The changing space of optimization.

LxMLS

Sara Hooker

Cohere For AI

Cohere For AI

Exploring the unknown, together.

Research

- Research Lab
- Publications
- Scholars Program

Open Science

- Cross-institutional collaborations
- Open science initiatives

Leadership

- AI Policy/Safety
- Technical Talks

1 Mission

Changing *where, how,* and by *whom* research is done.

4 Principles

Open Collaboration
Fundamental Research
Training and Development
Community

100+ papers published

60+ collaborating institutions

3 cohorts of research scholars

350K API Credits distributed

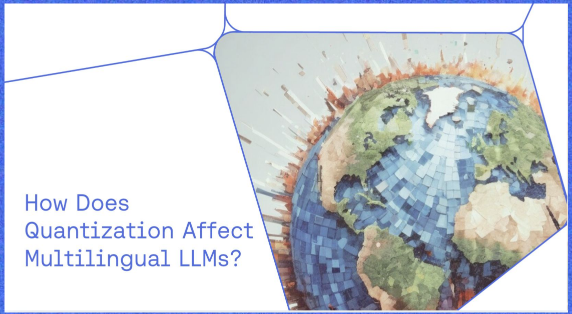
9 model releases

2.8M Aya model downloads

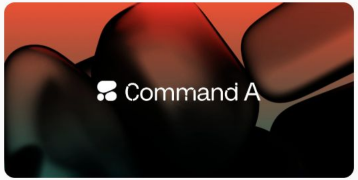
4500+ Members in our Open Science Community

from **130** Countries

Frontier AI lab, we release state-of-art models and regularly publish.



Our Models



MODEL WEIGHTS FOR DEMOCRATIZING RESEARCH ACCESS

Command A

DOWNLOAD THE MODEL ↗

MULTIMODAL ACCESSIBLE VLLM

Aya Vision - 8B

DOWNLOAD THE MODEL ↗

MULTIMODAL STATE OF THE ART VLLM

Aya Vision - 32B

DOWNLOAD THE MODEL ↗

STATE OF THE ART, ACCESSIBLE RESEARCH LLM

Aya Expanse - 8B

STATE OF THE ART RESEARCH LLM

Aya Expanse - 32B

MASSIVELY MULTILINGUAL RESEARCH LLM

Aya

MODEL WEIGHTS FOR DEMOCRATIZING RESEARCH ACCESS

C4AI Command R - 104B

DOWNLOAD THE MODEL ↗

MODEL WEIGHTS FOR DEMOCRATIZING RESEARCH ACCESS

C4AI Command R - 35B

DOWNLOAD THE MODEL ↗

MODEL WEIGHTS FOR DEMOCRATIZING RESEARCH ACCESS

Command R7B

DOWNLOAD THE MODEL ↗

1 Aya dataset

3 Aya Models

101 languages,

55 of which are completely new to any instruction-style dataset

117 countries

6 regional leads, 84 language ambassadors

>/

1,307 independent researchers

26,000 messages sent on discord

65,000+ annotations via the Aya UI

Cohere Labs

As of February 2025:

We've awarded **150+** grants

Totalling **$200,000+** so far

Across **20** countries →

Our impact exploring the unknown, together

21 research scholars, spanning 9 countries, 25 published papers

Cohere Labs

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.

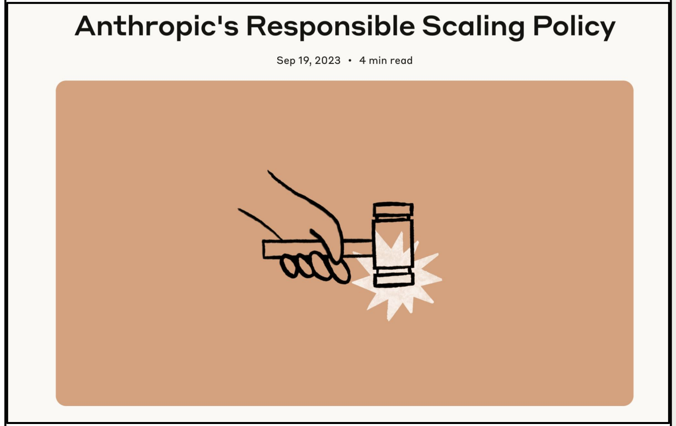
For most of the last two decades, a belief that most progress is scaling model size has prevailed: **“bigger is better.”**

Today, we will ask:

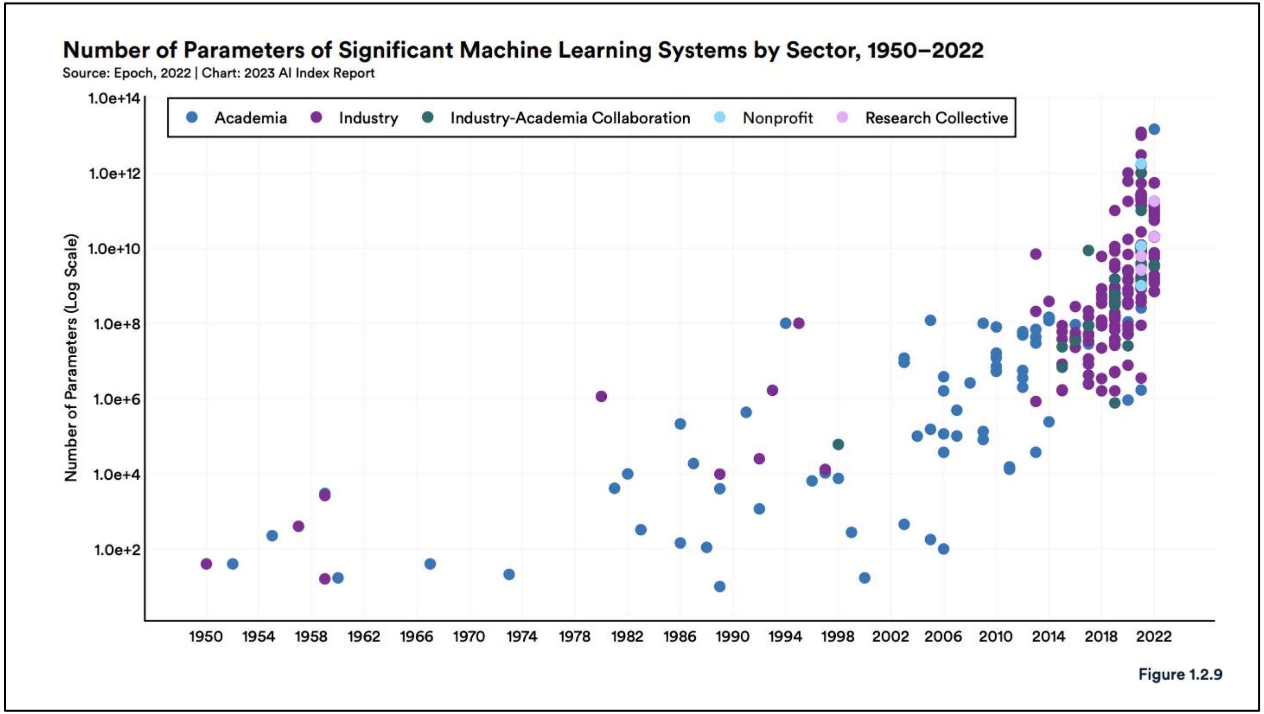
1) **Is bigger always better?**
2) **2)** How are our optimization space and tools are rapidly changing.

This will change **the nature of our field of AI research.**

The belief that "bigger has better" has shaped our ecosystem.

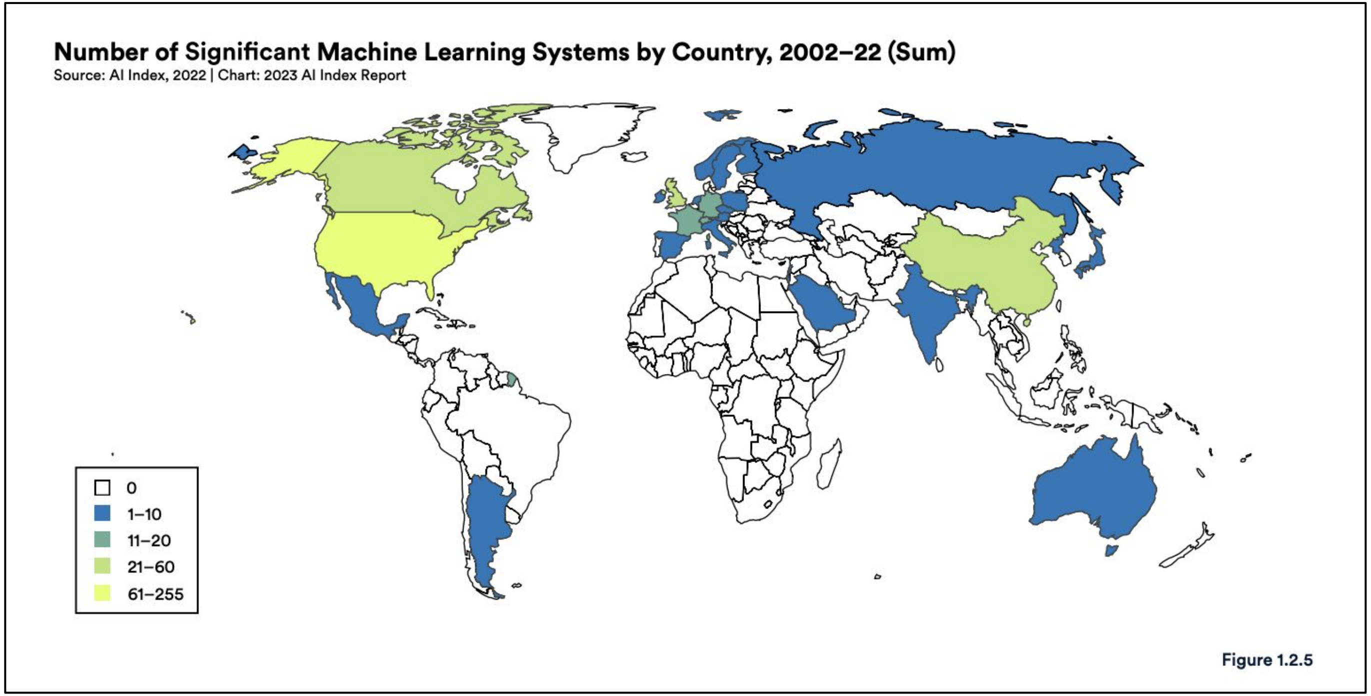


It has resulted in a shift of contributions from academic to industry research due to gaps in compute.



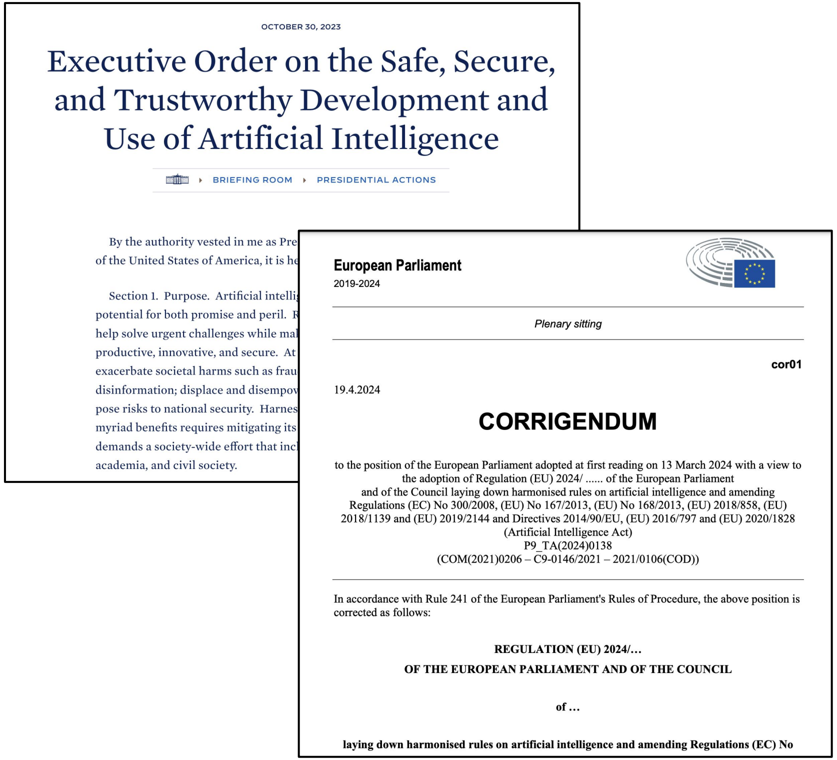
AI index report

Has determined who gets to participate and who doesn't.



AI index report, The Low resource double blind, Ahia et al.

And has even led to widespread adoption of policy where larger models are assumed to bring new inflection points of risk.


OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM ▸ PRESIDENTIAL ACTIONS

By the authority vested in me as Pre
of the United States of America, it is he

Section 1. Purpose. Artificial intelli
potential for both promise and peril. R
help solve urgent challenges while ma
productive, innovative, and secure. At
exacerbate societal harms such as frau
disinformation; displace and disempov
pose risks to national security. Harnes
myriad benefits requires mitigating its
demands a society-wide effort that inc
academia, and civil society.

European Parliament
2019-2024

Plenary sitting

cor01

19.4.2024

CORRIGENDUM

to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/...... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)
P9_TA(2024)0138
(COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))

In accordance with Rule 241 of the European Parliament's Rules of Procedure, the above position is corrected as follows:

REGULATION (EU) 2024/…

OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of …

laying down harmonised rules on artificial intelligence and amending Regulations (EC) No

Any model "*trained using a quantity of computing power greater than 10^26 integer or floating point operations.*" will be subject to more scrutiny.

Implicit is the idea that more compute results in a new inflection point of capabilities and hence risk.

Hooker 2024

So today I will ask a controversial question. **Is bigger always better?** We will cover a few things:

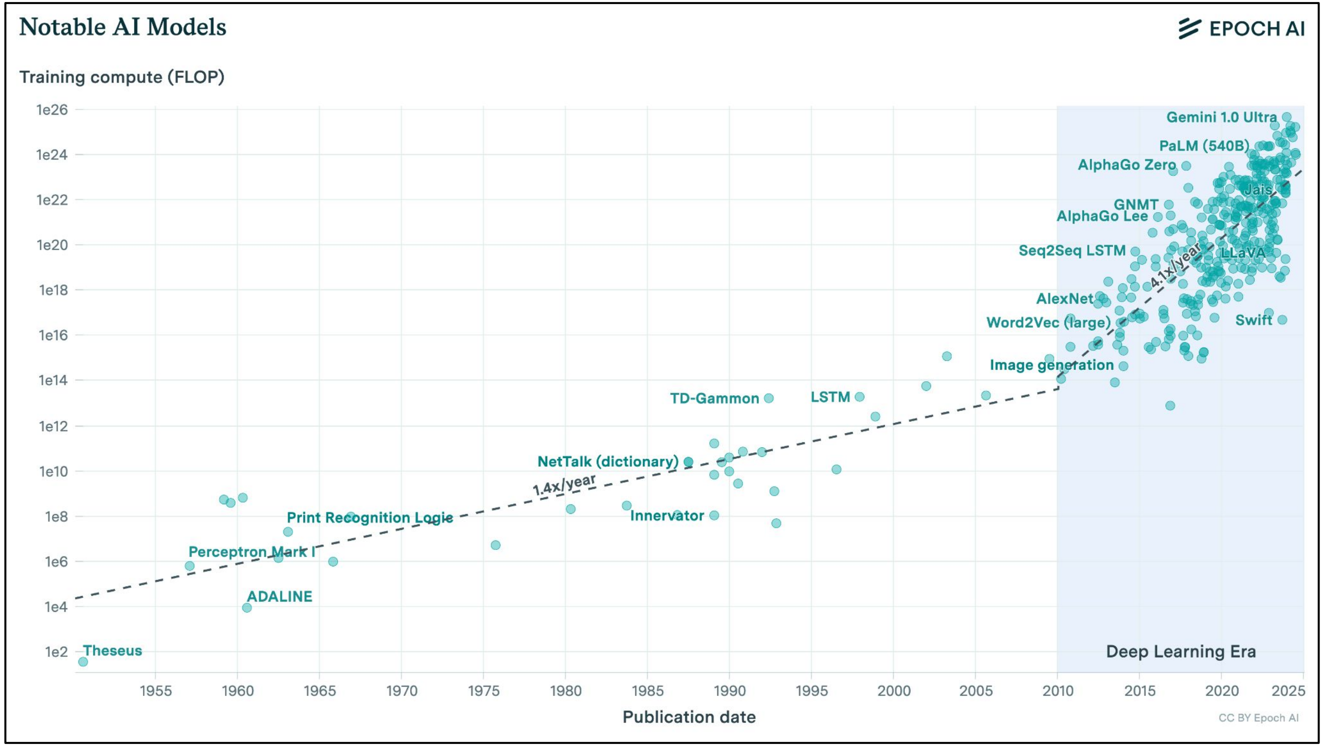
What do we gain when we scale?

What comes next: **gradient free performance gains.**

Open challenges and opportunities.

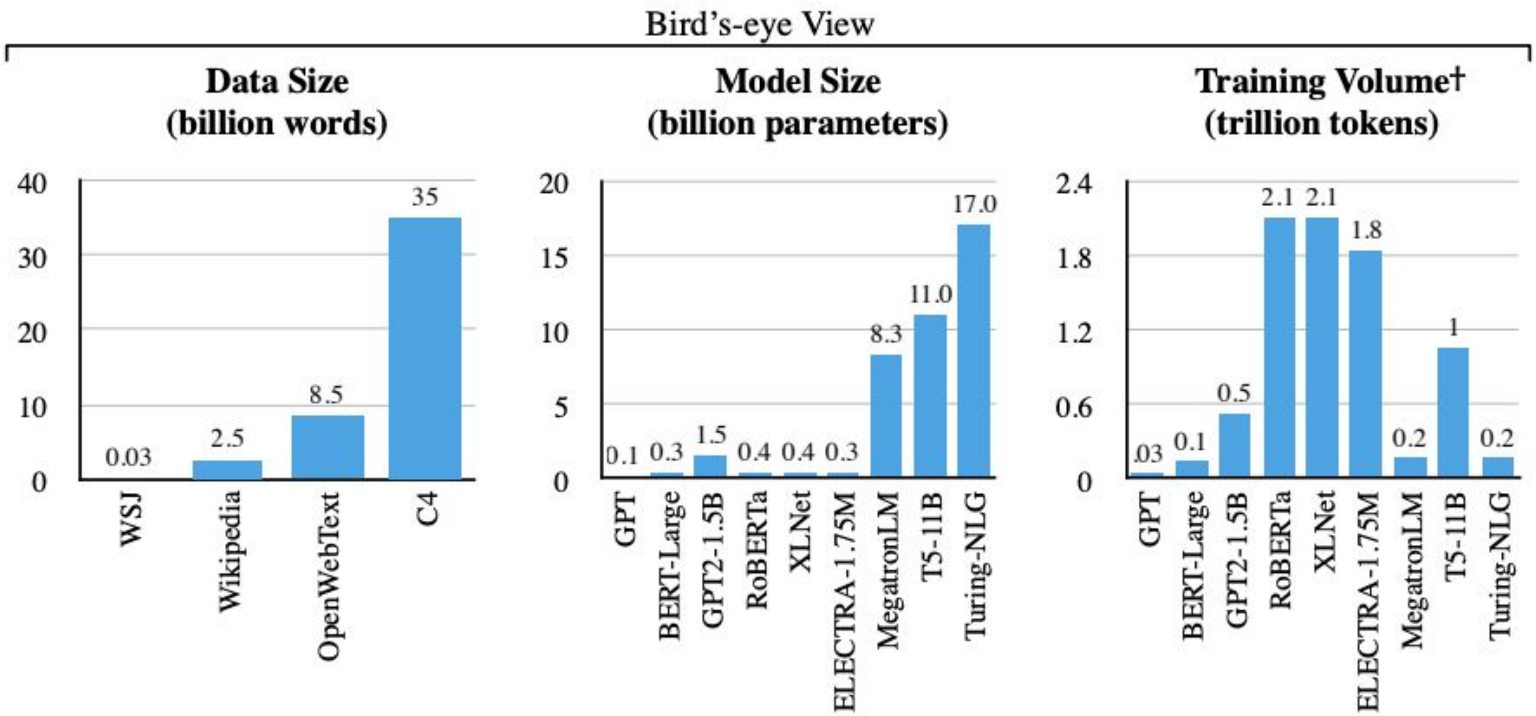
The role of model scale and data in recent breakthroughs

A “bigger is better” race in the amount of compute, parameters, data has gripped the field of machine learning.



