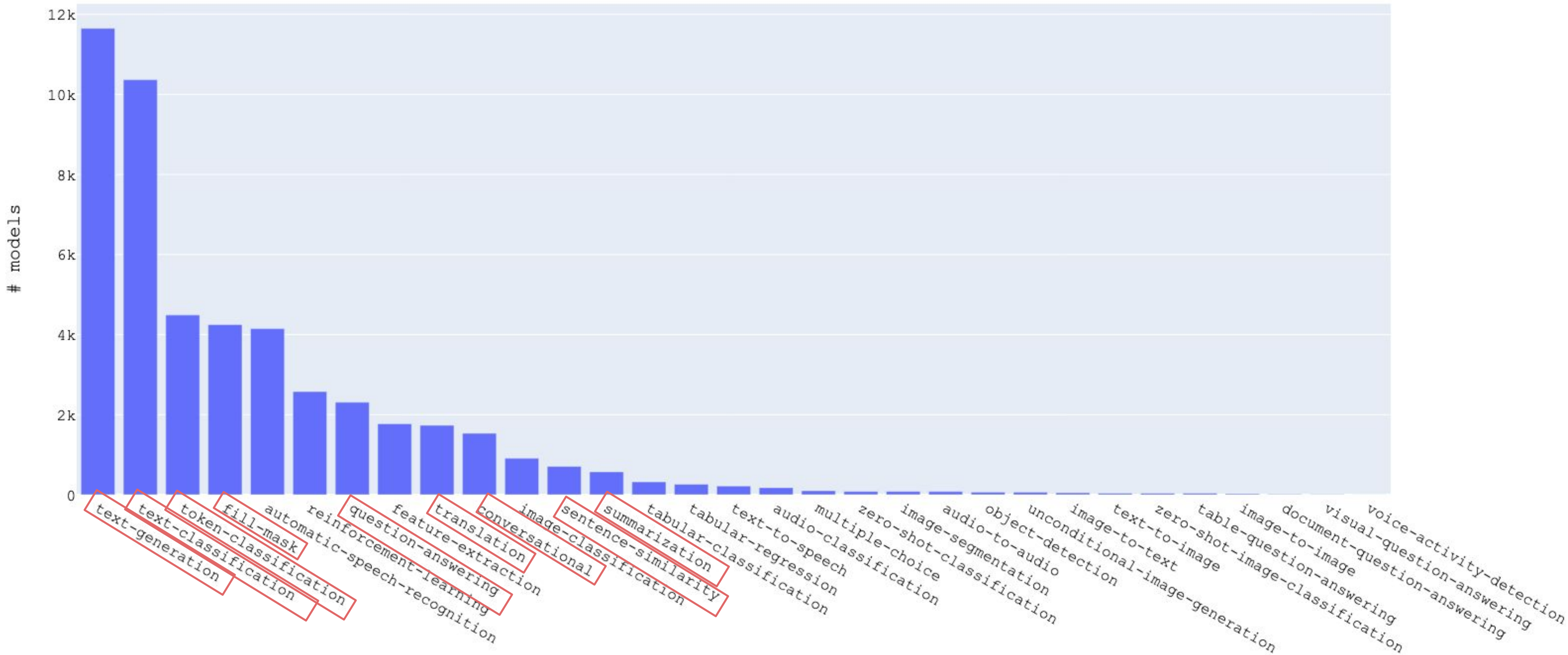
Distribution by task categories



NLP Modeling Landscape

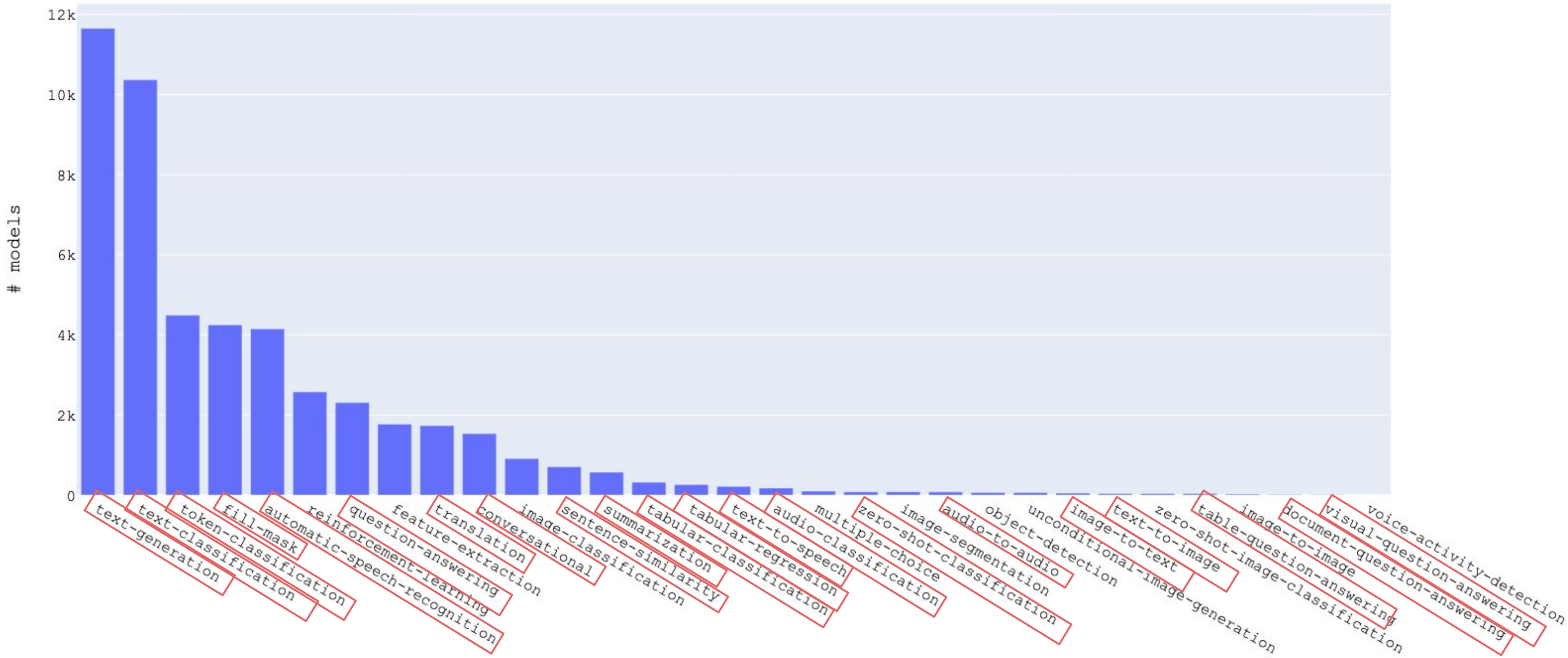
Approx 40% of the task categories are NLP

