

Natural Language Processing

Lecture 22: NLP for Low-Resource Languages

Some *slides & instruction ideas borrowed from:*
Greg Durret, Mohit Iyyer, & Mar'Aurelio Ranzato

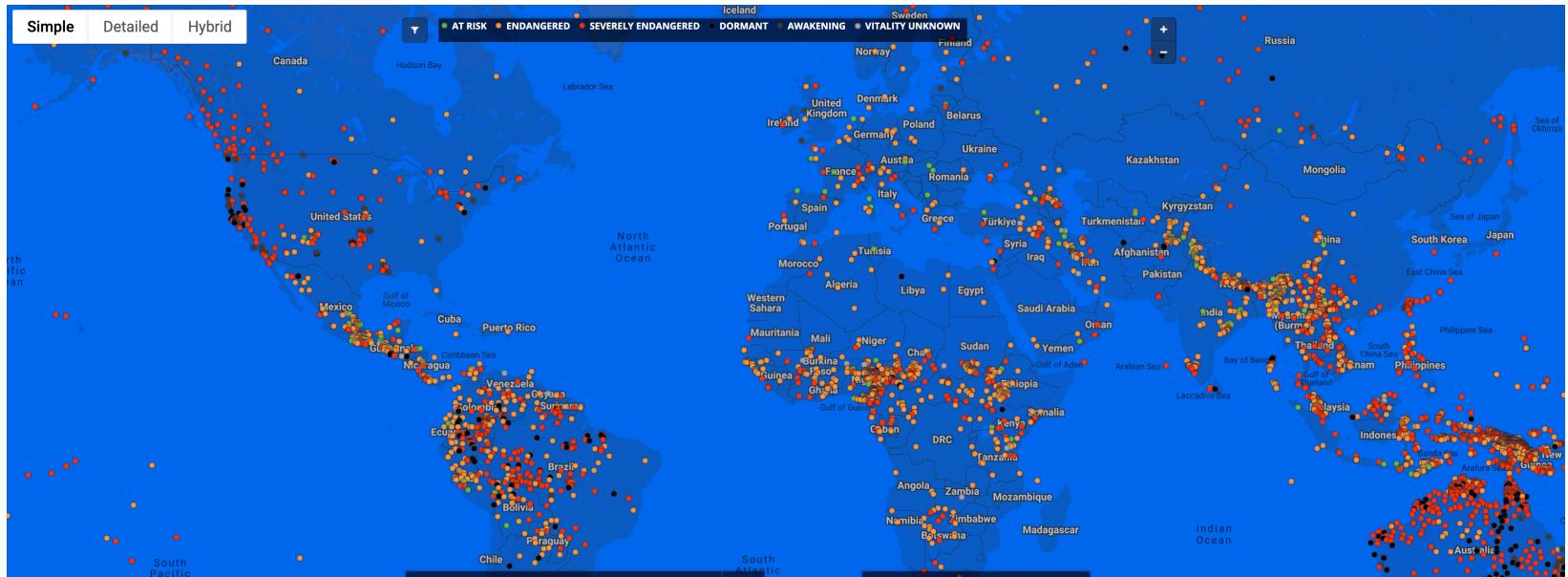
Logistics

- Homework 6 due this Thursday (April 18)
- AP2 and 259 Mid-project reports are being graded.
- AP3 is due April 26
- Tonight: NLP for low resource languages

So far . . .

- Mostly: NLP for English
- Other languages:
 - Machine Translation
 - Tokenization
 - Parsing & Semantics:
 - Universal Dependency Bank
 - FrameNet