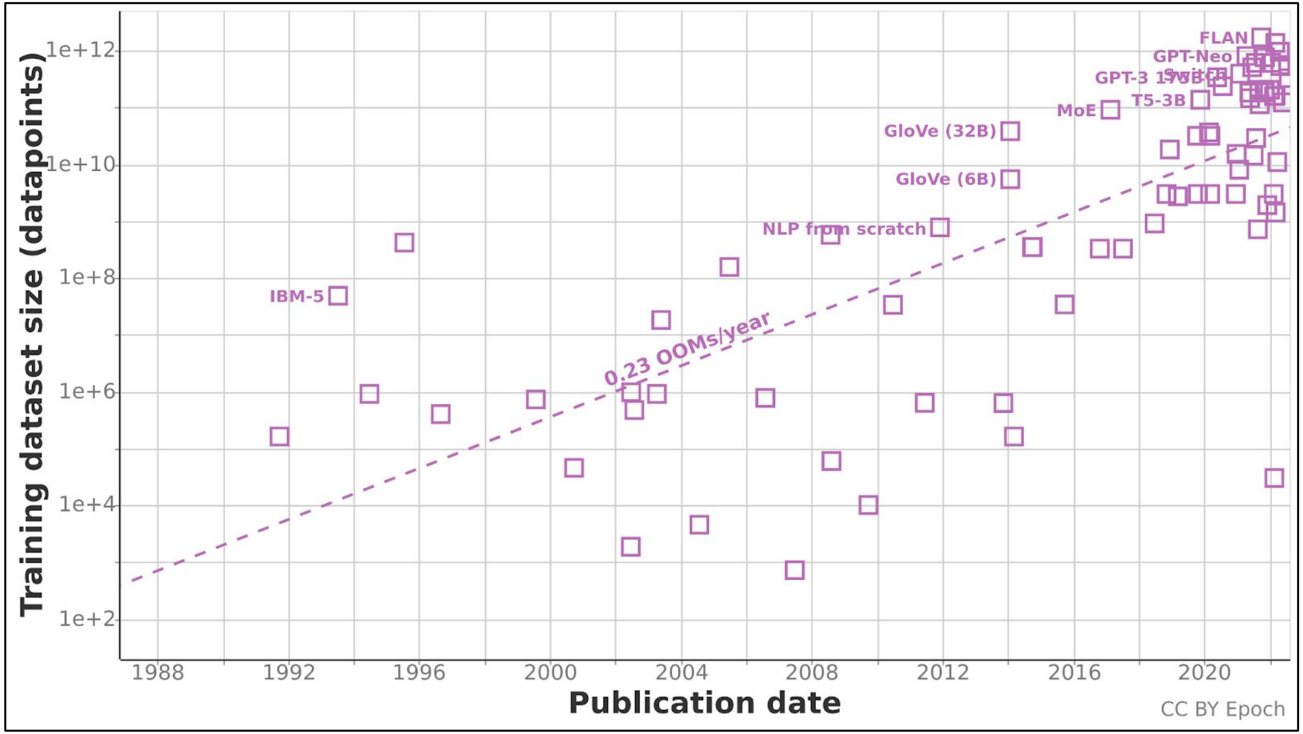
Epoch AI, 2024

This characterizes both vision and NLP tasks.



[Sharir et al. 2020]

And involves large increases in both model and dataset sizes:



Size of modern datasets over time.

Epoch AI, 2024

This is captured by Rich Sutton as the “bitter lesson”

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that ``brute force" search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

A similar pattern of research progress was seen in computer Go, only delayed by a further 20 years. Enormous initial efforts went into avoiding search by taking advantage of human knowledge, or of the special features of the game, but all those efforts proved irrelevant, or worse, once search was applied effectively at scale. Also important was the use of learning by self play to learn a value function (as it was in many other games and even in chess, although learning did not play a big role in the 1997 program that first beat a world champion). Learning by self play, and learning in general, is like search in that it enables massive computation to be brought to bear. Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research. In computer Go, as in computer chess, researchers' initial effort was directed towards utilizing human understanding (so that less search was needed) and only much later was much greater success had by embracing search and learning.

“... the only thing that matters in the long run is the leveraging of compute.”

In a punch to the ego of every computer scientist out there, what is being implied is that nothing in computer science history has worked as well as letting a model learn patterns for itself coupled with scaling the algorithm.

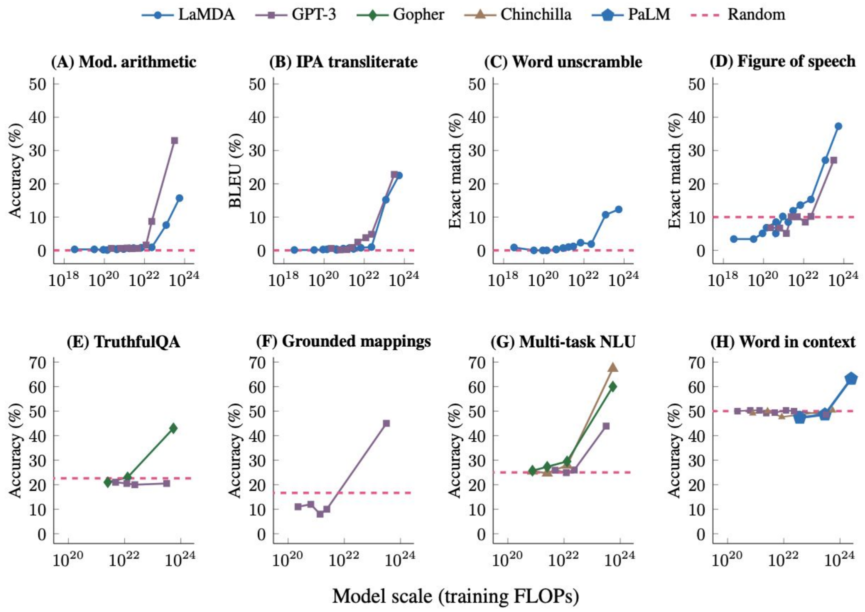
Is Sutton right?

There is an argument in favor of this approach:

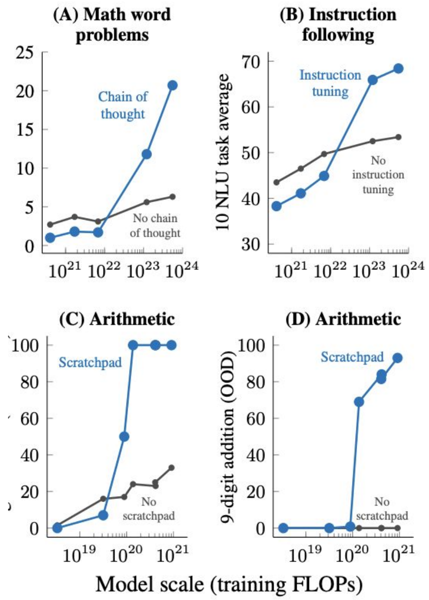
- Different regimes of capacity appear to allow for different generalization properties.
- It is very simple formula (throw more parameters at the model)

[Wei et al. 2022, Nakkiran et al. 2019, Petroni et al., Brown et al., , Adam et al.]

For example, instruction tuning only improves zero-shot performance on models above a certain size.



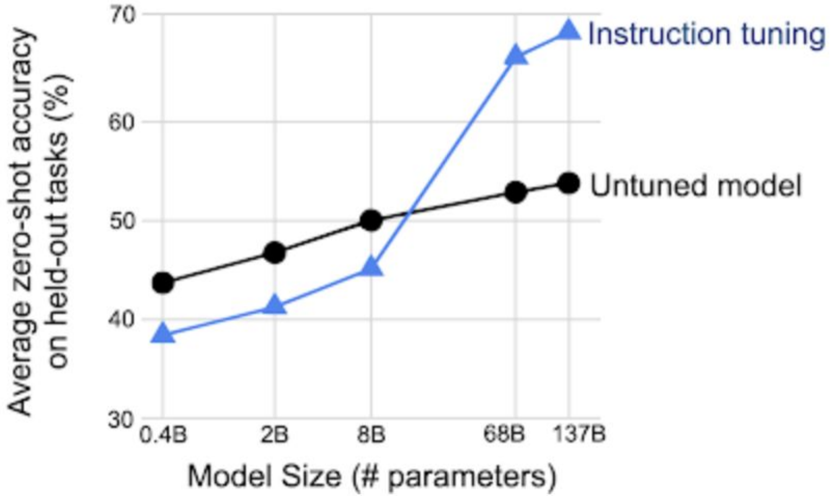
Few shot prompting performance improves with FLOPs.



Finetuning and few shot.

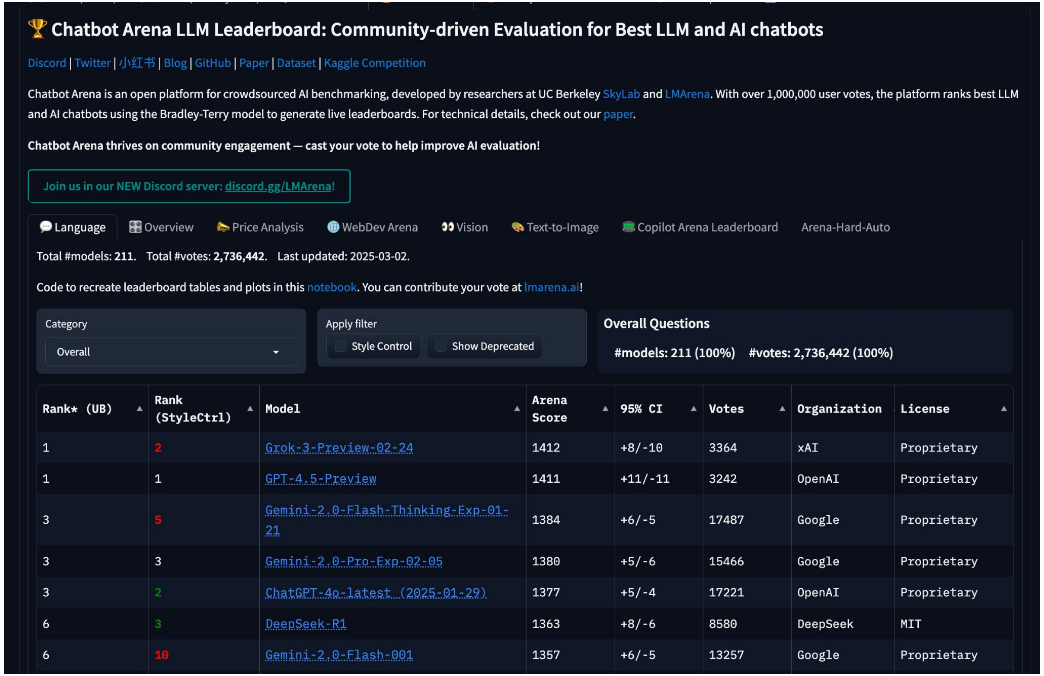
[Wei et al. 2022]

It also requires larger and larger models to take advantage of instruction fine-tuning.



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

Certainly if you looked at chatbot arena, it is very clear the largest models index higher.



🏆 Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots

Discord | Twitter | 小红书 | Blog | GitHub | Paper | Dataset | Kaggle Competition

Chatbot Arena is an open platform for crowdsourced AI benchmarking, developed by researchers at UC Berkeley SkyLab and LMArena. With over 1,000,000 user votes, the platform ranks best LLM and AI chatbots using the Bradley-Terry model to generate live leaderboards. For technical details, check out our paper.

Chatbot Arena thrives on community engagement — cast your vote to help improve AI evaluation!

Join us in our NEW Discord server: discord.gg/LMArena!

Language | Overview | Price Analysis | WebDev Arena | Vision | Text-to-Image | Copilot Arena Leaderboard | Arena-Hard-Auto

Total #models: 211. Total #votes: 2,736,442. Last updated: 2025-03-02.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control | Show Deprecated

Overall Questions

#models: 211 (100%) #votes: 2,736,442 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	2	Grok-3-Preview-02-24	1412	+8/-10	3364	xAI	Proprietary
1	1	GPT-4.5-Preview	1411	+11/-11	3242	OpenAI	Proprietary
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+6/-5	17487	Google	Proprietary
3	3	Gemini-2.0-Pro-Exp-02-05	1380	+5/-6	15466	Google	Proprietary
3	2	ChatGPT-4o-latest (2025-01-29)	1377	+5/-4	17221	OpenAI	Proprietary
6	3	DeepSeek-R1	1363	+8/-6	8580	DeepSeek	MIT
6	10	Gemini-2.0-Flash-001	1357	+6/-5	13257	Google	Proprietary

First model on the leaderboard with known parameter count is Deepseek-R1 – **685 billion parameters.**

Cohere For AI

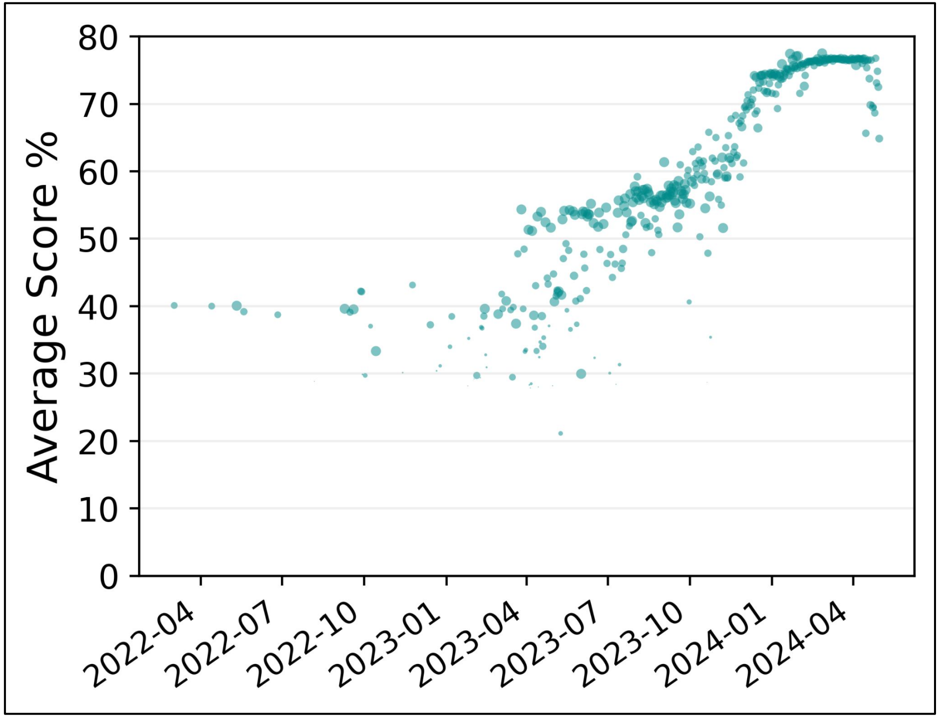
link

Scaling model size is still widely favored:

- More de-risked vs more difficult approaches of crafting new optimization techniques
- Fits into industry quarterly planning cycles – hard to justify deviating from the predictable path of gains.

However, a key limitation of blindly following "the bitter lesson" is that the relationship between model size and generalization is still not well understood.

Models at the same capacity have been getting far more performant over time.

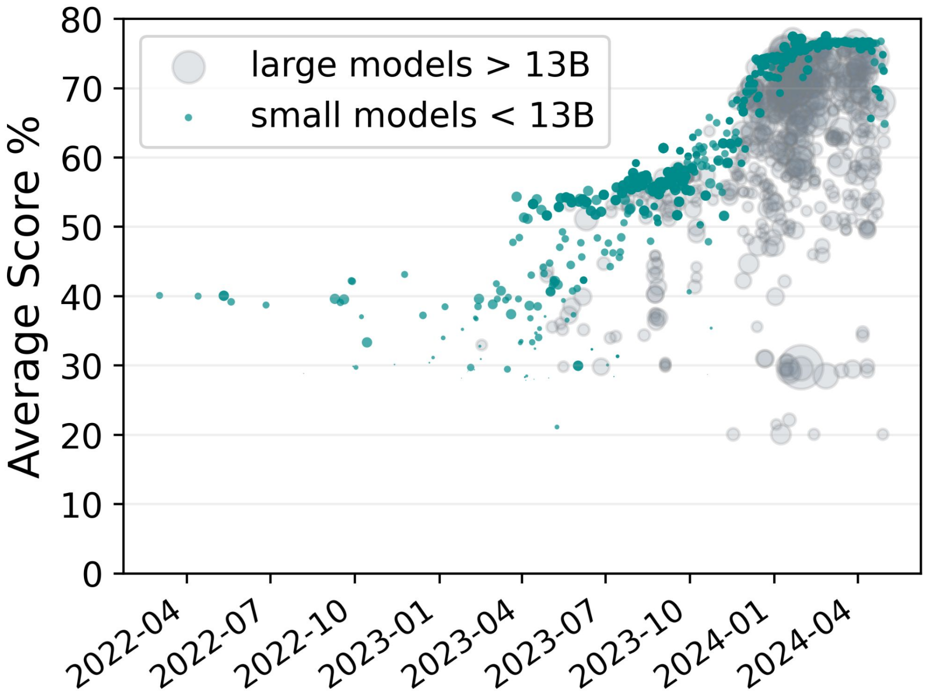


Models **under 13B** on the llm open leaderboard over time.

Cohere For AI

Hooker 2024

Smaller models frequently outperform far larger models.

