Understanding the role of data,scale and capacity in recent breakthroughs

LxMLS 2023

Sara Hooker - Cohere For AI

Cohere For AI

I lead Cohere For AI, a research lab that works on fundamental machine learning research.

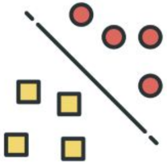
Open Science – contributing technical work to the open domain.

Full time research staff – focus on large language models and AI at scale.

Cohere For AI Research Programs – Scholars Program.

My research agenda to-date has focused on:

- Making large language models accessible by improving efficiency.
- Training models to fulfill multiple criteria



Model Compression - compact machine learning models to work in resource constrained environments.

Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

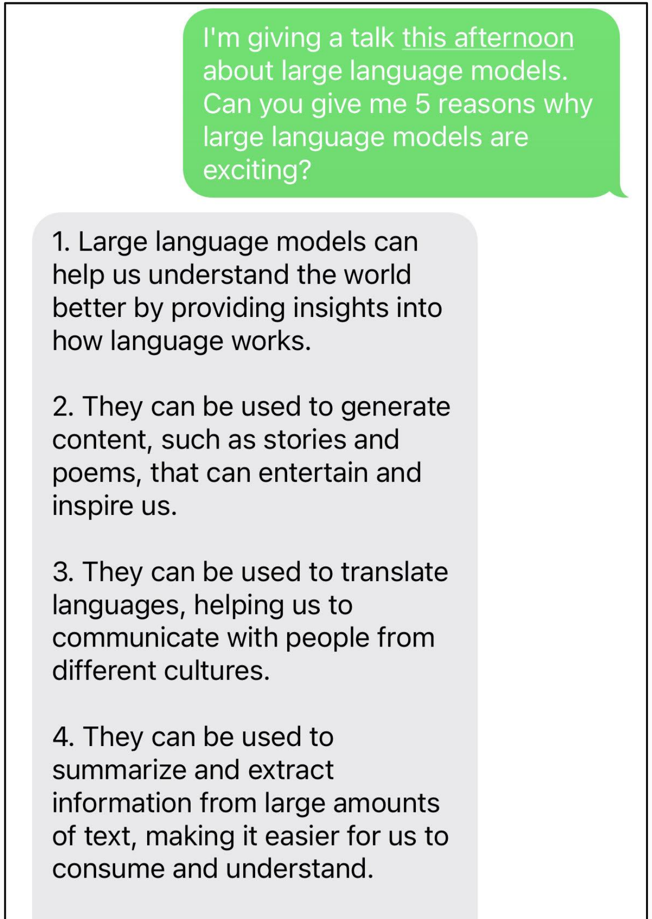
Model fragility and security - deploy secure models that protect user privacy.

Model Interpretability - reliable explanations for model behavior.

I currently work on designing large scale language models that are **efficient, multilingual, reliable and trustworthy.**

If any of these topics are interesting the talk, happy to discuss after the talk.

We are currently in a very exciting time for large language modeling research.



I'm giving a talk this afternoon about large language models. Can you give me 5 reasons why large language models are exciting?

1. Large language models can help us understand the world better by providing insights into how language works.

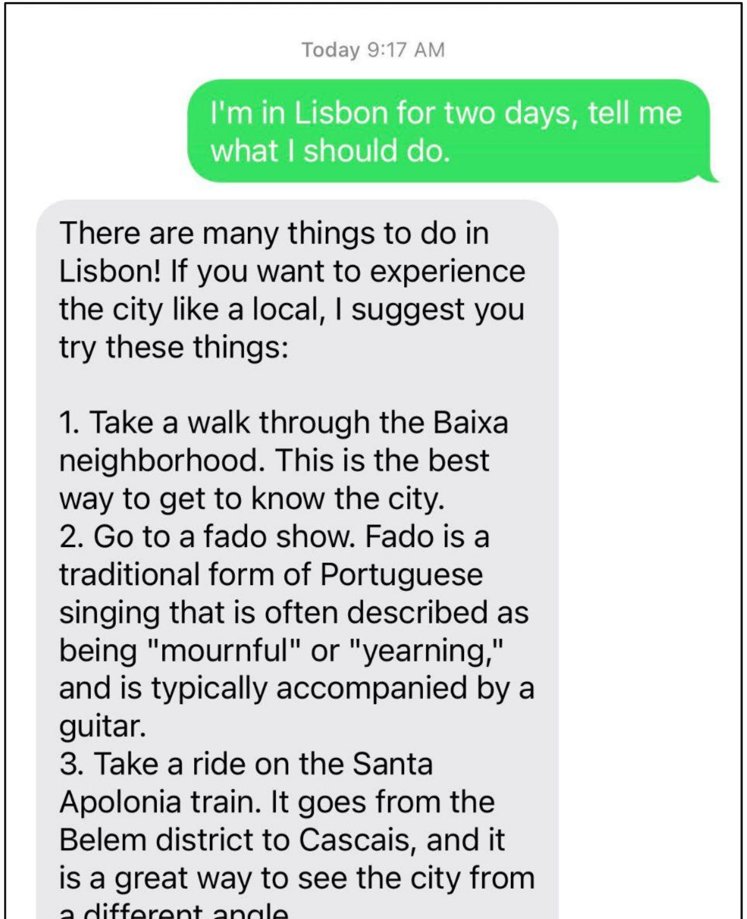
2. They can be used to generate content, such as stories and poems, that can entertain and inspire us.

3. They can be used to translate languages, helping us to communicate with people from different cultures.

4. They can be used to summarize and extract information from large amounts of text, making it easier for us to consume and understand.

Text tiny command:
US +1 (778) 718-8884
UK 447488883681

We are currently in a very exciting time for large language modeling research.



Today 9:17 AM

I'm in Lisbon for two days, tell me what I should do.

There are many things to do in Lisbon! If you want to experience the city like a local, I suggest you try these things:

1. Take a walk through the Baixa neighborhood. This is the best way to get to know the city.
2. Go to a fado show. Fado is a traditional form of Portuguese singing that is often described as being "mournful" or "yearning," and is typically accompanied by a guitar.
3. Take a ride on the Santa Apolonia train. It goes from the Belem district to Cascais, and it is a great way to see the city from a different angle

Text tiny command:
US +1 (778) 718-8884
UK 447488883681

While all of this is very exciting, in the time we have together today – I thought it might be fun to understand how we got here.

I want to introduce you to Joseph Weizenbaum – who made one of the chatbots called Eliza.

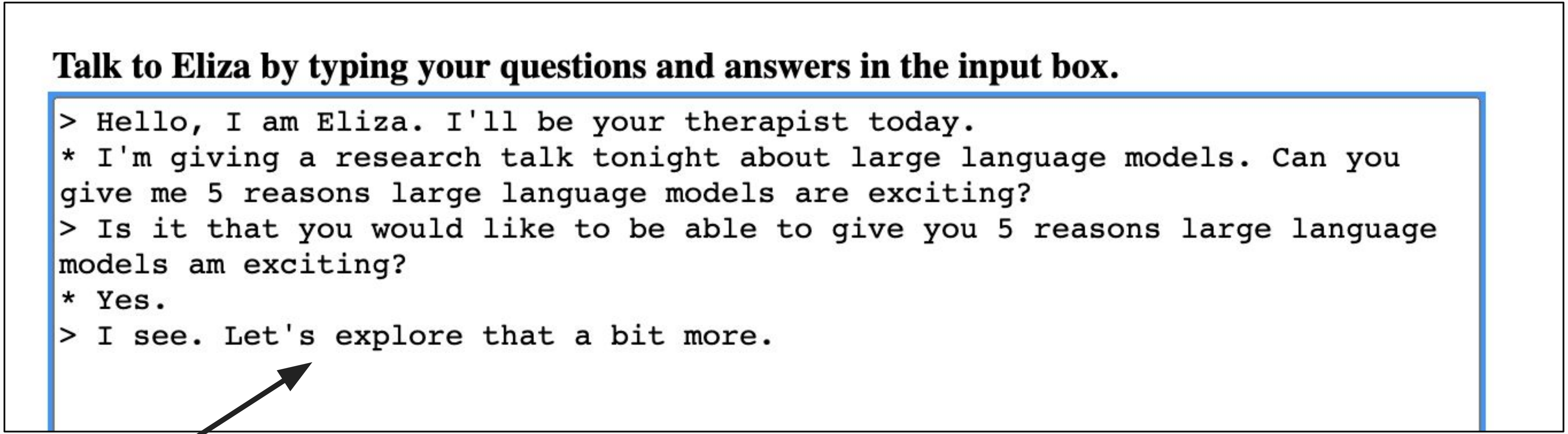
```
Welcome to
            EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
            EE      LL       II       ZZ   AA   AA
            EEEEE   LL       II     ZZZ    AAAAAAA
            EE      LL       II    ZZ      AA   AA
            EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weiz
This implementation by Norbert Landsteiner 2005.
```

```
ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something o
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
  ZA:  Is it important to you that your boyfriend
        ays I'm depressed much of the time.
        orry to hear that you are depressed.
        ue. I am unhappy.
         explain what made you unhappy ?
```

Eliza was based upon a popular form of therapy (rogerian therapy) — that involved repeating back to the patient what they just said.

Eliza used rules to determine the reply. There was no deep learning. You can see how this can fall of a cliff quickly.



Talk to Eliza by typing your questions and answers in the input box.

```
> Hello, I am Eliza. I'll be your therapist today.
* I'm giving a research talk tonight about large language models. Can you
give me 5 reasons large language models are exciting?
> Is it that you would like to be able to give you 5 reasons large language
models am exciting?
* Yes.
> I see. Let's explore that a bit more.
```

If Eliza couldn't identify a keyword or phrase, it uses a stock phrase, such as "Please go on." or “Let’s explore that a bit more.”

You can explore Eliza yourself at https://web.njit.edu/~ronkowit/eliza.html

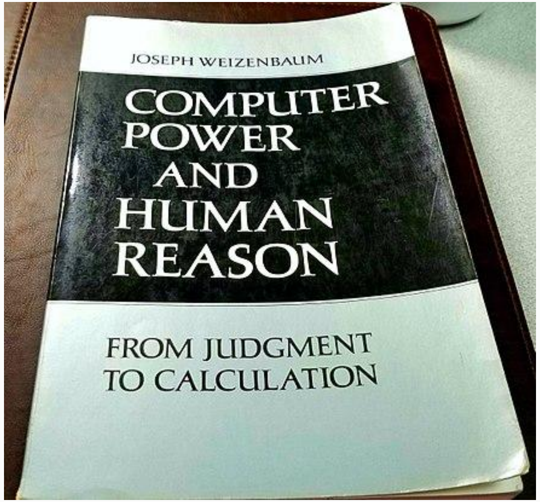
Nonetheless, many users of Eliza were convinced of ELIZA's intelligence and understanding.

```
Welcome to
                    EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
                    EE      LL       II        ZZ  AA   AA
                    EEEEE   LL       II      ZZZ   AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weiz
This implementation by Norbert Landsteiner 2005.
```

```
ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something o
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
  ZA:  Is it important to you that your boyfriend
         ays I'm depressed much of the time.
         orry to hear that you are depressed.
         ue. I am unhappy.
          explain what made you unhappy ?
```

Joseph became concerned when he caught his secretary using the program and telling it confidential things about her boyfriend.



What separates Eliza from our current chatbots? What allows for research breakthroughs.

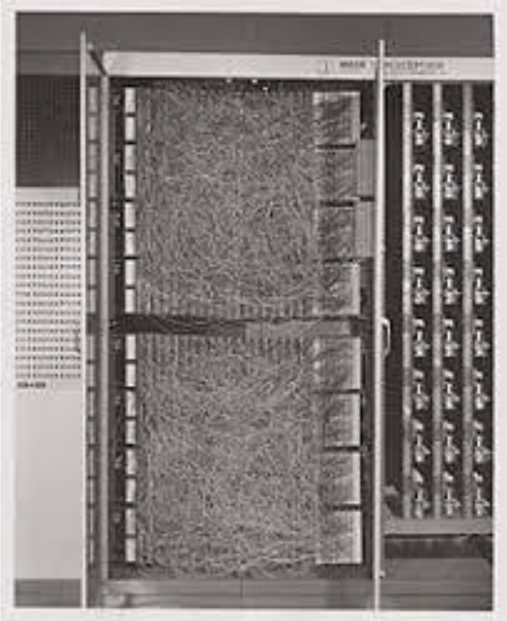
2012

1964

2017

For most of computer science history – two very different visions of how to make progress.

Rule based expert systems (ELIZA)

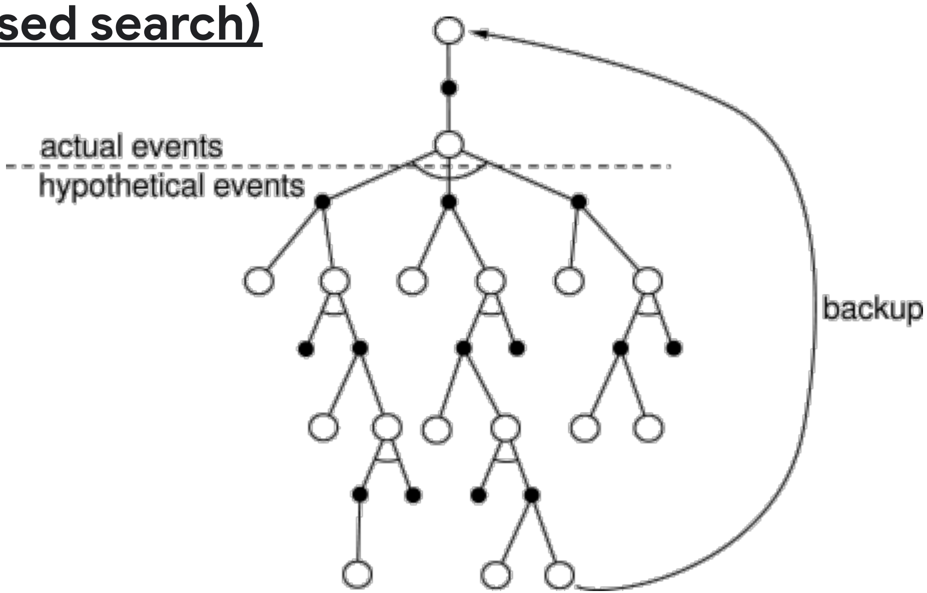


Deep Neural Network Approaches

Cohere For AI

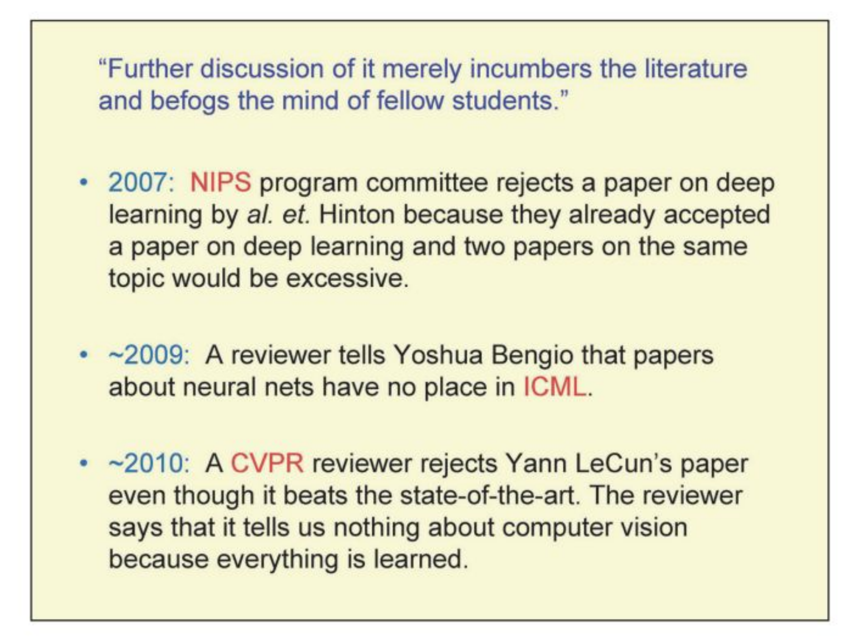
Rule based systems dominated mainstream research efforts from 1950-1980’s

Checkers (1956) Arthur Samuel (logic based search)



Wiki Link

Researchers working on deep neural networks were marginalized for several decades. This continued through 2012.

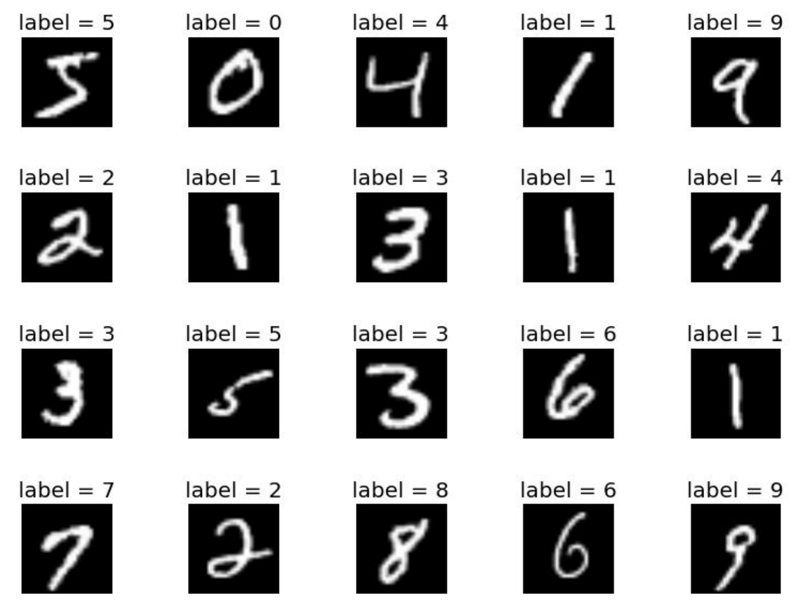


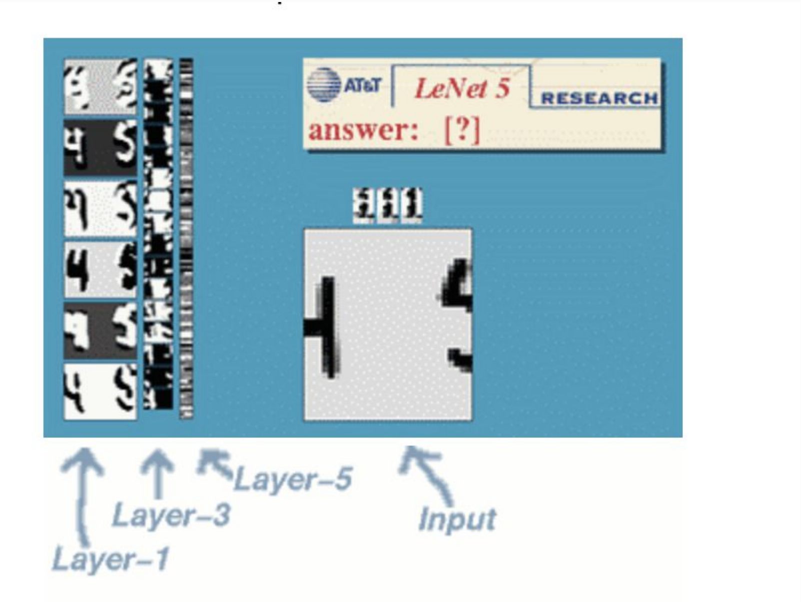
"Further discussion of it merely incumbers the literature and befogs the mind of fellow students."

- 2007: NIPS program committee rejects a paper on deep learning by *al. et.* Hinton because they already accepted a paper on deep learning and two papers on the same topic would be excessive.
- ~2009: A reviewer tells Yoshua Bengio that papers about neural nets have no place in ICML.
- ~2010: A CVPR reviewer rejects Yann LeCun's paper even though it beats the state-of-the-art. The reviewer says that it tells us nothing about computer vision because everything is learned.

Link here

During this time, only a few countries preserved funding for connectionist ideas.





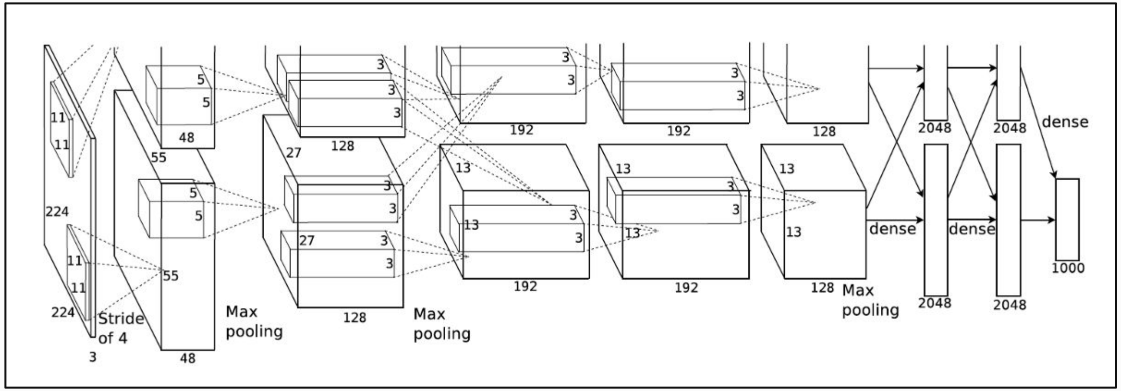
Breakthroughs in Digit Recognition (1998)

Image source: Yann Lecun site,CS Toronto, ImageNet.org

We know in hindsight that the breakthrough for deep neural networks was in 2012, when overnight AlexNet dominated ILSCVR.