Covering 78% of the models

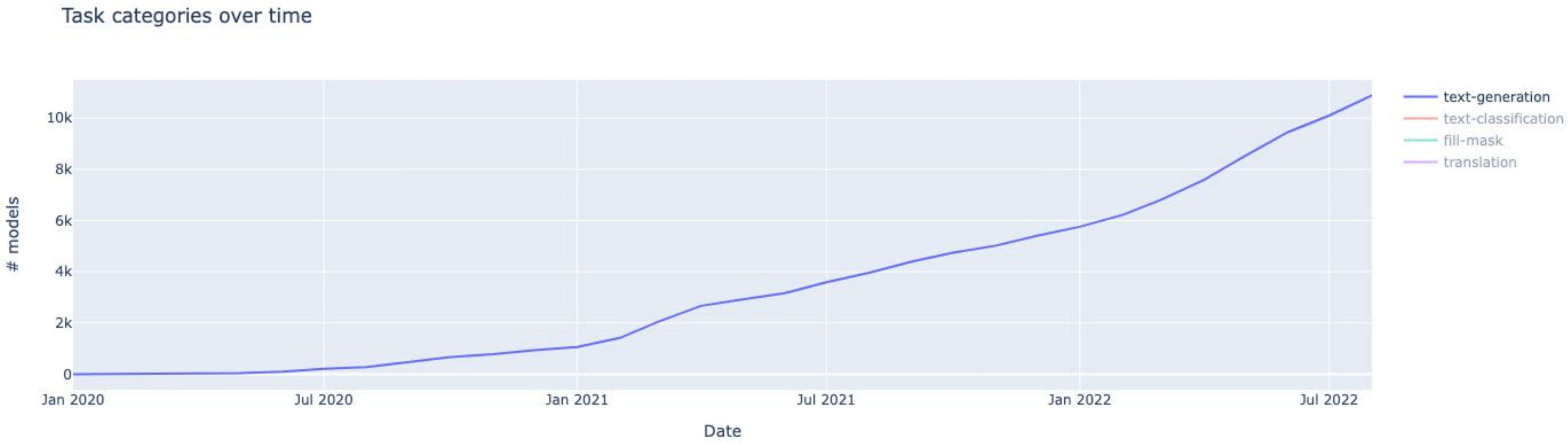


NLP Modeling Landscape

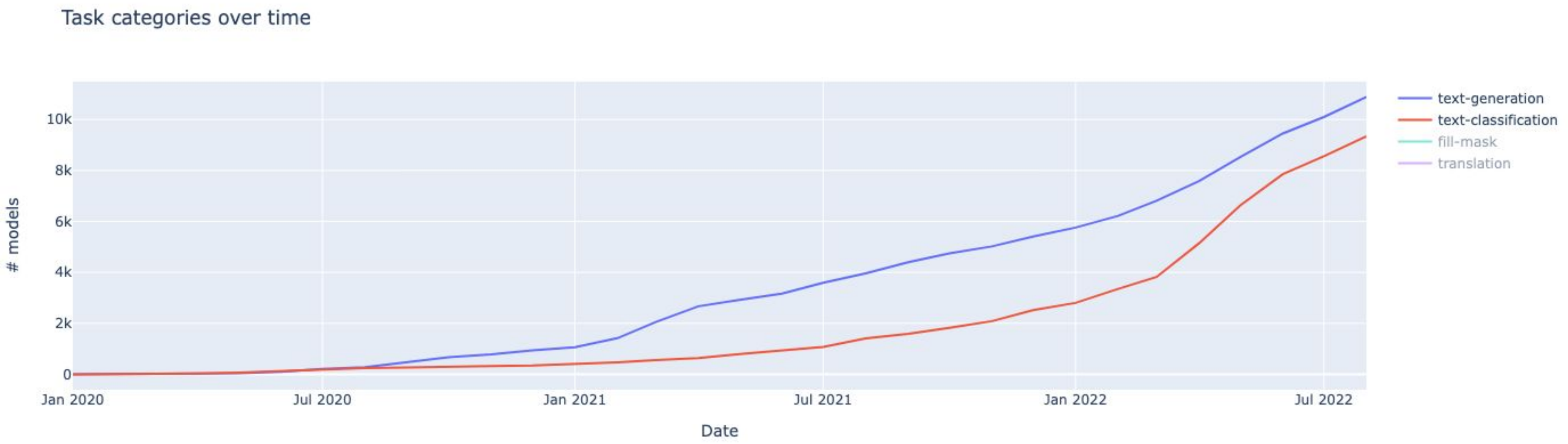
Coverage is 90% of models if we include speech and multimodal categories