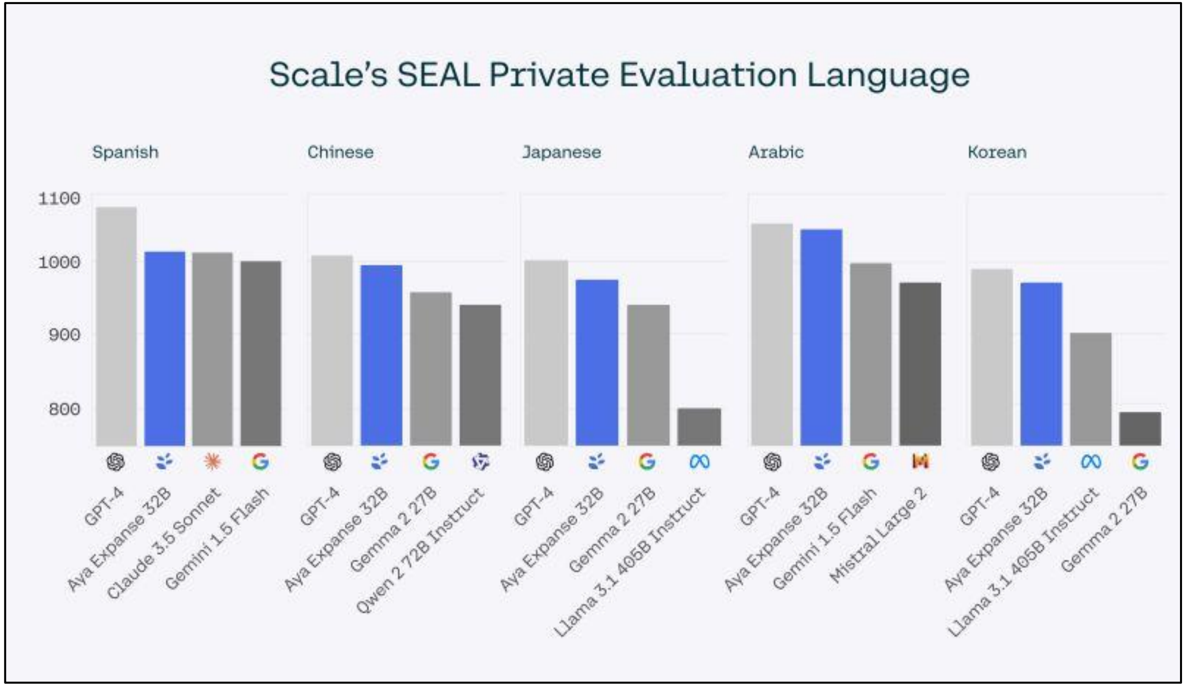


All models **over 13B** (grey) that underperform the best daily model **under 13B** submitted to the llm open leaderboard (green).

Cohere For AI

[Hooker 2024]

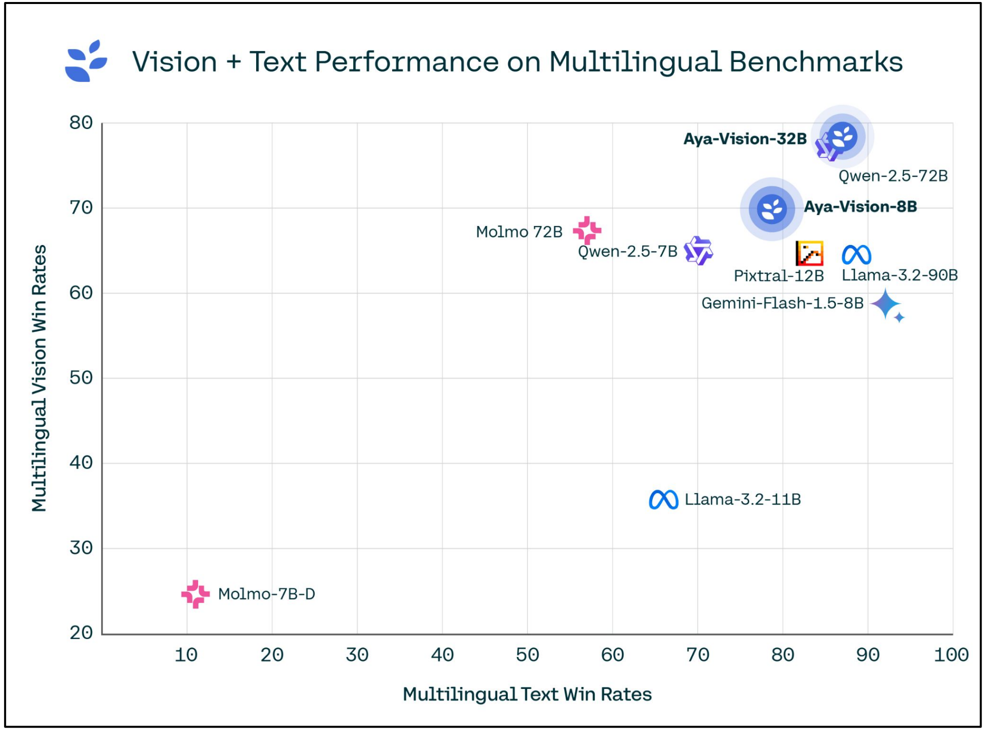
We also see this frequently in our own work. **Aya Expanse 32B** is our state-of-art multilingual and on Scale's private leaderboard (third party, no released test set) **outperforms drastically larger models including Claude, Mistral Large 2, & Llama 3.1 405B parameters.**



Cohere For AI

Scale Private Leaderboard

We recently released Aya Vision multilingual multimodal model 8B – which **outperforms llama-3.2- 90B and Molmo 72B** across languages spoken by 50% of the worlds population.



Aya Vision 8B outperforms models **11x its size** – llama 90B.

In fact, we observe a highly uncertain relationship between compute and performance.

In fact, we observe a highly uncertain relationship between compute and performance.

1) Data quality compensates for need for compute
2) Architecture plays a significant role in determining scalability
3) Post-training optimization reduces need for training time compute.
4) Diminishing returns to adding more weights.
5) Many redundancies between weights, most weights can be removed after training.
6) Majority of weights used to represent a small slice of overall distribution.

Data quality compensates for the need for compute.

Recent work finds smaller amounts of higher quality data removes the need for a larger model.

There is increasing evidence that efforts to better curate training corpus, including **deduping, pruning data and better quality synthetic data** can compensate for the need for larger networks and/or improve training dynamics.

	% train examples with dup in train	% train examples with dup in valid	% valid with dup in train
C4	3.04%	1.59%	4.60%
RealNews	13.63%	1.25%	14.35%
LM1B	4.86%	0.07%	4.92%
Wiki40B	0.39%	0.26%	0.72%

Table 2: The fraction of examples identified by NEARDUP as near-duplicates.

Lee et al. 2022

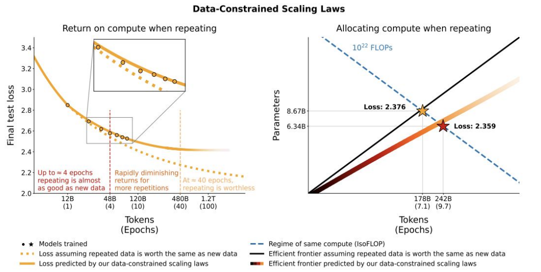


Figure 1: ***Return* and *Allocation* when repeating data.** *(Left):* Loss of LLMs (4.2B parameters) scaled on repeated data decays predictably (§6). *(Right):* To maximize performance when repeating, our data-constrained scaling laws and empirical data suggest training smaller models for more epochs in contrast to what assuming Chinchilla scaling laws [42] hold for repeated data would predict (§5).

Muennighoff et al. 2023

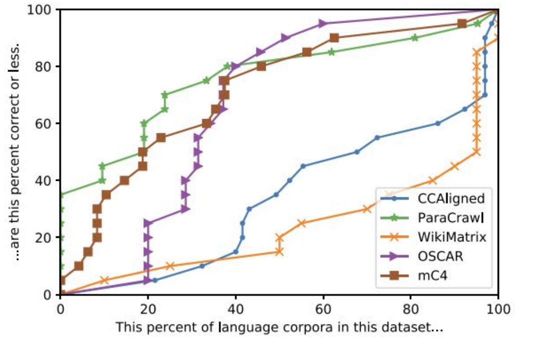
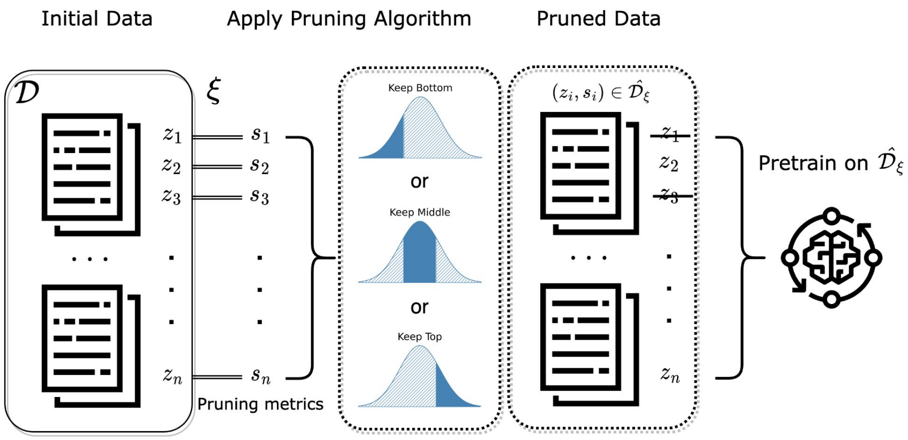


Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

Kreutzer at al. 2022

Our recent work focuses on effective data pruning for pretraining internet scale.



When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale

Max Marion *Cohere for AI* maxwell@cohere.com

Ahmet Üstün *Cohere for AI* ahmet@cohere.com

Luiza Pozzobon *Cohere for AI* luiza@cohere.com

Alex Wang *Cohere* alexwang@cohere.com

Marzieh Fadaee *Cohere for AI* marzieh@cohere.com

Sara Hooker *Cohere for AI* sarahooker@cohere.com

Abstract

Large volumes of text data have contributed significantly to the development of large language models (LLMs) in recent years. This data is typically acquired by scraping the internet, leading to pretraining datasets comprised of noisy web text. To date, efforts to prune these datasets down to a higher quality subset have relied on hand-crafted heuristics encoded as rule-based filters. In this work, we take a wider view and explore scalable estimates of data quality that can be used to systematically measure the quality of pretraining data. We perform a rigorous comparison at scale of the simple data quality estimator of perplexity, as well as more sophisticated and computationally intensive estimates of the Error L2-Norm and memorization. These metrics are used to rank and prune pretraining corpora, and we subsequently compare LLMs trained on these pruned datasets. Surprisingly, we find that the simple technique of perplexity outperforms our more computationally expensive scoring methods. We improve over our no-pruning baseline while training on as little as 30% of the original training dataset. Our work sets the foundation for unexplored strategies in automatically curating high quality corpora and suggests the majority of pretraining data can be removed while retaining performance.

[[Marion et al. 2023]]

We can improve over our no-pruning baseline **while training on as little as 30% of the original training dataset.**

When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale

Max Marion
Cohere for AI
maxwell@cohere.com

Ahmet Üstün
Cohere for AI
ahmet@cohere.com

Luiza Pozzobon
Cohere for AI
luiza@cohere.com

Alex Wang
Cohere
alexwang@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Large volumes of text data have contributed significantly to the development of large language models (LLMs) in recent years. This data is typically acquired by scraping the internet, leading to pretraining datasets comprised of noisy web text. To date, efforts to prune these datasets down to a higher quality subset have relied on hand-crafted heuristics encoded as rule-based filters. In this work, we take a wider view and explore scalable estimates of data quality that can be used to systematically measure the quality of pretraining data. We perform a rigorous comparison at scale of the simple data quality estimator of perplexity, as well as more sophisticated and computationally intensive estimates of the Error L2-Norm and memorization. These metrics are used to rank and

[[Marion et al. 2023]]

Data pruning is a valuable optimization at multiple stages of training pipeline – here we also show promising results in preference training.

We reduce instances of indecisive (or "tie") outcomes by up to 54% compared to a random sample when focusing on the top-20 percentile of prioritized instances.

This helps save valuable human feedback for the most important instances.

Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation

Meriem Boubdir
Cohere for AI
meri.boubdir@gmail.com

Edward Kim
Cohere
edward@cohere.com

Beyza Ermis
Cohere for AI
beyza@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Human evaluation is increasingly critical for assessing large language models, capturing linguistic nuances, and reflecting user preferences more accurately than traditional automated metrics. However, the resource-intensive nature of this type of annotation process poses significant challenges. The key question driving our work: *is it feasible to minimize human-in-the-loop feedback by prioritizing data instances which most effectively distinguish between models?* We evaluate several metric-based methods and find that these metrics enhance the efficiency of human evaluations by minimizing the number of required annotations, thus saving time and cost, while ensuring a robust performance evaluation. We show that our method is effective across widely used model families, reducing instances of indecisive (or "tie") outcomes by up to 54% compared to a random sample when focusing on the top-20 percentile of prioritized instances. This potential reduction in required human effort positions our approach as a valuable strategy in future large language model evaluations.

[[Boubdir et al. 2023]]

Relationship between weights and performance is not well understood.

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

Model	Parameters[a]	Features	Image Size	ImageNet Top-1 Accurac	
				Paper	Public Checkpoint[b]
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]

The looming question of diminishing returns has also impacted recent model launches.

Interconnects | Nathan Lambert
https://www.interconnects.ai

GPT-4.5: "Not a frontier model"? - by Nathan Lambert

Feb 28, 2025 — GPT-4.5 is a point on the graph that scaling is still coming, but trying to make sense of it in a day-by-day transition is hard.

The End of Scaling: GPT-4.5 and the Looming AI Winter | by Gabriel...

The transformer architecture that powers models like GPT-4.5 has been pushed to its limits, and the returns on further scaling have diminished to the point where they…

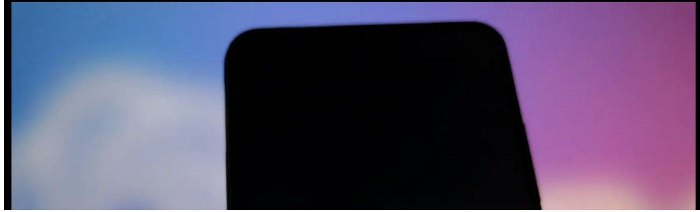
Analysis and Technology

Is OpenAI hitting a wall with huge and expensive GPT-4.5 model?

Some researchers think OpenAI's giant and expensive latest model is a sign that tech companies cannot keep making progress by continually scaling up

By Matthew Sparkes

28 February 2025



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

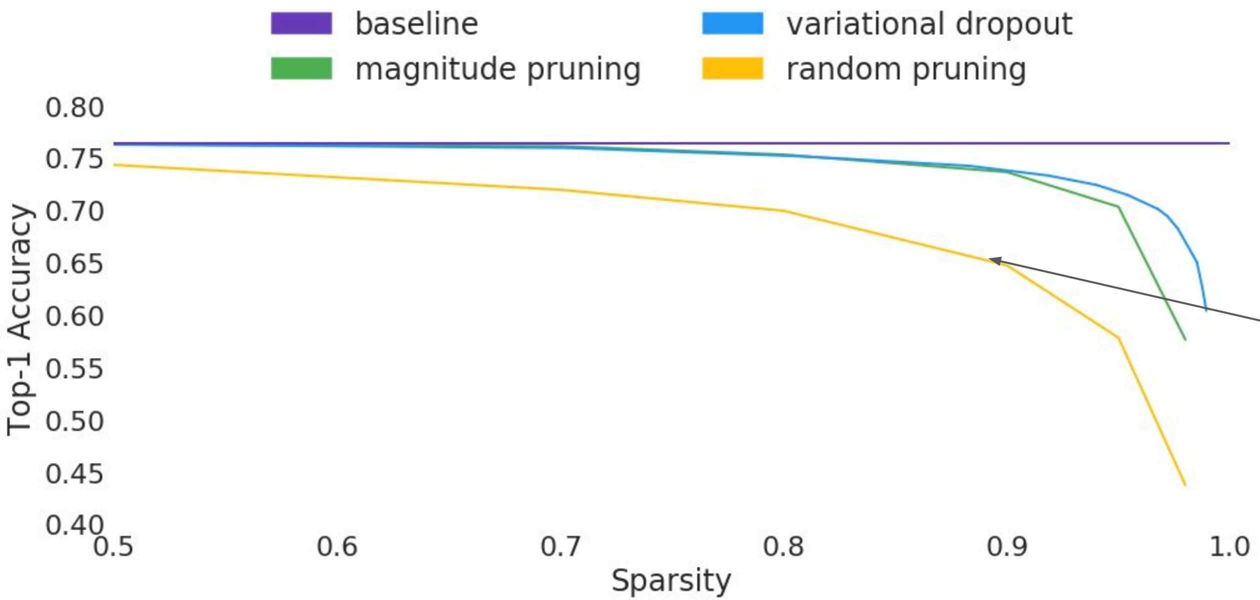
Abstract

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. 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]]

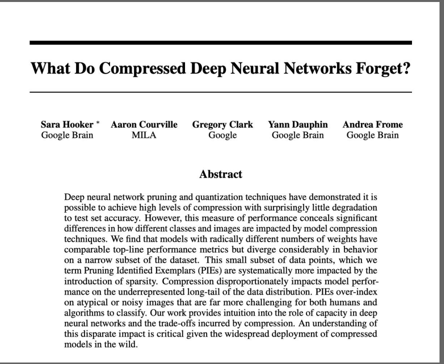
Empirical risk minimization means we optimize to reduce average error:

This means it takes more capacity or longer training to learn rare features.

Majority of features are learnt early in training. Despite this most of training focuses on long-tail.

Majority of features can be learnt using small models. Scaling of size primarily benefits small tiny part of distribution.

Work with colleagues over last 5 years has focused on understanding what is lost and gained as we vary model size.

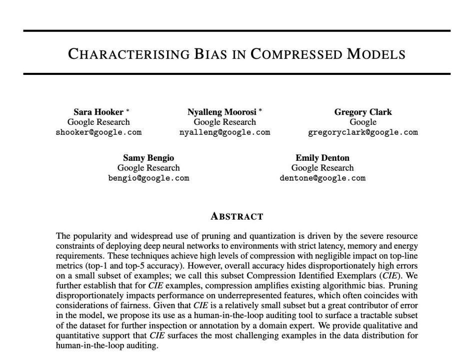


What Do Compressed Deep Neural Networks Forget?

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. Compression disproportionately impacts model performance on the underrepresented long-tail of the data distribution. PIEs over-index on atypical or noisy images that are far more challenging for both humans and algorithms to classify. Our work provides intuition into the role of capacity in deep neural networks and the trade-offs incurred by compression. An understanding of this disparate impact is critical given the widespread deployment of compressed models in the wild.



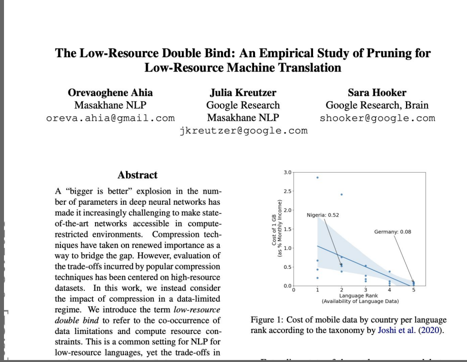
Characterising Bias in Compressed Models

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.