AlexNet achieved a top-5 error of 15.3%, more than 10.8 percentage points lower than that of the runner up.

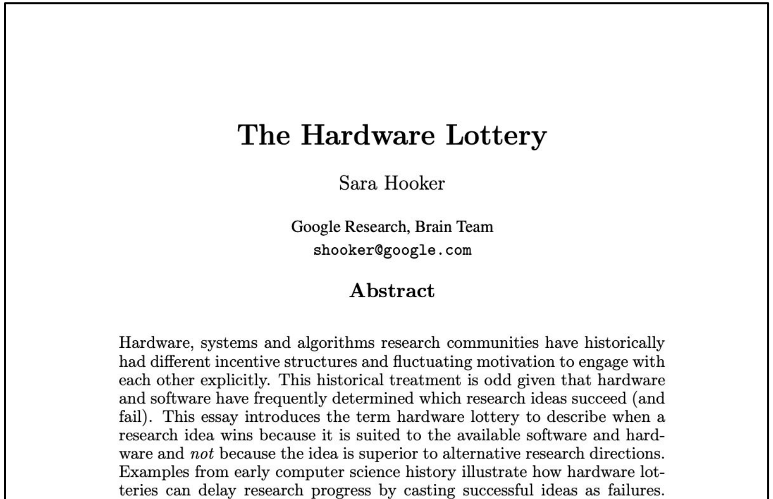
Everyone switched to convolutional neural networks overnight.



This breakthrough was in part due to a lucky fluke.

Hardware designed for video games was repurposed for training deep neural networks.

- GPUs developed for video games
- Good at parallelizing matrix multiplies



The Hardware Lottery

Sara Hooker

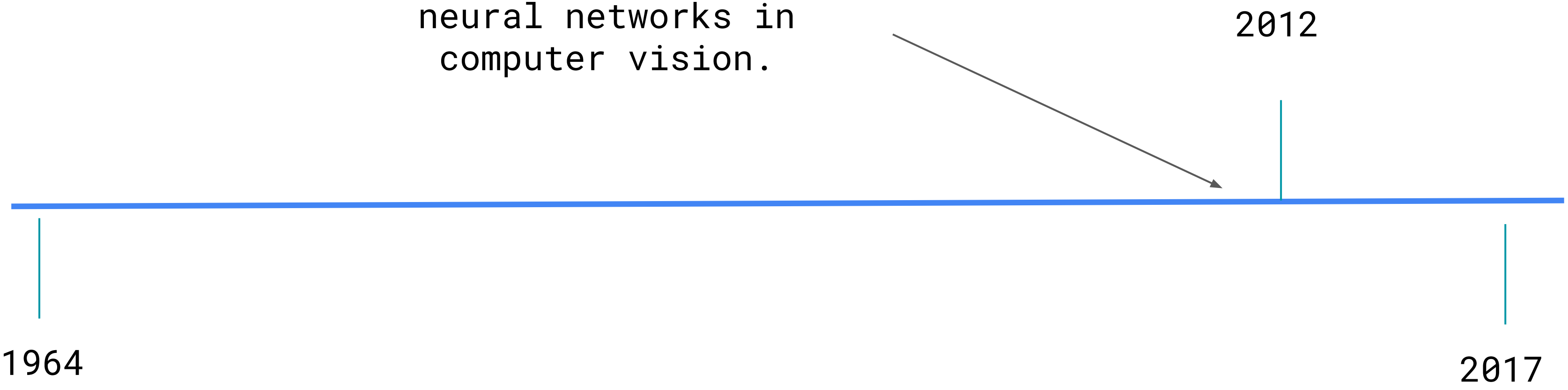
Google Research, Brain Team
shooker@google.com

Abstract

Hardware, systems and algorithms research communities have historically had different incentive structures and fluctuating motivation to engage with each other explicitly. This historical treatment is odd given that hardware and software have frequently determined which research ideas succeed (and fail). This essay introduces the term hardware lottery to describe when a research idea wins because it is suited to the available software and hardware and *not* because the idea is superior to alternative research directions. Examples from early computer science history illustrate how hardware lotteries can delay research progress by casting successful ideas as failures.

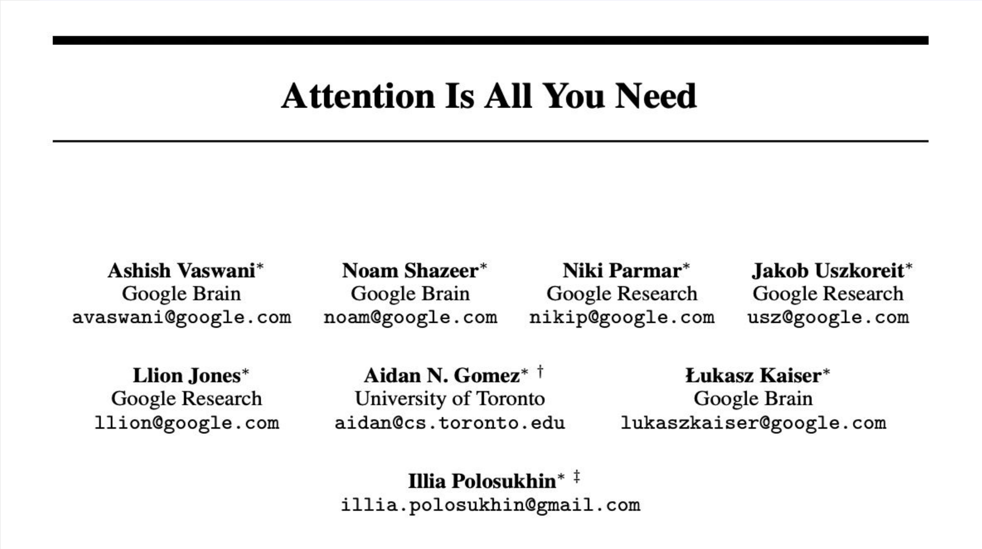
Link here

Overnight in 2012, everyone switched to deep neural networks.



Perseverance over decades led to the breakthrough of deep neural networks in computer vision.

2017: **The Transformer**, the culmination of a rich history of language modelling



Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

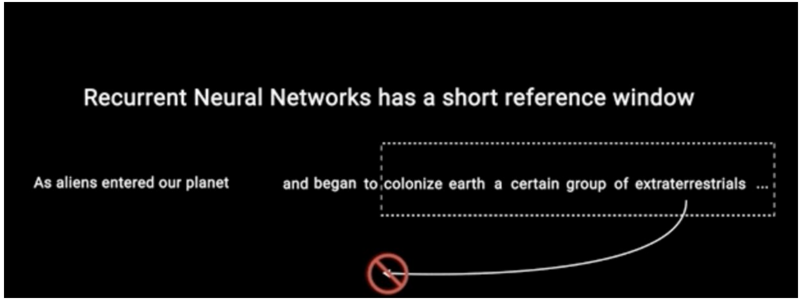
Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*, †]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

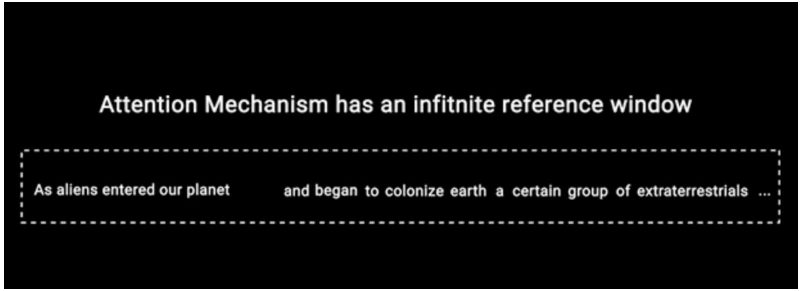
Illia Polosukhin[*, ‡]
illia.polosukhin@gmail.com

Transformers were a breakthrough at modelling longer text dependencies. Overnight everyone abandoned LSTMs and switched to transformers.



Recurrent Neural Networks has a short reference window

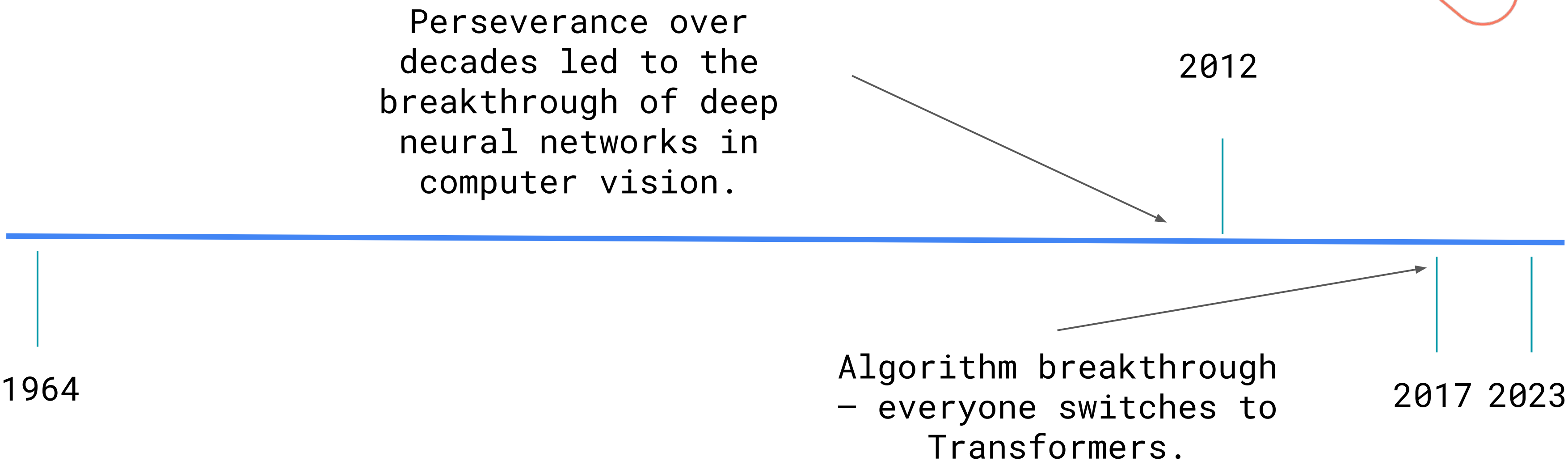
As aliens entered our planet and began to colonize earth a certain group of extraterrestrials ...



Attention Mechanism has an infitnite reference window

As aliens entered our planet and began to colonize earth a certain group of extraterrestrials ...

Transformers are the shared building block of almost all large language models.



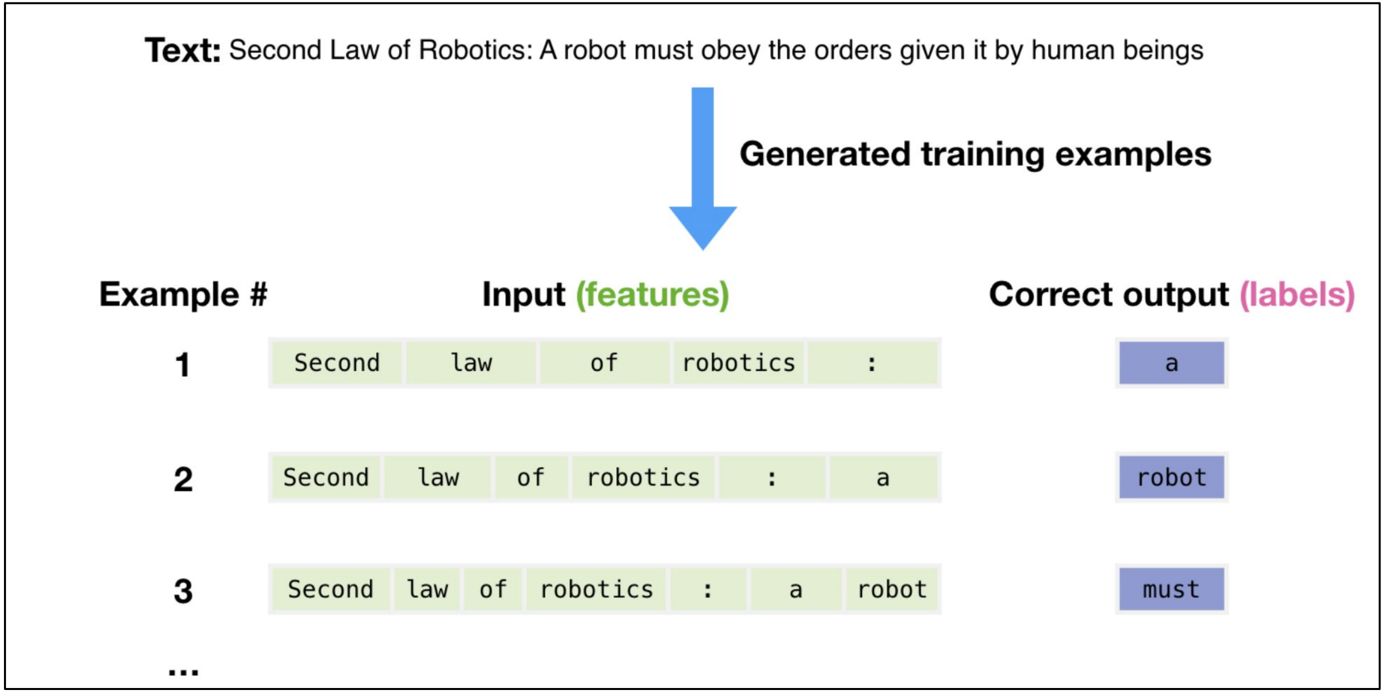
Perseverance over decades led to the breakthrough of deep neural networks in computer vision.

Algorithm breakthrough – everyone switches to Transformers.

2017-2023: The great acceleration.

Pretraining on larger and larger datasets in an unsupervised fashion.

Step 1: Unsupervised pre-training of a transformer model on a massive web crawled dataset (i.e. train on the internet).



Text: Second Law of Robotics: A robot must obey the orders given it by human beings

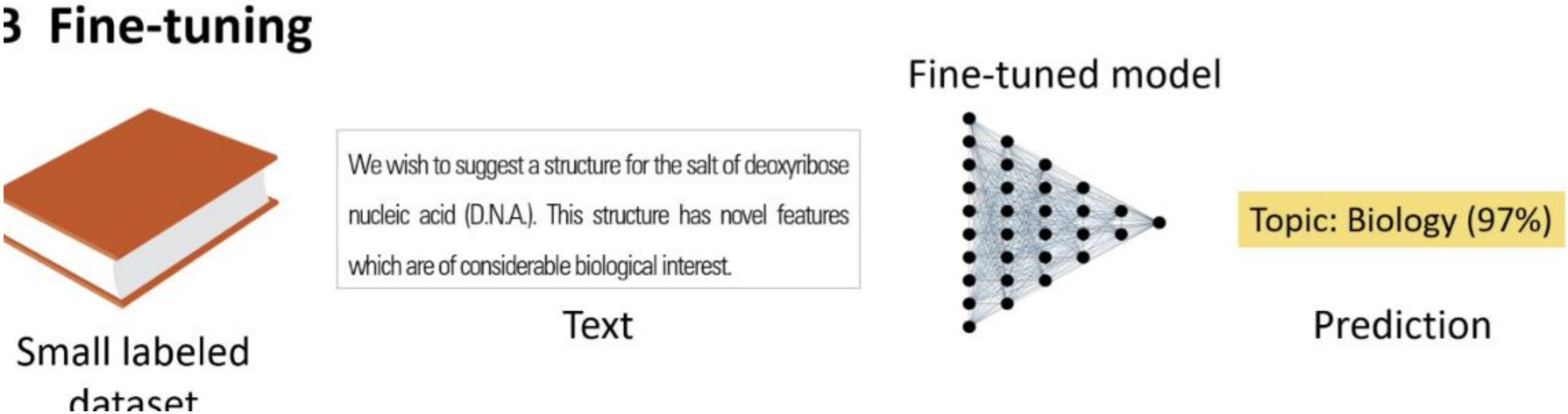
Generated training examples

Example #	**Input (features)**	**Correct output (labels)**
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		

https://jalammar.github.io/how-gpt3-works-visualizations-animations/

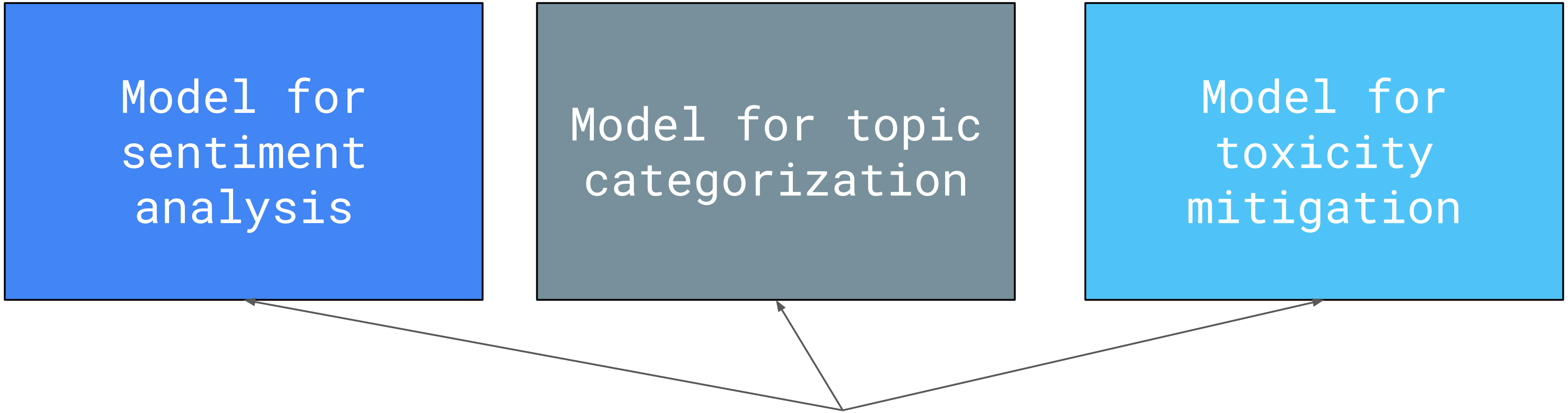
Why do we pretrain?

Changes in optimization strategy – previously, we would finetuning the model one task at a time.



We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

Changes in optimization. This meant we would end up with dedicated independent models for each downstream task:



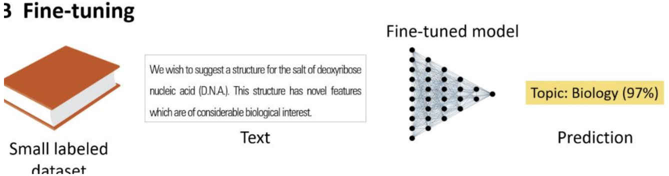
Model for sentiment analysis

Model for topic categorization

Model for toxicity mitigation

One custom model per task

Changed to multi-task fine-tuning. Moving to a single global model – train on multiple tasks at once.