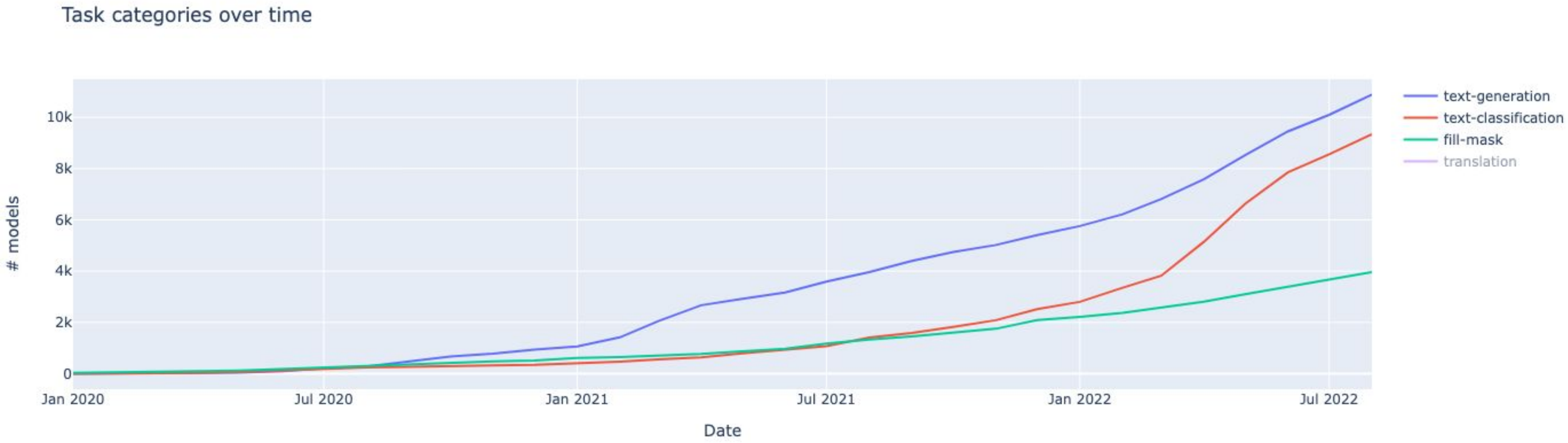
NLP Task Categories



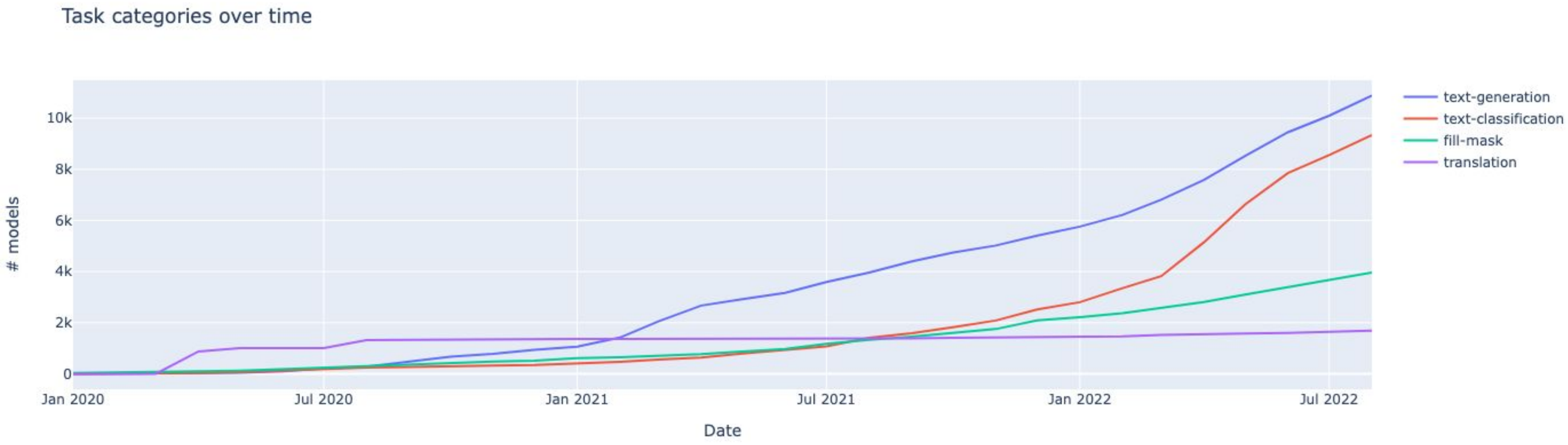
NLP Task Categories



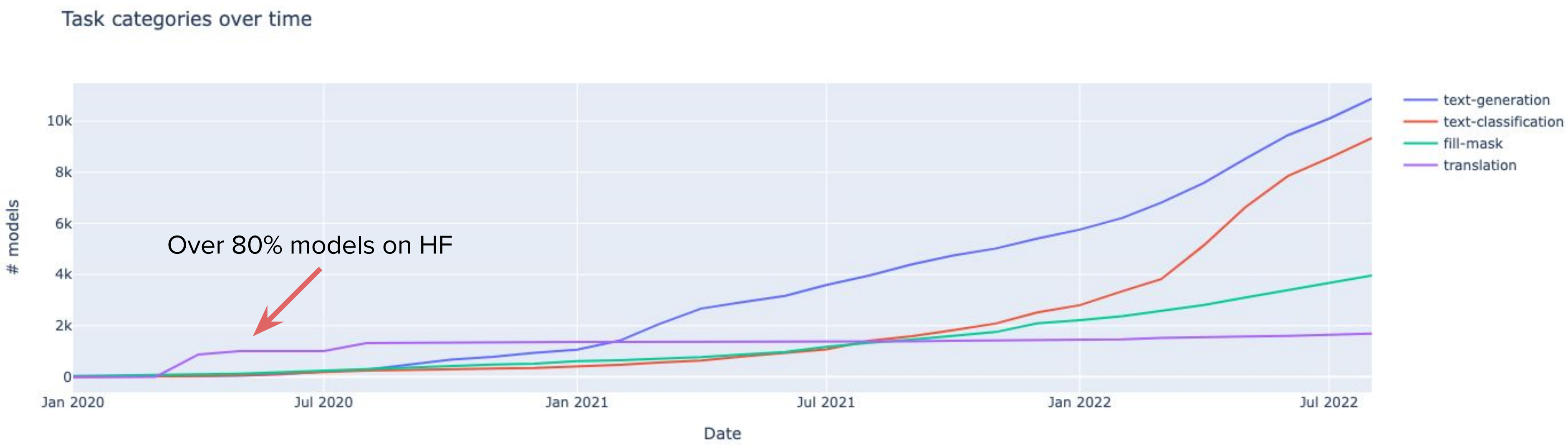
NLP Task Categories



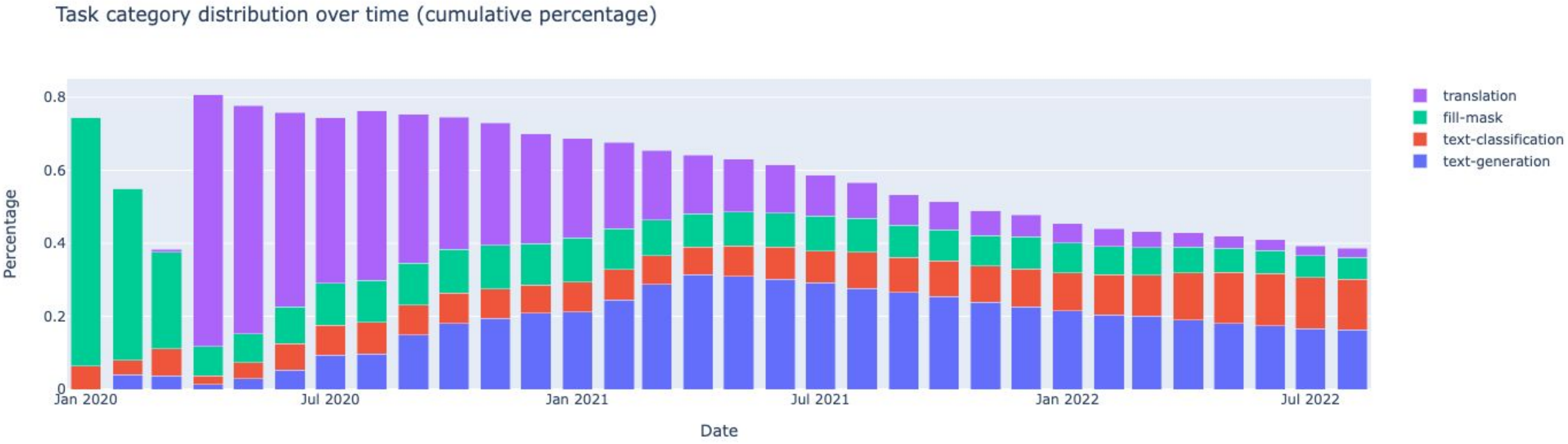
NLP Task Categories



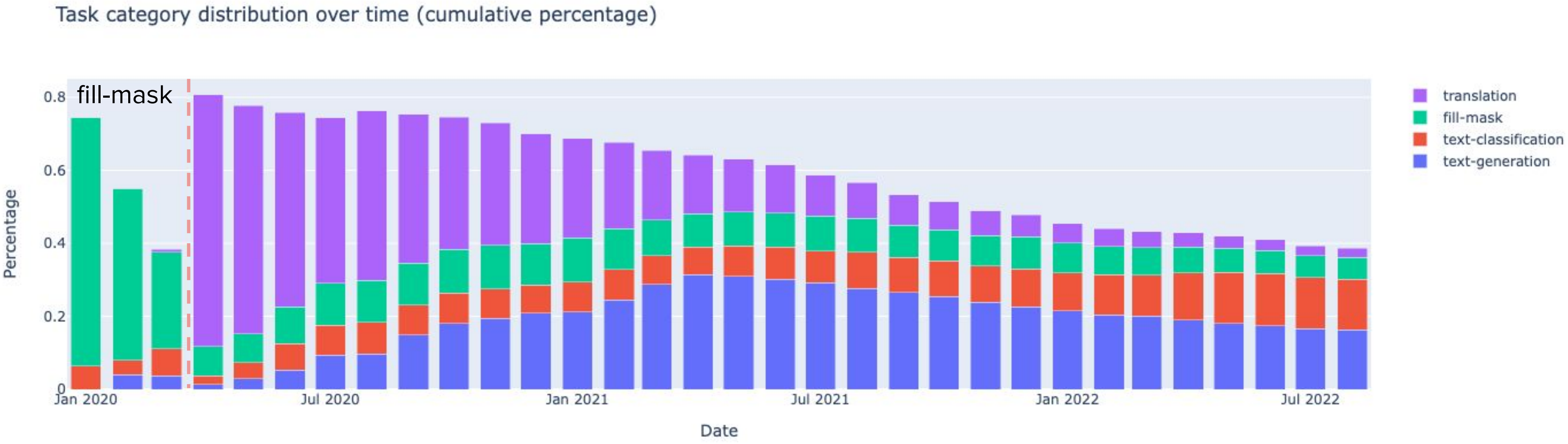
NLP Task Categories