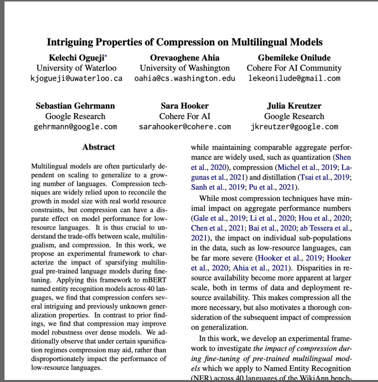
The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation

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

Abstract

A "bigger is better" explosion in the number of parameters in deep neural networks has made it increasingly challenging to make state-of-the-art networks accessible in compute-restricted environments. Compression techniques have taken on renewed importance as a way to bridge the gap. However, evaluation of the trade-offs incurred by popular compression techniques has been centered on high-resource datasets. In this work, we instead consider the impact of compression in a data-limited regime. We introduce the term *low-resource double bind* to refer to the co-occurrence of data limitations and compute resource constraints. This is a common setting for NLP for low-resource languages, yet the trade-offs in

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



Intriguing Properties of Compression on Multilingual Models

Kelechi Ogueji* University of Waterloo kjogueji@uwaterloo.ca · Orevaoghene Ahia University of Washington oahia@cs.washington.edu · Gbemileke Onilude Cohere For AI Community lekeonilude@gmail.com

Sebastian Gehrmann Google Research gehrmann@google.com · Sara Hooker Cohere For AI sarahooker@cohere.com · Julia Kreutzer Google Research jkreutzer@google.com

Abstract

Multilingual models are often particularly dependent on scaling to generalize to a growing number of languages. Compression techniques are widely relied upon to reconcile the growth in model size with real world resource constraints, but compression can have a disparate effect on model performance for low-resource languages. It is thus crucial to understand the trade-offs between scale, multilingualism, and compression. In this work, we propose an experimental framework to characterize the impact of sparsifying multilingual pre-trained language models during fine-tuning. Applying this framework to mBERT named entity recognition models across 40 languages, we find that compression confers several intriguing and previously unknown generalization properties. In contrast to prior findings, we find that compression may improve model robustness over dense models. We additionally observe that under certain sparsification regimes compression may aid, rather than disproportionately impact the performance of low-resource languages.

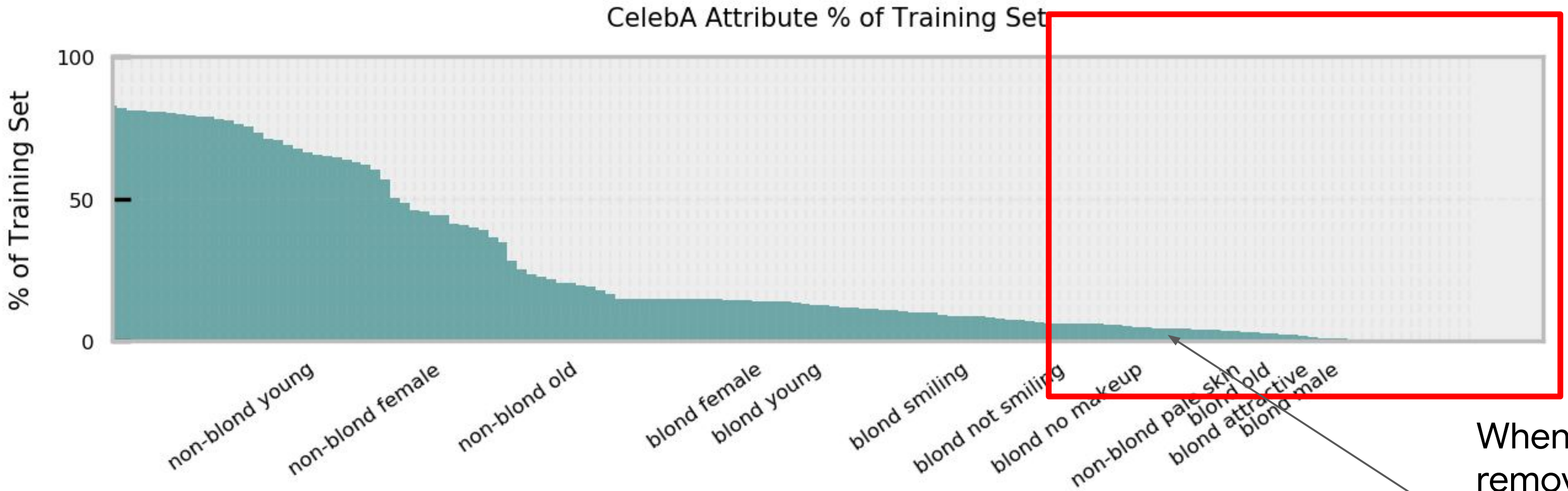
while maintaining comparable aggregate performance are widely used, such as quantization (Shen et al., 2020), compression (Michel et al., 2019; Lagunas et al., 2021) and distillation (Tsai et al., 2019; Sanh et al., 2019; Pu et al., 2021).

While most compression techniques have minimal impact on aggregate performance numbers (Gale et al., 2019; Li et al., 2020; Hou et al., 2020; Chen et al., 2021; Bai et al., 2020; ab Tessera et al., 2021), the impact on individual sub-populations in the data, such as low-resource languages, can be far more severe (Hooker et al., 2019; Hooker et al., 2020; Ahia et al., 2021). Disparities in resource availability become more apparent at larger scale, both in terms of data and deployment resource availability. This makes compression all the more necessary, but also motivates a thorough consideration of the subsequent impact of compression on generalization.

In this work, we develop an experimental framework to investigate *the impact of compression during fine-tuning of pre-trained multilingual models* which we apply to Named Entity Recognition (NER) across 40 languages of the WikiAnn bench-

[[Hooker et al. 2019, Hooker, Moorosi et al, 2020, Ahia et al. 2021, Ogueji et al. 2022, Marchisio 2024]]

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.

Cohere For AI

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

It is worth emphasizing this finding: We lose the long-tail when we remove the majority of all training weights.

Put differently, we are using the majority of our weights to encode a useful representation for a small fraction of our training distribution.

0 %

90 %

Overparameterized
Dense Model

Model with 90%
weights removed

Google

When we scale models, we are paying an enormous cost to learn a small slice of the distribution (noisy and atypical examples).

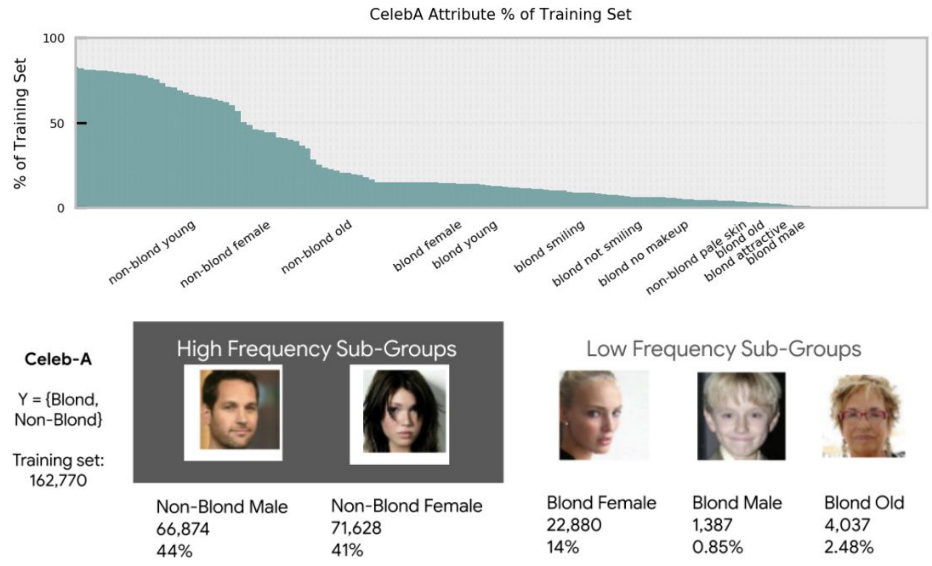
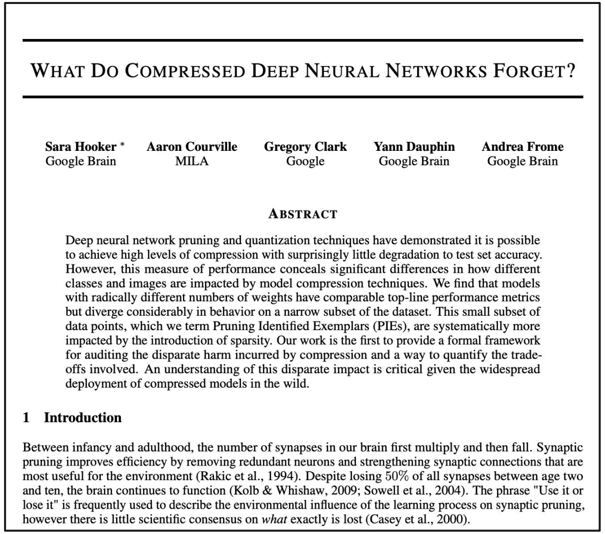


Figure 1: Most natural image datasets exhibit a long-tail distribution with an unequal frequency of attributes in the training data. Below each attribute sub-group in CelebA, we report the share of training set and total frequency count.



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

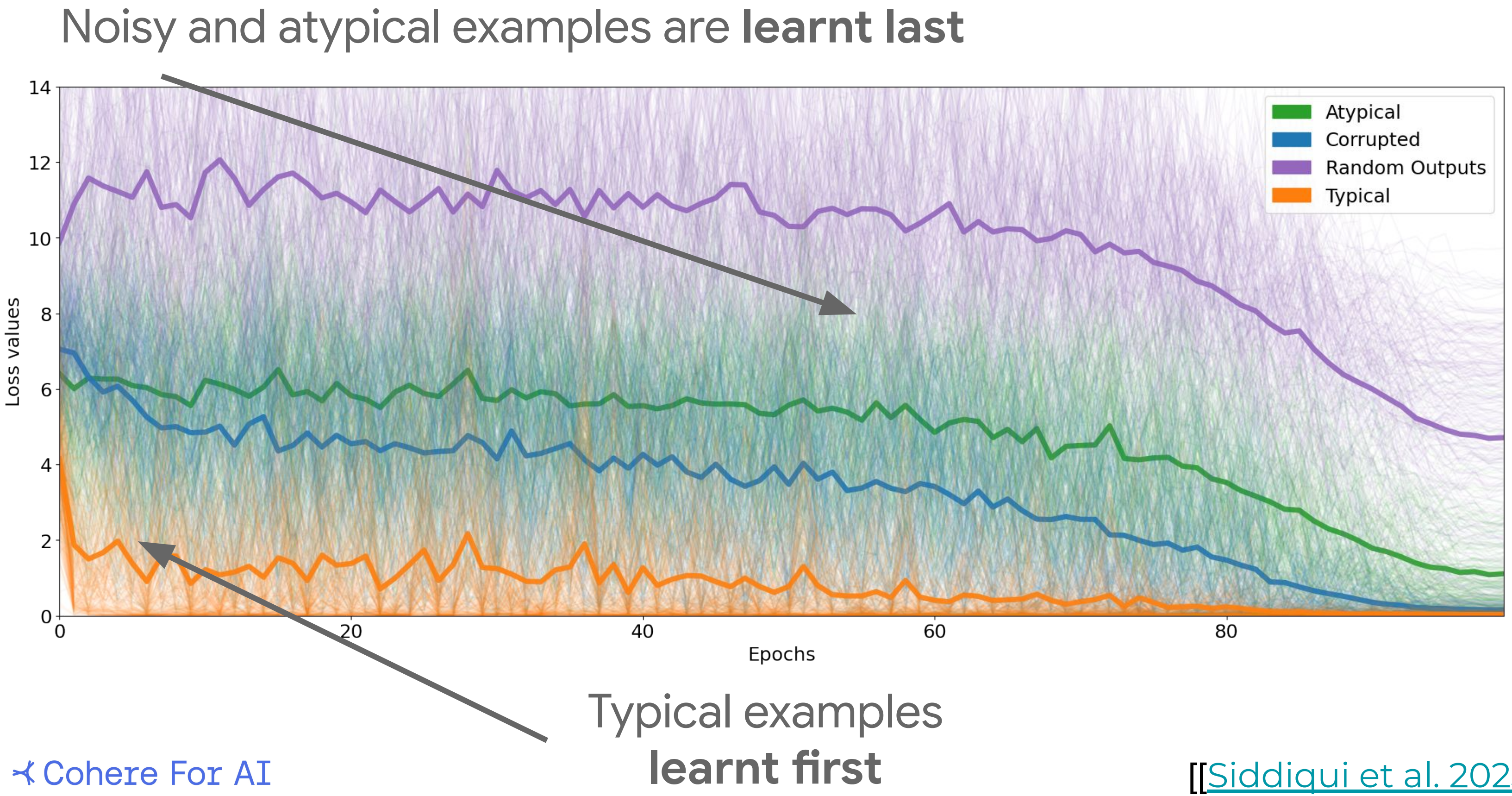
Deep neural network pruning and quantization techniques have demonstrated it is possible to achieve high levels of compression with surprisingly little degradation to test set accuracy. However, this measure of performance conceals significant differences in how different classes and images are impacted 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.

1 Introduction

Between infancy and adulthood, the number of synapses in our brain first multiply and then fall. Synaptic pruning improves efficiency by removing redundant neurons and strengthening synaptic connections that are most useful for the environment (Rakic et al., 1994). Despite losing 50% of all synapses between age two and ten, the brain continues to function (Kolb & Whishaw, 2009; Sowell et al., 2004). The phrase "Use it or lose it" is frequently used to describe the environmental influence of the learning process on synaptic pruning, however there is little scientific consensus on *what* exactly is lost (Casey et al., 2000).

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

5. Most of training time is spent learning rare examples. High frequency examples are learnt early on and don't require much training time.



Cohere For AI

[[Siddiqui et al. 2023]]

So where do we end up?

I **may** have
convinced you
that we are now in
a period of
decreasing returns
to compute.

So where do we end up?

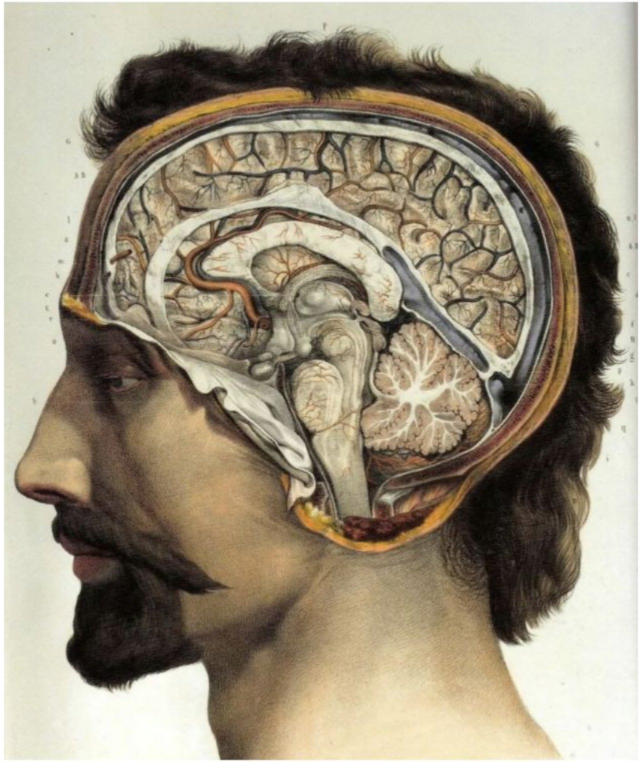
I **may** have convinced you that we are now in a period of decreasing returns to compute.

Regardless of whether you are convinced that transformers are saturated, I hope I have convinced you that our current trajectory is extremely expensive. **We pay a lot to learn infrequent and rare features.**

Point of comparison: our Brain is incredibly energy efficient.

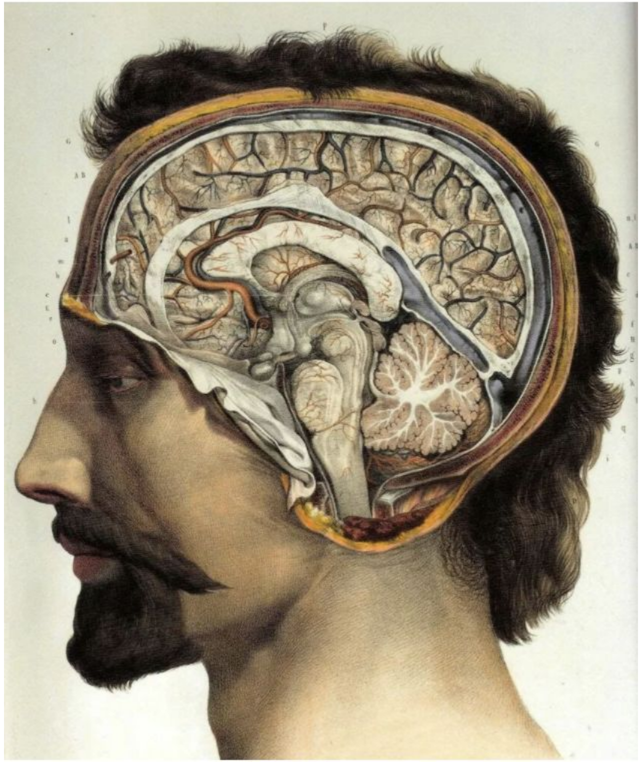
Has over 85 billion neurons but runs on the energy equivalent of an electric shaver

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



Some aspects of what we do with deep neural networks is painfully inefficient.

- We do not have adaptive compute. Typically we see all examples same amount of time during training.
- Global updates mean all prior information is erased.
- Empirical risk minimization means while we optimize for average performance, it takes considerable more compute to model rare or infrequent artefacts.



So where do we end up?

I **may** have convinced you that we are now in a period of decreasing returns to compute.

Regardless of whether you are convinced that transformers are saturated, I hope I have convinced you that our current trajectory is extremely expensive. **We pay a lot to learn infrequent and rare features.**

So – that prompts the question of what comes next.

If scaling model size is slowly dying – what is our biggest lever of progress?

“What we have before us are
some breathtaking opportunities
disguised as insoluble problems”
John Gardner, 1965.

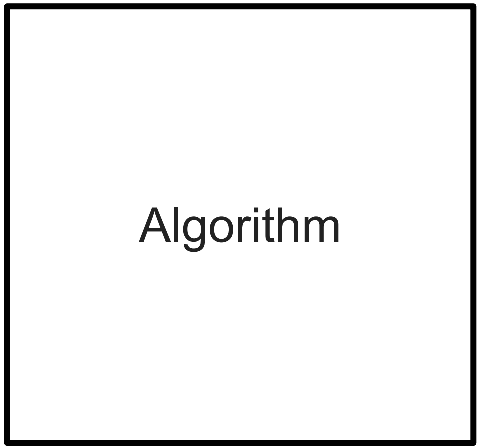
Modern computer science as a field has only existed for the last 75 years.

2025

1950

Cohere For AI

Our pursuit as a field has been centered around optimizing the algorithm.



Now, we are an interesting time where our tools for optimization are finding new spaces.

Optimization in the data space.

Algorithm

Now, we are an interesting time where our tools for optimization are finding new spaces.

Optimization in the data space.

Algorithm

Inference time scaling.

Gradient free performance boosts.

Optimization in the data space

– for the first time it is scalable to steer the data space towards properties we care about.

Promising directions for optimizing in the data space to make better use of capacity:

1

Data pruning,
Weighting
Data Arbitrage

"Spending more capacity on the data points we care about"

2

Synthetic data

"Steering dataset generation using 'on-the-fly' objectives"

1.