We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

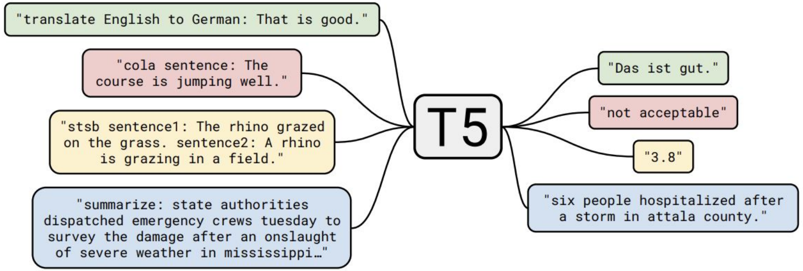
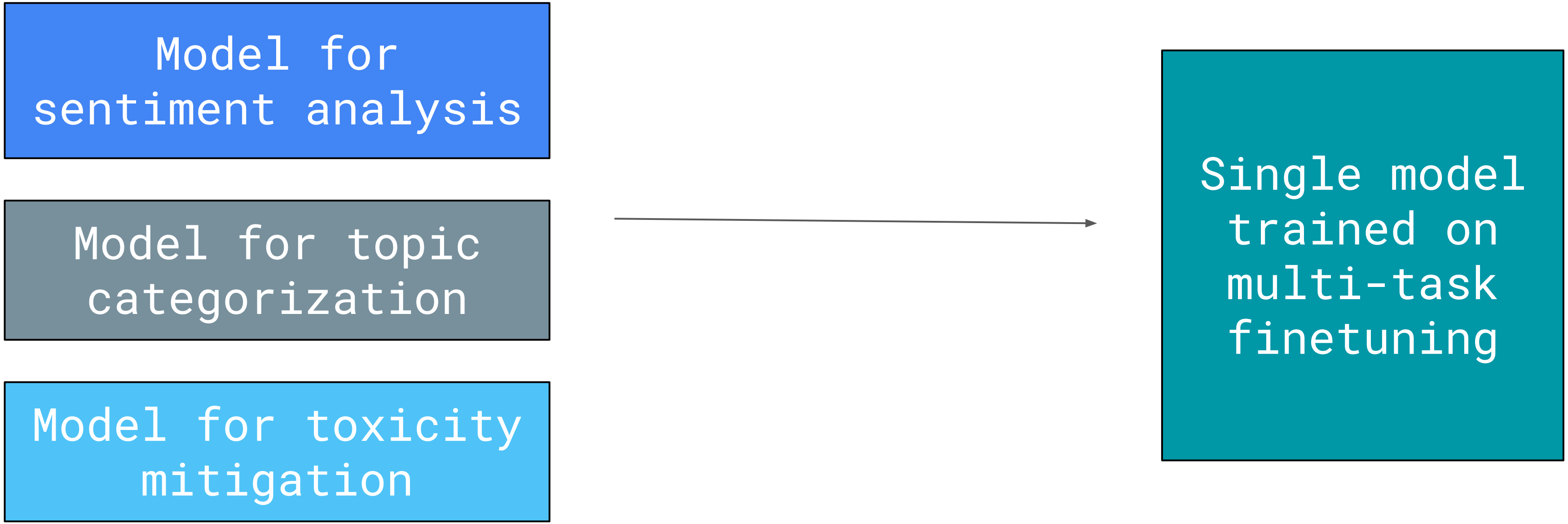

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

Finetuning on a single task

Finetuning on many different tasks

Why is this a big deal – it transitions from having custom models for each task to having a single task-general model that can perform a lot of tasks, which only require zero or few examples



Model for sentiment analysis

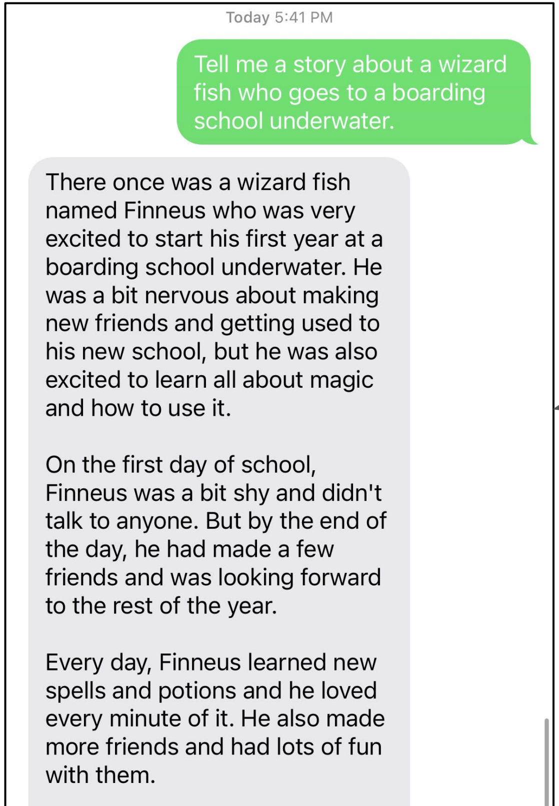
Model for topic categorization

Model for toxicity mitigation

Single model trained on multi-task finetuning

The role of data in recent breakthroughs

There has been renewed interest in the types of data that lead to powerful zero shot abilities – **performing well on completely unseen tasks.**



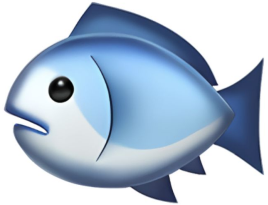
Today 5:41 PM

Tell me a story about a wizard fish who goes to a boarding school underwater.

There once was a wizard fish named Finneus who was very excited to start his first year at a boarding school underwater. He was a bit nervous about making new friends and getting used to his new school, but he was also excited to learn all about magic and how to use it.

On the first day of school, Finneus was a bit shy and didn't talk to anyone. But by the end of the day, he had made a few friends and was looking forward to the rest of the year.

Every day, Finneus learned new spells and potions and he loved every minute of it. He also made more friends and had lots of fun with them.

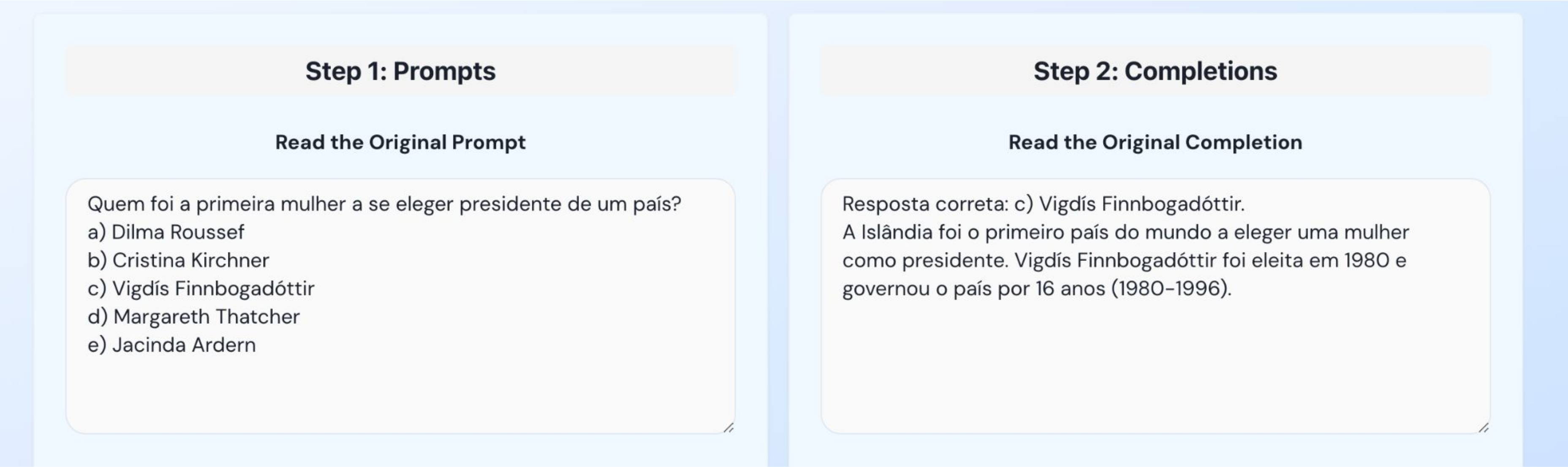


Text tiny Command +1 (778) 718-8884

It turns out two ingredients have been particularly important at leading to breakthroughs in performance on zero and few shot tasks:

1. **Structuring multi-task finetuning data as questions and answers.**

1. Structuring multi-task fine-tuning data as questions and answers.



Step 1: Prompts

Read the Original Prompt

Quem foi a primeira mulher a se eleger presidente de um país?
a) Dilma Roussef
b) Cristina Kirchner
c) Vigdís Finnbogadóttir
d) Margareth Thatcher
e) Jacinda Ardern

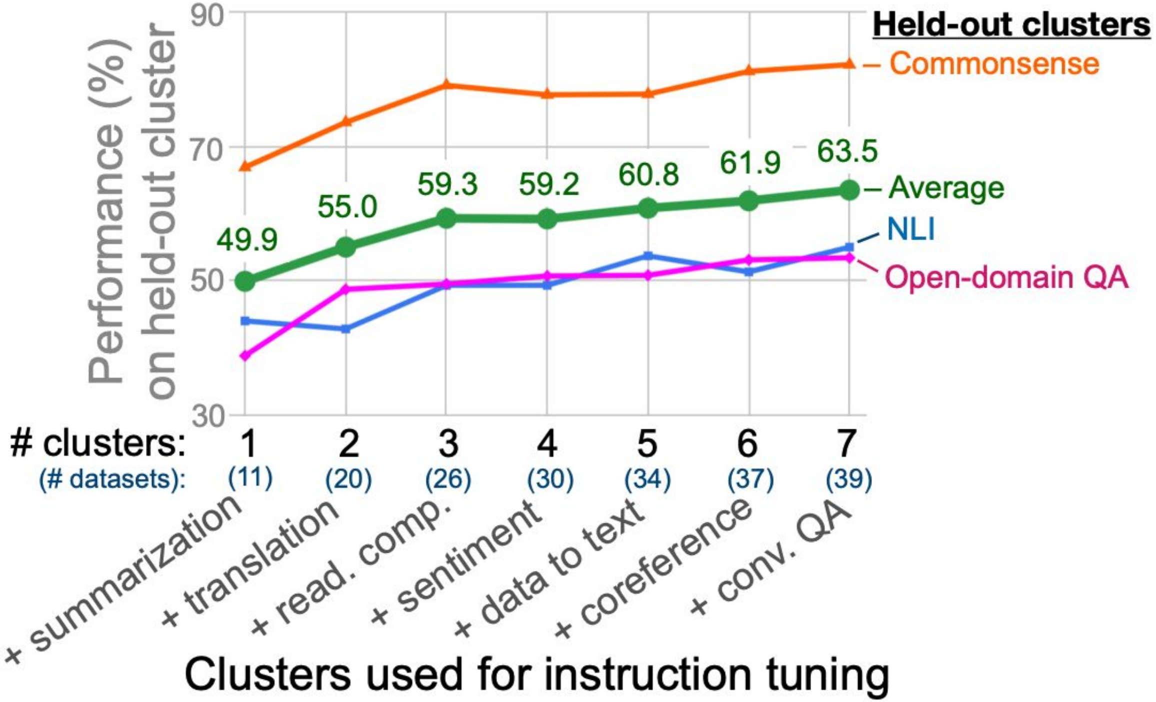
Step 2: Completions

Read the Original Completion

Resposta correta: c) Vigdís Finnbogadóttir.
A Islândia foi o primeiro país do mundo a eleger uma mulher como presidente. Vigdís Finnbogadóttir foi eleita em 1980 e governou o país por 16 anos (1980-1996).

Cohere For AI

This combination – of multitask training and instruction style improves zero shot performance.

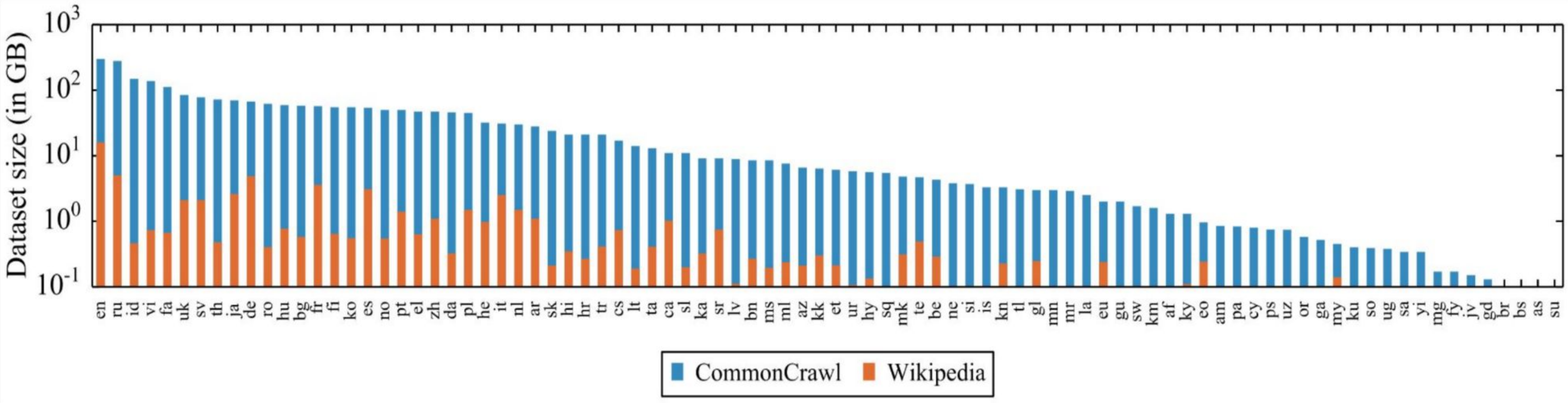


Finetuned language models are zero-shot learners (Wei et al., 2021).

Cohere For AI

Zero shot is particularly helpful for data limited regimes. Why?

Zero shot is particularly helpful for data limited regimes.

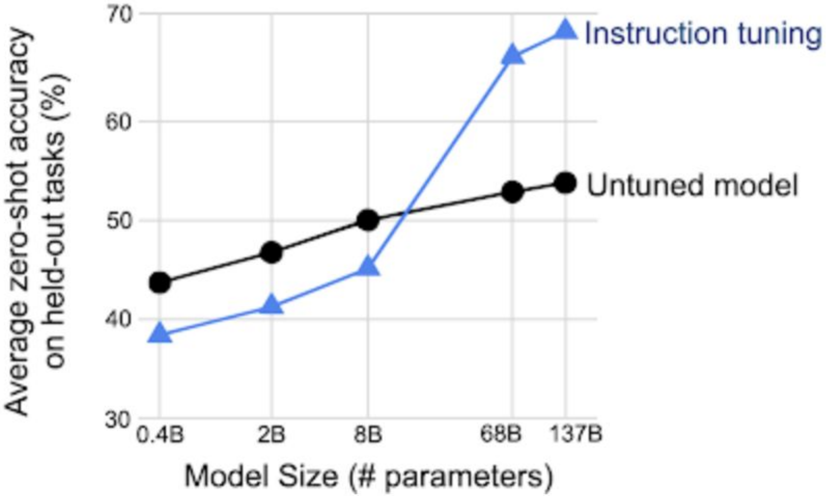


- Data limited regimes struggle to realize gains of fine-tuning.
- Fine-tuning large language models can be expensive — would be great if a model generalized to a task out of the box.

ACL Keynote, Conneau et al.

Cohere For AI

However, it also requires larger and larger models to take advantage of instruction tuning (partly explaining our race to ever larger models).



Instruction tuning only improves performance on unseen tasks for models of certain size.

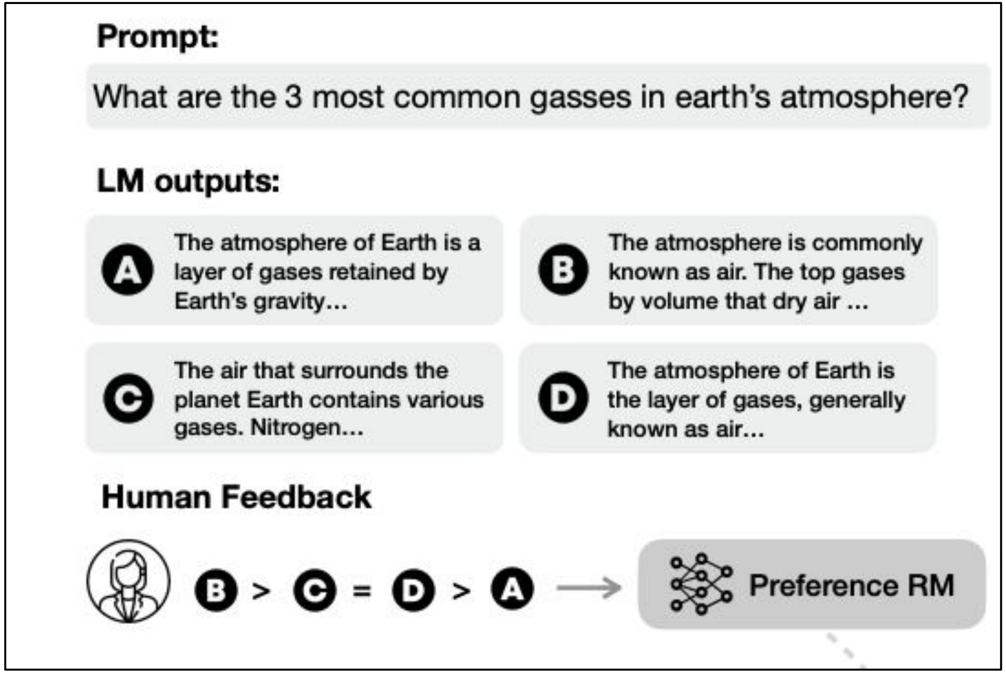
Cohere For AI

link

It turns out two ingredients have been particularly important at leading to breakthroughs in chatbot performance:

1. **Structuring multi-task finetuning data as questions and answers.**
2. **Integrating human feedback about preferences**

Which one do you prefer?



Prompt:

What are the 3 most common gasses in earth's atmosphere?

LM outputs:

- A: The atmosphere of Earth is a layer of gases retained by Earth's gravity...
- B: The atmosphere is commonly known as air. The top gases by volume that dry air ...
- C: The air that surrounds the planet Earth contains various gases. Nitrogen...
- D: The atmosphere of Earth is the layer of gases, generally known as air...

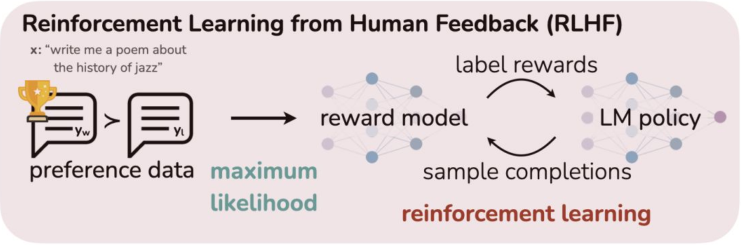
Human Feedback

Gather feedback -> train model to align with feedback.

Upweight answers that humans find more meaningful.

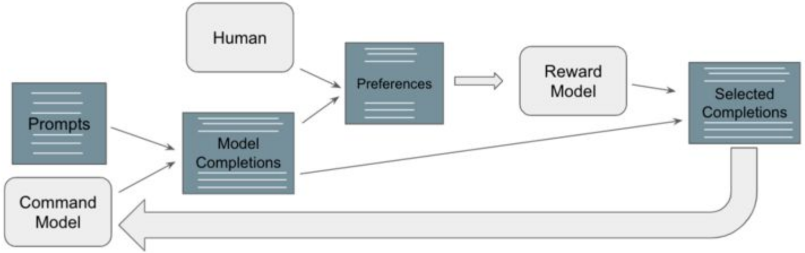
Emerging area of optimization - not a clear winner

w Reward Model:

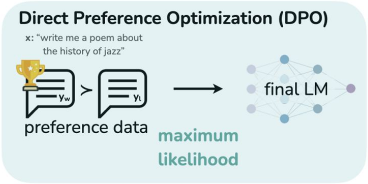


$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y\mid x)\;||\;\pi_{\mathrm{ref}}(y\mid x)\right]$$

w Synthetic & Good only Data

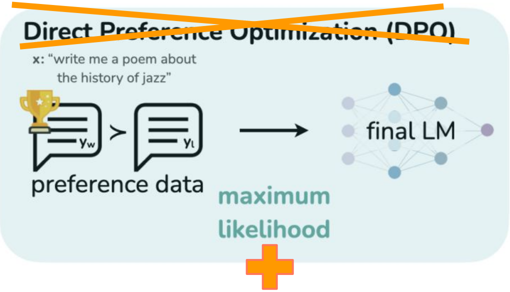


w/o Reward Model:


Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
>
y_l
preference data
maximum likelihood
final LM

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\mathrm{ref}}(y_l\mid x)}\right)\right]$$

Sequence Likelihood Calibration from Human Feedback (SLiC-HF)

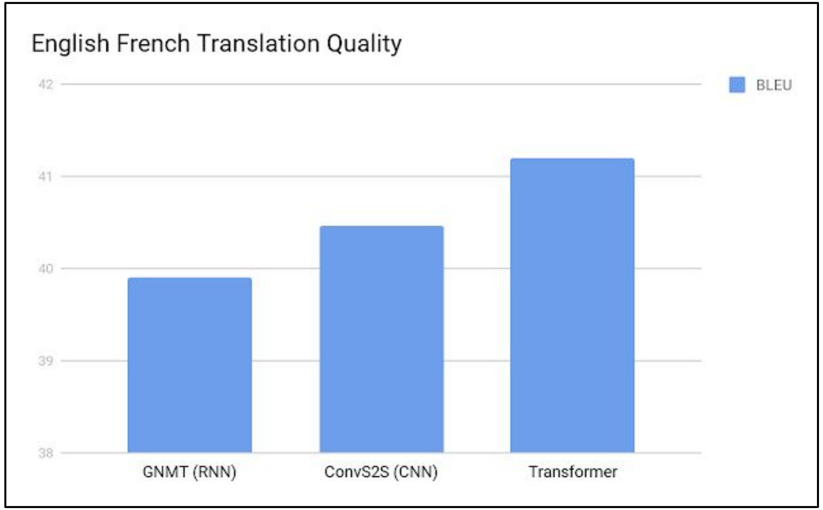

~~Direct Preference Optimization (DPO)~~
x: "write me a poem about the history of jazz"
y_w
>
y_l
preference data
maximum likelihood
final LM
+

Margin Triplet Contrastive Loss

$$\mathcal{L}(\theta) = max(0, \delta - \log P_\theta(\mathbf{y}^+|\mathbf{x}) + \log P_\theta(\mathbf{y}^-|\mathbf{x})) - \lambda\log P_\theta(\mathbf{y}_{ref}|\mathbf{x})$$

In summary, 2017-2023 has been a rollercoaster ride.