Languages of the World

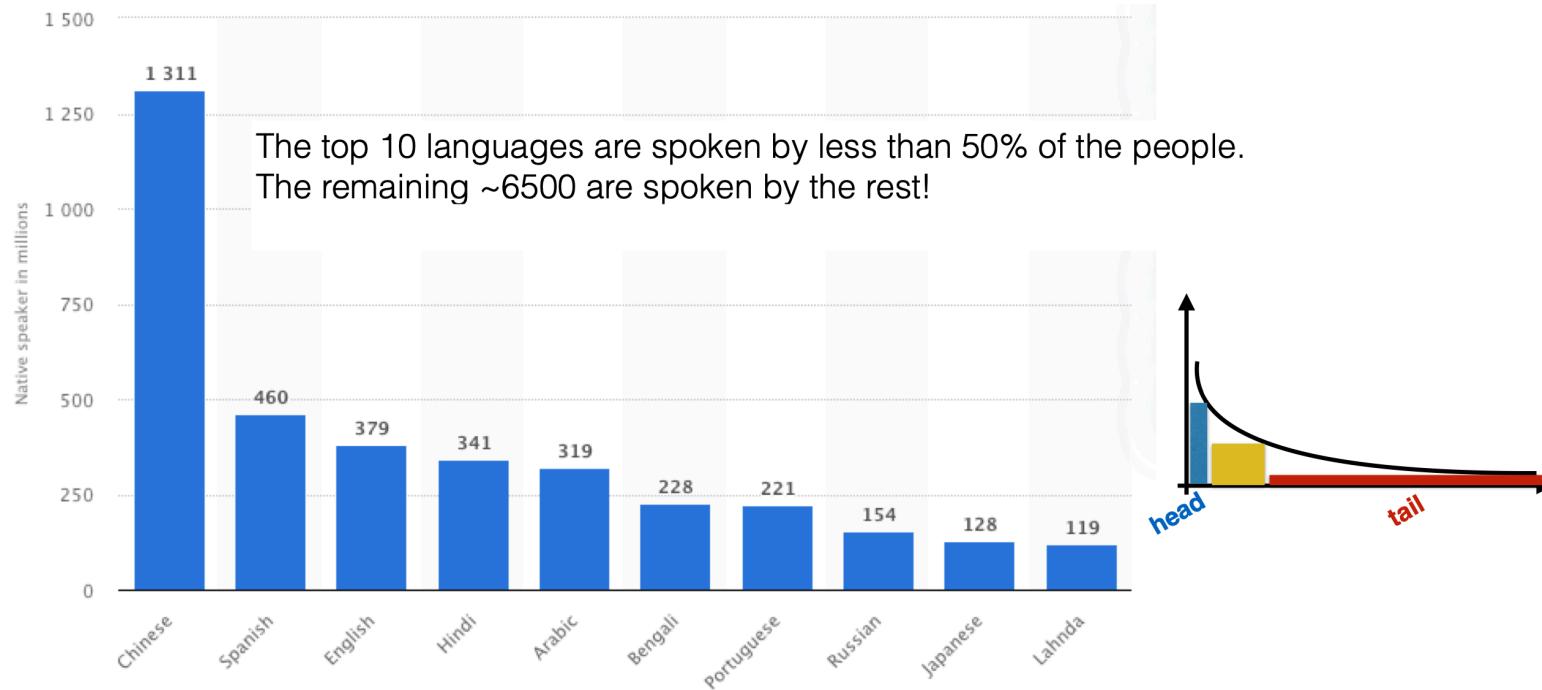


<https://endangeredlanguages.com/>

Languages of the World

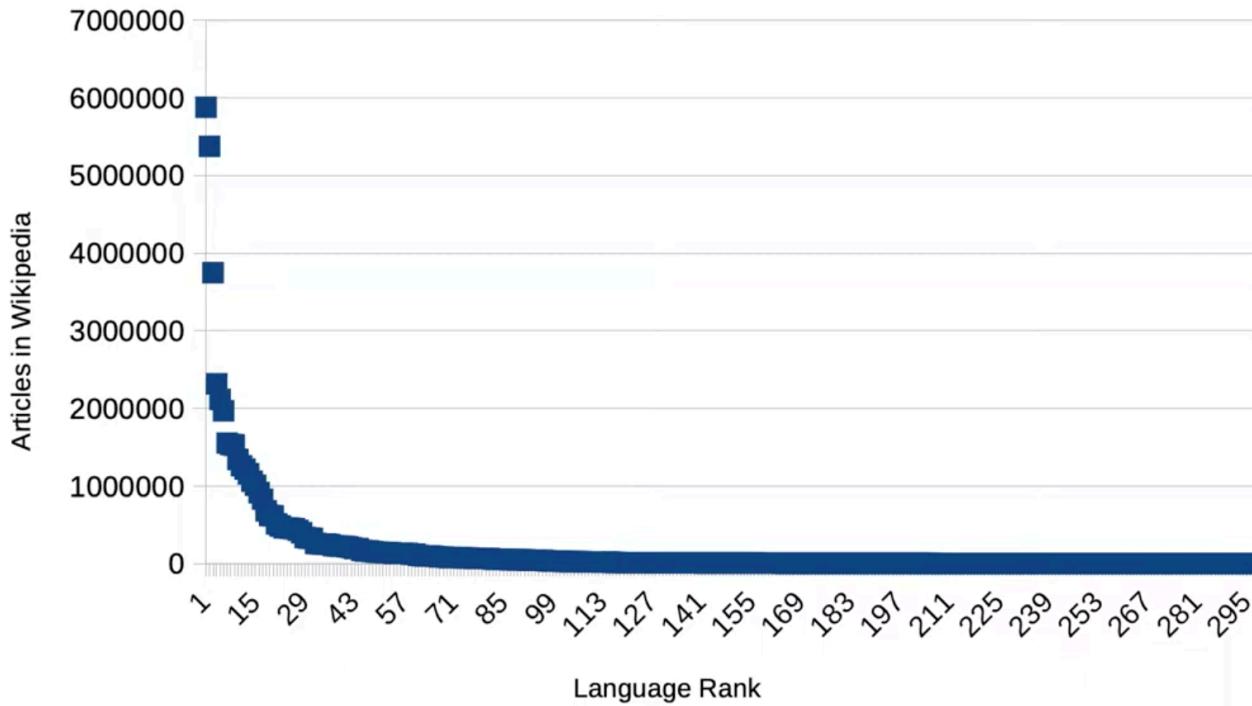
- 6500+ languages around the world
- ~70% of the world don't speak English.
- Only 10%- of the world are native English speakers.

NLP Ethics: Exclusion of the underprivileged

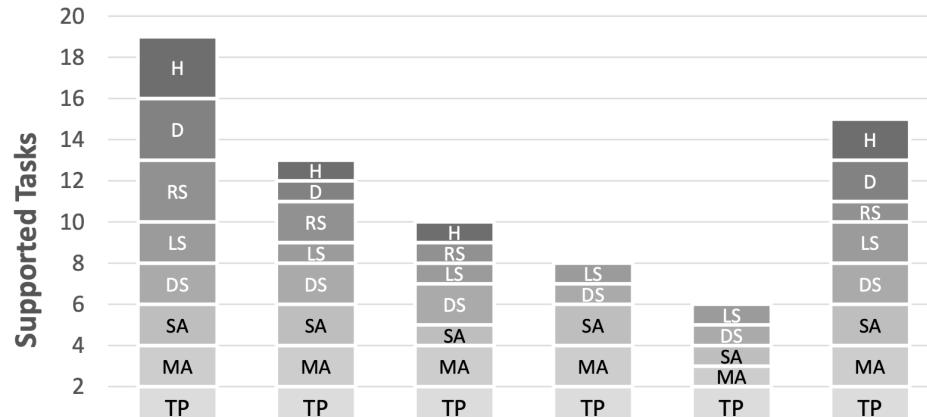


<https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/>

Data



NLP Beyond English



Language (Speakers in million) above / Tasks below

- H: Higher-level NLP applications
- D: Discourse
- RS: Relational semantics
- LS: Lexical semantics
- DS: Distributional semantics
- SA: Syntactic analysis
- MA: Morphological analysis
- TP: Text processing