Data pruning or Weighting

“Spending more capacity on the data points we care about”

Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation

Meriem Boubdir, *Cohere for AI*, meri.boubdir@gmail.com
Edward Kim, *Cohere*, edward@cohere.com
Beyza Ermis, *Cohere for AI*, beyza@cohere.com
Marzieh Fadaee, *Cohere for AI*, marzieh@cohere.com
Sara Hooker, *Cohere for AI*, sarahooker@cohere.com

Abstract

Human evaluation is increasingly critical for assessing large language models, capturing linguistic nuances, and reflecting user preferences more accurately than traditional automated metrics. However, the resource-intensive nature of this type of annotation process poses significant challenges. The key question driving our work: *is it feasible to minimize human-in-the-loop feedback by prioritizing data instances which most effectively distinguish between models?* We evaluate several metric-based methods and find that these metrics enhance the efficiency of human evaluations by minin
robust p
families,
sample w
required
evaluatic

Does your data spark joy? Performance gains from domain upsampling at the end of training

Cody Blakeney*, Mansheej Paul*, Brett W. Larsen*, Sean Owen, and Jonathan Frankle

Databricks Mosaic Research

Abstract

Pretraining datasets for large language models (LLMs) have grown to trillions of tokens composed of large amounts of CommonCrawl (CC) web scrape along with smaller, domain-specific datasets. It is expensive to understand the impact of these domain-specific datasets on model capabilities as training at large FLOP scales is required to reveal significant changes to difficult and emergent benchmarks. Given the increasing cost of experimenting with pretraining data, how does one determine the optimal balance between the diversity in general web scrapes and the information density of domain specific data? In this work, we show how to leverage the smaller domain specific datasets by upsampling them relative to CC at the end of training to drive performance improvements on difficult benchmarks. This simple technique allows us to improve up to 6.90 pp on MMLU, 8.26 pp on GSM8K, and 6.17 pp on HumanEval relative to the base data mix for a 7B model trained for 1 trillion (T) tokens, thus rivaling Llama-2 (7B)—a model trained for twice as long. We experiment with ablating the duration of domain upsampling from 5% to 30% of training and find that 10% to 20% percent is optimal for navigating the tradeoff between general language modeling capabilities and targeted benchmarks. We also use domain upsampling to characterize at scale the utility of individual datasets for improving various benchmarks by removing them during this final phase of training. This tool opens up the ability to experiment with the impact of different pretraining datasets at scale, but at an order of magnitude lower cost compared to full pretraining runs.

When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale

Max Marion, *Cohere for AI*, maxwell@cohere.com
Ahmet Üstün, *Cohere for AI*, ahmet@cohere.com
Luiza Pozzobon, *Cohere for AI*, luiza@cohere.com
Alex Wang, *Cohere*, alexwang@cohere.com
Marzieh Fadaee, *Cohere for AI*, marzieh@cohere.com
Sara Hooker, *Cohere for AI*, sarahooker@cohere.com

Abstract

Large volumes of text data have contributed significantly to the development of large language models (LLMs) in recent years. This data is typically acquired by scraping the internet, leading to pretraining datasets comprised of noisy web text. To date, efforts to prune these datasets down to a higher quality subset have relied on hand-crafted heuristics encoded as rule-based filters. In this work, we take a wider view and explore scalable estimates of data quality that can be used to systematically measure the quality of pretraining data. We perform a rigorous comparison at scale of the simple data quality estimator of perplexity, as well as more sophisticated and computationally intensive estimates of the Error L2-Norm and memorization. These metrics are used to rank and prune pretraining corpora, and we subsequently compare LLMs trained on these pruned datasets. Surprisingly, we find that the simple technique of perplexity outperforms our more computationally expensive scoring methods. We improve over our no-pruning baseline while training on as little as 30% of the original training dataset. Our work sets the foundation for unexplored strategies in automatically curating high quality corpora and suggests the majority of pretraining data can be removed while retaining performance.

Critical Learning Periods: Leveraging Early Training Dynamics for Efficient Data Pruning

Everlyn Asiko Chimoto[1,2,3] Jay Gala[1,5] Orevaoghene Ahia[1,6]
Julia Kreutzer[7] Bruce A. Bassett[2,4] Sara Hooker[7]

[1]Cohere For AI Community [2]University of Cape Town, South Africa
[3]African Institute for Mathematical Sciences [4]South African Astronomical Observatory
[5]Mohamed bin Zayed University of Artificial Intelligence
[6]University of Washington [7]Cohere For AI

Abstract

Neural Machine Translation models are extremely data and compute-hungry. However, not all data points contribute equally to model training and generalization. Data pruning to remove the low-value data points has the benefit of drastically reducing the compute budget without a significant drop in model performance. In this paper, we propose a new data pruning technique: *Checkpoints Across Time* (*CAT*), that leverages early model training dynamics to identify the most relevant data points for model performance. We benchmark *CAT* against several data pruning techniques including COMET-QE, LASER and LaBSE. We find that *CAT* outperforms the benchmarks on Indo-European languages on multiple test sets. When applied to English-German, English-French and English-Swahili translation tasks, *CAT* achieves comparable performance to using the full dataset, while pruning up to 50% of training data. We inspect the data points that *CAT* selects and find that it tends to favor longer sentences and sentences with unique or rare words.

[[Boubdir et al. 2023,Marion et al. 2023, Blakeney et al. 2024, Chimoto et al. 2024]]

Much of our recent work over the last two years has focused on data pruning, prioritization of examples.

When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale

Max Marion
Cohere for AI
maxwell@cohere.com

Ahmet Üstün
Cohere for AI
ahmet@cohere.com

Luiza Pozzobon
Cohere for AI
luiza@cohere.com

Alex Wang
Cohere
alexwang@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Large volumes of text data have contributed significantly to the development of large language models (LLMs) in recent years. This data is typically acquired by scraping the internet, leading to pretraining datasets comprised of noisy web text. To date, efforts to prune these datasets down to a higher quality subset have relied on hand-crafted heuristics encoded as rule-based filters. In this work, we take a wider view and explore scalable estimates of data quality that can be used to systematically measure the quality of pretraining data. We perform a rigorous comparison at scale of the simple data quality estimator of perplexity, as well as more sophisticated and computationally intensive estimates of the Error L2-Norm and memorization. These metrics are used to rank and prune pretraining corpora, and we subsequently compare LLMs trained on these pruned datasets. Surprisingly, we find that the simple technique of perplexity outperforms our more computationally expensive scoring methods. We improve over our no-pruning baseline while training on as little as 30% of the original training dataset. Our work sets the foundation for unexplored strategies in automatically curating high quality corpora and suggests the majority of pretraining data can be removed while retaining performance.

Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning

Shivalika Singh♦[1], Freddie Vargus♦[1], Daniel D'souza♦[1], Börje F. Karlsson♦[2], Abinaya Mahendiran♦[1], Wei-Yin Ko♦[3], Herumb Shandilya♦[1], Jay Patel[4], Deividas Mataciunas[1], Laura O'Mahony[5], Mike Zhang[6], Ramith Hettiarachchi[7], Joseph Wilson[8], Marina Machado[3], Luisa Souza Moura[3], Dominik Krzemiński[1], Hakimeh Fadaei[1], Irem Ergün[3], Ifeoma Okoh[1], Aisha Alaagib[1], Oshan Mudannayake[1], Zaid Alyafeai[9], Vu Minh Chien[1], Sebastian Ruder[3], Surya Guthikonda[1], Emad A. Alghamdi[10], Sebastian Gehrmann[11], Niklas Muennighoff[1], Max Bartolo[3], Julia Kreutzer[12], Ahmet Üstün[12], Marzieh Fadaee[12], and Sara Hooker[12]

[1]Cohere For AI Community, [2]Beijing Academy of Artificial Intelligence, [3]Cohere, [4]Binghamton University, [5]University of Limerick, [6]IT University of Copenhagen, [7]MIT, [8]University of Toronto, [9]King Fahd University of Petroleum and Minerals, [10]King Abdulaziz University, ASAS.AI, [11]Bloomberg LP, [12]Cohere For AI

Corresponding authors: Shivalika Singh <shivalikasingh95@gmail.com>, Marzieh Fadaee <marzieh@cohere.com>, Sara Hooker <sarahooker@cohere.com>

Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation

Meriem Boubdir
Cohere for AI
meri.boubdir@gmail.com

Edward Kim
Cohere
edward@cohere.com

Beyza Ermis
Cohere for AI
beyza@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Abstract

Human evaluation is increasingly critical for assessing large language models, capturing linguistic nuances, and reflecting user preferences more accurately than traditional automated metrics. However, the resource-intensive nature of this type of annotation process poses significant challenges. The key question driving our work: *is it feasible to minimize human-in-the-loop feedback by prioritizing data instances which most effectively distinguish between models?* We evaluate several metric-based methods and find that these metrics enhance the efficiency of human evaluations by minimizing the number of required annotations, thus saving time and cost, while ensuring a robust performance evaluation. We show that our method is effective across widely used model families, reducing instances of indecisive (or "tie") outcomes by up to 54% compared to a random sample when focusing on the top-20 percentile of prioritized instances. This potential reduction in required human effort positions our approach as a valuable strategy in future large language model evaluations.

Pretraining Scale

[[Marion et al. 2023]]

Instruction Finetuning Pruning and Dataset Weighting

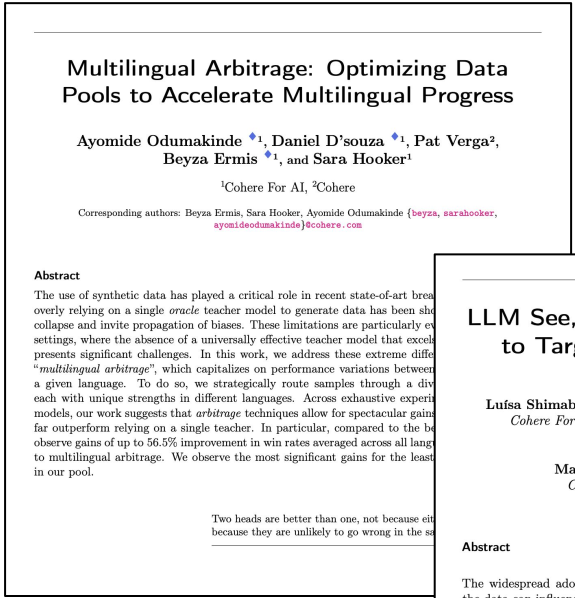
[[Singh et al. 2023]]

Prioritizing human annotation

[[Boubdir et al. 2023]]

2. Moving away from static datasets.

"Optimize in the data-space to steer on-the-fly towards desirable properties."



Multilingual Arbitrage: Optimizing Data Pools to Accelerate Multilingual Progress

Ayomide Odumakinde ♦1, Daniel D'souza ♦1, Pat Verga2, Beyza Ermis ♦1, and Sara Hooker1

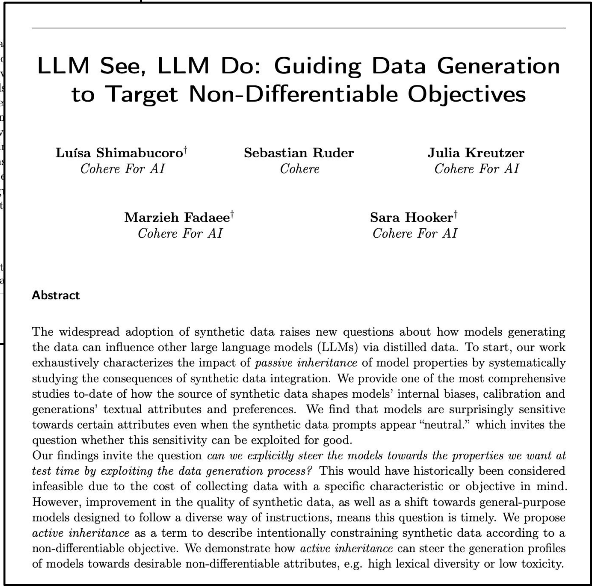
1Cohere For AI, 2Cohere

Corresponding authors: Beyza Ermis, Sara Hooker, Ayomide Odumakinde {beyza, sarahooker, ayomideodumakinde}@cohere.com

Abstract

The use of synthetic data has played a critical role in recent state-of-art brea overly relying on a single *oracle* teacher model to generate data has been sho collapse and invite propagation of biases. These limitations are particularly e settings, where the absence of a universally effective teacher model that excels presents significant challenges. In this work, we address these extreme diffe "*multilingual arbitrage*", which capitalizes on performance variations between a given language. To do so, we strategically route samples through a div each with unique strengths in different languages. Across exhaustive experi models, our work suggests that *arbitrage* techniques allow for spectacular gains far outperform relying on a single teacher. In particular, compared to the be observe gains of up to 56.5% improvement in win rates averaged across all lang to multilingual arbitrage. We observe the most significant gains for the least in our pool.

Two heads are better than one, not because eit because they are unlikely to go wrong in the sa



LLM See, LLM Do: Guiding Data Generation to Target Non-Differentiable Objectives

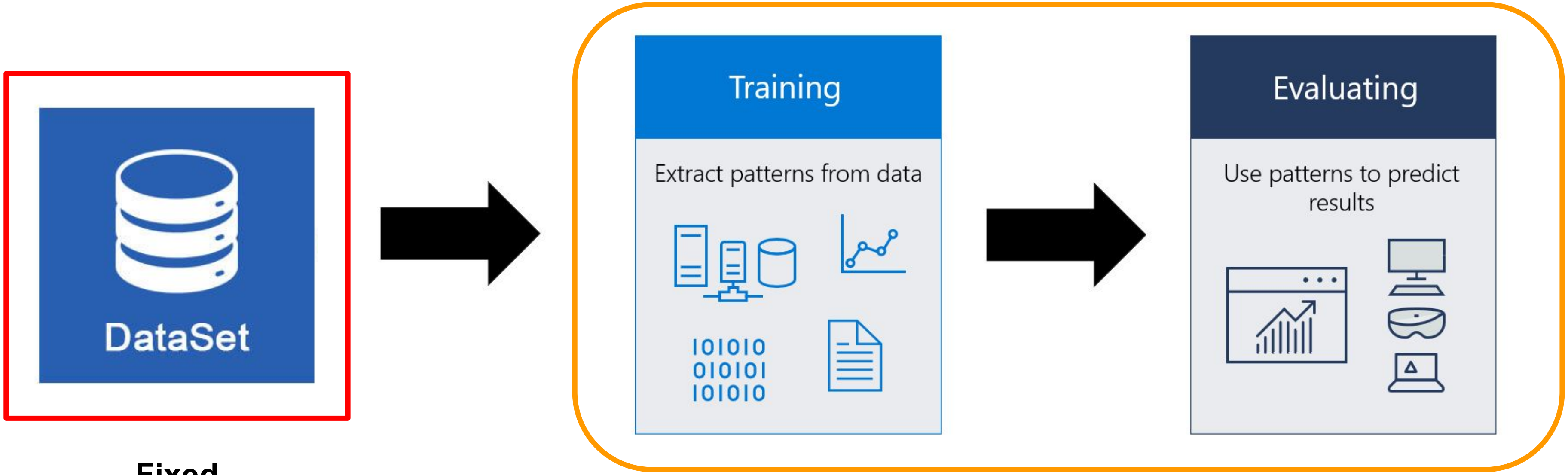
Luísa Shimabucoro† Cohere For AI; Sebastian Ruder Cohere; Julia Kreutzer Cohere For AI; Marzieh Fadaee† Cohere For AI; Sara Hooker† Cohere For AI

Abstract

The widespread adoption of synthetic data raises new questions about how models generating the data can influence other large language models (LLMs) via distilled data. To start, our work exhaustively characterizes the impact of *passive inheritance* of model properties by systematically studying the consequences of synthetic data integration. We provide one of the most comprehensive studies to-date of how the source of synthetic data shapes models' internal biases, calibration and generations' textual attributes and preferences. We find that models are surprisingly sensitive towards certain attributes even when the synthetic data prompts appear "neutral." which invites the question whether this sensitivity can be exploited for good.
Our findings invite the question *can we explicitly steer the models towards the properties we want at test time by exploiting the data generation process?* This would have historically been considered infeasible due to the cost of collecting data with a specific characteristic or objective in mind. However, improvement in the quality of synthetic data, as well as a shift towards general-purpose models designed to follow a diverse way of instructions, means this question is timely. We propose *active inheritance* as a term to describe intentionally constraining synthetic data according to a non-differentiable objective. We demonstrate how *active inheritance* can steer the generation profiles of models towards desirable non-differentiable attributes, e.g. high lexical diversity or low toxicity.

[[Odumakinde et al 2024, Shimabucoro et al 2024.]]

ML researchers have historically treated as data to be fixed, something to be worked around rather than something they can control.



Fixed
(dataset creators are usually distinct from dataset consumers)

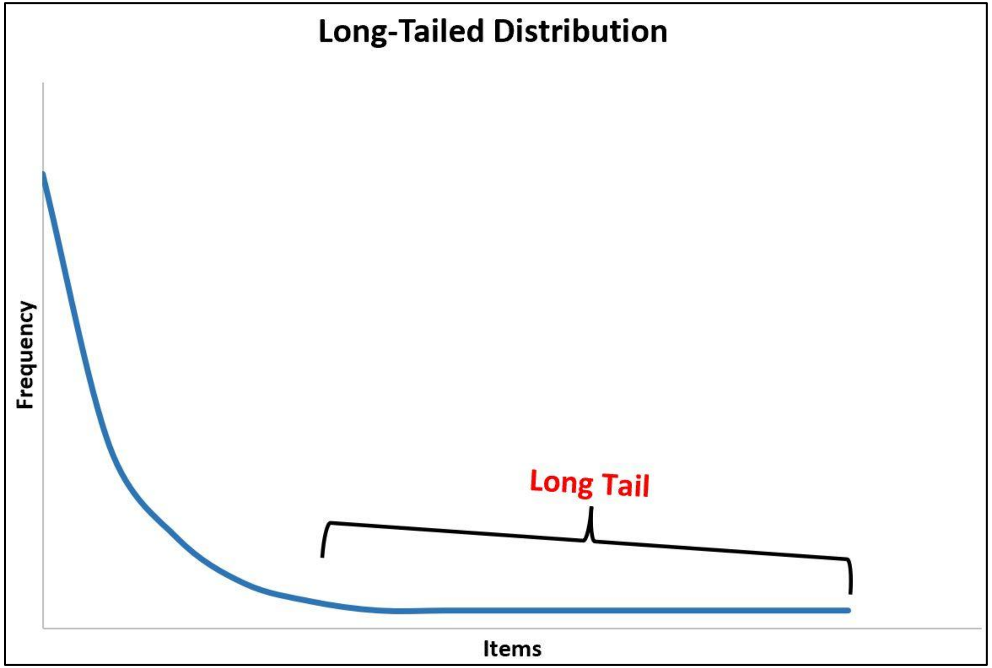
Focus of the ML community

This also meant we were stuck with the quality of datasets collected.