The original transformer was applied to machine learning transformation



2023 – breakthroughs in almost all language tasks

Conversation

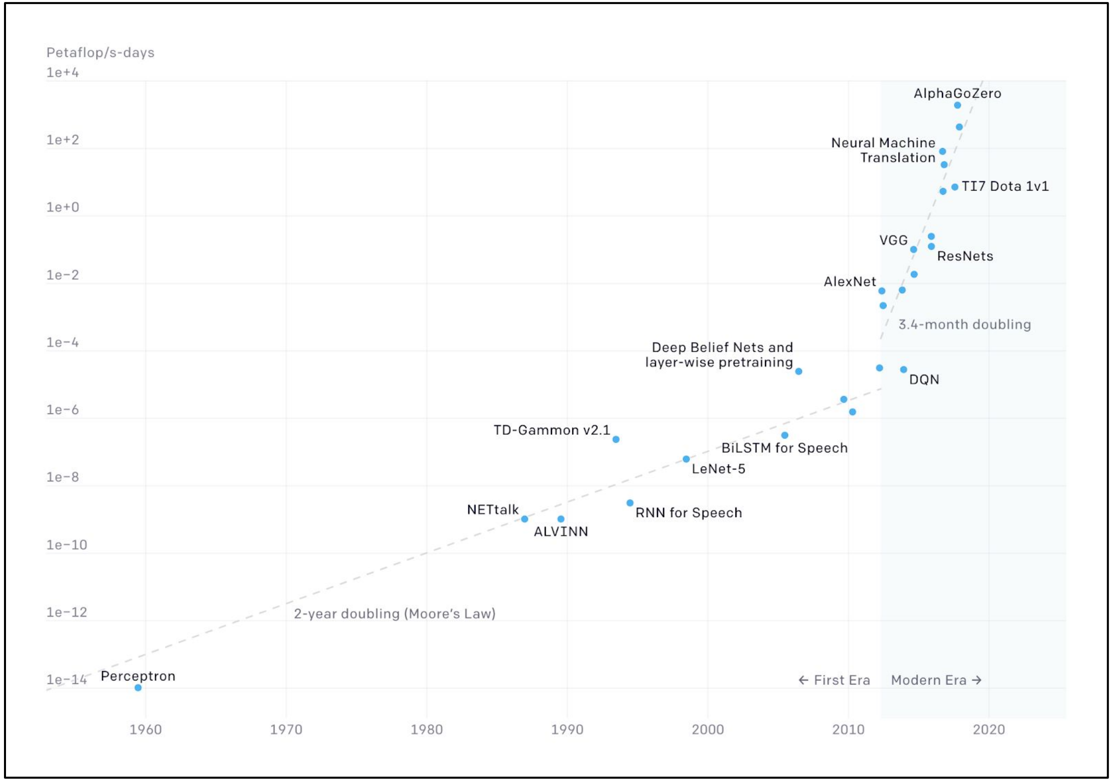
Generation

Classification

So where are we now. Open challenges I am currently thinking about.

Challenge 1: The renewed urgency for efficiency.

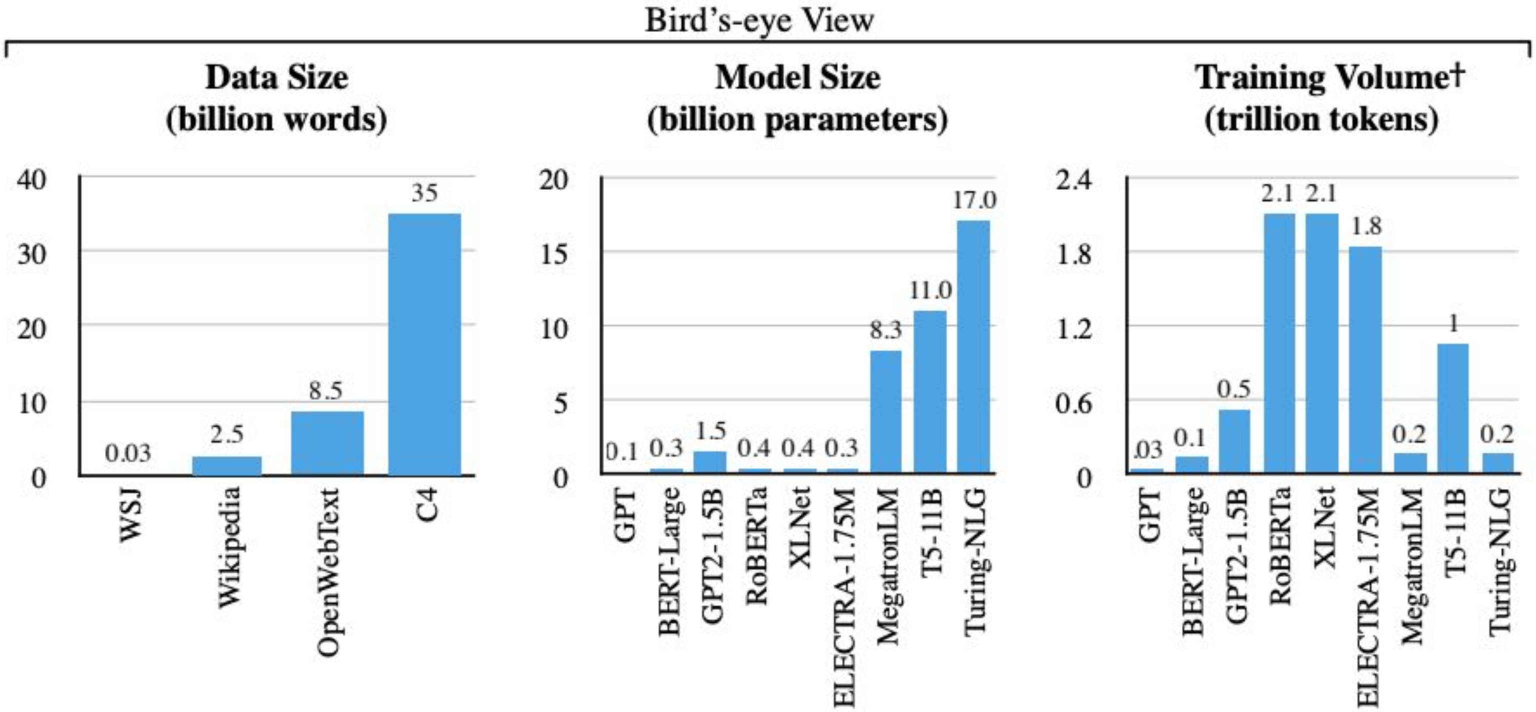
A rapid growth in the size of networks. “Bigger is better” race in the number of parameters.



Cohere For AI

Canziani et al., 2016, Open AI 2019

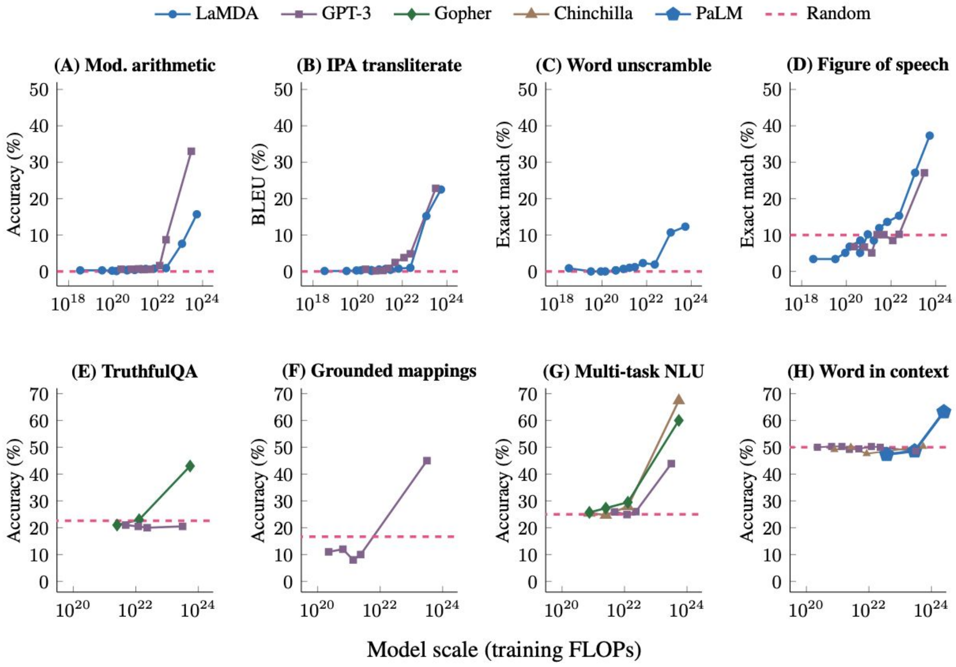
This characterizes both vision and NLP tasks.



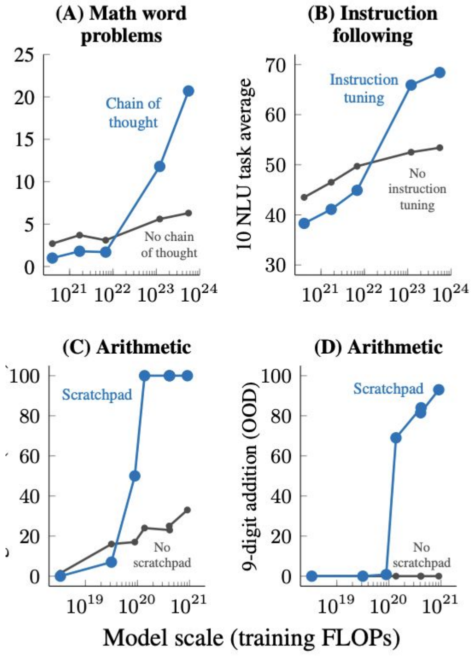
Cohere For AI

Link here [Sharir et al. 2020]

Different regimes of scale appear to induce emergent abilities – model performance which cannot be predicted from smaller models.



Few shot prompting performance



Finetuning and few shot.

[Wei et al. 2022]

Cohere For AI

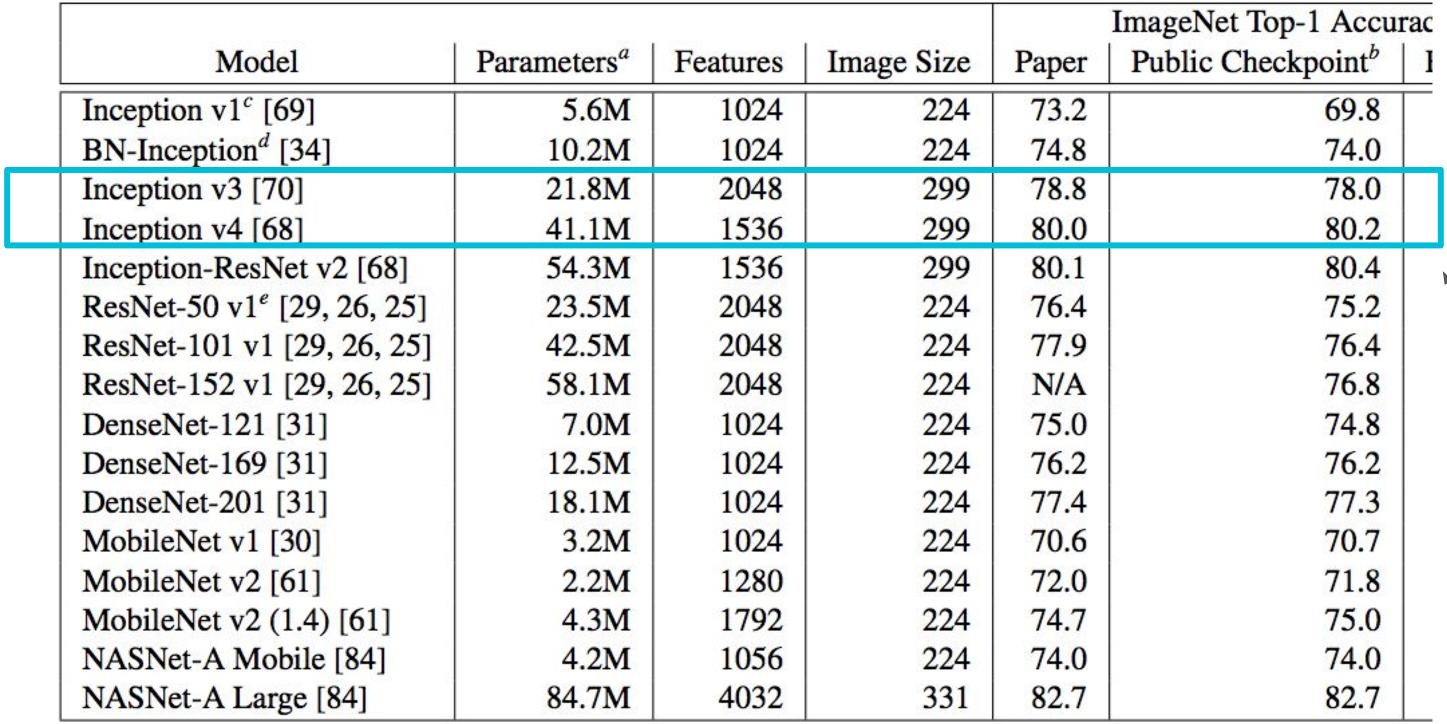
A key limitation of this approach:

Relationship between weights and generalization properties is not well understood.

Why do we need so many weights in the first place?

1) Diminishing returns to adding more weights.
2) Many redundancies between weights
3) Intriguing relationship with data quantity and quality.
4) We can remove most weights after training.
5) Most weights necessary for representing a small fraction of the dataset.

1. **Diminishing returns** to adding parameters. Millions of parameters are needed to **eek** out additional gains.

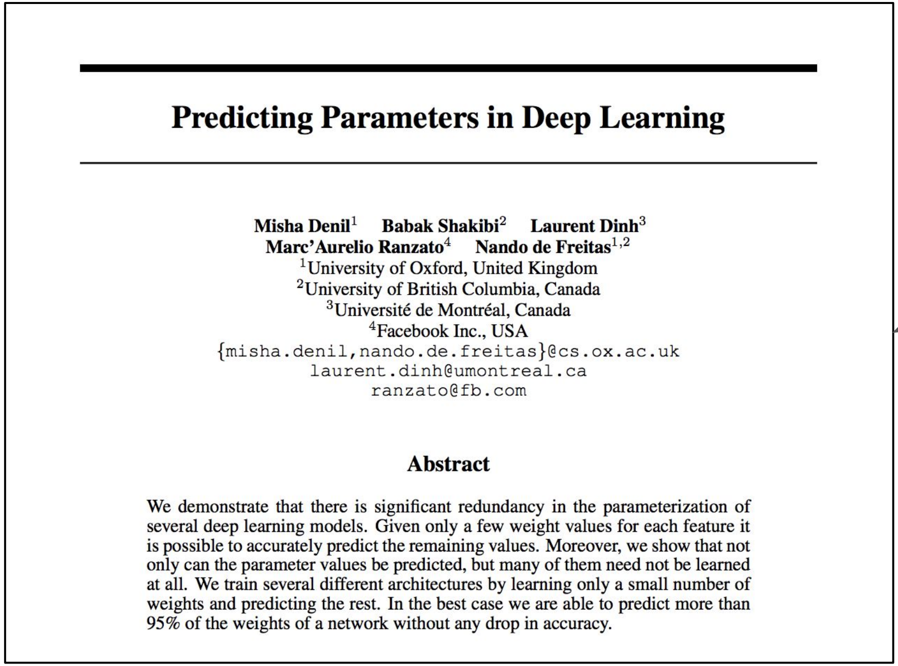


				ImageNet Top-1 Accurac		
Model	Parameters[a]	Features	Image Size	Paper	Public Checkpoint[b]	I
Inception v1[c] [69]	5.6M	1024	224	73.2	69.8	
BN-Inception[d] [34]	10.2M	1024	224	74.8	74.0	
Inception v3 [70]	21.8M	2048	299	78.8	78.0	
Inception v4 [68]	41.1M	1536	299	80.0	80.2	
Inception-ResNet v2 [68]	54.3M	1536	299	80.1	80.4	
ResNet-50 v1[e] [29, 26, 25]	23.5M	2048	224	76.4	75.2	
ResNet-101 v1 [29, 26, 25]	42.5M	2048	224	77.9	76.4	
ResNet-152 v1 [29, 26, 25]	58.1M	2048	224	N/A	76.8	
DenseNet-121 [31]	7.0M	1024	224	75.0	74.8	
DenseNet-169 [31]	12.5M	1024	224	76.2	76.2	
DenseNet-201 [31]	18.1M	1024	224	77.4	77.3	
MobileNet v1 [30]	3.2M	1024	224	70.6	70.7	
MobileNet v2 [61]	2.2M	1280	224	72.0	71.8	
MobileNet v2 (1.4) [61]	4.3M	1792	224	74.7	75.0	
NASNet-A Mobile [84]	4.2M	1056	224	74.0	74.0	
NASNet-A Large [84]	84.7M	4032	331	82.7	82.7	

Almost double the amount of weights for a gain in 2% points.

Table: Kornblith et al., 2018 [Kaplan + 2020]

2. Redundancies Between Weights



Predicting Parameters in Deep Learning

Misha Denil[1] **Babak Shakibi**[2] **Laurent Dinh**[3]
Marc'Aurelio Ranzato[4] **Nando de Freitas**[1,2]
[1]University of Oxford, United Kingdom
[2]University of British Columbia, Canada
[3]Université de Montréal, Canada
[4]Facebook Inc., USA
`{misha.denil,nando.de.freitas}@cs.ox.ac.uk`
`laurent.dinh@umontreal.ca`
`ranzato@fb.com`

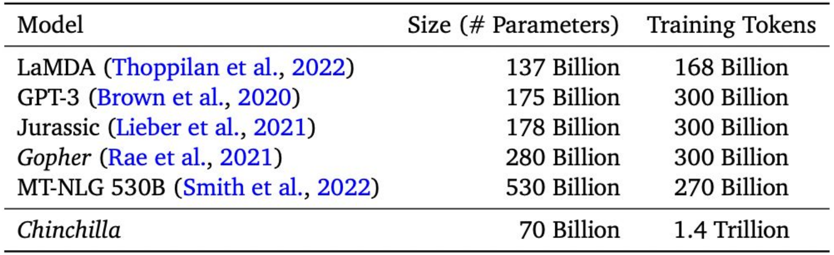
Abstract

We demonstrate that there is significant redundancy in the parameterization of several deep learning models. Given only a few weight values for each feature it is possible to accurately predict the remaining values. Moreover, we show that not only can the parameter values be predicted, but many of them need not be learned at all. We train several different architectures by learning only a small number of weights and predicting the rest. In the best case we are able to predict more than 95% of the weights of a network without any drop in accuracy.

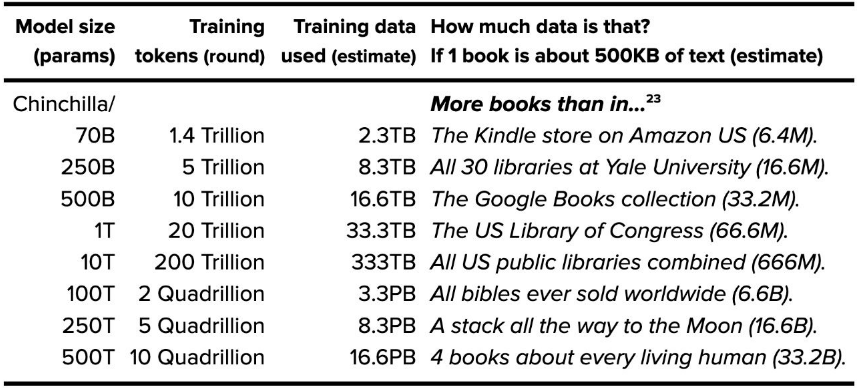
Denil et al. find that a small set of weights can be used to predict 95% of weights in the network.

[[Denil et al., 2014]]

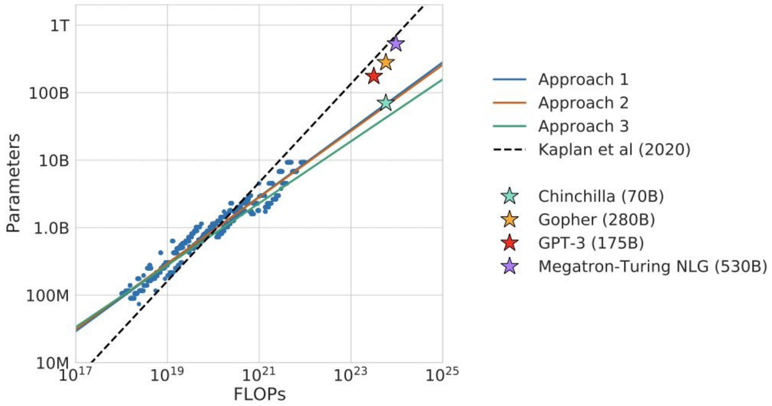
3. Intriguing relationship between weights and data.



Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

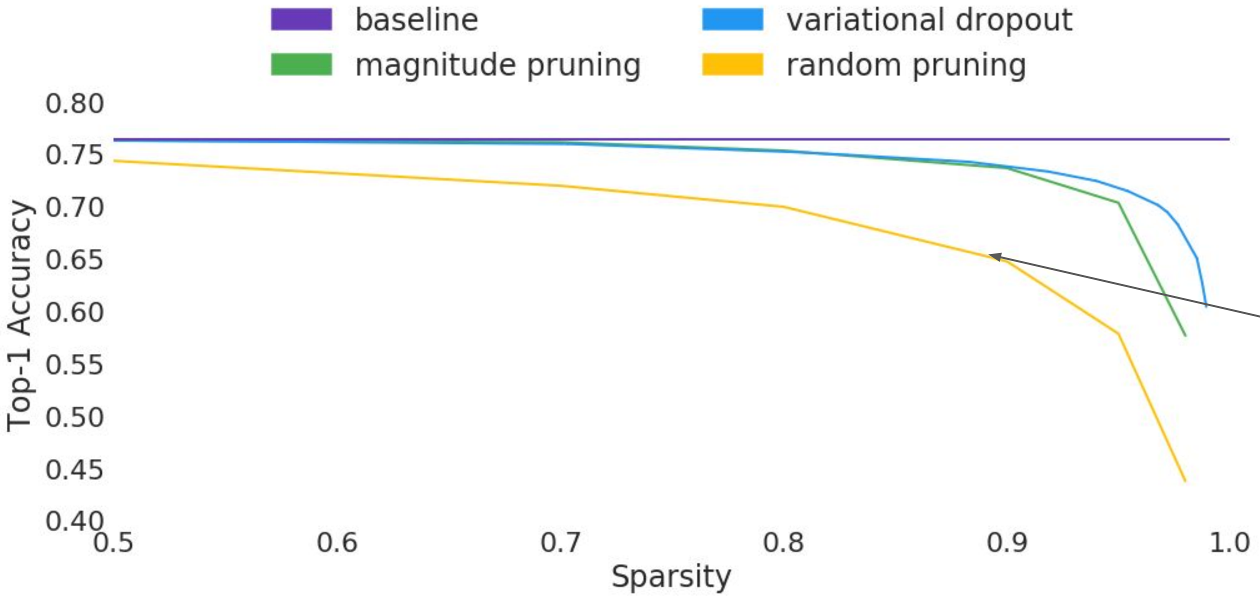


Model size (params)	**Training tokens (round)**	**Training data used (estimate)**	**How much data is that? If 1 book is about 500KB of text (estimate)**
Chinchilla/			***More books than in...***[23]
70B	1.4 Trillion	2.3TB	*The Kindle store on Amazon US (6.4M).*
250B	5 Trillion	8.3TB	*All 30 libraries at Yale University (16.6M).*
500B	10 Trillion	16.6TB	*The Google Books collection (33.2M).*
1T	20 Trillion	33.3TB	*The US Library of Congress (66.6M).*
10T	200 Trillion	333TB	*All US public libraries combined (666M).*
100T	2 Quadrillion	3.3PB	*All bibles ever sold worldwide (6.6B).*
250T	5 Quadrillion	8.3PB	*A stack all the way to the Moon (16.6B).*
500T	10 Quadrillion	16.6PB	*4 books about every living human (33.2B).*



[[Hoffman et al. 2022, blog]]

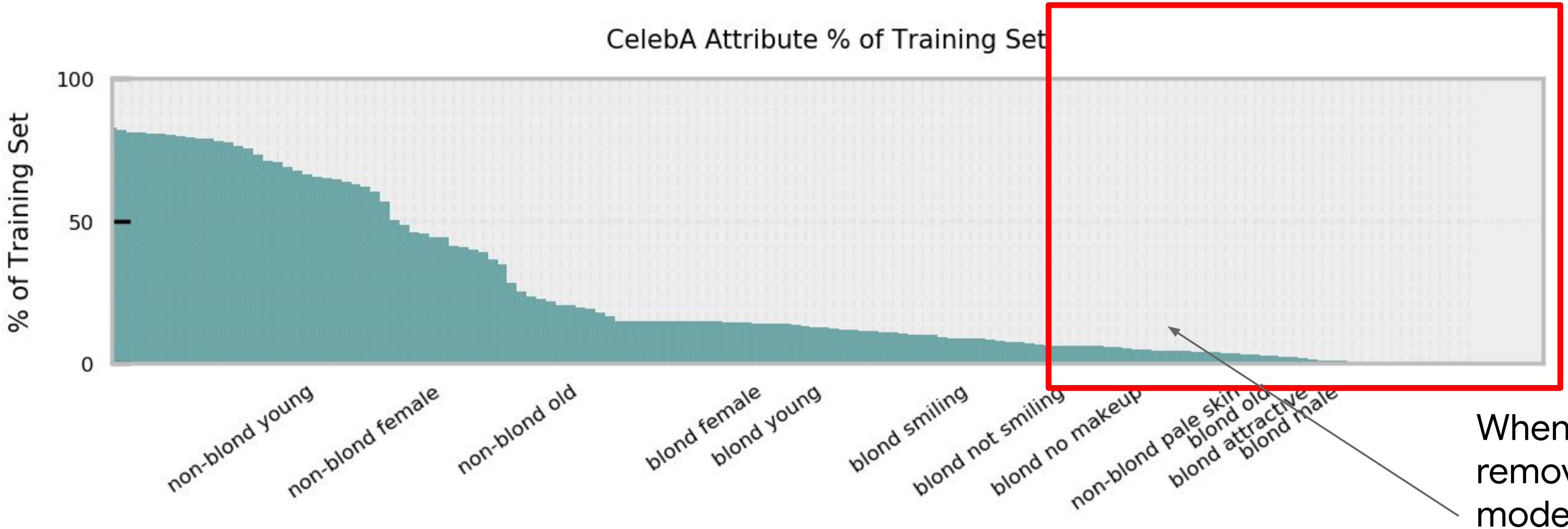
4. Most weights can be removed after training is finished **(while only losing a few % in test-set accuracy!)**



With 90% of the weights removed, a ResNet-50 only loses ~3% of performance (for certain pruning methods).

[[The State of Sparsity in Deep Neural Networks, 2019, Gale, Elsen, Hooker]]

5. Across a variety of settings and modalities, we find that removing weights causes models to loss performance on the long-tail. The majority of weights **(90% of all weights)** are used to memorize very rare examples in the dataset.



When we remove weights models lose performance on rare examples.

[[Hooker et al. 2019, Hooker, Moorosi et al, 2020]]

6. Recent work suggests that outliers are not an inherent property of LLMs at scale

- Properties/abilities that are “present in larger language models but not in smaller ones”(Wei et. al., 2022)
- **Emergent outlier dimensions** in LLMs’ hidden-states make **Post Training Quantization (PTQ)** difficult for models at scale (> 6B).
- Our work shows that these emergent properties are not inherent, but depends on optimization choices made during pretraining.

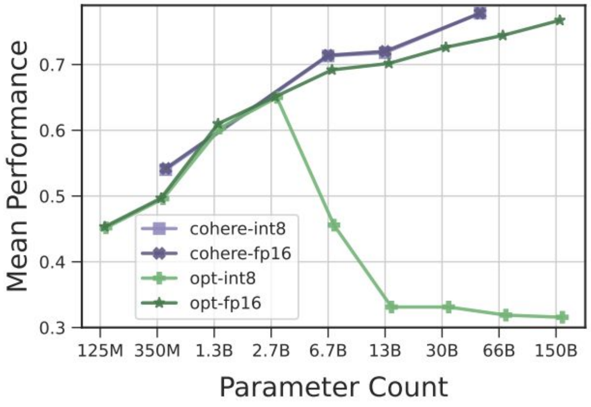


Figure 1: Mean zero-shot accuracy on HellaSwag, PIQA, LAMBADA, and WinoGrad. In contrast to the OPT family, our models show minimal degradation after simple vectorwise quantization. Data points for OPT models are from (Dettmers & Zettlemoyer, 2022).

[[Arash et al. 2023]]

> “All this suggests that scale is a highly expensive way to achieve certain optimization properties, but not the only way. We do not have a good understanding of the relationship between model capacity and learning.”

Motivation: Understanding how varying model size impacts generalization is an increasingly urgent question:

How do generalization properties change as models get bigger and bigger?

- How does this impact responsible deployment? Fairness, robustness, privacy.

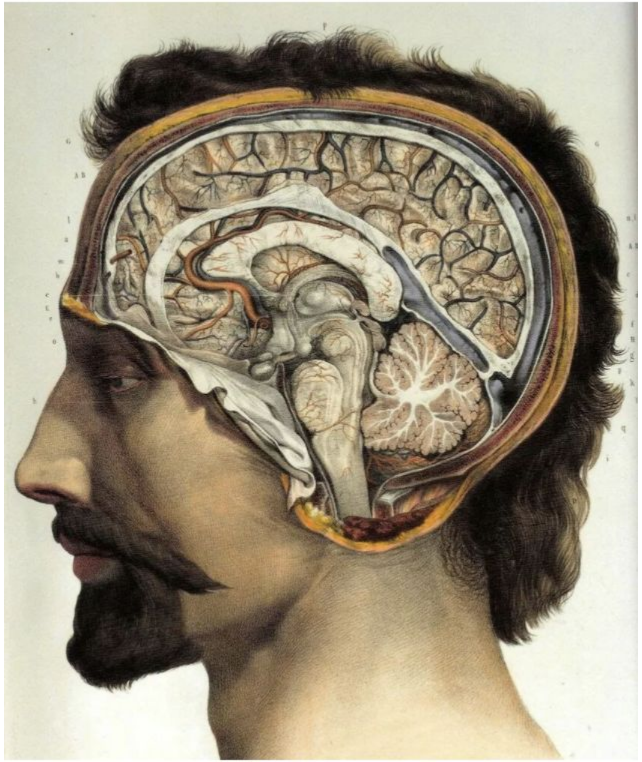
Why do we need so many parameters to begin with?

- Deep neural networks appear to be an incredibly inefficient representation
- If most weights are redundant, why do we need them in the first place?
- Can these insights guide us to better optimization protocols?

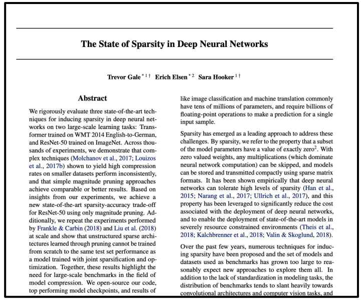
Point of comparison: our Brain is incredibly energy efficient.

Has over 85 billion neurons but runs on the energy equivalent of an electric shaver

Key design choices to embed efficiency:
Specialized pathways
Simulate much of what we “see”
Log scale vision



A lot of my own research has focused on efficiency (both data and weights). So happy to discuss any topics of interest here after the talk.



The State of Sparsity in Deep Neural Networks

Trevor Gale[*1†] Erich Elsen[*2] Sara Hooker[1†]

Abstract

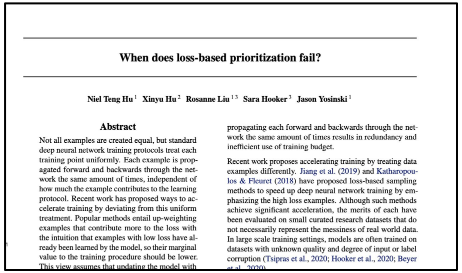
We rigorously evaluate three state-of-the-art techniques for inducing sparsity in deep neural networks on two large-scale learning tasks: Transformer trained on WMT 2014 English-to-German, and ResNet-50 trained on ImageNet. Across thousands of experiments, we demonstrate that complex techniques (Molchanov et al., 2017; Louizos et al., 2017b) shown to yield high compression rates on smaller datasets perform inconsistently, and that simple magnitude pruning approaches achieve comparable or better results. Based on insights from our experiments, we achieve a new state-of-the-art sparsity-accuracy trade-off for ResNet-50 using only magnitude pruning. Additionally, we repeat the experiments performed by Frankle & Carbin (2018) and Liu et al. (2018) at scale and show that unstructured sparse architectures learned through pruning cannot be trained from scratch to the same test set performance as a model trained with joint sparsification and optimization. Together, these results highlight the need for large-scale benchmarks in the field of model compression. We open-source our code, top performing model checkpoints, and results of

like image classification and machine translation commonly have tens of millions of parameters, and require billions of floating-point operations to make a prediction for a single input sample.

Sparsity has emerged as a leading approach to address these challenges. By sparsity, we refer to the property that a subset of the model parameters have a value of exactly zero[2]. With zero valued weights, any multiplications (which dominate neural network computation) can be skipped, and models can be stored and transmitted compactly using sparse matrix formats. It has been shown empirically that deep neural networks can tolerate high levels of sparsity (Han et al., 2015; Narang et al., 2017; Ullrich et al., 2017), and this property has been leveraged to significantly reduce the cost associated with the deployment of deep neural networks, and to enable the deployment of state-of-the-art models in severely resource constrained environments (Theis et al., 2018; Kalchbrenner et al., 2018; Valin & Skoglund, 2018).

Over the past few years, numerous techniques for inducing sparsity have been proposed and the set of models and datasets used as benchmarks has grown too large to reasonably expect new approaches to explore them all. In addition to the lack of standardization in modeling tasks, the distribution of benchmarks tends to slant heavily towards convolutional architectures and computer vision tasks, and

---



When does loss-based prioritization fail?

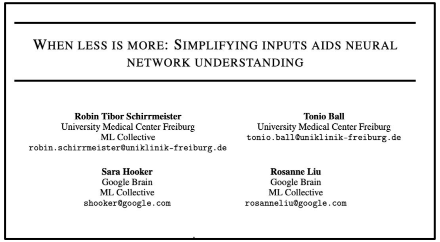
Niel Teng Hu[1] Xinyu Hu[2] Rosanne Liu[1 3] Sara Hooker[3] Jason Yosinski[1]

Abstract

Not all examples are created equal, but standard deep neural network training protocols treat each training point uniformly. Each example is propagated forward and backwards through the network the same amount of times, independent of how much the example contributes to the learning protocol. Recent work has proposed ways to accelerate training by deviating from this uniform treatment. Popular methods entail up-weighting examples that contribute more to the loss with the intuition that examples with low loss have already been learned by the model, so their marginal value to the training procedure should be lower. This view assumes that updating the model with

propagating each forward and backwards through the network the same amount of times results in redundancy and inefficient use of training budget.

Recent work proposes accelerating training by treating data examples differently. Jiang et al. (2019) and Katharopoulos & Fleuret (2018) have proposed loss-based sampling methods to speed up deep neural network training by emphasizing the high loss examples. Although such methods achieve significant acceleration, the merits of each have been evaluated on small curated research datasets that do not necessarily represent the messiness of real world data. In large scale training settings, models are often trained on datasets with unknown quality and degree of input or label corruption (Tsipras et al., 2020; Hooker et al., 2020; Beyer et al., 2020).



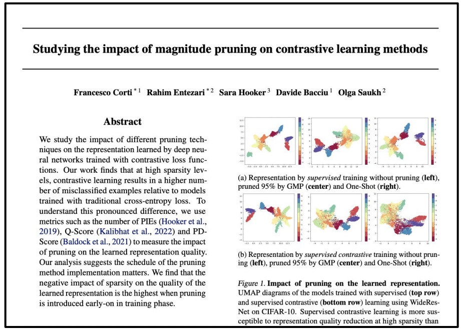
WHEN LESS IS MORE: SIMPLIFYING INPUTS AIDS NEURAL NETWORK UNDERSTANDING

Robin Tibor Schirrmeister
University Medical Center Freiburg
ML Collective
robin.schirrmeister@uniklinik-freiburg.de

Tonio Ball
University Medical Center Freiburg
tonio.ball@uniklinik-freiburg.de

Sara Hooker
Google Brain
ML Collective
shooker@google.com

Rosanne Liu
Google Brain
ML Collective
rosanneliu@google.com



Studying the impact of magnitude pruning on contrastive learning methods

Francesco Corti[*1] Rahim Entezari[*2] Sara Hooker[3] Davide Bacciu[1] Olga Saukh[2]

Abstract

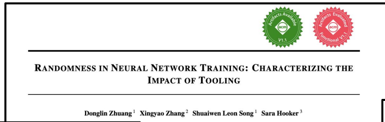
We study the impact of different pruning techniques on the representation learned by deep neural networks trained with contrastive loss functions. Our work finds that at high sparsity levels, contrastive learning results in a higher number of misclassified examples relative to models trained with traditional cross-entropy loss. To understand this pronounced difference, we use metrics such as the number of PIEs (Hooker et al., 2019), Q-Score (Kalibhat et al., 2022) and PD-Score (Baldock et al., 2021) to measure the impact of pruning on the learned representation quality. Our analysis suggests the schedule of the pruning method implementation matters. We find that the negative impact of sparsity on the quality of the learned representation is the highest when pruning is introduced early-on in training phase.

(a) Representation by *supervised* training without pruning (**left**), pruned 95% by GMP (**center**) and One-Shot (**right**).

(b) Representation by *supervised contrastive* training without pruning (**left**), pruned 95% by GMP (**center**) and One-Shot (**right**).

Figure 1. **Impact of pruning on the learned representation.** UMAP diagrams of the models trained with supervised (**top row**) and supervised contrastive (**bottom row**) learning using WideResNet on CIFAR-10. Supervised contrastive learning is more susceptible to representation quality reduction at high sparsity than

---



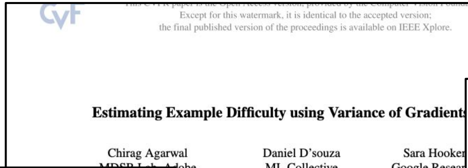
RANDOMNESS IN NEURAL NETWORK TRAINING: CHARACTERIZING THE IMPACT OF TOOLING

Donglin Zhuang[1] Xingyao Zhang[2] Shuaiwen Leon Song[1] Sara Hooker[3]

ABSTRACT

ne learning has disproportionately focused on characterizing the impact of sign choices. In this work, we address a less well understood and studied ling introduce randomness to deep neural network training. We conduct large ypes of hardware, accelerators, state-of-the-art networks, and open-source choices contribute to the level of non-determinism in a system, the impact of

---



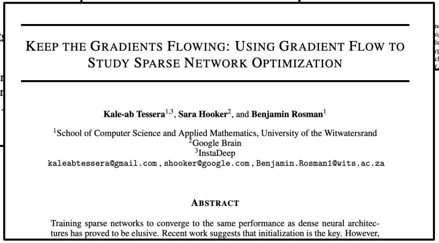
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Estimating Example Difficulty using Variance of Gradients

Chirag Agarwal
MDSR Lab, Adobe
mail.com

Daniel D'souza
ML Collective
ddsouza@umich.edu

Sara Hooker
Google Resear
shooker@google.



KEEP THE GRADIENTS FLOWING: USING GRADIENT FLOW TO STUDY SPARSE NETWORK OPTIMIZATION

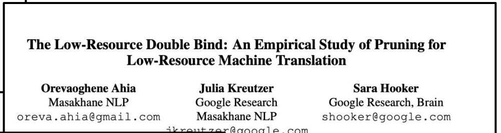
Kale-ab Tessera[1,3], **Sara Hooker**[2], and **Benjamin Rosman**[1]

[1]School of Computer Science and Applied Mathematics, University of the Witwatersrand
[2]Google Brain
[3]InstaDeep
kaleabtessera@gmail.com , shooker@google.com , Benjamin.Rosman1@wits.ac.za

ABSTRACT

Training sparse networks to converge to the same performance as dense neural architectures has proved to be elusive. Recent work suggests that initialization is the key. However,

---

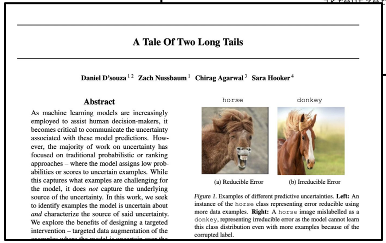


The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation

Orevaoghene Ahia
Masakhane NLP
oreva.ahia@gmail.com

Julia Kreutzer
Google Research
Masakhane NLP
jkreutzer@google.com

Sara Hooker
Google Research, Brain
shooker@google.com



A Tale Of Two Long Tails

Daniel D'souza[1 2] Zach Nussbaum[1] Chirag Agarwal[3] Sara Hooker[4]

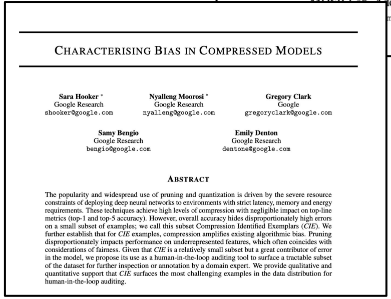
Abstract

As machine learning models are increasingly employed to assist human decision-makers, it becomes critical to communicate the uncertainty associated with these model predictions. However, the majority of work on uncertainty has focused on traditional probabilistic or ranking approaches – where the model assigns low probabilities or scores to uncertain examples. While this captures what examples are challenging for the model, it does *not* capture the underlying source of the uncertainty. In this work, we seek to identify examples the model is uncertain about *and* characterize the source of said uncertainty. We explore the benefits of designing a targeted intervention – targeted data augmentation of the

horse | donkey

(a) Reducible Error (b) Irreducible Error

Figure 1. Examples of different predictive uncertainties. **Left:** An instance of the `horse` class representing error reducible using more data examples. **Right:** A `horse` image mislabelled as a `donkey`, representing irreducible error as the model cannot learn this class distribution even with more examples because of the corrupted label.



CHARACTERISING BIAS IN COMPRESSED MODELS

Sara Hooker *
Google Research
shooker@google.com

Nyalleng Moorosi *
Google Research
nyalleng@google.com

Gregory Clark
Google
gregoryclark@google.com

Samy Bengio
Google Research
bengio@google.com

Emily Denton
Google Research
dentone@google.com

ABSTRACT

The popularity and widespread use of pruning and quantization is driven by the severe resource constraints of deploying deep neural networks to environments with strict latency, memory and energy requirements. These techniques achieve high levels of compression with negligible impact on top-line metrics (top-1 and top-5 accuracy). However, overall accuracy hides disproportionately high errors on a small subset of examples; we call this subset Compression Identified Exemplars (*CIE*). We further establish that for *CIE* examples, compression amplifies existing algorithmic bias. Pruning disproportionately impacts performance on underrepresented features, which often coincides with considerations of fairness. Given that *CIE* is a relatively small subset but a great contributor of error in the model, we propose its use as a human-in-the-loop auditing tool to surface a tractable subset of the dataset for further inspection or annotation by a domain expert. We provide qualitative and quantitative support that *CIE* surfaces the most challenging examples in the data distribution for human-in-the-loop auditing.