NLP for low resource languages

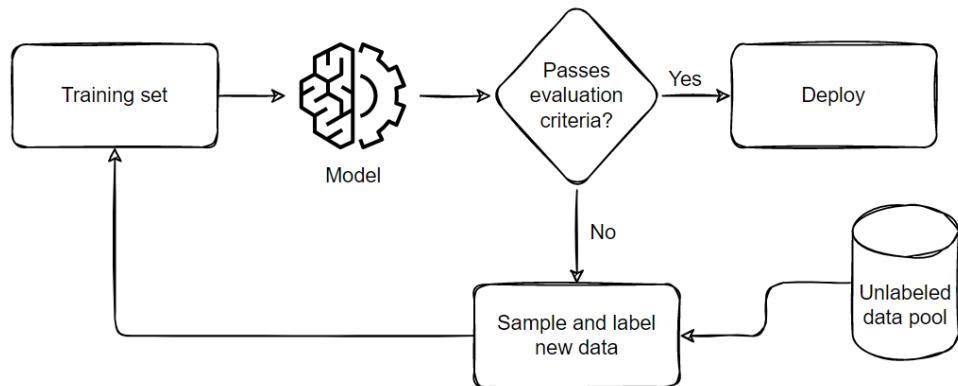
- 310 languages that have at least 1M speakers each (Eberhard et al 2019)
- **Goal:** supporting tech development  increasing participation in a digital world
- The low-resource setting can be applied for non main-stream domains of high resource languages too.
- Bender rule: clarifying the language of focus in publications.

Generating Additional Data

- Shortage of labeled data for supervised learning is the most prevalent challenge
 - Annotation with Active Learning
 - Data Augmentation
 - Cross-lingual projection

Annotation by Active Learning

- Optimizing the new annotation iteratively



https://keras.io/examples/nlp/active_learning_review_classification/

Data Augmentation

- Expand your data by augmenting the (small) existing ones.



A boy is holding a bat.

*computer vision
augmentation*



A boy is holding a bat.

A boy is holding a bat.
Ein Junge hält einen Schläger.

*translation
augmentation*

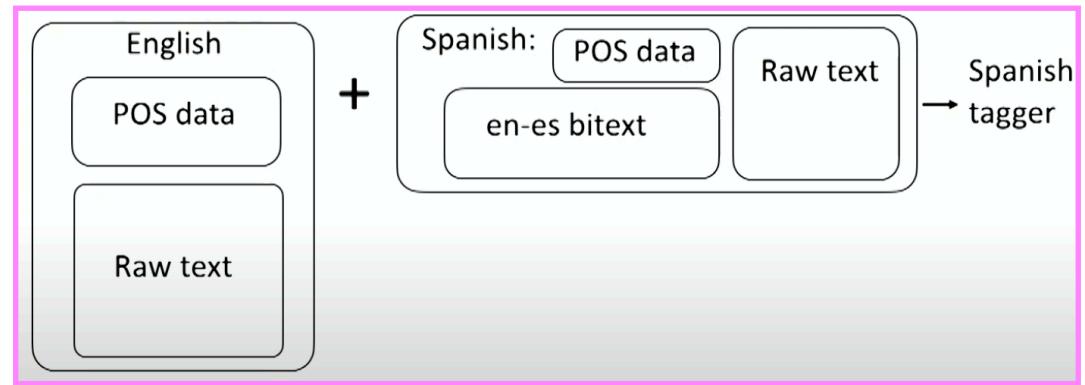
A boy is holding a **backpack**.
Ein Junge hält einen *Rucksack*.

Challenge: Scaling can result in noisy data.

Fadaee et al 2017

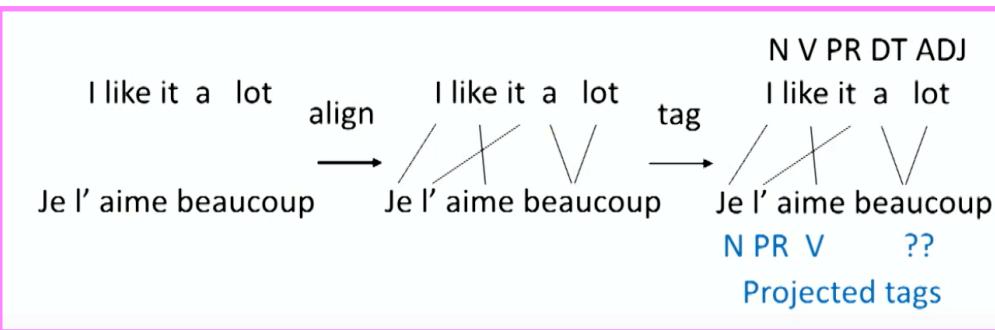
Weak Supervision

- Leveraging from MT data to create labeled data for other tasks.