MS CELEB dataset

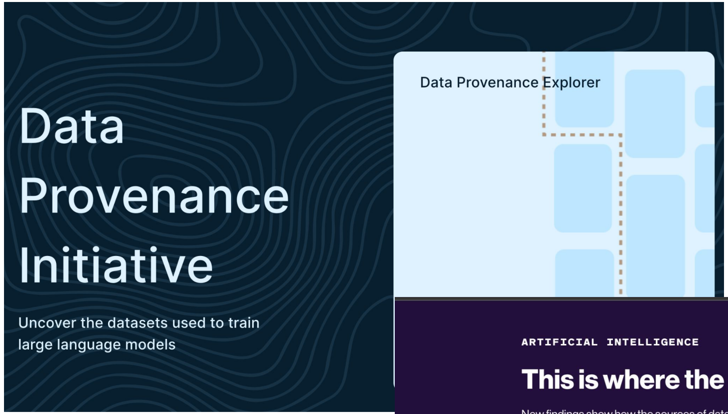
Most of machine learning has been built around the assumption that we sample IID from the underlying distribution we want to model.

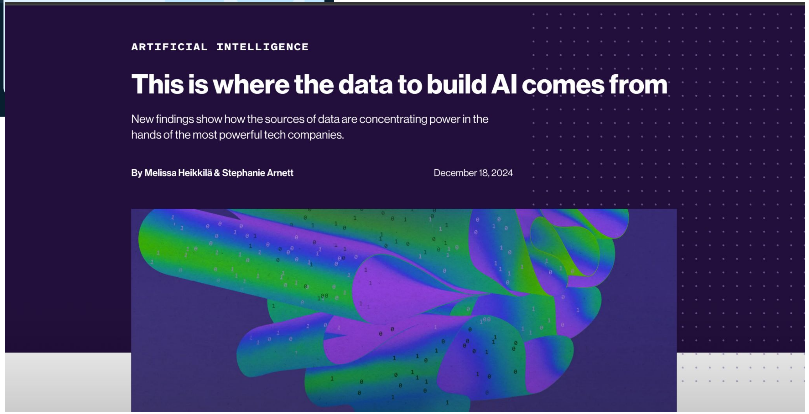
However, this is highly inefficient – because it means we have to wade through a lot of frequent examples before we start to learn the rare examples.



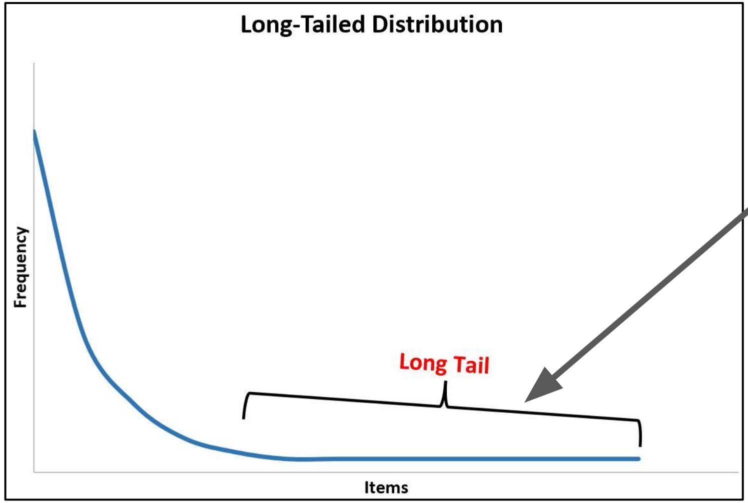
However, we are in the midst of one of the most profound paradigm shifts. Advances in synthetic data make it much more interesting to imagine the data space as malleable.

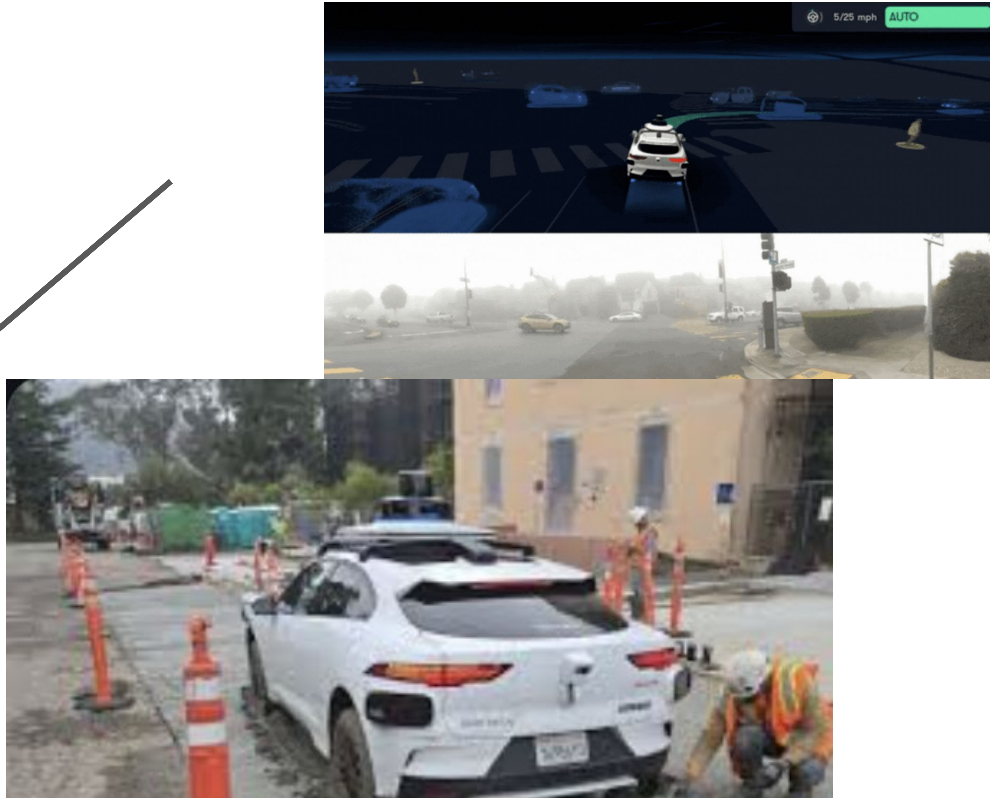
1. Large scale annotation from llms allow for more malleable annotation categories.





Targeted synthetic data creation allows us to oversample from parts of the distribution we deem important but isn't well represented in a random collected sample.





Now we can start to optimize and steer in the data space. We have done significant work on this over the last year – we call this "active inheritance."

Can we explicitly steer the models towards the properties we want at test time by exploiting the data generation process?

LLM See, LLM Do: Guiding Data Generation to Target Non-Differentiable Objectives

Luísa Shimabucoro[†] *Cohere For AI*

Sebastian Ruder *Cohere*

Julia Kreutzer *Cohere For AI*

Marzieh Fadaee[†] *Cohere For AI*

Sara Hooker[†] *Cohere For AI*

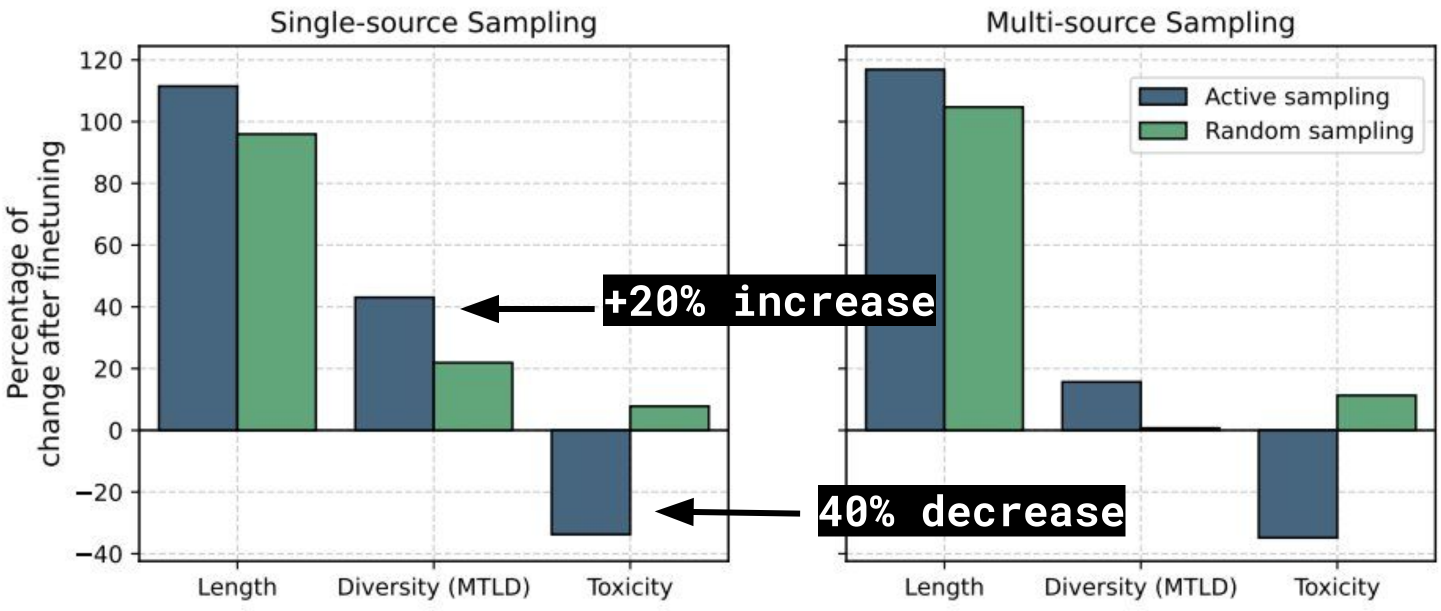
Abstract

The widespread adoption of synthetic data raises new questions about how models generating the data can influence other large language models (LLMs) via distilled data. To start, our work exhaustively characterizes the impact of *passive inheritance* of model properties by systematically studying the consequences of synthetic data integration. We provide one of the most comprehensive studies to-date of how the source of synthetic data shapes models' internal biases, calibration and generations' textual attributes and preferences. We find that models are surprisingly sensitive towards certain attributes even when the synthetic data prompts appear "neutral." which invites the question whether this sensitivity can be exploited for good.

Our findings invite the question *can we explicitly steer the models towards the properties we want at test time by exploiting the data generation process?* This would have historically been considered infeasible due to the cost of collecting data with a specific characteristic or objective in mind. However, improvement in the quality of synthetic data, as well as a shift towards general-purpose models designed to follow a diverse way of instructions, means this question is timely. We propose *active inheritance* as a term to describe intentionally constraining synthetic data according to a

Shimabucoro et al. 2024

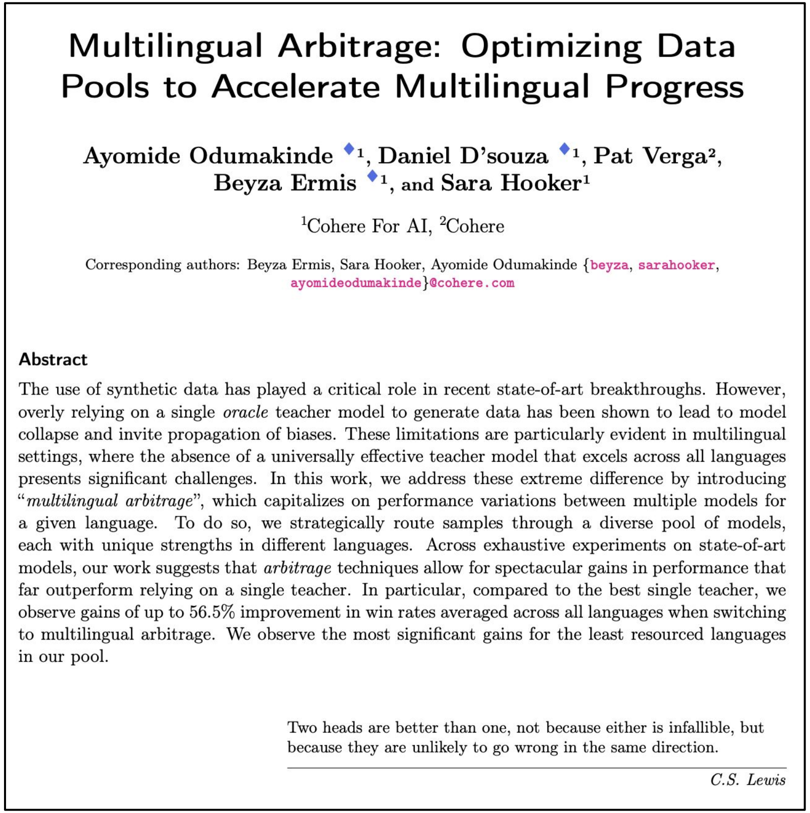
Our recent work show significant gains when we explicitly steer data generations toward non-differentiable properties (toxicity, length).



Cohere For AI

Shimabucoro et al. 2024

We also show that we can dramatically improve performance by targeting to pick the best teacher model for parts of the distribution we care about.



Multilingual Arbitrage: Optimizing Data Pools to Accelerate Multilingual Progress

Ayomide Odumakinde ♦[1], Daniel D'souza ♦[1], Pat Verga[2], Beyza Ermis ♦[1], and Sara Hooker[1]

[1]Cohere For AI, [2]Cohere

Corresponding authors: Beyza Ermis, Sara Hooker, Ayomide Odumakinde {beyza, sarahooker, ayomideodumakinde}@cohere.com

Abstract

The use of synthetic data has played a critical role in recent state-of-art breakthroughs. However, overly relying on a single *oracle* teacher model to generate data has been shown to lead to model collapse and invite propagation of biases. These limitations are particularly evident in multilingual settings, where the absence of a universally effective teacher model that excels across all languages presents significant challenges. In this work, we address these extreme difference by introducing "*multilingual arbitrage*", which capitalizes on performance variations between multiple models for a given language. To do so, we strategically route samples through a diverse pool of models, each with unique strengths in different languages. Across exhaustive experiments on state-of-art models, our work suggests that *arbitrage* techniques allow for spectacular gains in performance that far outperform relying on a single teacher. In particular, compared to the best single teacher, we observe gains of up to 56.5% improvement in win rates averaged across all languages when switching to multilingual arbitrage. We observe the most significant gains for the least resourced languages in our pool.

Two heads are better than one, not because either is infallible, but because they are unlikely to go wrong in the same direction.

C.S. Lewis

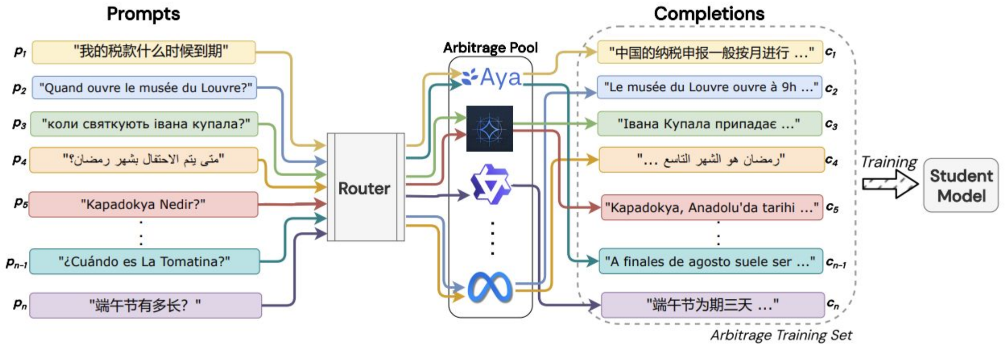


Figure 1: **Overview of *Multilingual Arbitrage*.** Instead of relying on a single "oracle" teacher, multilingual arbitrage re-frames the distillation problem as learning how to optimize sampling for a desired part of the data distribution from an ensemble of teachers.

avoids mode collapse - leveraging pools of models with different strengths to compose data distribution.

[Ayomide Odumakinde et al. 2025](#)

What can we do when we don’t allow for any gradient updates?

Increasingly, optimization has moved post training.

What are some optimization approaches that don’t require gradient updates but greatly improve model performance?

A profound shift in how we optimize is underway. We are in an era where we can learn “on-the-fly” – and adapt models based upon immediate context.

Gradient free performance boosts.

Changes model itself:

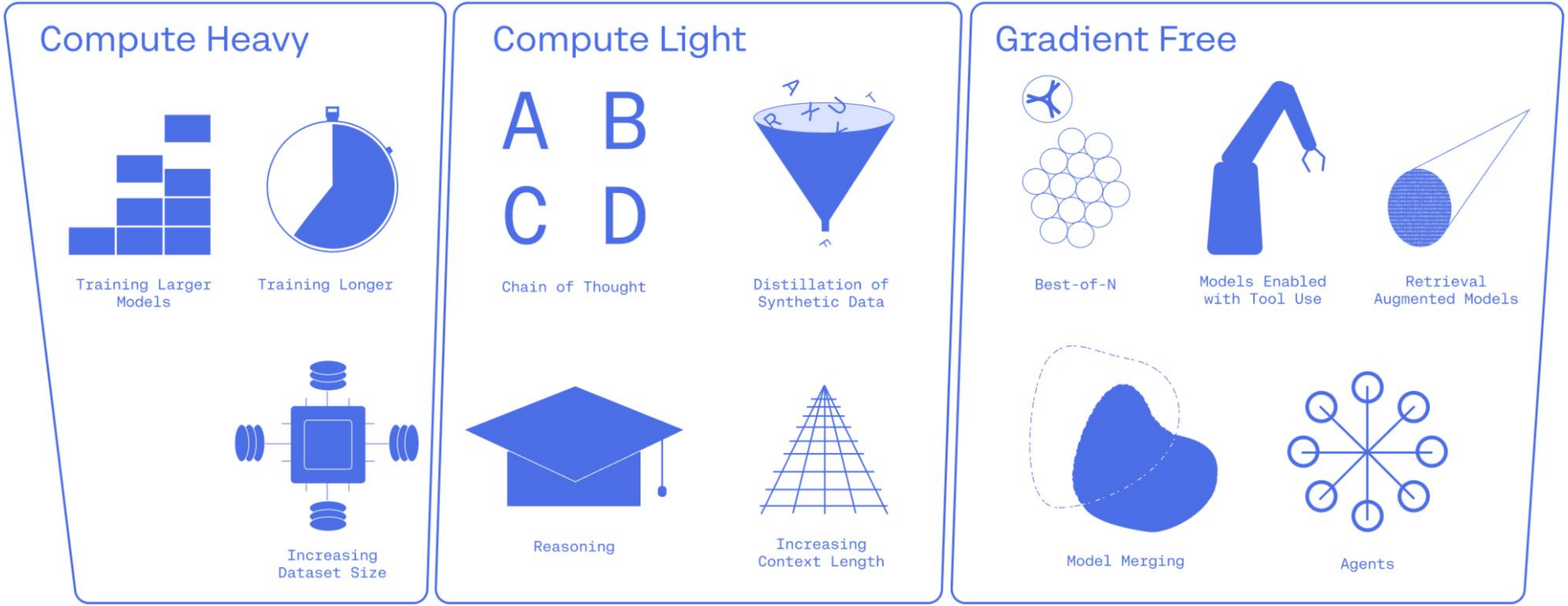
Merging
RAG

Navigates search space of solution:

Inference scaling

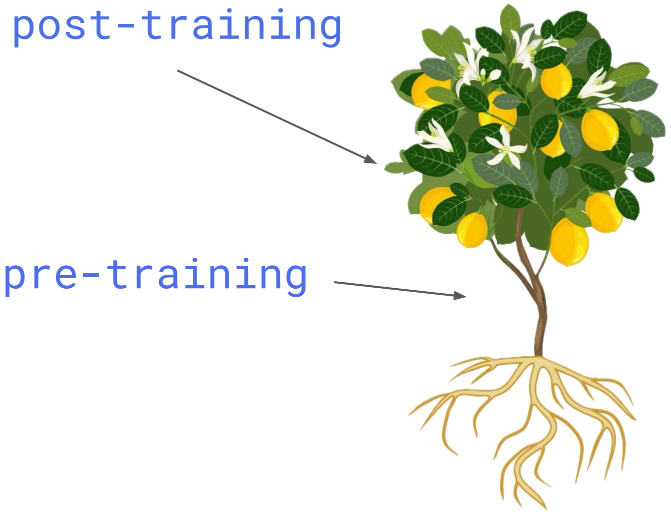
Conditions response in-place to immediate feedback

Many techniques which add large boosts to performance do not require **any additional gradient updates**.