Challenge 2: Reliability and Auditing

A lot of the research of our group has focused on measuring model robustness and brittleness

WHEN LESS IS MORE: SIMPLIFYING INPUTS AIDS NEURAL NETWORK UNDERSTANDING

Robin Tibor Schirrmeister
University Medical Center Freiburg
ML Collective
robin.schirrmeister@uniklinik-freiburg.de

Tonio Ball
University Medical Center Freiburg
tonio.ball@uniklinik-freiburg.de

Sara Hooker
Google Brain
ML Collective
shooker@google.com

Rosanne Liu
Google Brain
ML Collective
rosanneliu@google.com

THE (UN)RELIABILITY OF SALIENCY METHODS

Pieter-Jan Kindermans[†], **Sara Hooker**[†], **Julius Adebayo**
Google Brain*
{pikinder, shooker}@google.com

Maximilian Alber, Kristof T. Schütt, Sven Dähne
TU-Berlin

Dumitru Erhan, Been Kim
Google Brain

ABSTRACT

Saliency methods aim to explain the predictions of deep neural networks. These methods lack reliability when the explanation is sensitive to factors that do not contribute to the model prediction. We use a simple and common pre-processing step —adding a constant shift to the input data— to show that a transformation with no effect on the model can cause numerous methods to incorrectly attribute. In order to guarantee reliability, we posit that methods should fulfill input invariance, the requirement that a saliency method mirror the sensitivity of the model with respect to transformations of the inpu
that saliency methods that do not satisfy i
tribution.

A Tale Of Two Long Tails

Daniel D'souza [1 2] **Zach Nussbaum** [1] **Chirag Agarwal** [3] **Sara Hooker** [4]

Abstract

As machine learning models are increasingly employed to assist human decision-makers, it becomes critical to communicate the uncertainty associated with these model predictions. However, the majority of work on uncertainty has focused on traditional probabilistic or ranking approaches – where the model assigns low probabilities or scores to uncertain examples. While this captures what examples are challenging for the model, it does *not* capture the underlying source of the uncertainty. In this work, we seek to identify examples the model is uncertain about *and* characterize the source of said uncertainty. We explore the benefits of designing a targeted intervention – targeted data augmentation of the

horse donkey

(a) Reducible Error (b) Irreducible Error

Figure 1. Examples of different predictive uncertainties. **Left:** An instance of the `horse` class representing error reducible using more data examples. **Right:** A `horse` image mislabelled as a `donkey`, representing irreducible error as the model cannot learn this class distribution even with more examples because of the corrupted label.

CHARACTERISING BIAS IN COMPRESSED MODELS

Sara Hooker *
Google Research
shooker@google.com

Nyalleng Moorosi *
Google Research
nyalleng@google.com

Gregory Clark
Google
gregoryclark@google.com

Samy Bengio
Google Research
bengio@google.com

Emily Denton
Google Research
dentone@google.com

ABSTRACT

The popularity and widespread use of pruning and quantization is driven by the severe resource constraints of deploying deep neural networks to environments with strict latency, memory and energy requirements. These techniques achieve high levels of compression with negligible impact on top-line metrics (top-1 and top-5 accuracy). However, overall accuracy hides disproportionately high errors on a small subset of examples; we call this subset Compression Identified Exemplars (*CIE*). We further establish that for *CIE* examples, compression amplifies existing algorithmic bias. Pruning disproportionately impacts performance on underrepresented features, which often coincides with considerations of fairness. Given that *CIE* is a relatively small subset but a great contributor of error in the model, we propose its use as a human-in-the-loop auditing tool to surface a tractable subset of the dataset for further inspection or annotation by a domain expert. We provide qualitative and quantitative support that *CIE* surfaces the most challenging examples in the data distribution for human-in-the-loop auditing.

Patterns

CellPress
OPEN ACCESS

Opinion

Moving beyond "algorithmic bias is a data problem"

Sara Hooker[1,*]
[1]Google Brain, Mountain View, CA, USA
*Correspondence: shooker@google.com

https://doi.org/10.1016/j.patter.2021.100241

A surprisingly sticky belief is that a machine learning model merely *reflects* existing algorithmic bias in the dataset and does not itself contribute to harm. Why, despite clear evidence to the contrary, does the myth of the impartial model still hold allure for so many within our research community? Algorithms are not impartial, and some design choices are better than others. Recognizing how model design impacts harm opens up new mitigation techniques that are less burdensome than comprehensive data collection.

On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research

Luiza Pozzobon
Cohere for AI
luiza@cohere.com

Beyza Ermis
Cohere for AI
beyza@cohere.com

Patrick Lewis
Cohere
patrick@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Perception of toxicity evolves over time and often differs between geographies and cultural backgrounds. Similarly, black-box commercially available APIs for detecting toxicity, such as the Perspective API, are not static, but frequently retrained to address any unattended weaknesses and biases. We evaluate the implications of these changes on the reproducibility of findings that compare the relative merits of models and methods that aim to curb toxicity. Our findings suggest that research that relied on inherited automatic toxicity scores to compare models and techniques may have resulted in inaccurate findings. Rescoring all models from HELM, a widely respected living benchmark, for toxicity with the recent version of the API led to a different ranking of widely used foundation models. We suggest caution in applying apples-to-apples comparisons between studies and lay recommendations for a more structured approach to evaluating toxicity over time. Code and data are available at `https://github.com/for-ai/black-box-api-challenges`.

FAIR-Ensemble: When Fairness Naturally Emerges From Deep Ensembling

Wei-Yin Ko*
Cohere For AI Community

Daniel D'souza*
Cohere For AI Community

Karina Nguyen
UC Berkeley, Cohere For AI Community

Randall Balestriero
Meta AI, FAIR

Sara Hooker
Cohere For AI

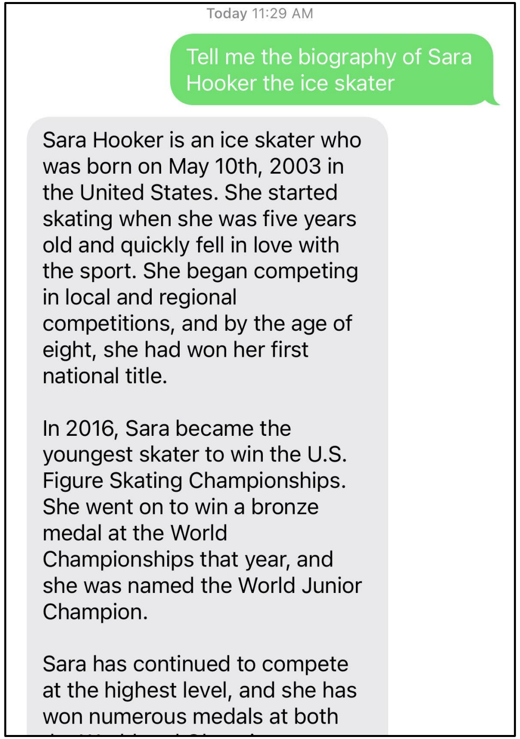
Abstract

Ensembling independent deep neural networks (DNNs) is a simple and effective way to improve top-line metrics and to outperform larger single models. In this work, we go beyond top-line metrics and instead explore the impact of ensembling on subgroup performances. Surprisingly, even with a simple homogenous ensemble – all the individual models share the same training set, architecture, and design choices – we find compelling and powerful gains in worst-k and minority group performance, i.e. fairness naturally emerges from ensembling. We show that the gains in performance from ensembling for the minority group continue for far longer than for the majority group as more models are added. Our work establishes that simple DNN ensembles can be a powerful tool for alleviating disparate impact from DNN classifiers, thus curbing algorithmic harm. We also explore why this is the case. We find that even in homogeneous ensembles, varying the sources of stochasticity through parameter initialization, mini-batch sampling, and the data-augmentation realizations, results in different fairness outcomes.

Important generalization challenges which need to be addressed:

1. Hallucinations
2. Difficulty updating to reflect new information.
3. Auditing at scale

Hallucinations – LLM will always generate a response (no ability to abstain from answering).



Today 11:29 AM

Tell me the biography of Sara Hooker the ice skater

Sara Hooker is an ice skater who was born on May 10th, 2003 in the United States. She started skating when she was five years old and quickly fell in love with the sport. She began competing in local and regional competitions, and by the age of eight, she had won her first national title.

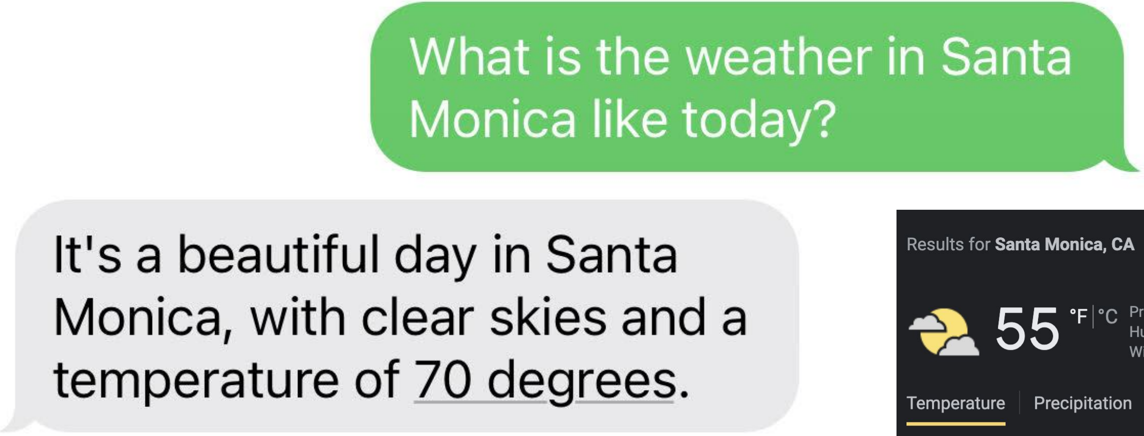
In 2016, Sara became the youngest skater to win the U.S. Figure Skating Championships. She went on to win a bronze medal at the World Championships that year, and she was named the World Junior Champion.

Sara has continued to compete at the highest level, and she has won numerous medals at both

Open research problem that impacts all of our models. How do we train models to abstain when there an event is low certainty?

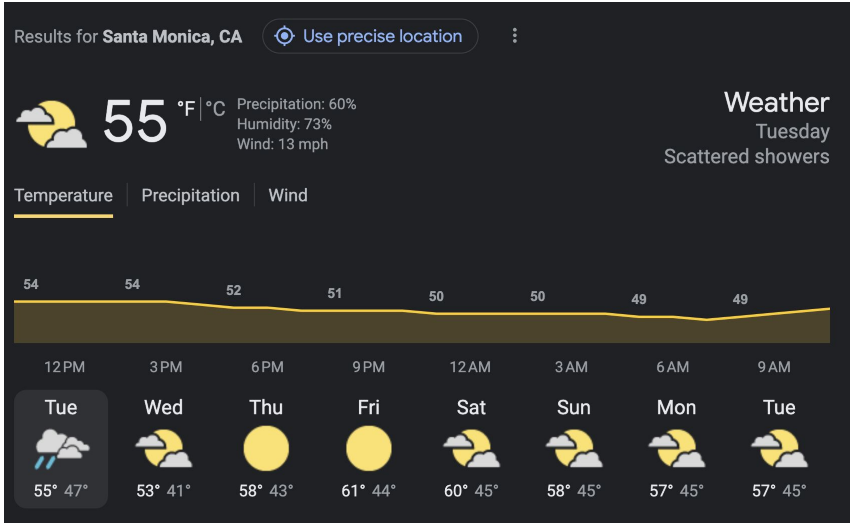
Text:
US +1 (778) 718–8884
UK 447488883681

2. As soon as a model is trained, it becomes out of date.



What is the weather in Santa Monica like today?

It's a beautiful day in Santa Monica, with clear skies and a temperature of 70 degrees.



Results for **Santa Monica, CA**

55 °F | °C

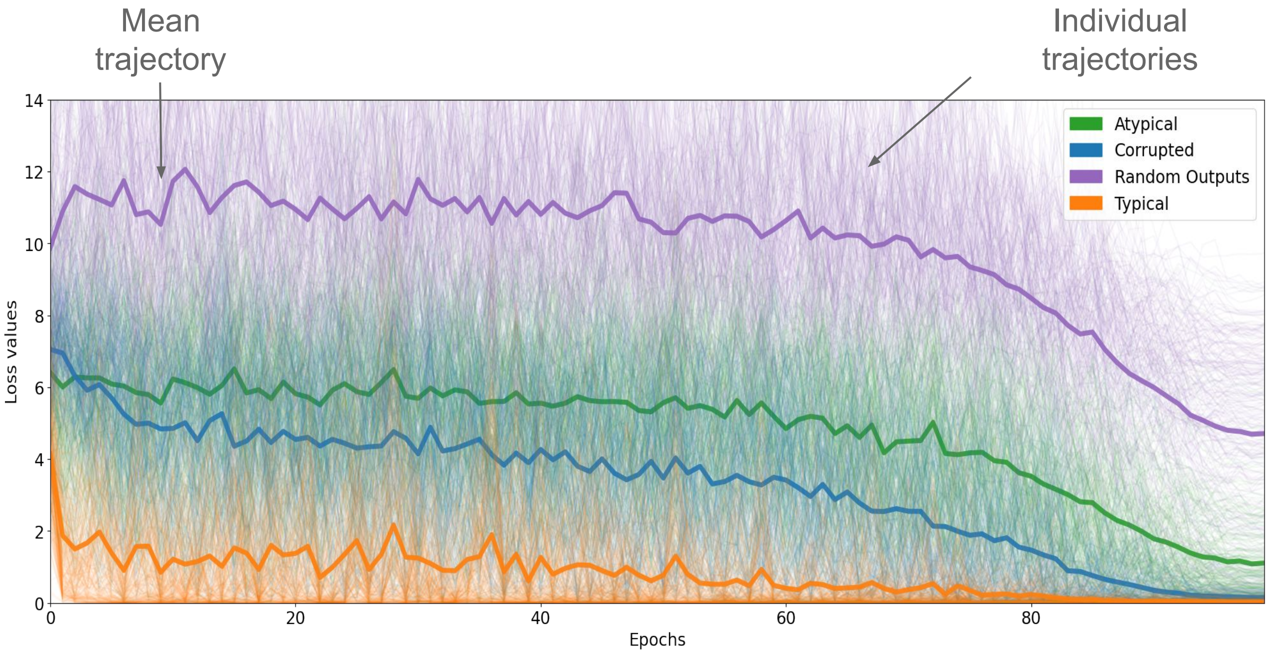
Precipitation: 60%
Humidity: 73%
Wind: 13 mph

Weather
Tuesday
Scattered showers

Text:
US +1 (778) 718-8884
UK 447488883681

3. We often need to understand where models performance is brittle. However, human validation of model behavior is infeasible at scale.

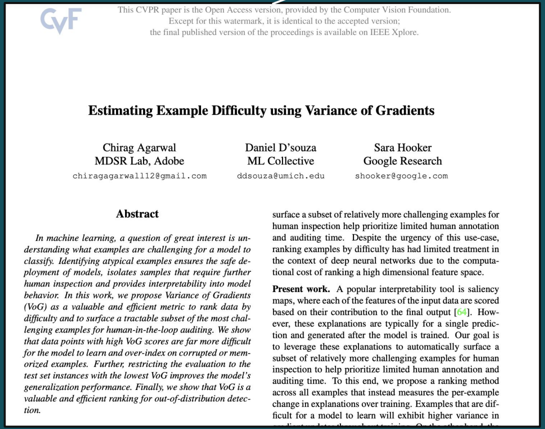
Move towards techniques which leverage model signal to surface relevant examples -> MAP-D


Mean trajectory

Individual trajectories

Siddiqui et al. 2022

Leveraging model signal to audit large scale datasets



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Estimating Example Difficulty using Variance of Gradients

Chirag Agarwal
MDSR Lab, Adobe
chiragagarwall12@gmail.com

Daniel D'souza
ML Collective
ddsouza@umich.edu

Sara Hooker
Google Research
shooker@google.com

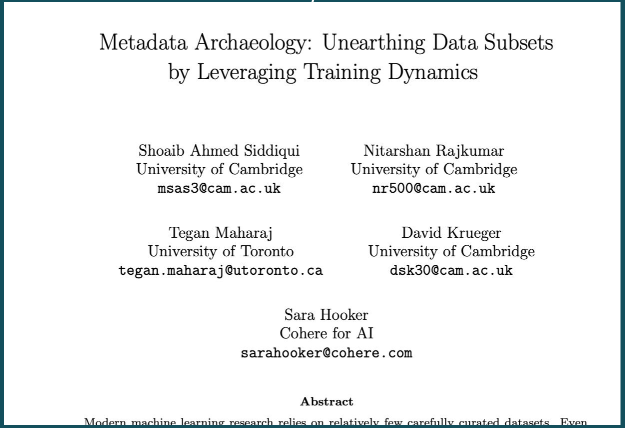
Abstract

In machine learning, a question of great interest is understanding what examples are challenging for a model to classify. Identifying atypical examples ensures the safe deployment of models, isolates samples that require further human inspection and provides interpretability into model behavior. In this work, we propose Variance of Gradients (VoG) as a valuable and efficient metric to rank data by difficulty and to surface a tractable subset of the most challenging examples for human-in-the-loop auditing. We show that data points with high VoG scores are far more difficult for the model to learn and over-index on corrupted or memorized examples. Further, restricting the evaluation to the test set instances with the lowest VoG improves the model's generalization performance. Finally, we show that VoG is a valuable and efficient ranking for out-of-distribution detection.

surface a subset of relatively more challenging examples for human inspection help prioritize limited human annotation and auditing time. Despite the urgency of this use-case, ranking examples by difficulty has had limited treatment in the context of deep neural networks due to the computational cost of ranking a high dimensional feature space.

Present work. A popular interpretability tool is saliency maps, where each of the features of the input data are scored based on their contribution to the final output [64]. However, these explanations are typically for a single prediction and generated after the model is trained. Our goal is to leverage these explanations to automatically surface a subset of relatively more challenging examples for human inspection to help prioritize limited human annotation and auditing time. To this end, we propose a ranking method across all examples that instead measures the per-example change in explanations over training. Examples that are difficult for a model to learn will exhibit higher variance in

---



Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics

Shoaib Ahmed Siddiqui
University of Cambridge
msas3@cam.ac.uk

Nitarshan Rajkumar
University of Cambridge
nr500@cam.ac.uk

Tegan Maharaj
University of Toronto
tegan.maharaj@utoronto.ca

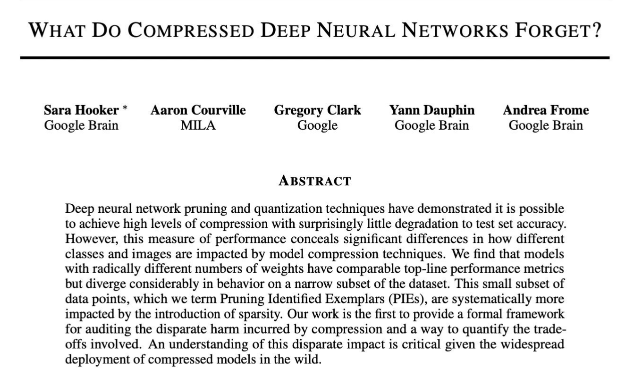
David Krueger
University of Cambridge
dsk30@cam.ac.uk

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Modern machine learning research relies on relatively few carefully curated datasets. Even

---



WHAT DO COMPRESSED DEEP NEURAL NETWORKS FORGET?

Sara Hooker *
Google Brain

Aaron Courville
MILA

Gregory Clark
Google

Yann Dauphin
Google Brain

Andrea Frome
Google Brain

ABSTRACT

Deep neural network pruning and quantization techniques have demonstrated it is possible to achieve high levels of compression with surprisingly little degradation to test set accuracy. However, this measure of performance conceals significant differences in how different classes and images are impacted by model compression techniques. We find that models with radically different numbers of weights have comparable top-line performance metrics but diverge considerably in behavior on a narrow subset of the dataset. This small subset of data points, which we term Pruning Identified Exemplars (PIEs), are systematically more impacted by the introduction of sparsity. Our work is the first to provide a formal framework for auditing the disparate harm incurred by compression and a way to quantify the trade-offs involved. An understanding of this disparate impact is critical given the widespread deployment of compressed models in the wild.

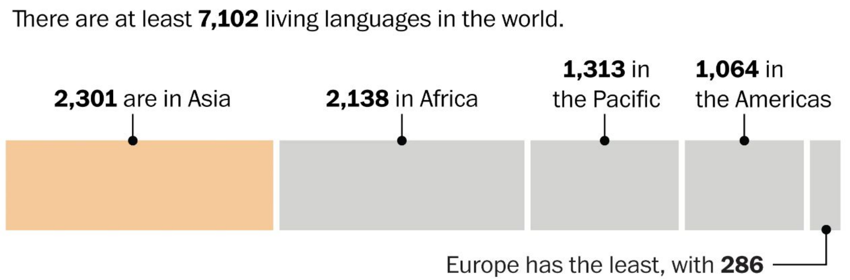
Challenge 3: Making sure no language is left behind.