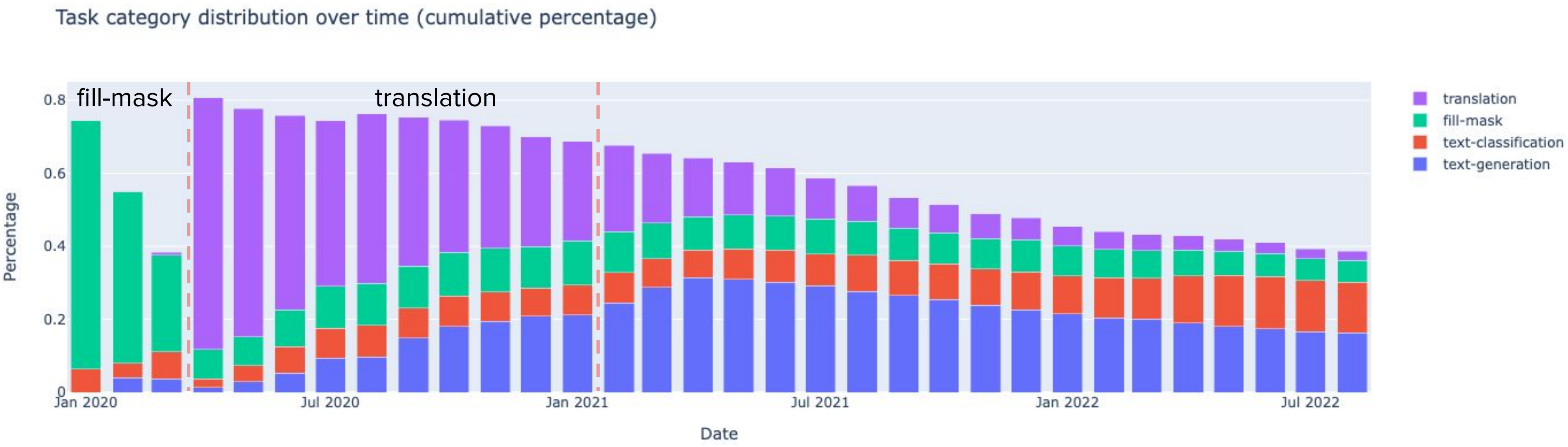
NLP Task Categories



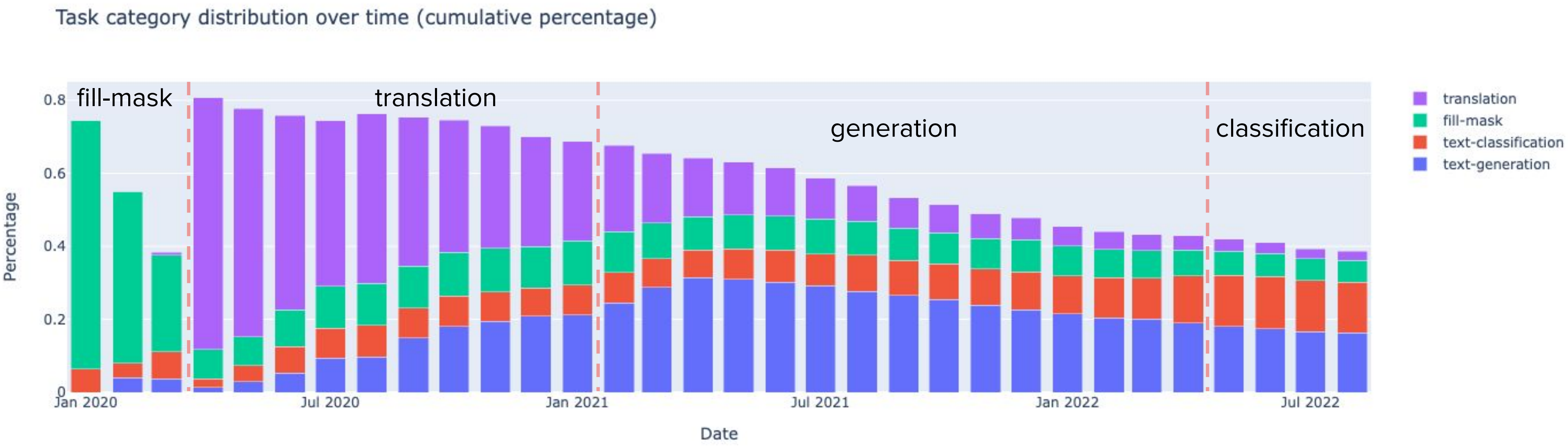
NLP Task Categories



NLP Task Categories

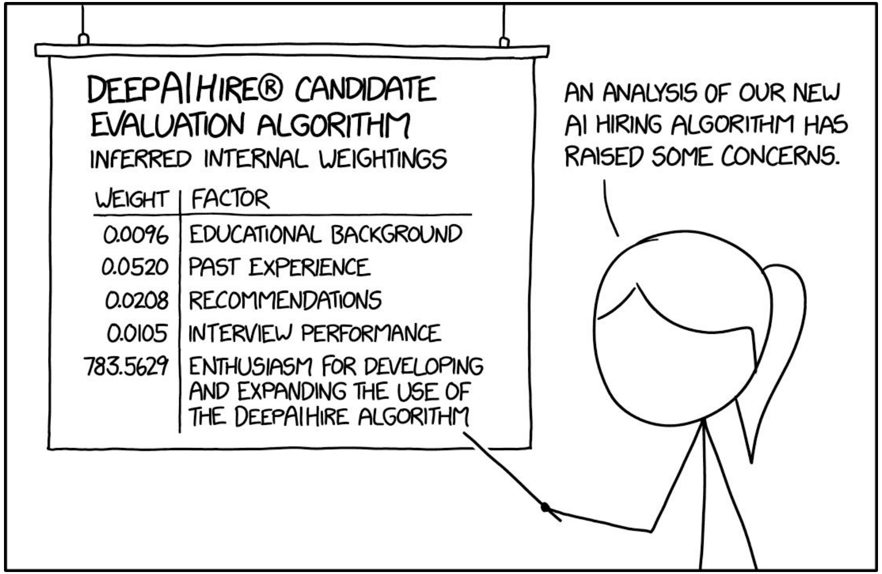


NLP Task Categories



NLP Modeling Landscape

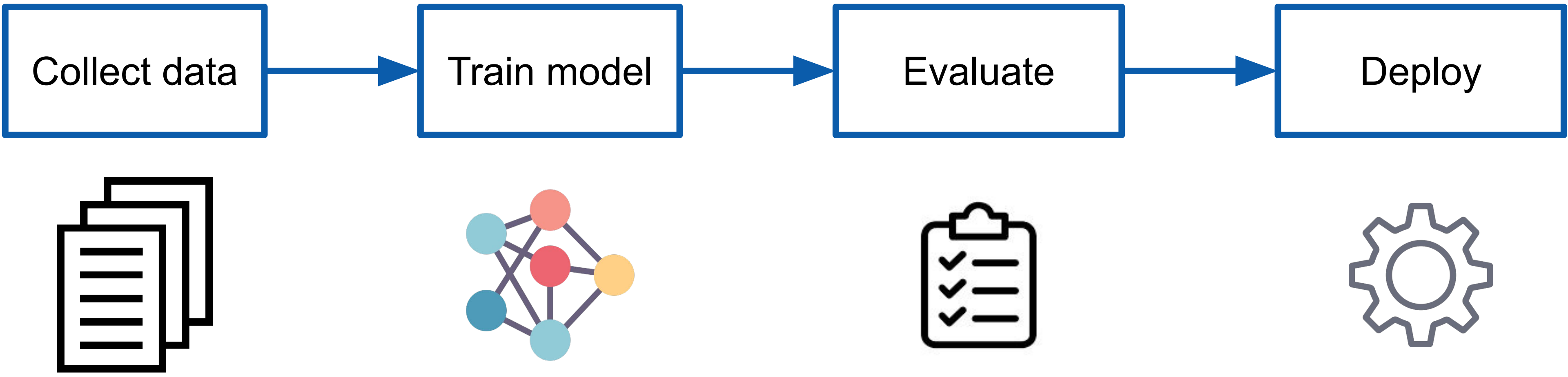
How can we learn more about the models?