Cohere Labs

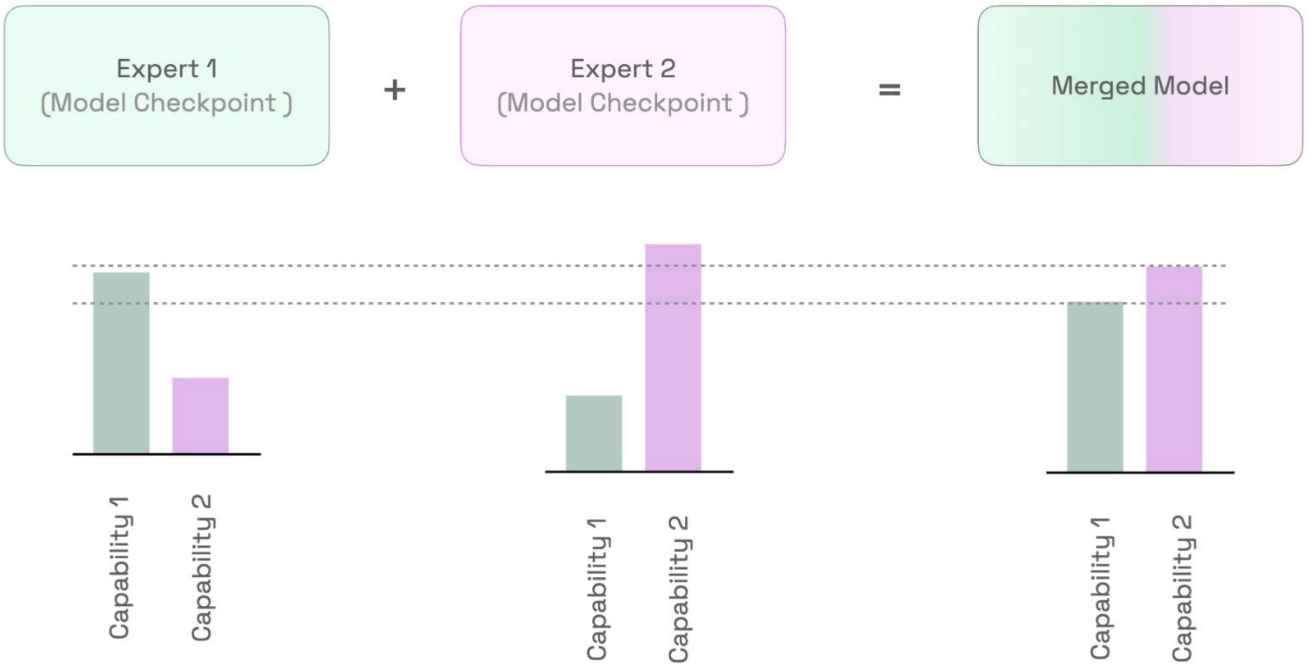
You can think of merging as bonsai grafting – you can target inheriting certain capabilities from a pool of models.



Pre-Training	Post-Training

Pre-Training	Post-Training

Model merging combines two or more neural networks into a single model by combining the weights. No gradient updates are needed.



Merging requires no additional training, and often preserves performance while introducing new capabilities.

Mix Data or Merge Models? Optimizing for Diverse Multi-Task Learning

Aakanksha[1], Arash Ahmadian[1,2], Seraphina Goldfarb-Tarrant[2], Beyza Ermis[1], Marzieh Fadaee[1], and Sara Hooker[1]

[1]Cohere For AI, [2]Cohere

Corresponding authors: Aakanksha, Marzieh Fadaee, Sara Hooker {aakanksha, marzieh, sarahooker}@cohere.com

Abstract

Large Language Models (LLMs) have been adopted and deployed worldwide for a broad variety of applications. However, ensuring their safe use remains a significant challenge. Preference training and safety measures often overfit to harms prevalent in Western-centric datasets, and safety protocols frequently fail to extend to multilingual settings. In this work, we explore model merging in a diverse multi-task setting, combining safety and general-purpose tasks within a multilingual context. Each language introduces unique and varied learning challenges across tasks. We find that objective-based merging is more effective than mixing data, with improvements of up to 8% and 10% in general performance and safety respectively. We also find that language-based merging is highly effective — by merging monolingually fine-tuned models, we achieve a 4% increase in general performance and 7% reduction in harm across all languages on top of the data mixtures method using the same available data. Overall, our comprehensive study of merging approaches provides a useful framework for building strong and safe multilingual models.

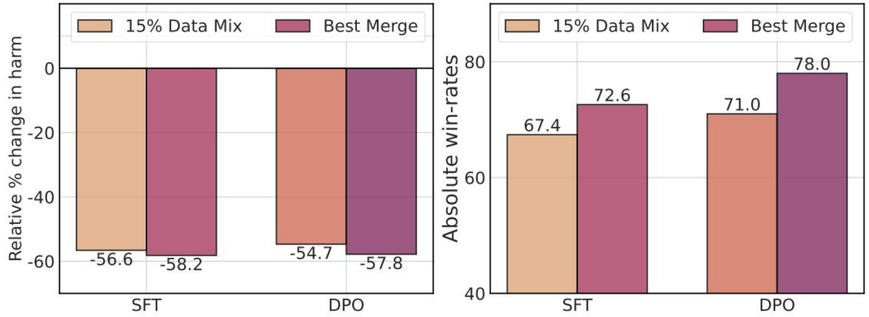


Figure 2: *Mixing versus merging*: Safety and general performance of a *15% Safety Mix* model (§2.2) against SLERP merging, which emerges as the best method for balancing trade-offs, for both SFT and DPO based checkpoints. Lower is better for (a) and higher is better for (b). Both metrics are measured with respect to the Aya 23 base model.

Aakanksha et al. 2024

Aya Vision extends multimodal performance to multilingual.

Aya Vision: Advancing the Frontier of Multilingual Multimodality

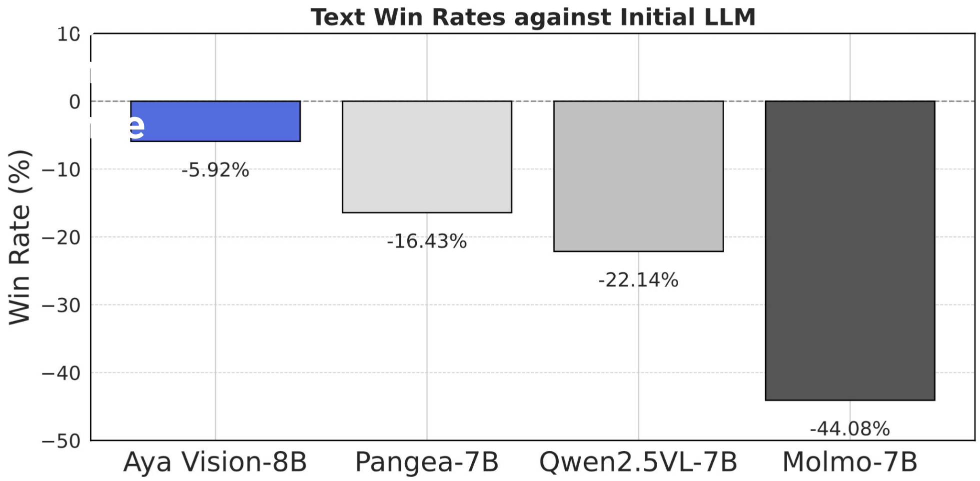
Saurabh Dash[★1], Yiyang Nan[★1], John Dang[1], Arash Ahmadian[1,2], Shivalika Singh[1], Madeline Smith[1], Bharat Venkitesh[2], Vlad Shmyhlo[2], Viraat Aryabumi[2], Walter Beller-Morales[2], Jeremy Pekmez[2], Jason Ozuzu[2], Pierre Richemond[2], Acyr Locatelli[2], Nick Frosst[2], Phil Blunsom[2], Aidan Gomez[2], Ivan Zhang[2], Marzieh Fadaee[1], Manoj Govindassamy[2], Sudip Roy[2], Matthias Gallé[♦1], Beyza Ermis[♦1], Ahmet Üstün[♦1], and Sara Hooker[♦1]

[1]Cohere Labs, [2]Cohere

Corresponding authors: {saurabh, olivernan, matthias, beyza, ahmet, sarahooker}@cohere.com

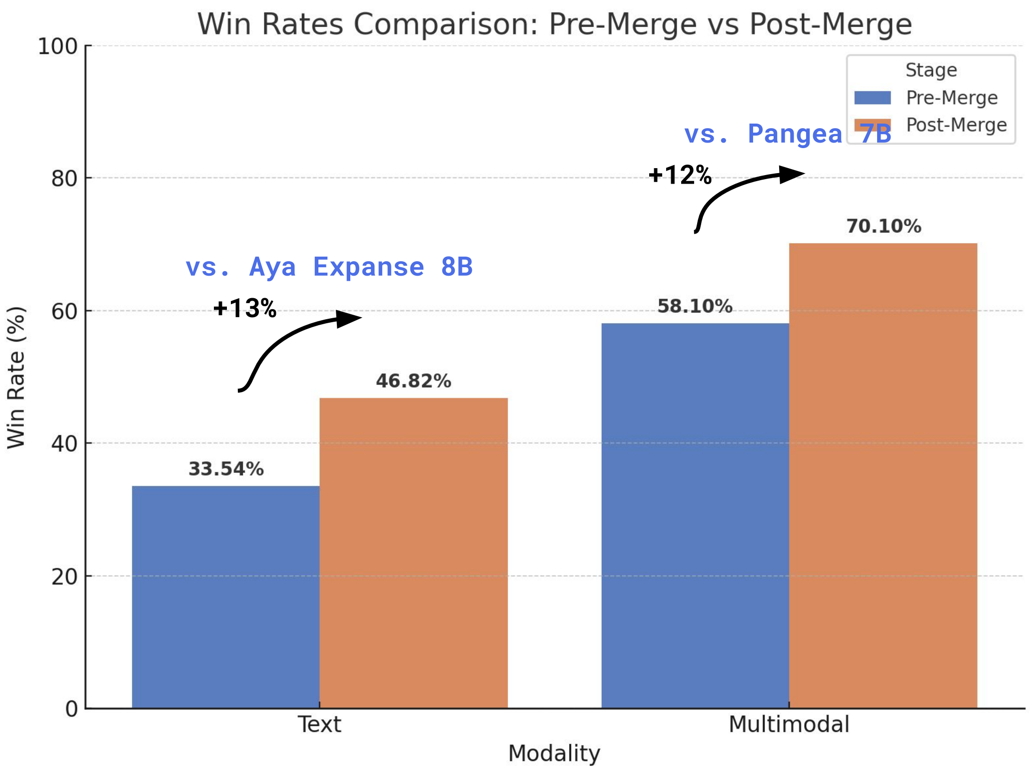
Abstract

Building multimodal language models is fundamentally challenging: it requires aligning vision and language modalities, curating high-quality instruction data, and avoiding the degradation of existing text-only capabilities once vision is introduced. These difficulties are further magnified in the multilingual setting, where the need for multimodal data in different languages exacerbates existing data scarcity, machine translation often distorts meaning, and catastrophic forgetting is more pronounced. To address the aforementioned challenges, we introduce novel techniques spanning both data and modeling. First, we develop a synthetic annotation framework that curates high-quality, diverse multilingual multimodal instruction data, enabling Aya Vision models to produce natural, human-preferred responses to multimodal inputs across many languages. Complementing this, we propose a cross-modal model merging technique that mitigates catastrophic forgetting, effectively preserving text-only capabilities while simultaneously enhancing multimodal generative performance. Aya-Vision-8B achieves best-in-class performance compared to strong multimodal models such as Qwen-2.5-VL-7B, Pixtral-12B, and even much larger Llama-3.2-90B-Vision. We



Just finetuning results in large text degradation.

We can avoid degradation by merging – add in new capabilities without compromising existing performance.



Merging not only boosts text win-rates **but also vision win-rates!!!**

There are considerable benefits and simplicity to merging – for inheriting desirable capabilities while preserving existing behavior.

Command A:
An Enterprise-Ready Large Language Model

Abstract

In this report we desc
excel at real-world en
with support for 23 la
of the range perform
grounding and tool u
a decentralised traini
also include results fo
Weights for both mo
training pipeline and
and public benchmar

Aya Vision: Advancing the Frontier of Multilingual Multimodality

Saurabh Dash[★1], Yiyang Nan[★1],
Shivalika Singh[1], Madeline
Vlad Shmyhlo[2], Viraat Aryab
Jeremy Pekmez[2], Pierre Richemon
Phil Blunsom[2], Aidan Gomez[2],
Manoj Govindassamy[2], Sud
Beyza Ermis[◆1], Ahmet Üs

[1]Cohere Lab

Corresponding authors: {saurabh, olivernan, matthias

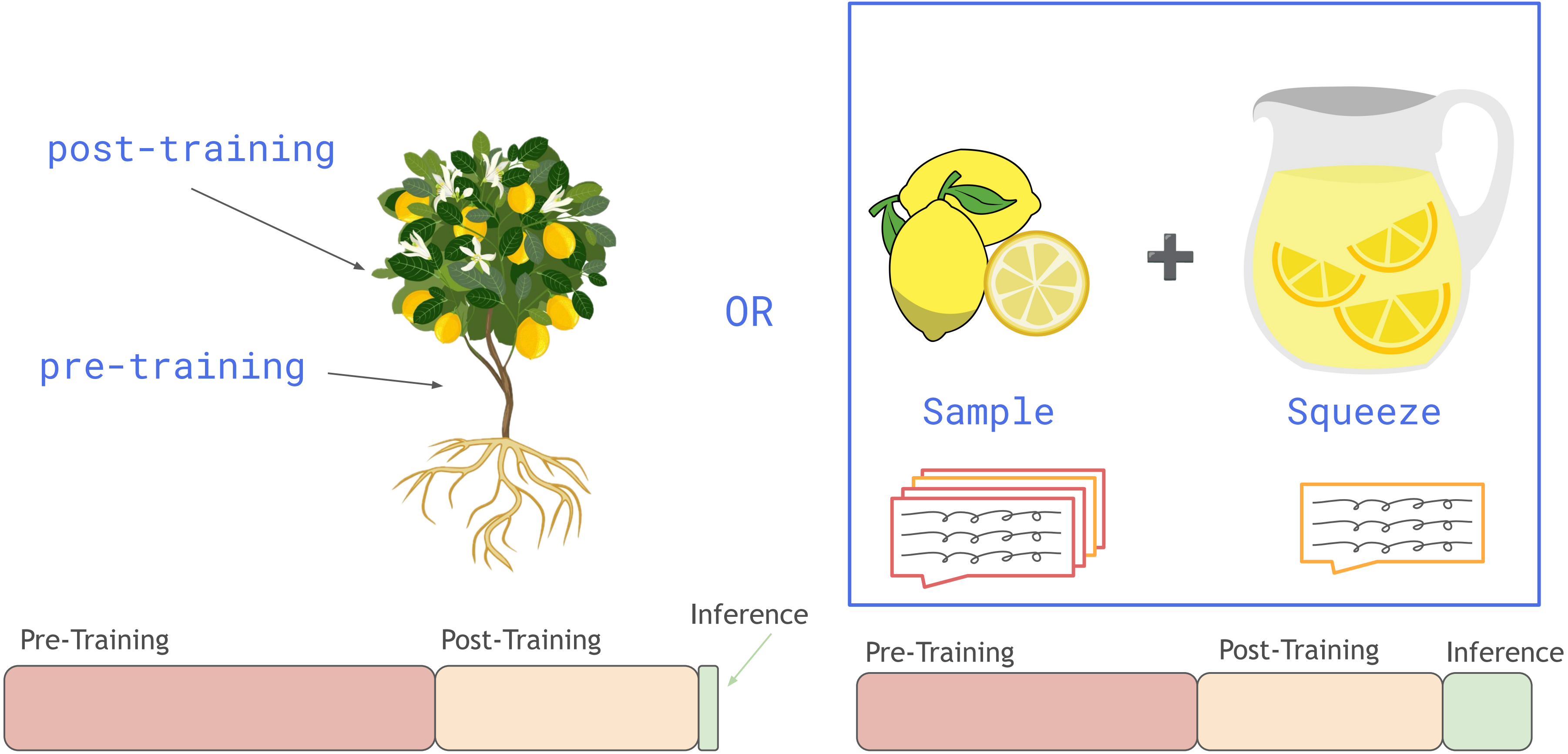
Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier

John Dang[◆1], Shivalika Singh[◆1], Daniel D'souza[◆1],
Arash Ahmadian[◆1], Alejandro Salamanca[1], Madeline Smith[1],
Aidan Peppin[1], Sungjin Hong[2], Manoj Govindassamy[2],
Terrence Zhao[2], Sandra Kublik[2], Meor Amer[2], Viraat Aryabumi[2],
Jon Ander Campos[2], Yi-Chern Tan[2], Tom Kocmi[2], Florian Strub[2],
Nathan Grinsztajn[2], Yannis Flet-Berliac[2], Acyr Locatelli[2],
Hangyu Lin[2], Dwarak Talupuru[2], Bharat Venkitesh[2],
David Cairuz[2], Bowen Yang[2], Tim Chung[2], Wei-Yin Ko[2],
Sylvie Shang Shi[2], Amir Shukayev[2], Sammie Bae[2],
Aleksandra Piktus[2], Roman Castagné[2], Felipe Cruz-Salinas[2],
Eddie Kim[2], Lucas Crawhall-Stein[2], Adrien Morisot[2], Sudip Roy[2],
Phil Blunsom[2], Ivan Zhang[2], Aidan Gomez[2], Nick Frosst[1,2],
Marzieh Fadaee[★1], Beyza Ermis[★1], Ahmet Üstün[★1],
and Sara Hooker[★1]

[1]Cohere For AI, [2]Cohere

Corresponding authors: Ahmet Üstün <ahmet@cohere.com>, Sara Hooker <sarahooker@cohere.com>

Inference time techniques spend more time on selecting which fruit to pick, and how to squeeze (combine) the best fruit.



With inference compute, you spend a fraction of the compute during pre-training but see large gains.