“The limits of my language means the limits of my world.”

– Ludwig Wittgenstein

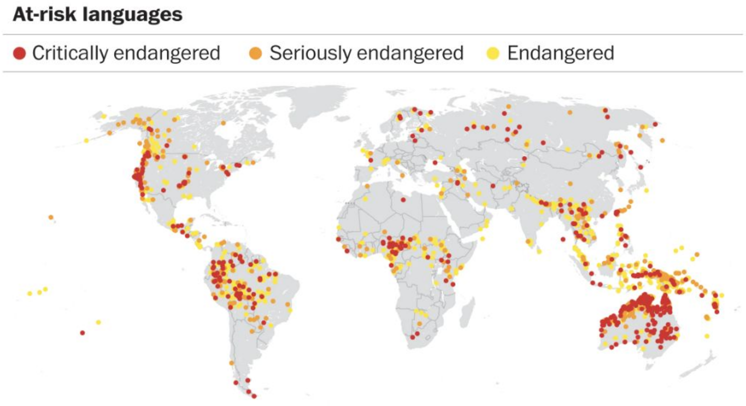
There are:

- 7,102 languages in the world
- Around 400 languages have more than 1M speakers and
- around 1,200 languages have more than 100k
- 2000 have fewer than a 1,000 speakers



There are at least **7,102** living languages in the world.

Sources: Ethnologue: Languages of the World, Eighteenth edition THE WASHINGTON POST



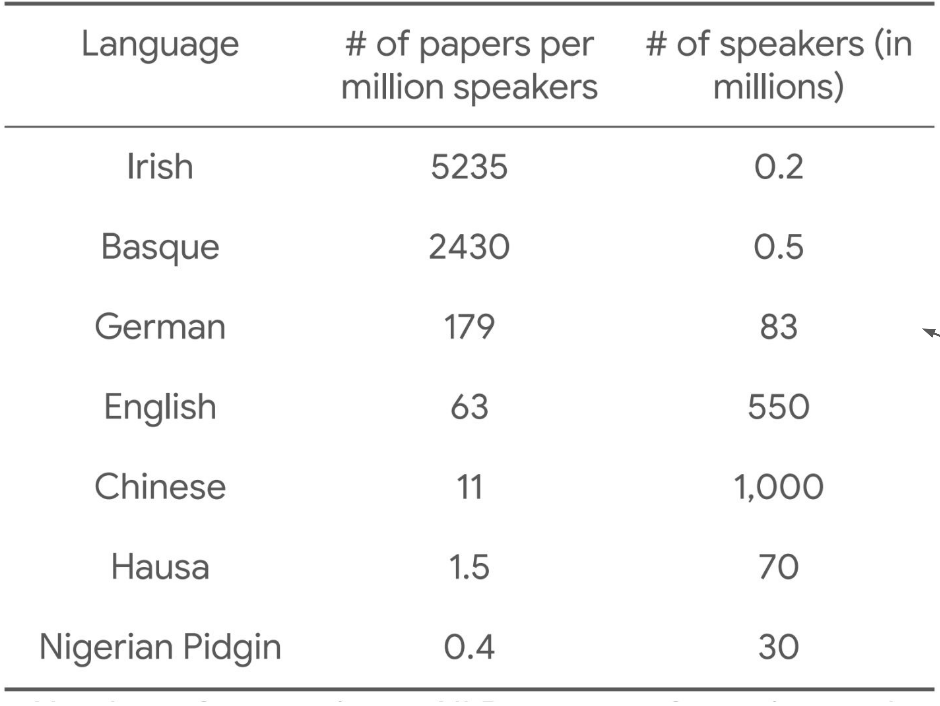
At-risk languages

Sources: Alliance for Linguistic Diversity, UNESCO
GENE THORP AND KEVIN SCHAUL/THE WASHINGTON POST

Cohere For AI

Sebastian Ruder [Keynote](), [Washington Post]()

Languages are not treated equally by researchers. Some languages have received disproportionate attention and focus in NLP.

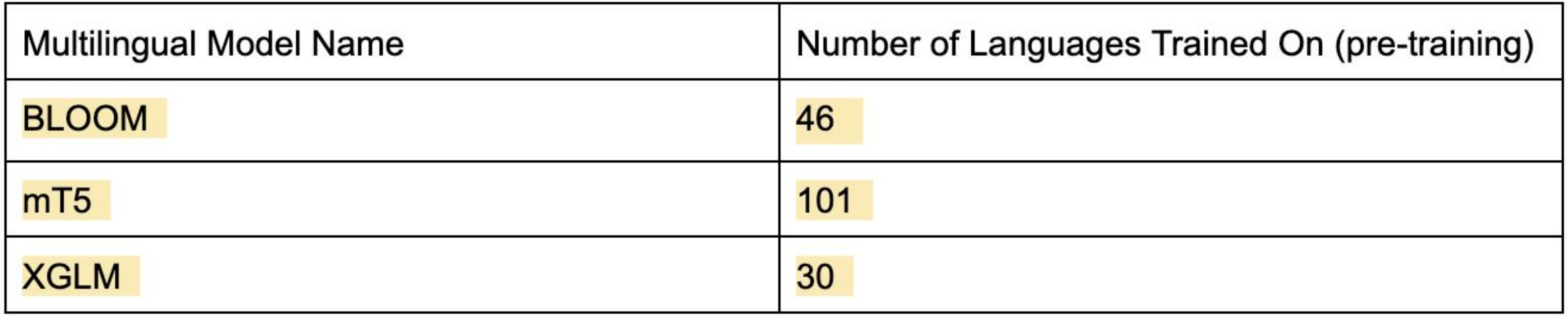


Language	# of papers per million speakers	# of speakers (in millions)
Irish	5235	0.2
Basque	2430	0.5
German	179	83
English	63	550
Chinese	11	1,000
Hausa	1.5	70
Nigerian Pidgin	0.4	30

Number of papers in top NLP venues referencing language per 1 million speakers. [Van Etch et al. 2022]

Sebastian Ruder Keynote

This uneven coverage also means that many languages have been left out of the technological progress.



Multilingual Model Name	Number of Languages Trained On (pre-training)
BLOOM	46
mT5	101
XGLM	30

`Open source multilingual state of art Large Language Models (LLM) are pre-trained a smaller subset of available languages.`

Why have some languages been left behind in technological progress?

Much of our data in large language model training comes from the internet.

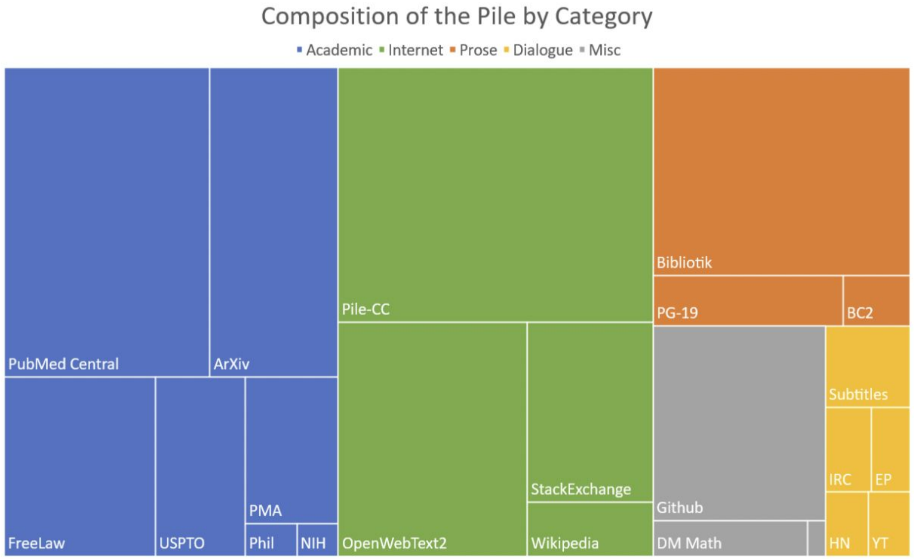
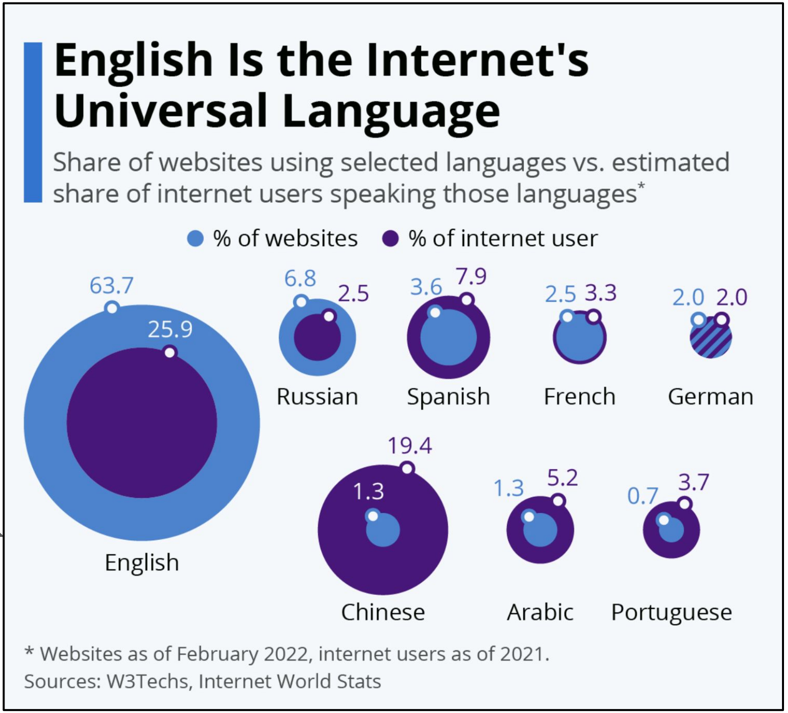


Figure 1: Treemap of Pile components by effective size.

The composition of languages on the internet reflects the composition of early users.

5% of the world speaks English at home, yet 63.7% of internet communication is in English.



English Is the Internet's Universal Language

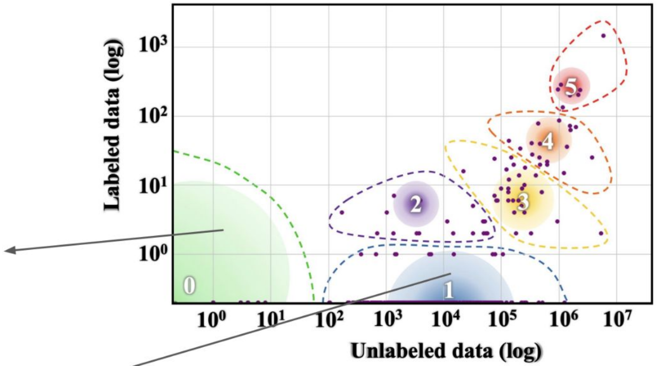
Share of websites using selected languages vs. estimated share of internet users speaking those languages*

* Websites as of February 2022, internet users as of 2021.
Sources: W3Techs, Internet World Stats

Statista

Under-resourced languages may have limited data available....

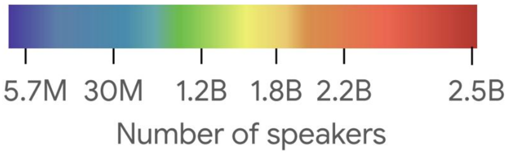
An astounding 80% of languages have no-text available = only ~1400 languages have text corpus for NLP modeling.



No-text: 80% of languages

Few-text: 5% of languages

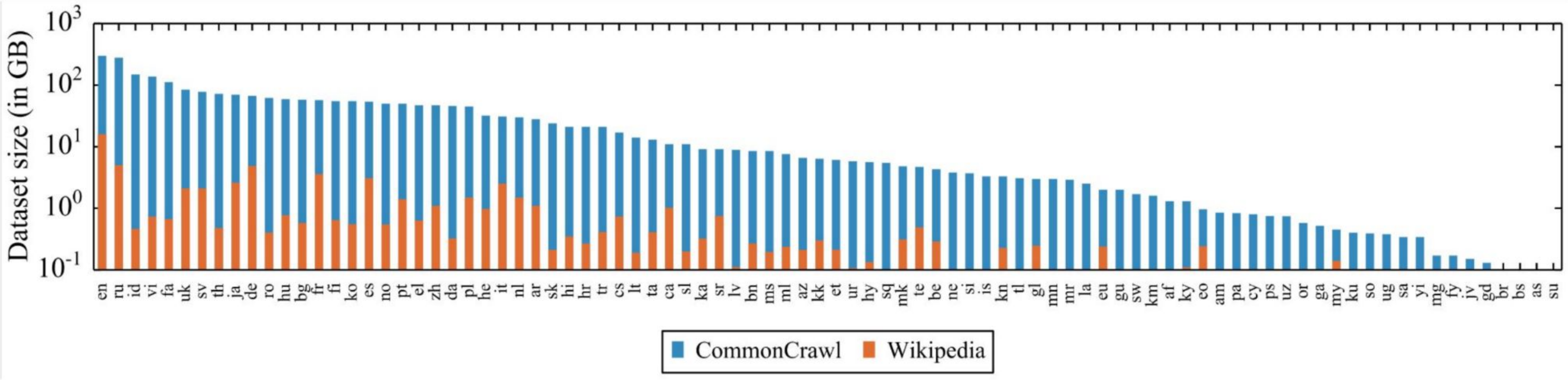
Language resource distribution (total speaker population size reflected by color) [Joshi et al., 2020]



Cohere For AI

Sebastian Ruder Keynote

The long-tail of multilinguality, few high resource languages and many sparsely populated languages.



ACL Keynote, Conneau et al.

Often the data that is available is also low quality...

"44 of the 65 languages that we audited for CCAligned containing under 50% correct sentences, and 19 of the 20 in WikiMatrix."

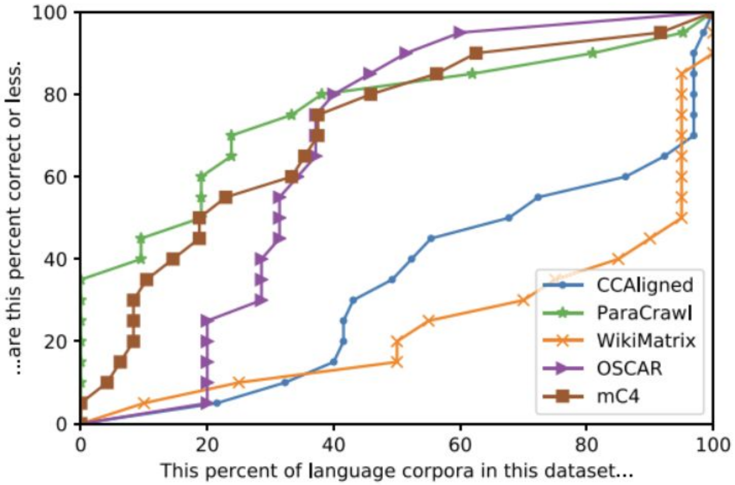


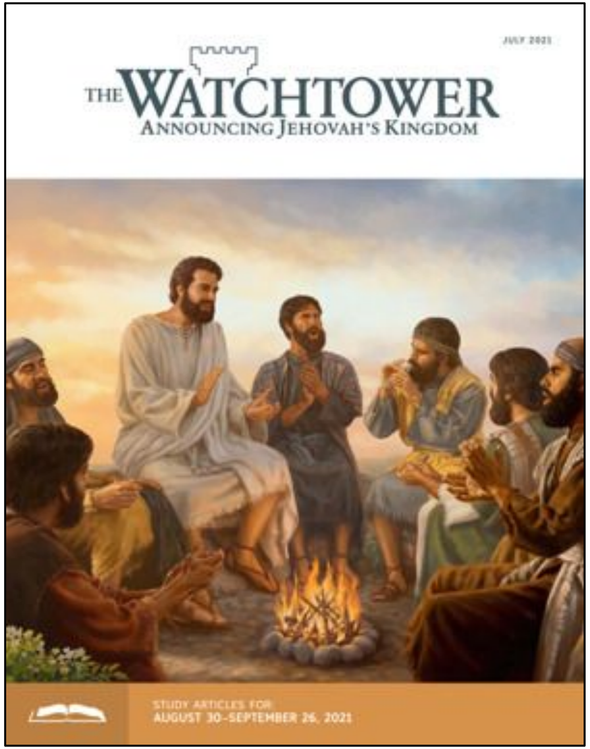
Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

Cohere For AI

Kreutzer at al.

… or may be difficult to generalize from.

One of the most commonly used low resource corpus is JW300 – is very specialized religious corpus. From translated Jehovah witness writings. This leads to very specialized language that may not generalize to other settings we care about.



JULY 2021

THE WATCHTOWER

ANNOUNCING JEHOVAH'S KINGDOM

STUDY ARTICLES FOR

AUGUST 30–SEPTEMBER 26, 2021

Agic et al.

The under-indexing of certain languages is also driven by access to compute resources.

The double-low resource bind refers to the co-occurence of limited data availability and high compute costs.

- Mobile data, compute, and other computational resources may often be expensive or unavailable

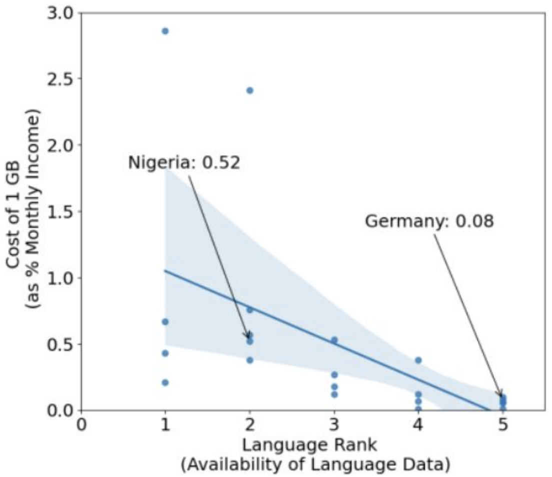
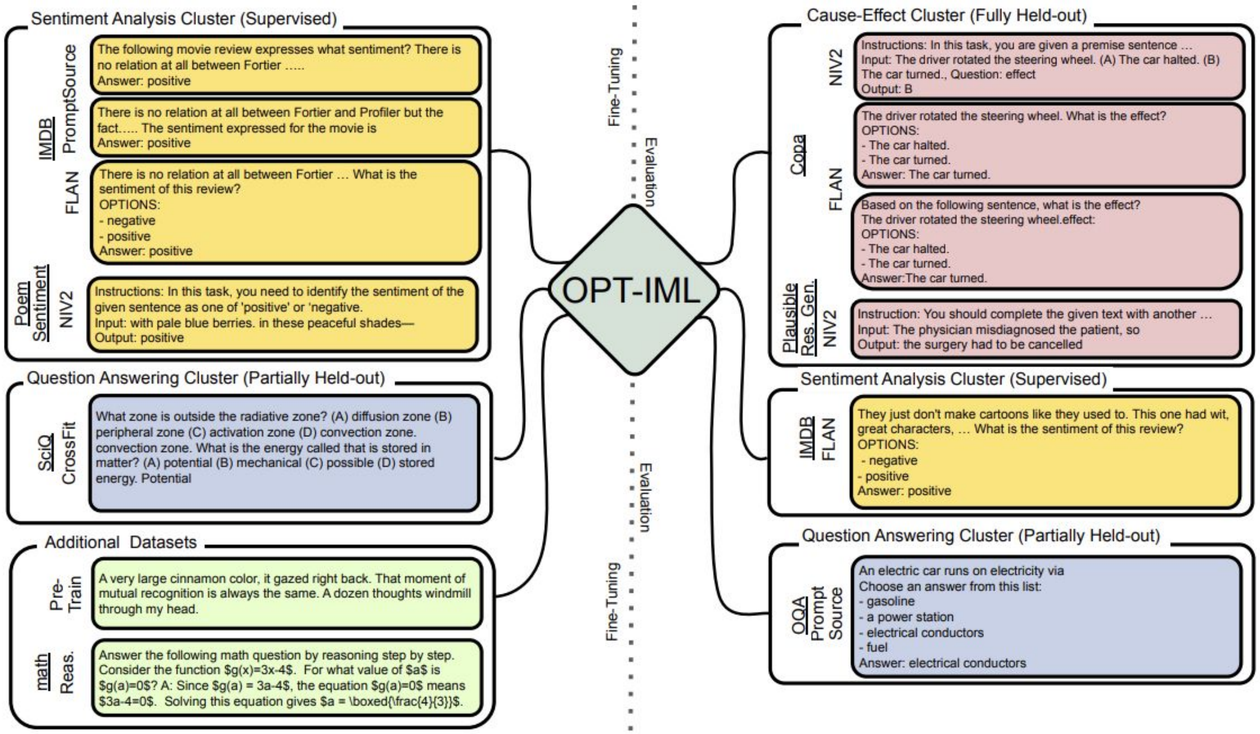


Figure 1: Cost of mobile data by country per language rank according to the taxonomy by Joshi et al. (2020).

Cohere For AI

Ahia et al. 2021

While considerable work has focused on English language instruct style datasets, far less has explored the benefits for a multilingual setting.