Model documentation!

Image credit: xkcd

Model Documentation



✔ Dataset

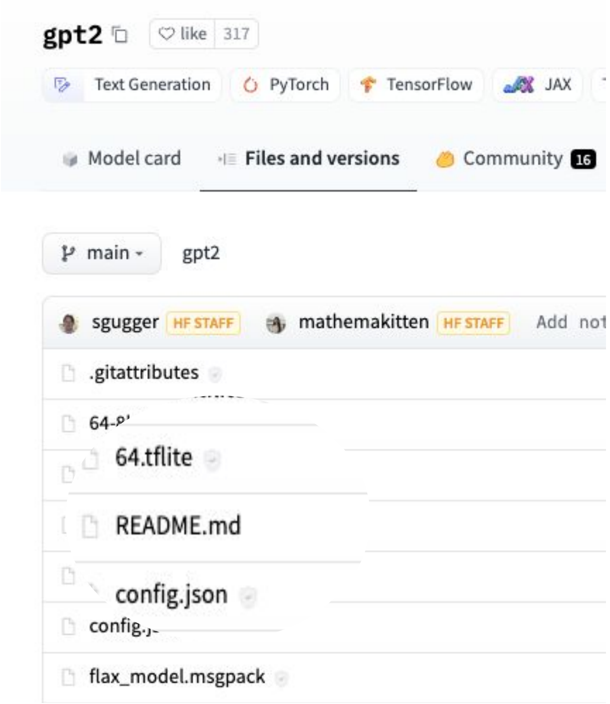
✔ Training
✔ Environmental impact

✔ Evaluation
✔ Limitations

✔ How to use
✔ Intended uses

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
```

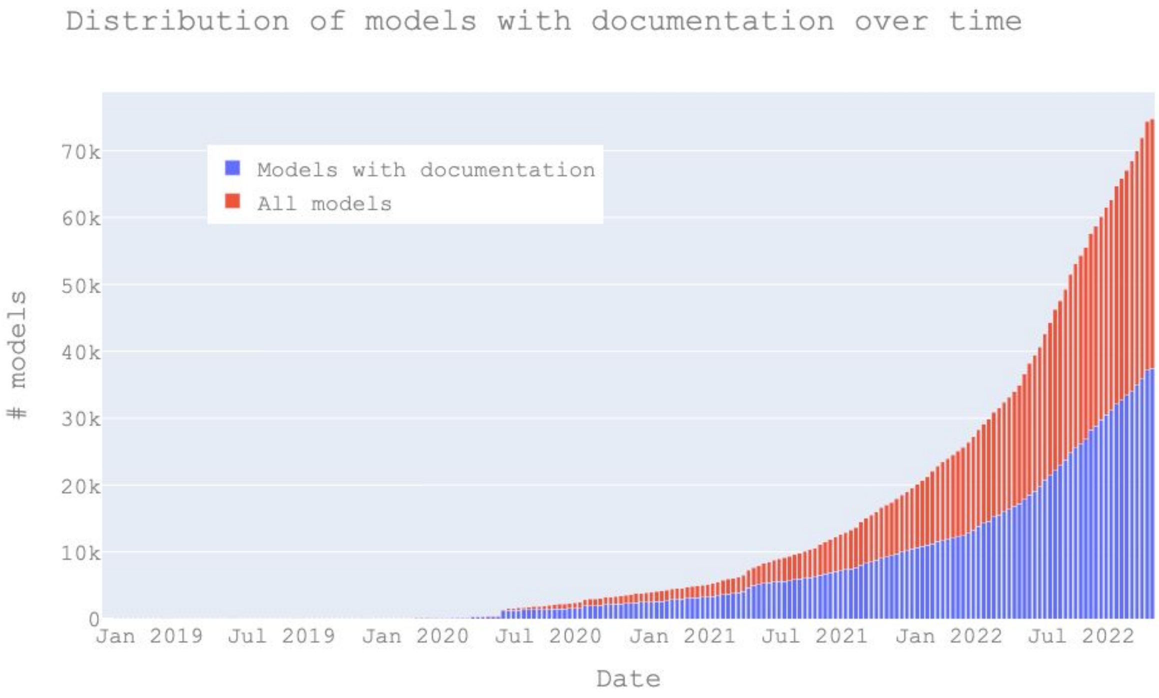
Model Documentation for GPT2

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

What do developers document about models?

Distribution of sections in model cards

What do developers document about models?

Distribution of sections in model cards

What do developers document about models?

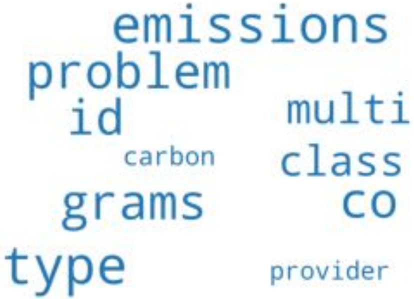
Distribution of length of sections in model cards

What do developers document about models?

Distribution of length of sections in model cards

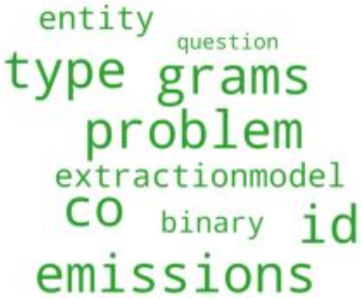
Impact section

Topic 0



Topic 1

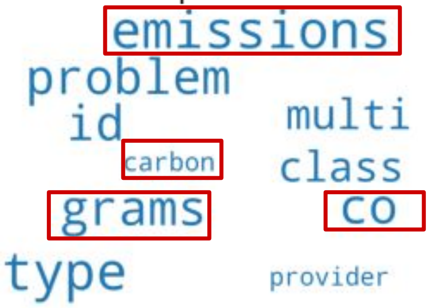
Topic 2



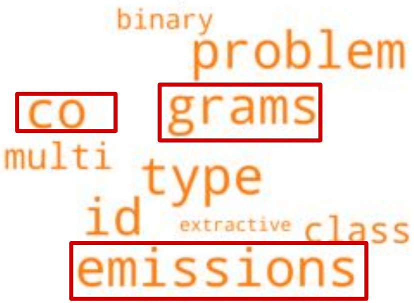
Topic 3

Impact section

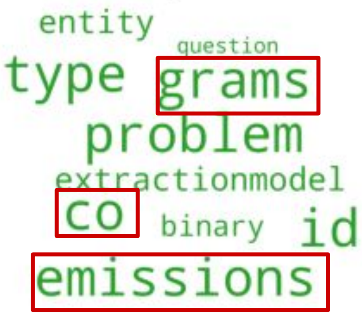
Topic 0



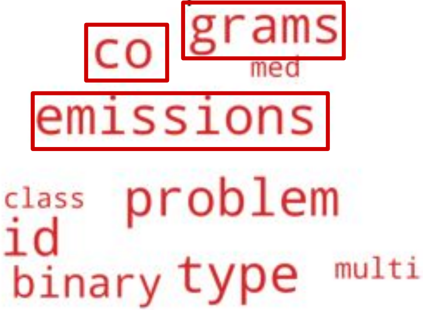
Topic 1



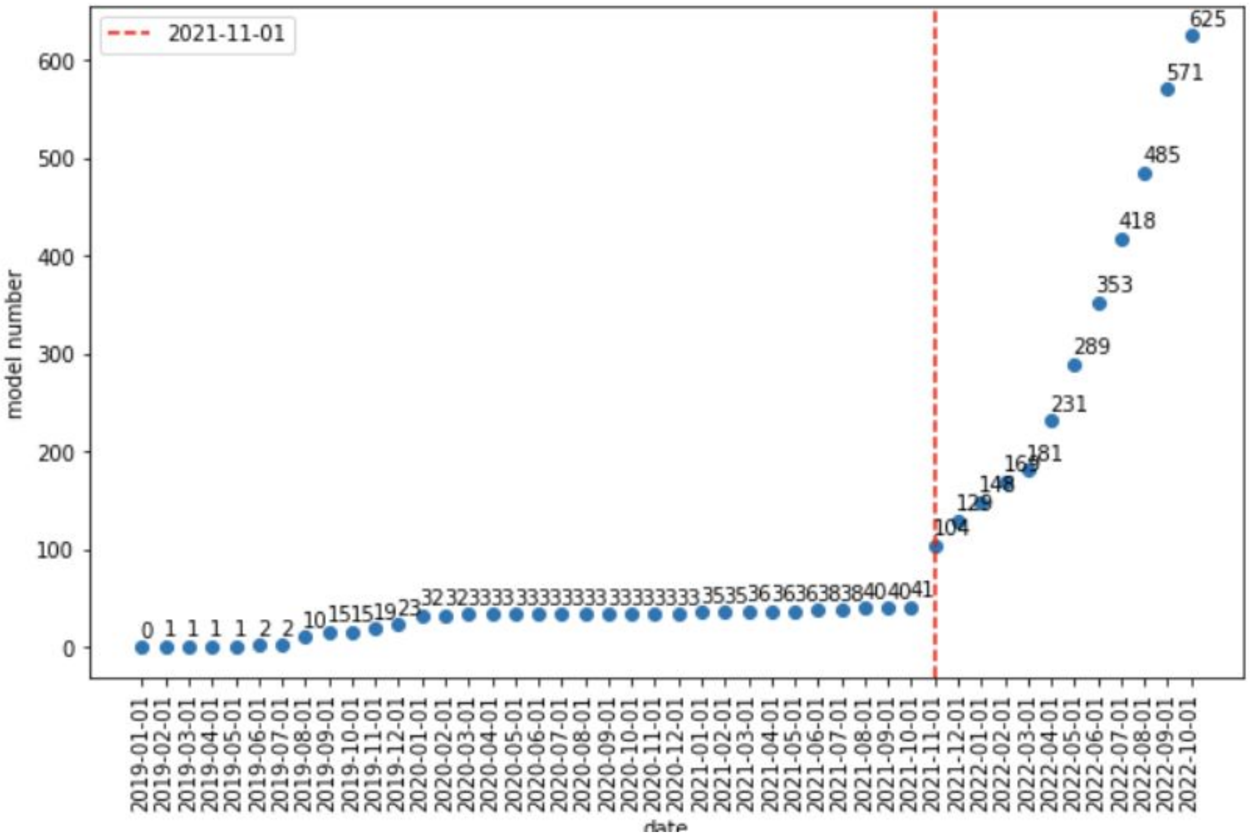
Topic 2



Topic 3



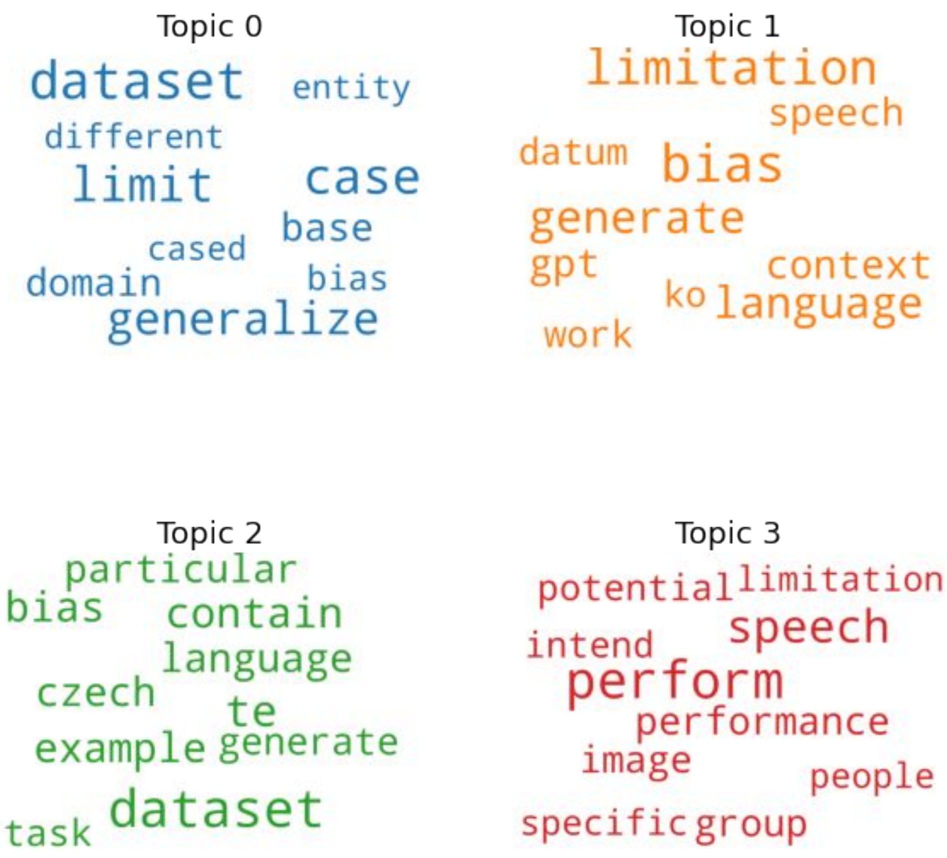
Impact section



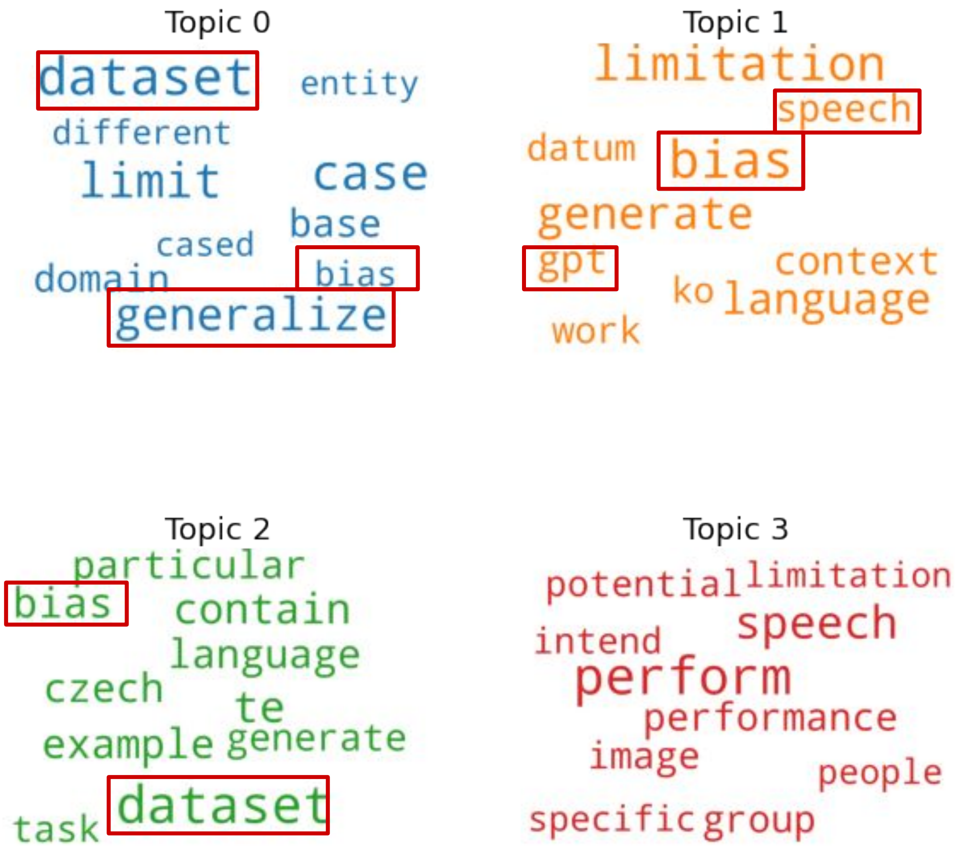
What do developers document about models?

Distribution of length of sections in model cards