Cohere Labs

When Life Gives You Samples

The Benefits of Scaling up Inference Compute for Multilingual LLMs

Ammar Khairi[★1], Daniel D'souza[1], Ye Shen[2], Julia Kreutzer[♦1], and Sara Hooker[♦1]

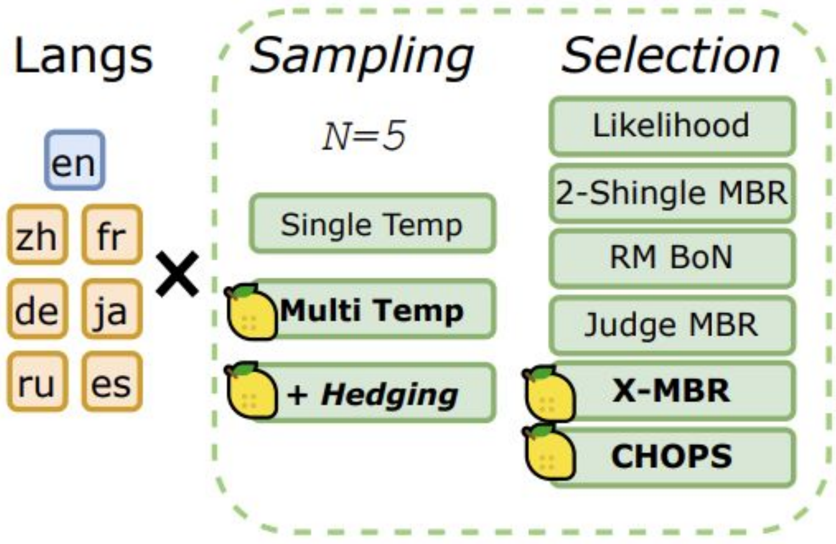
[1]Cohere Labs, [2]Cohere

Corresponding authors: {ammar, juliakreutzer, sarahooker}@cohere.com

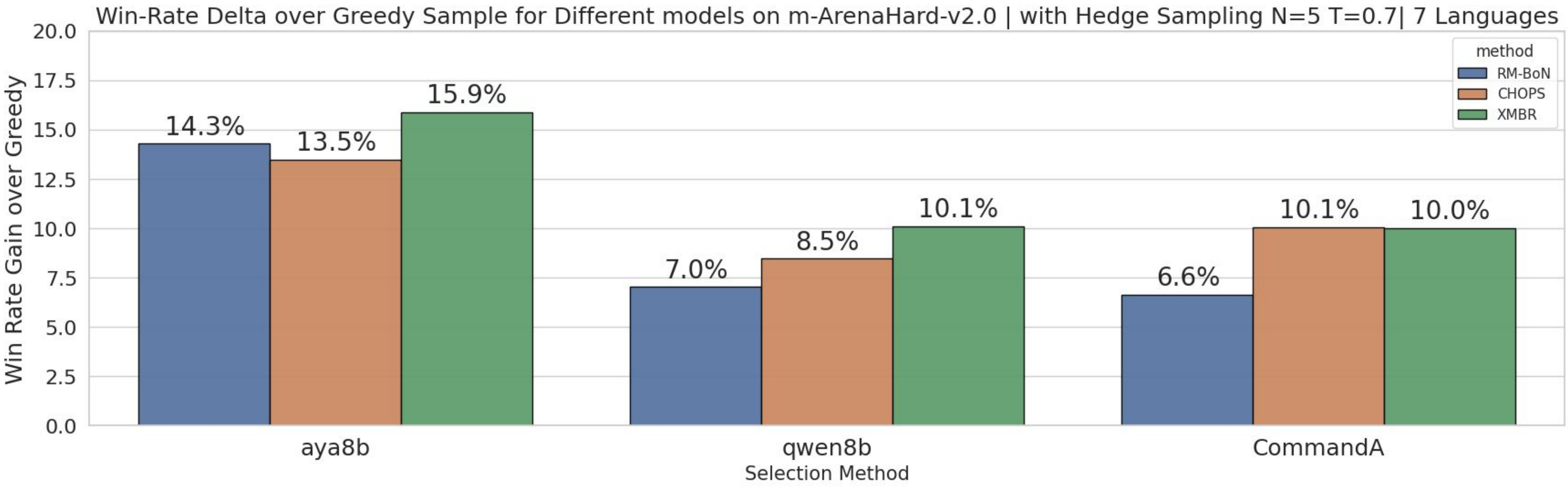
Abstract

Recent advancements in large language models (LLMs) have shifted focus toward scaling inference-time compute—improving performance without retraining the model. A common approach is to sample multiple outputs in parallel, and select one of these as the final output. However, work to date has focused on English and a handful of domains such as math and code. In contrast, we are most interested in techniques that generalize across open-ended tasks, formally verifiable tasks, and across languages. In this work, we study how to robustly scale inference-time compute for open-ended generative tasks in a multilingual, multi-task setting.

Our findings show that both sampling strategy—based on temperature variation—and selection



By strategically sampling up to 5 samples, 10%+ winrates on state-of-art Command-A 100 billion parameter model.



Khairi et al, 2025

Controllability and continuous learning techniques aim to remove the burden from the user, by inferring from the context and automatically adapting the model.

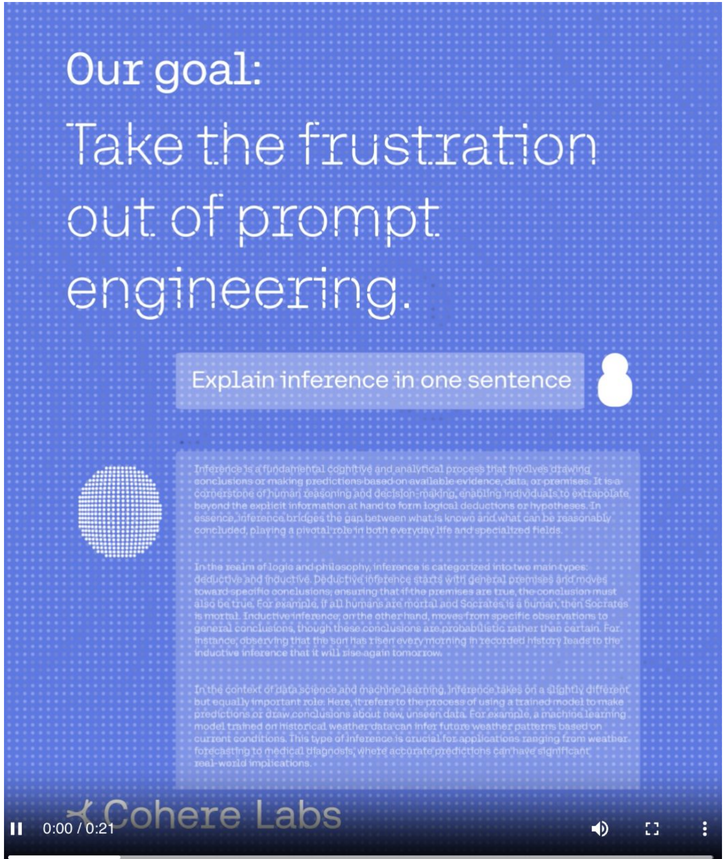
Avoid everyone have to be an expert prompt engineer.

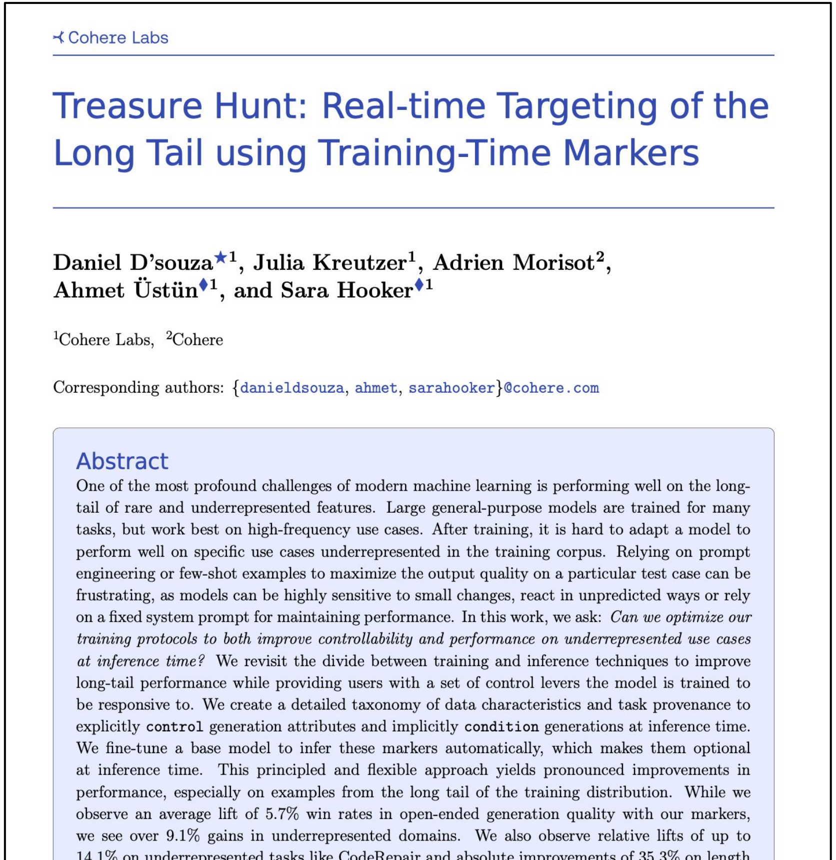
Respond to user feedback real-time.

Continuously adapt to new inputs.

Post-Training

One of our goals: make expert prompting a hack of the past.





Cohere Labs

Treasure Hunt: Real-time Targeting of the Long Tail using Training-Time Markers

Daniel D'souza★[1], **Julia Kreutzer**[1], **Adrien Morisot**[2], **Ahmet Üstün**♦[1], **and Sara Hooker**♦[1]

[1]Cohere Labs, [2]Cohere

Corresponding authors: {danieldsouza, ahmet, sarahooker}@cohere.com

Abstract

One of the most profound challenges of modern machine learning is performing well on the long-tail of rare and underrepresented features. Large general-purpose models are trained for many tasks, but work best on high-frequency use cases. After training, it is hard to adapt a model to perform well on specific use cases underrepresented in the training corpus. Relying on prompt engineering or few-shot examples to maximize the output quality on a particular test case can be frustrating, as models can be highly sensitive to small changes, react in unpredicted ways or rely on a fixed system prompt for maintaining performance. In this work, we ask: *Can we optimize our training protocols to both improve controllability and performance on underrepresented use cases at inference time?* We revisit the divide between training and inference techniques to improve long-tail performance while providing users with a set of control levers the model is trained to be responsive to. We create a detailed taxonomy of data characteristics and task provenance to explicitly `control` generation attributes and implicitly `condition` generations at inference time. We fine-tune a base model to infer these markers automatically, which makes them optional at inference time. This principled and flexible approach yields pronounced improvements in performance, especially on examples from the long tail of the training distribution. While we observe an average lift of 5.7% win rates in open-ended generation quality with our markers, we see over 9.1% gains in underrepresented domains. We also observe relative lifts of up to 14.1% on underrepresented tasks like CodeRepair and absolute improvements of 35.3% on length

We predict a complex taxonomy of treasure markers, which guides the model to higher performance.

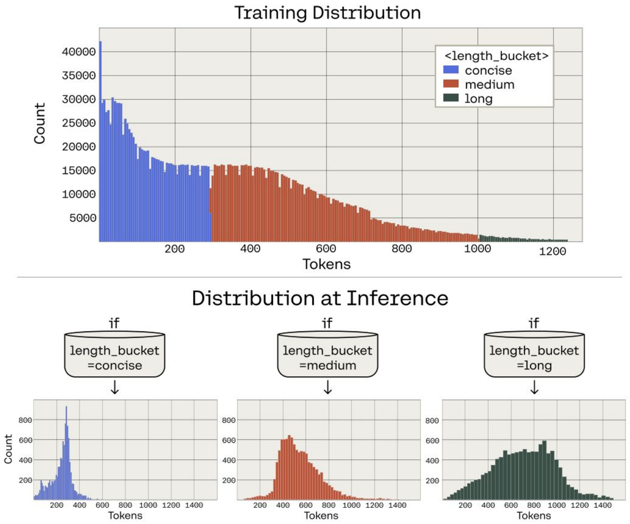
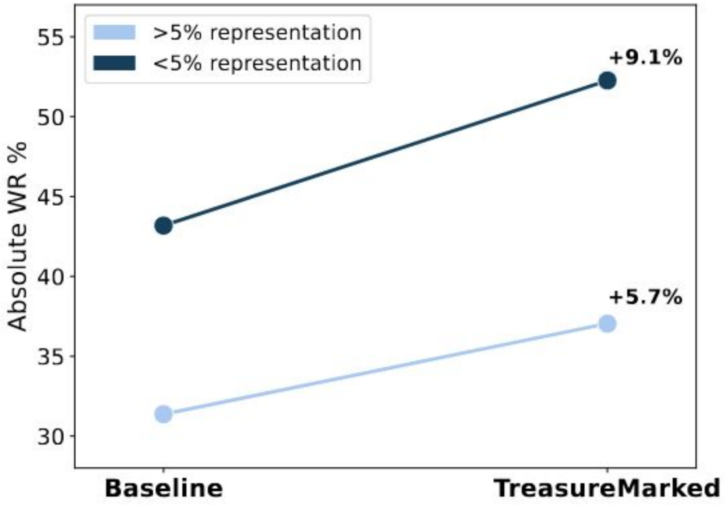


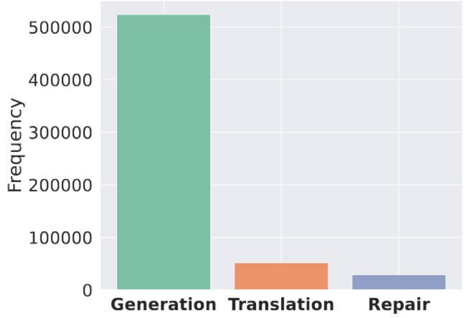
Figure 1: **Tapping into Distributions**: **(above)** illustrates the representation of various length buckets in the training distribution. **(below)** demonstrates the flexibility of the marker intervention on the mArena Hard test distribution. By modifying the <length_bucket>..</length_bucket> marker, the model can effectively tap into diverse training distributions, even for underrepresented length buckets.

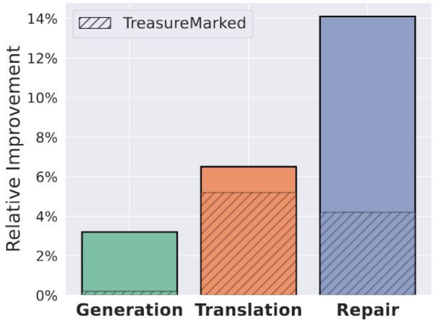
Prefix conditioning allows for more controllability at inference time.

ArenaHard Win Rates

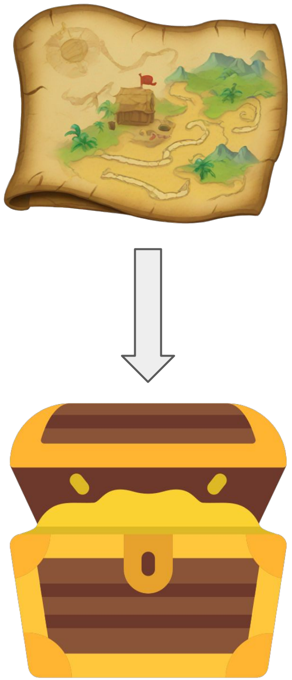


Coding



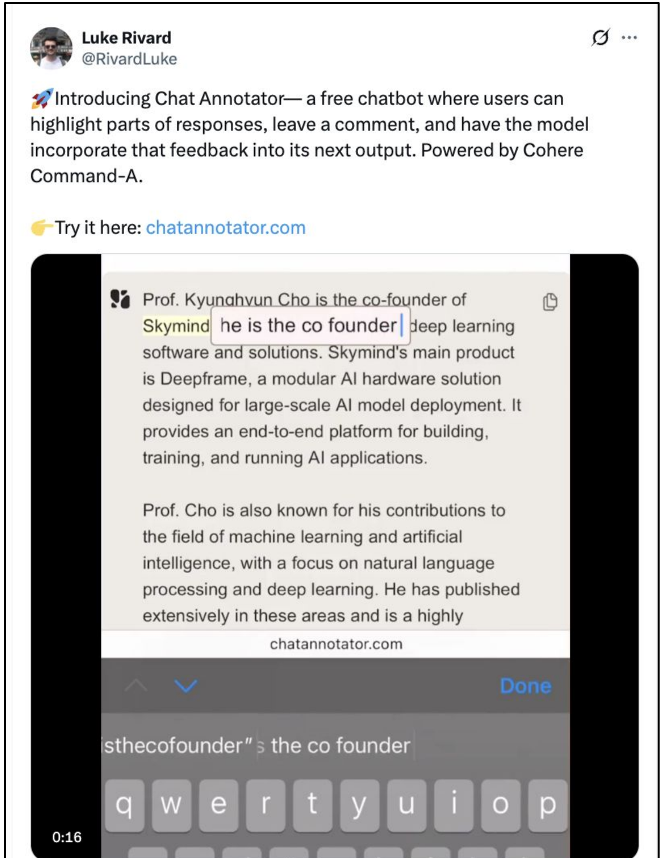


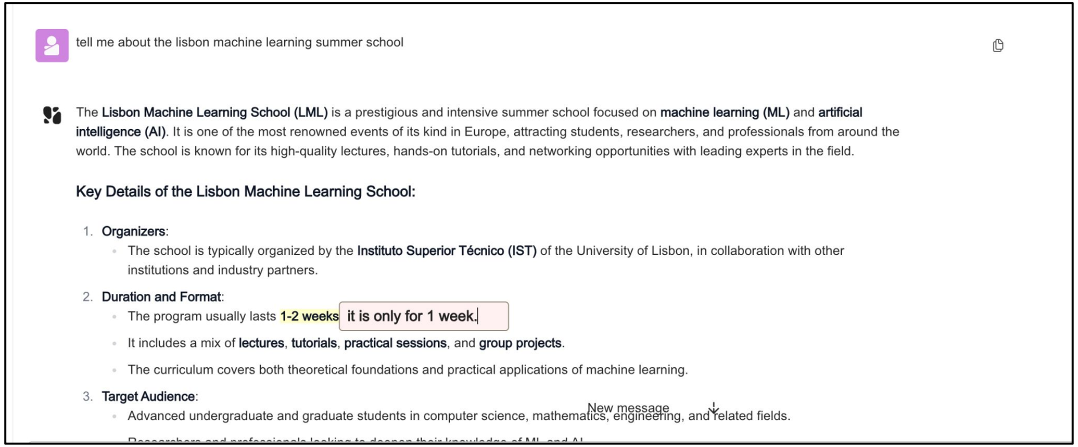
Map Markers at Training



Uncover Gains at Inference

Co-design of both model and interface.





So where does that leave us.

On a big picture level – gradient free improvements are also more similar to our own intelligence.

- Our intelligence is not individual, but collective.
- While our brain develops over the course of our lifetime, human intelligence is ever more collective and expanded based upon dynamic pooling of knowledge.
- Coordination of our intelligence does not require global updates, it is extremely cheap. It is driven by our societal ties.

It is very possible that the next breakthrough will require a fundamentally different way of modelling the world

with a different combination of hardware, software and algorithm.

Model scale is the least interesting idea to throw at a problem. Increasingly, we should justify additional complexity and bend scaling curves by focusing on efficiency.

1950's — **2012** — **What next?**

Our wider space for optimization will further amplify the divide between academia and industry. If intelligence is about interaction and continuous adaptation, control of the environment matters.

At the very least – an expanded optimization space makes our lives a lot more interesting. There is a lot of fun to be had over the next 10 years.

Let's open up for questions and discussion.

Feel free to reach out:

sarahookr@gmail.com