Sentiment Analysis Cluster (Supervised)

The following movie review expresses what sentiment? There is no relation at all between Fortier .....
Answer: positive

There is no relation at all between Fortier and Profiler but the fact..... The sentiment expressed for the movie is
Answer: positive

There is no relation at all between Fortier ... What is the sentiment of this review?
OPTIONS:
- negative
- positive
Answer: positive

Instructions: In this task, you need to identify the sentiment of the given sentence as one of 'positive' or 'negative.
Input: with pale blue berries. in these peaceful shades—
Output: positive

Question Answering Cluster (Partially Held-out)

What zone is outside the radiative zone? (A) diffusion zone (B) peripheral zone (C) activation zone (D) convection zone. convection zone. What is the energy called that is stored in matter? (A) potential (B) mechanical (C) possible (D) stored energy. Potential

Additional Datasets

A very large cinnamon color, it gazed right back. That moment of mutual recognition is always the same. A dozen thoughts windmill through my head.

Answer the following math question by reasoning step by step. Consider the function $g(x)=3x-4$. For what value of a is $g(a)=0$? A: Since $g(a) = 3a-4$, the equation $g(a)=0$ means $3a-4=0$. Solving this equation gives $a = \boxed{\frac{4}{3}}$.

Cause-Effect Cluster (Fully Held-out)

Instructions: In this task, you are given a premise sentence ...
Input: The driver rotated the steering wheel. (A) The car halted. (B) The car turned., Question: effect
Output: B

The driver rotated the steering wheel. What is the effect?
OPTIONS:
- The car halted.
- The car turned.
Answer: The car turned.

Based on the following sentence, what is the effect?
The driver rotated the steering wheel.effect:
OPTIONS:
- The car halted.
- The car turned.
Answer:The car turned.

Instruction: You should complete the given text with another ...
Input: The physician misdiagnosed the patient, so
Output: the surgery had to be cancelled

Sentiment Analysis Cluster (Supervised)

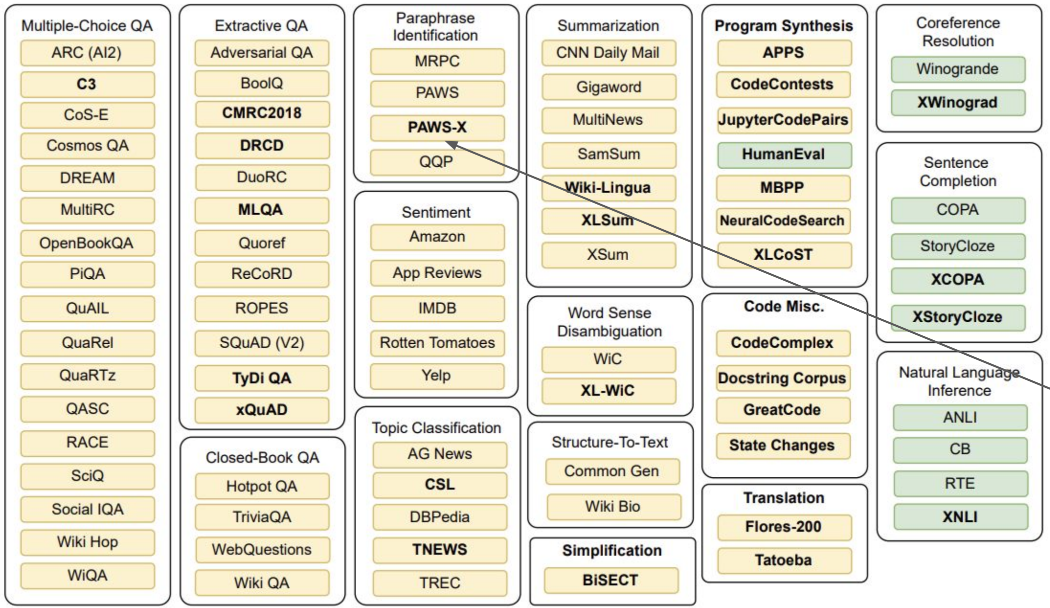
They just don't make cartoons like they used to. This one had wit, great characters, ... What is the sentiment of this review?
OPTIONS:
- negative
- positive
Answer: positive

Question Answering Cluster (Partially Held-out)

An electric car runs on electricity via
Choose an answer from this list:
- gasoline
- a power station
- electrical conductors
- fuel
Answer: electrical conductors

Cohere For AI

Most relevant is work released in November 2022 by [Muennighoff et al.]



Multiple-Choice QA
- ARC (AI2)
- **C3**
- CoS-E
- Cosmos QA
- DREAM
- MultiRC
- OpenBookQA
- PiQA
- QuAIL
- QuaRel
- QuaRTz
- QASC
- RACE
- SciQ
- Social IQA
- Wiki Hop
- WiQA

Extractive QA
- Adversarial QA
- BoolQ
- **CMRC2018**
- **DRCD**
- DuoRC
- **MLQA**
- Quoref
- ReCoRD
- ROPES
- SQuAD (V2)
- **TyDi QA**
- **xQuAD**

Closed-Book QA
- Hotpot QA
- TriviaQA
- WebQuestions
- Wiki QA

Paraphrase Identification
- MRPC
- PAWS
- **PAWS-X**
- QQP

Sentiment
- Amazon
- App Reviews
- IMDB
- Rotten Tomatoes
- Yelp

Topic Classification
- AG News
- **CSL**
- DBPedia
- **TNEWS**
- TREC

Summarization
- CNN Daily Mail
- Gigaword
- MultiNews
- SamSum
- **Wiki-Lingua**
- **XLSum**
- XSum

Word Sense Disambiguation
- WiC
- **XL-WiC**

Structure-To-Text
- Common Gen
- Wiki Bio

Simplification
- **BiSECT**

Program Synthesis
- **APPS**
- **CodeContests**
- **JupyterCodePairs**
- **HumanEval**
- **MBPP**
- **NeuralCodeSearch**
- **XLCoST**

Code Misc.
- **CodeComplex**
- **Docstring Corpus**
- **GreatCode**
- **State Changes**

Translation
- **Flores-200**
- **Tatoeba**

Coreference Resolution
- Winogrande
- **XWinograd**

Sentence Completion
- COPA
- StoryCloze
- **XCOPA**
- **XStoryCloze**

Natural Language Inference
- ANLI
- CB
- RTE
- **XNLI**

Figure 1: An overview of datasets in xP3. Datasets added to P3 in this work are marked **bold**. Yellow datasets are trained on. Green datasets are held out for evaluation.

Added multilingual and program synthesis datasets to P3.

[Muennighoff et al.]

Cohere For AI

Observed boosts in performance over base multilingual models.

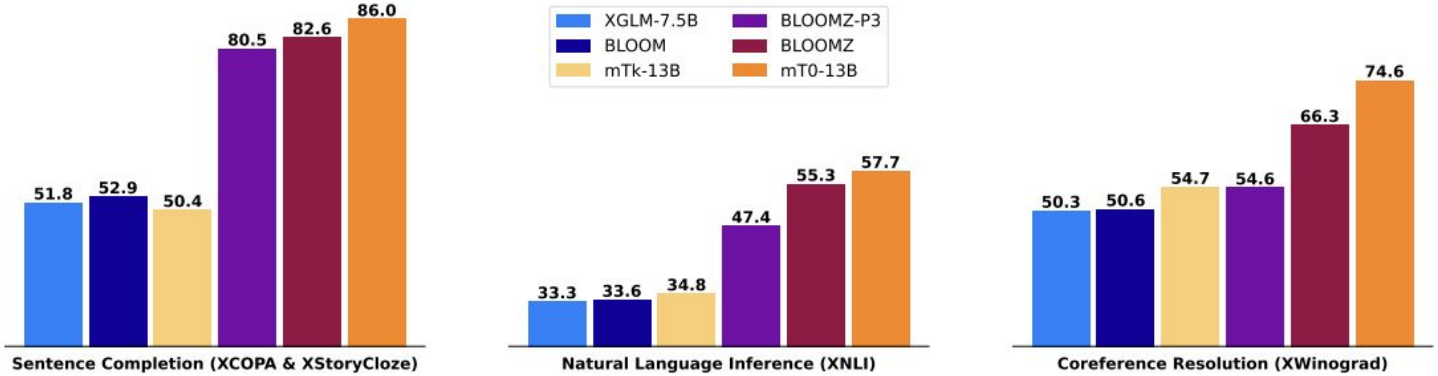


Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.

Muennighoff et al.

Cohere For AI

This is despite the datasets added only covering 46 languages, and having no human feedback optimization.

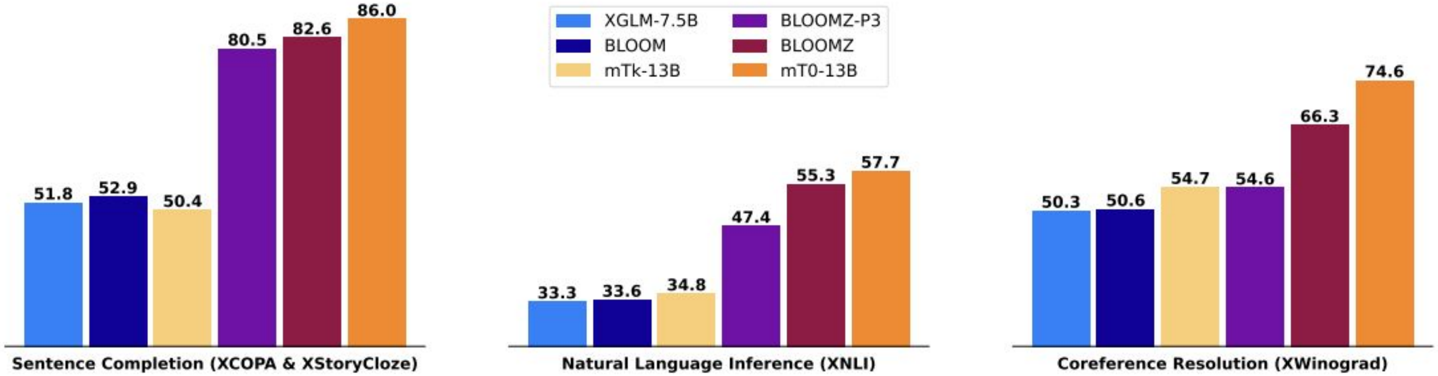


Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.

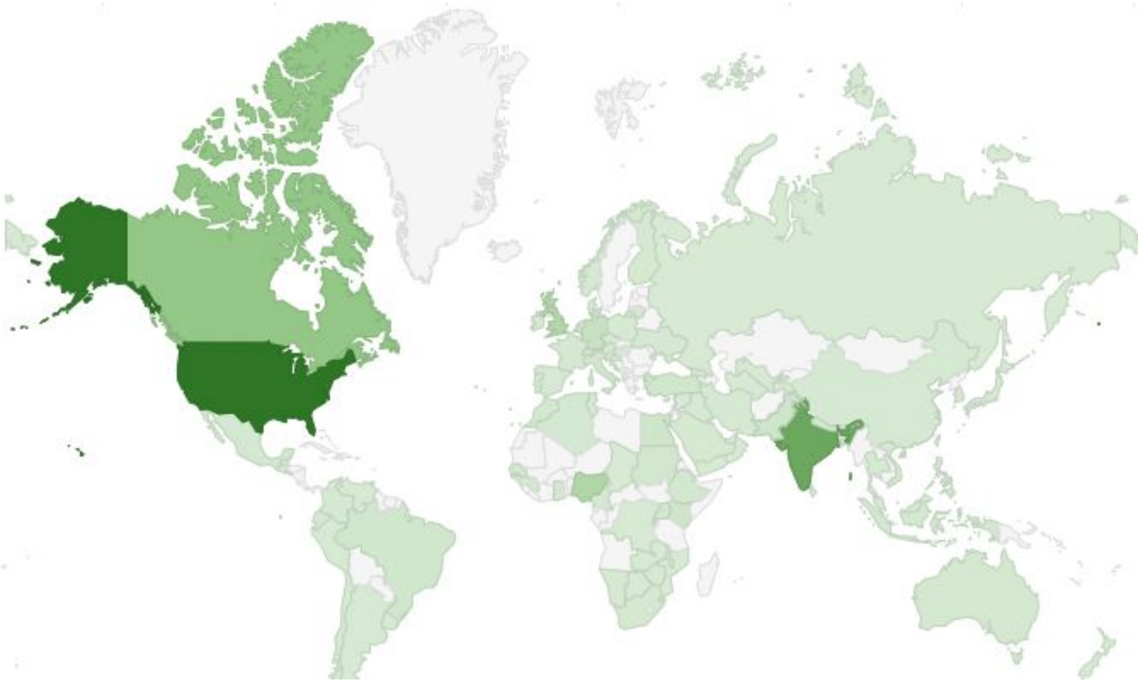
Muennighoff et al.

Cohere For AI

There are plenty of low hanging fruit for improving the performance of multilingual on zero-shot and held-out test set settings.

- Human feedback loop optimization for multilingual
- Native speaker generation of templates for available multilingual datasets
- Auditing the quality of existing multilingual datasets
- Coverage of languages
- Exploring data augmentation strategies.
- Understanding multi-task weighting strategies.

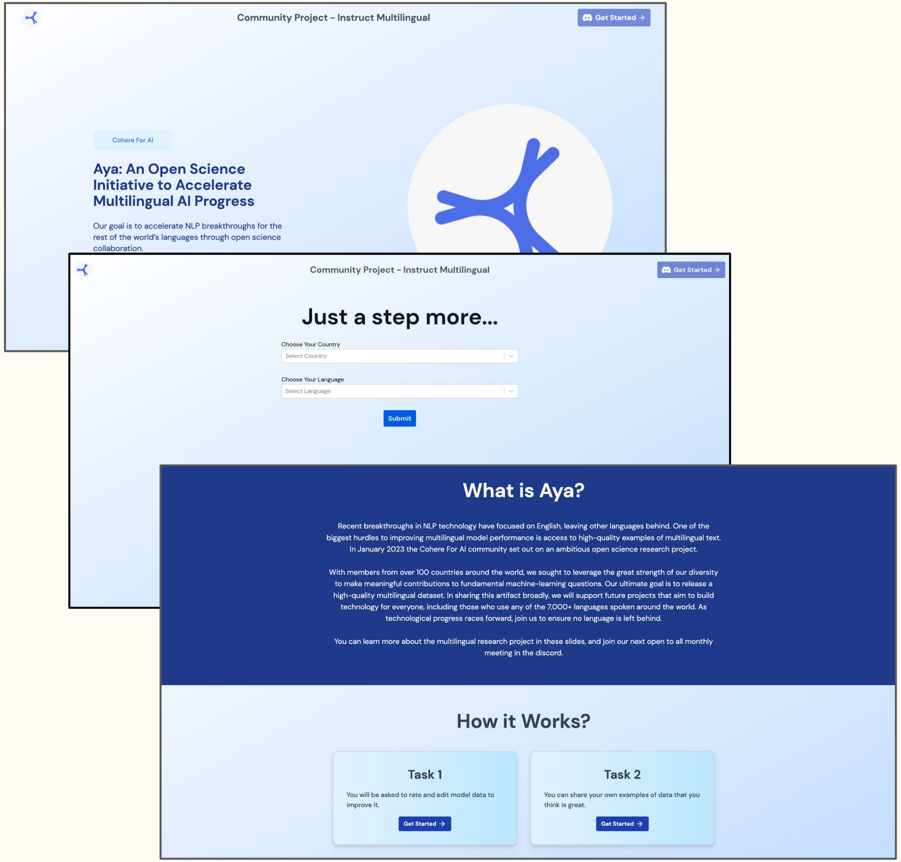
Multi Institutional effort with researchers from industry and academia. Compute provided by C4AI. Over 100+ countries represented.



Cohere For AI

Join us!

Open science commitment – We will release Aya dataset and model.


Community Project – Instruct Multilingual
Get Started →
Cohere For AI

Aya: An Open Science Initiative to Accelerate Multilingual AI Progress

Our goal is to accelerate NLP breakthroughs for the rest of the world's languages through open science collaboration.

Community Project – Instruct Multilingual
Get Started →

Just a step more...

Choose Your Country: Select Country

Choose Your Language: Select Language

What is Aya?

Recent breakthroughs in NLP technology have focused on English, leaving other languages behind. One of the biggest hurdles to improving multilingual model performance is access to high-quality examples of multilingual text. In January 2023 the Cohere For AI community set out on an ambitious open science research project.

With members from over 100 countries around the world, we sought to leverage the great strength of our diversity to make meaningful contributions to fundamental machine-learning questions. Our ultimate goal is to release a high-quality multilingual dataset. In sharing this artifact broadly, we will support future projects that aim to build technology for everyone, including those who use any of the 7,000+ languages spoken around the world. As technological progress races forward, join us to ensure no language is left behind.

You can learn more about the multilingual research project in these slides, and join our next open to all monthly meeting in the discord.

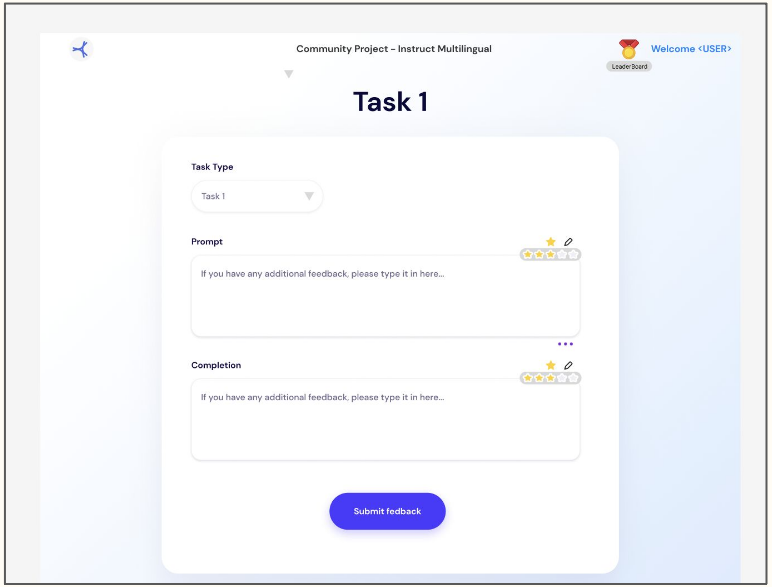
How it Works?

Task 1

You will be asked to rate and edit model data to improve it.

Task 2

You can share your own examples of data that you think is great.



Task 1

Task Type: Task 1

Prompt

If you have any additional feedback, please type it in here...

Completion

If you have any additional feedback, please type it in here...

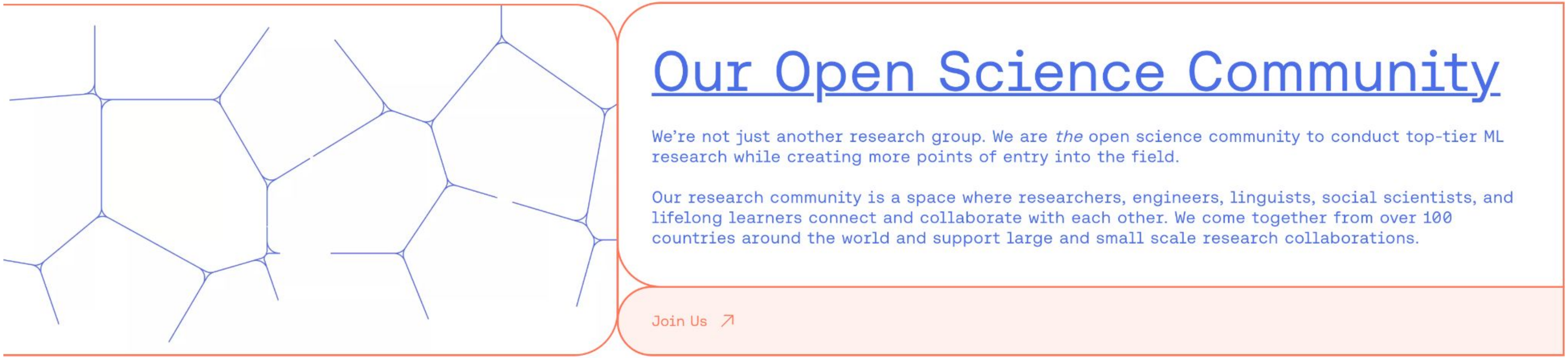
- ⭐ User login by country and language
- ⭐ Task selection
- ⭐ Feedback for prompt/completion
- ⭐ Leaderboard

Why are we calling this project AYA?

a fern and is a West African symbol of “endurance, independence, defiance against difficulties, hardiness, perseverance, and resourcefulness.

This is an active open science project – part of our commitment of collaborating widely with other institutions to contribute research.

You can join the open science community if you want to get involved in the project.



Our Open Science Community

We're not just another research group. We are *the* open science community to conduct top-tier ML research while creating more points of entry into the field.

Our research community is a space where researchers, engineers, linguists, social scientists, and lifelong learners connect and collaborate with each other. We come together from over 100 countries around the world and support large and small scale research collaborations.

Where does that leave us?

Exciting time to do research:

- Recent breakthroughs have unlocked more interesting problems
- Role of data has again begun to take on renewed importance
- Underexplored directions in multilingual, efficiency (both algorithmic, hardware<>algorithm co-design, in the data space).
- Important reliability issues - hallucinations, ability to update models efficiently.

Let’s open up to discussion and questions.

Questions?

Feel free to reach out:

sarahooker@cohere.com

Cohere For AI