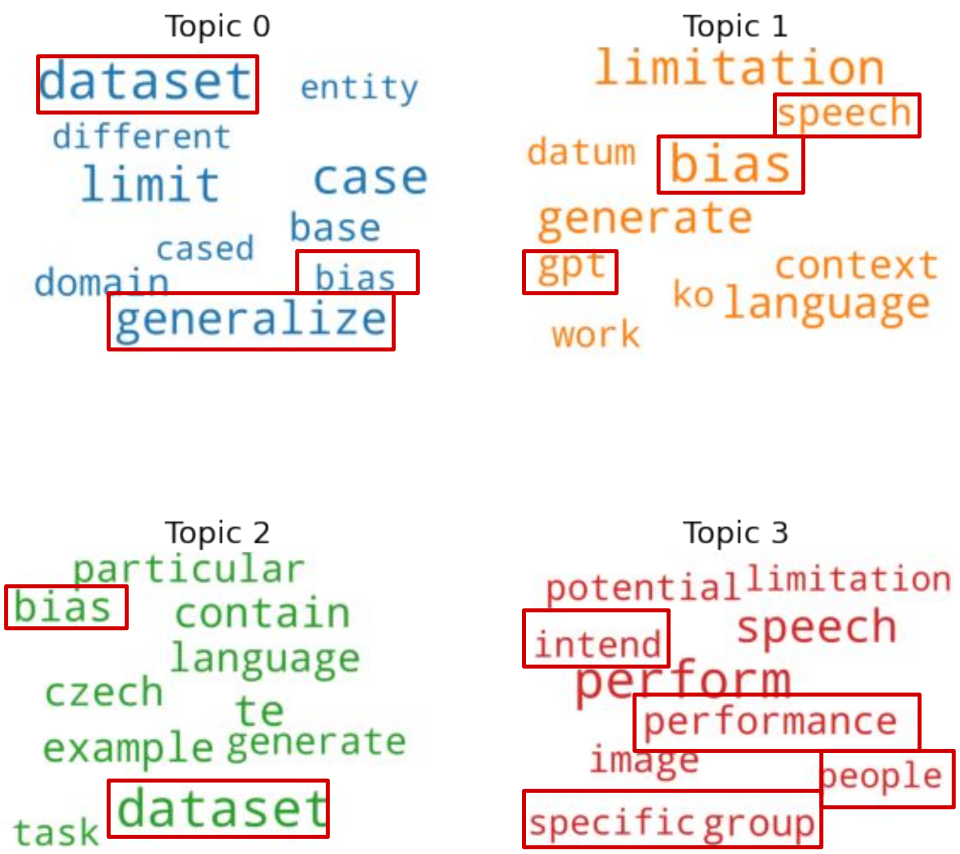
Limitation section



Limitation section



Limitation section



How has model documentation evolved?

Observation: Model documentation has evolved

How has model documentation evolved?

Observation: Model documentation has evolved

Goal: Use word embeddings to capture change in content

How has model documentation evolved?

Observation: Model documentation has evolved

Goal: Use word embeddings to capture change in content

Steps:

1. Train a word2vec model for each year

How has model documentation evolved?

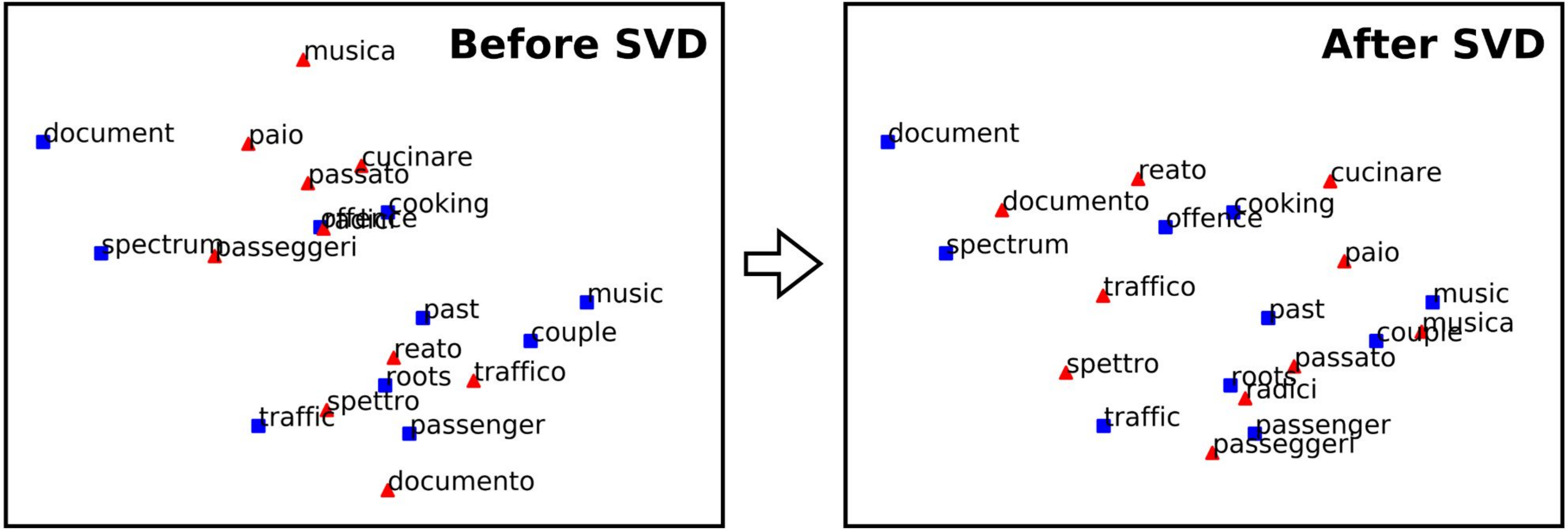
Observation: Model documentation has evolved

Goal: Use word embeddings to capture change in content

Steps:

1. Train a word2vec model for each year
2. Align the vocabulary (so same word can be compared across years) (Hamilton et al., 2016)

How has model documentation evolved?



Smith et al, 2017

How has model documentation evolved?

Observation: Model documentation has evolved

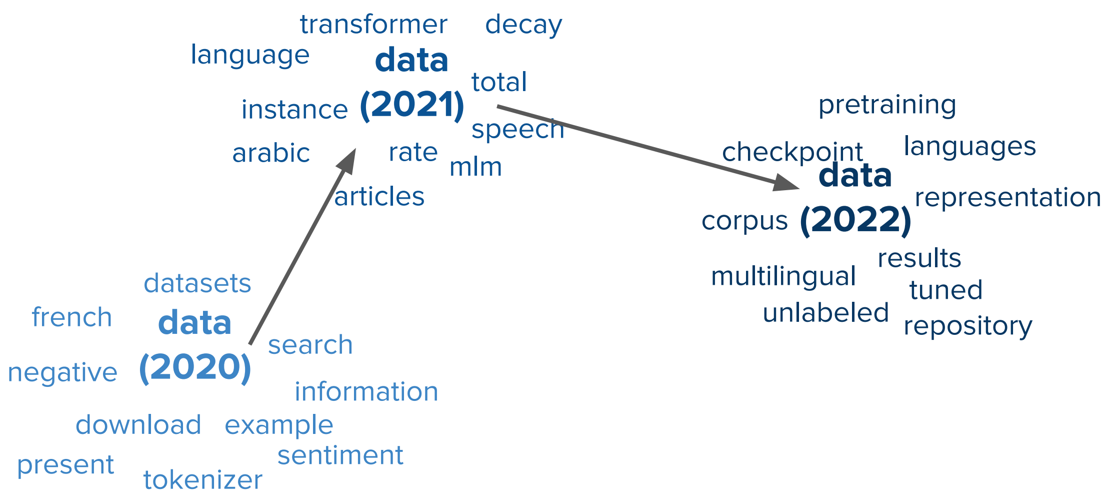
Goal: Use word embeddings to capture change in content

Steps:

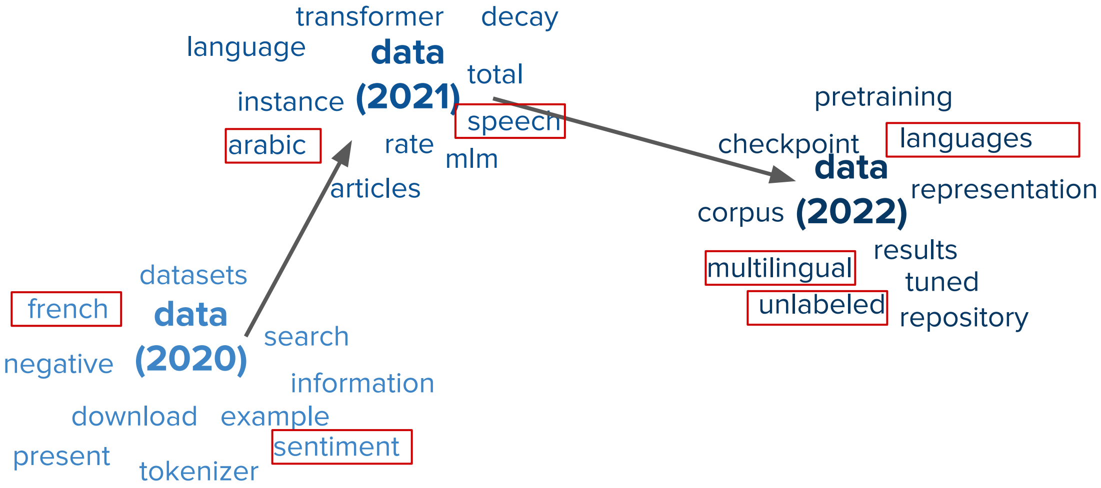
1. Train a word2vec model for each year
2. Align the embeddings (so same word can be compared across years) (Hamilton et al., 2016)
3. Compare nearest neighbors or pairwise similarity of vectors

$$s^{(t)}(w_i, w_j) = cos\text{-}sim(\mathbf{w}_i^{(t)}, \mathbf{w}_j^{(t)}), \text{for t} \in \{2020, 2021, 2022\}$$

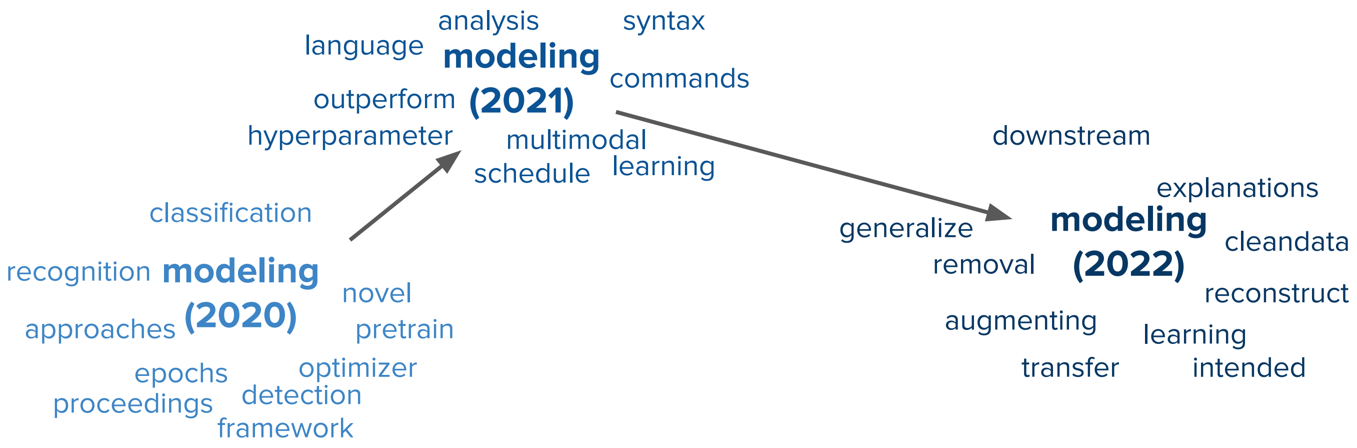
How has model documentation evolved?



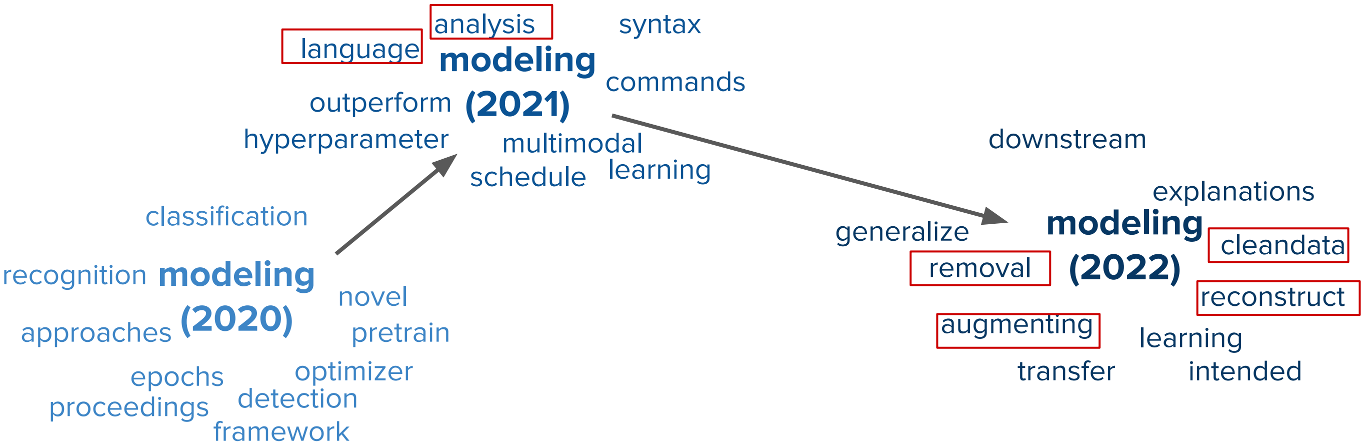
How has model documentation evolved?



How has model documentation evolved?



How has model documentation evolved?



Pairwise similarity between word vectors

Similarity to the word *“evaluate”*

Word	2020	2021	2022	Spearman correlation (ρ)
fairness	-0.1	0.0	0.99	0.9
biased	-0.01	0.99	0.99	0.86
humans	0.0	0.99	0.99	0.86
statistically	0.99	0.99	-0.02	-0.86

Pairwise similarity between word vectors

Similarity to the word *“evaluate”*

Word	2020	2021	2022	Spearman correlation (ρ)
fairness	-0.1	0.0	0.99	0.9
biased	-0.01	0.99	0.99	0.86
humans	0.0	0.99	0.99	0.86
statistically	0.99	0.99	-0.02	-0.86

Pairwise similarity between word vectors

Similarity to the word *“training”*

Word	2020	2021	2022	Spearman correlation (ρ)
harmful	-0.01	0.99	0.99	0.86
early	0.0	0.99	0.99	0.86
reproducible	0.99	0.02	0.99	0.0
statistically	0.99	0.99	-0.04	-0.87

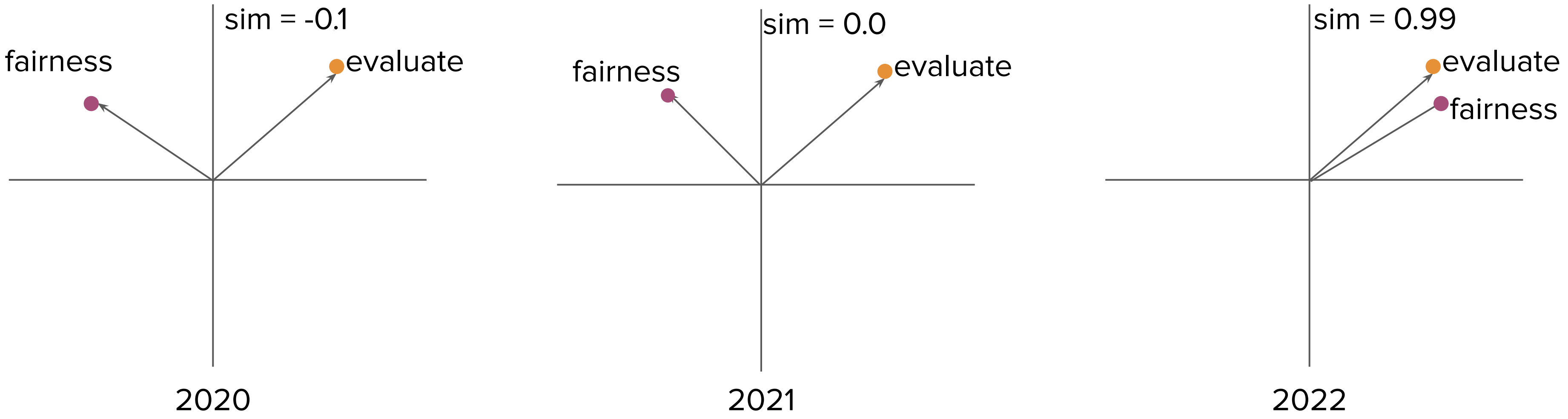
Pairwise similarity between word vectors

Similarity to the word *“training”*

Word	2020	2021	2022	Spearman correlation (ρ)
harmful	-0.01	0.99	0.99	0.86
early	0.0	0.99	0.99	0.86
reproducible	0.99	0.02	0.99	0.0
statistically	0.99	0.99	-0.04	-0.87

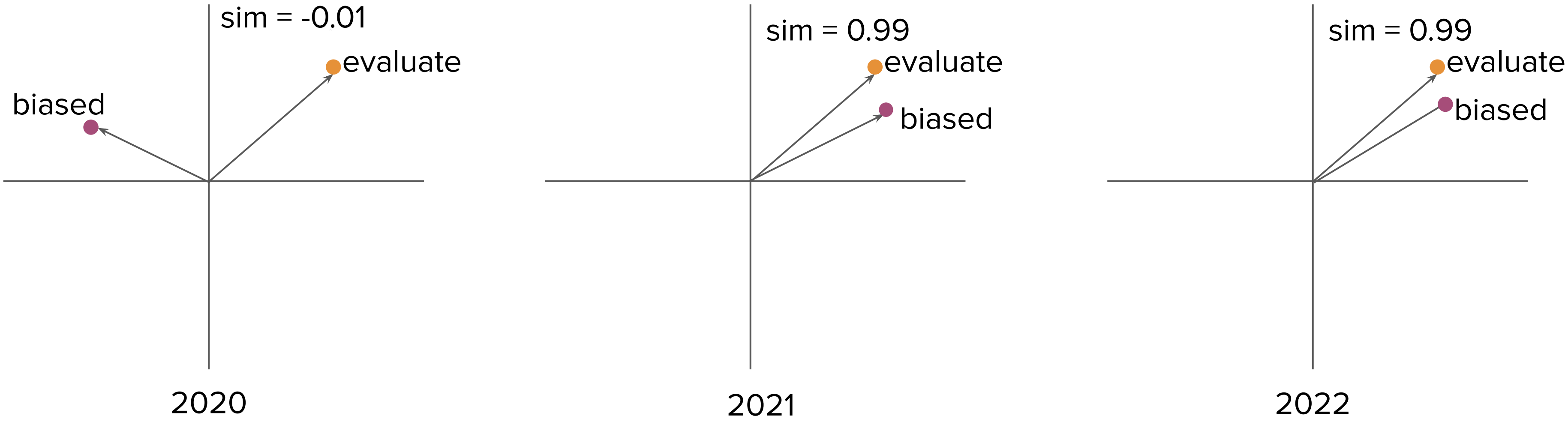
Relation to the word “evaluate”

Spearman correlation ρ= 0.9



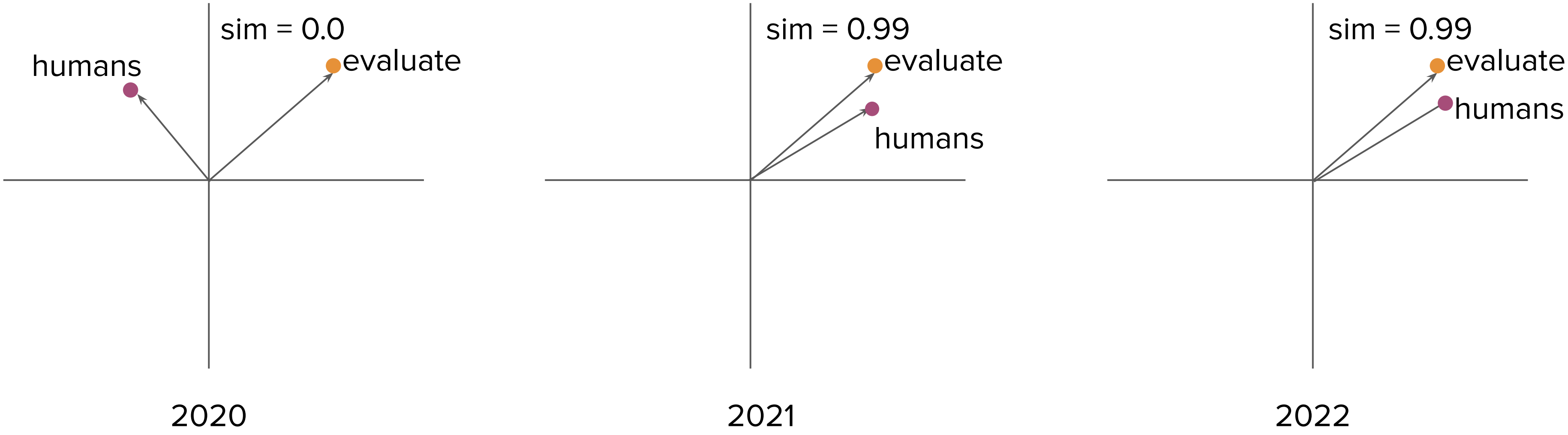
Relation to the word “evaluate”

Spearman correlation ρ= 0.86



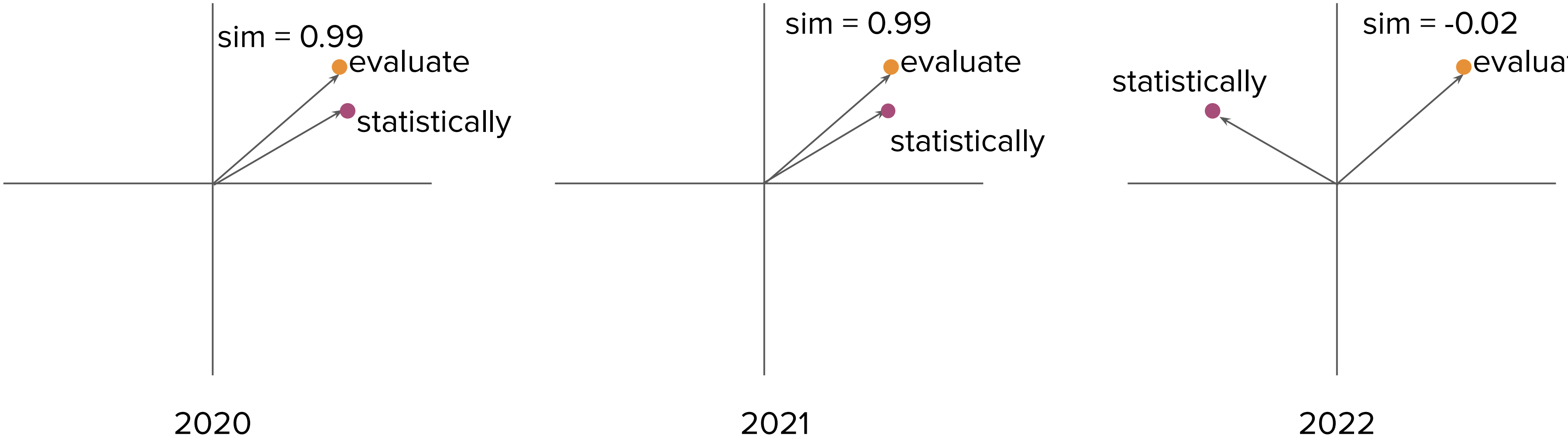
Relation to the word "evaluate"

Spearman correlation ρ= 0.86



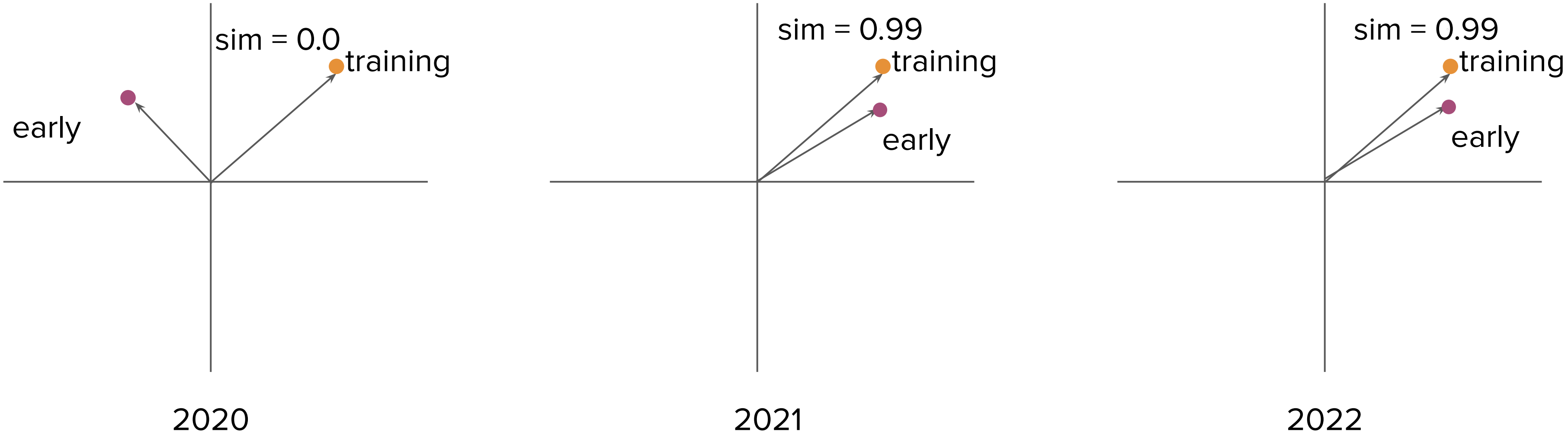
Relation to the word “evaluate”

Spearman correlation ρ= -0.86



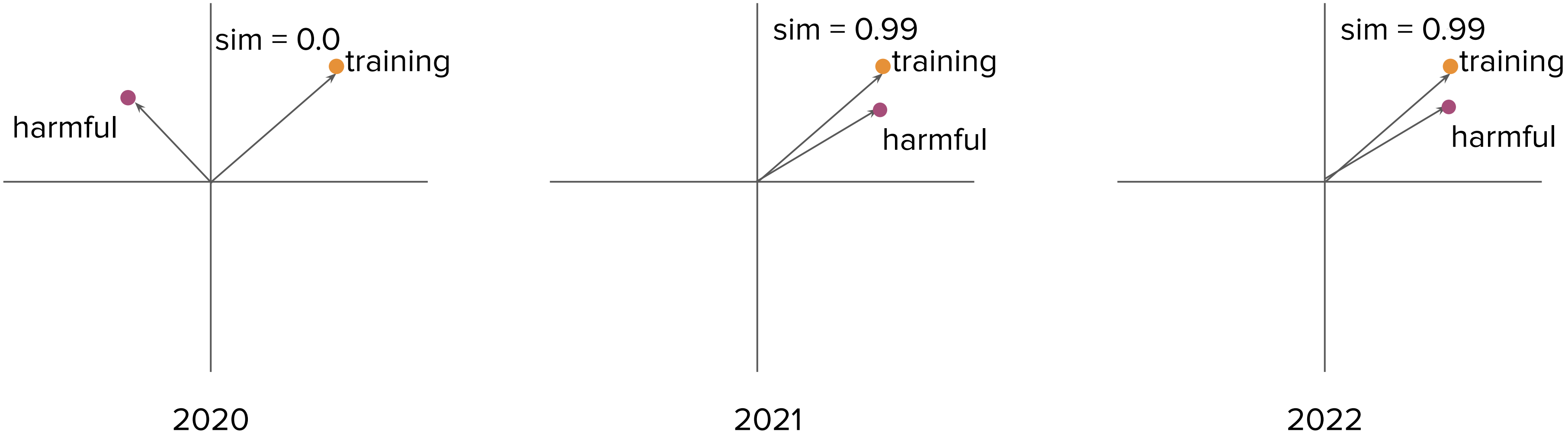
Relation to the word “training”

Spearman correlation ρ= 0.86



Relation to the word “training”

Spearman correlation ρ= 0.86



Takeaways

Although there is an exponential growth in NLP models, they are dominated by a few task categories.

The dominant NLP task categories show seasonal patterns

Model documentation has evolved from model-centric to data-centric*

Thanks for listening

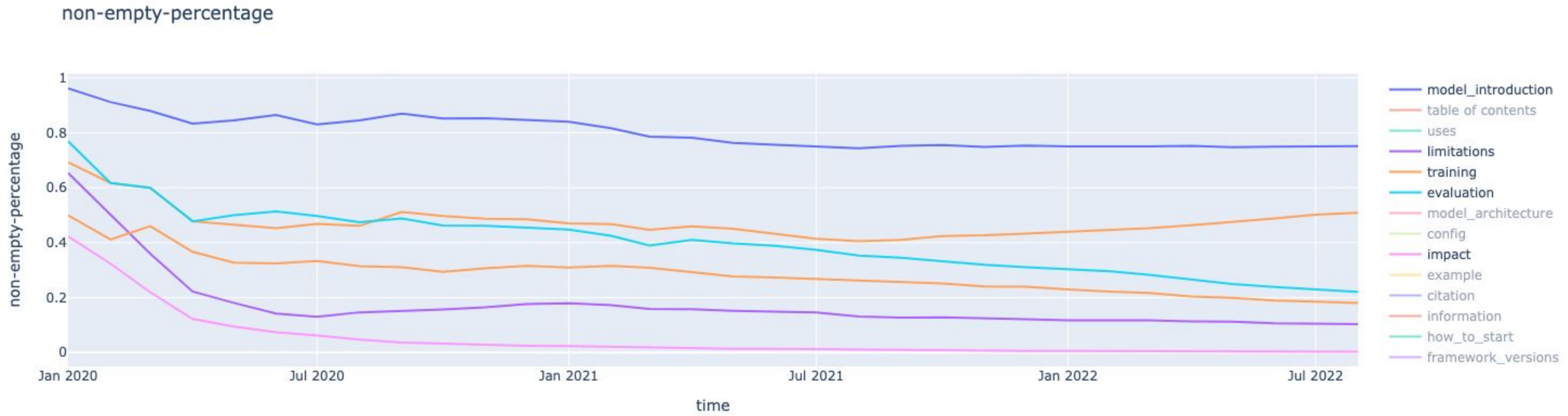
Collaborators

Weixin Liang
(Stanford)

Xinyu Yang
(ZJU)

Meg Mitchell
(Hugging Face)

James Zou
(Stanford)



non-empty-percentage
non-empty-percentage
time
0
0.2
0.4
0.6
0.8
1
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
model_introduction
table of contents
uses
limitations
training
evaluation
model_architecture
config
impact
example
citation
information
how_to_start
framework_versions