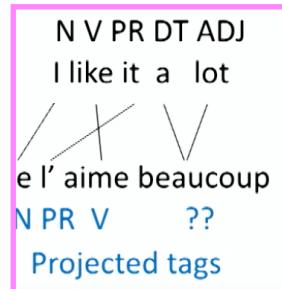
Cross Lingual Projection

- Use word alignments to project the labels across.
- Partially noisy data, better than no data.



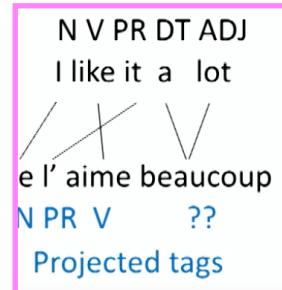
Cross Lingual Projection

- Use machine translation (and its word/phrase alignments) to project the labels across.
- Partially noisy data, better than no data.



Cross Lingual Projection

- Use machine translation (and its word/phrase alignments) to project the labels across.
- Partially noisy data, better than no data.



Challenge: Availability of parallel data/MT

Transfer Learning

- A lot of neural-based methodologies for dense representation and modeling are supposedly language agnostic.
- Word-piece tokenization, Byte-pair-encoding, etc. address a lot of **morphological differences** —> pre-trained embedding for 270+ languages
- Monolingual BERT has been applied successfully to many languages

Transfer Learning

- A lot of neural-based methodologies for dense representation and modeling are supposedly language agnostic.
- Word-piece tokenization, Byte-pair-encoding, etc. address a lot of **morphological differences** —> pre-trained embedding for 270+ languages
- Monolingual BERT has been applied successfully to many languages

Challenge: Availability and diversity of unlabeled data for low resource languages.
Word embeddings quality can vary.

Transfer Learning

- A lot of neural-based methodology for dense representation and modeling are language agnostic
- Word-piece tokenization, Byte-pair-encoding, etc. address a lot of **morphological differences** —> pre-trained embedding for 270+ languages
- Monolingual BERT has been applied successfully to many languages
- What about pre-training a shared pre-trained model?
 - **Multi-lingual models**

Multilingual Models

- Combining data into one multilingual model
 - Multilingual BERT, XLM-RoBERTa

Cross-lingual Zero Shot Learning

- **Goal:** We have labeled data for task **X** in **high resource language**. We want a model for task **X** in a **low resource language**.
- **Idea:** Leverage the resources for the high resource language
-

Cross-lingual Zero Shot Learning

- **Goal:** We have labeled data for task **X** in **high resource language**. We want a model for task **X** in a **low resource language**.
- **Idea:** Leverage the resources for the high resource language.
- **Zero-shot:** Fine-tune the multilingual backbone with the task X with the high resource language data (and flexible prompts/instructions) towards generalizing for the low resource languages.
 - NER (lin et al, 2019), reading comprehension (Hsu et al 2019), Parsing (Muller et al 2020)
 - Few shot: Add small set (10-100) of low-resource labeled data

Transfer Learning

- Low resource languages in multi-lingual pre-trained language models.
- **Challenge:** Availability of diversity of data for low resource languages
 - Word embedding quality can vary a lot.

Optimization Under Extreme Constraints: The tale of 101 languages

UC-Berkeley NLP INFO 159/259
Sara Hooker

Aya at a glance.

Cohere For AI

Exploring the unknown, together.

Research

- Research Lab
- Publications
- Scholars Program

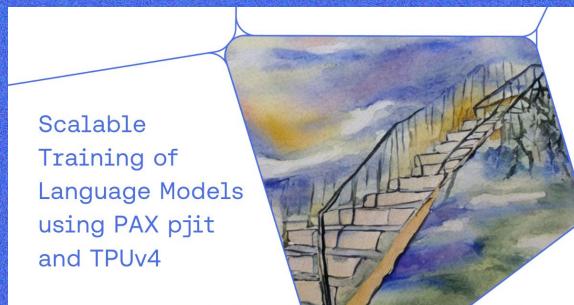
Open Science

- Cross-institutional collaborations
- Open science initiatives

Forum

- Fireside Chats
- Technical Talks
- Guest Series
- AI Policy/Safety

Fundamental research on critical areas like efficiency, LLMs at scale, safety, hardware/software interaction.



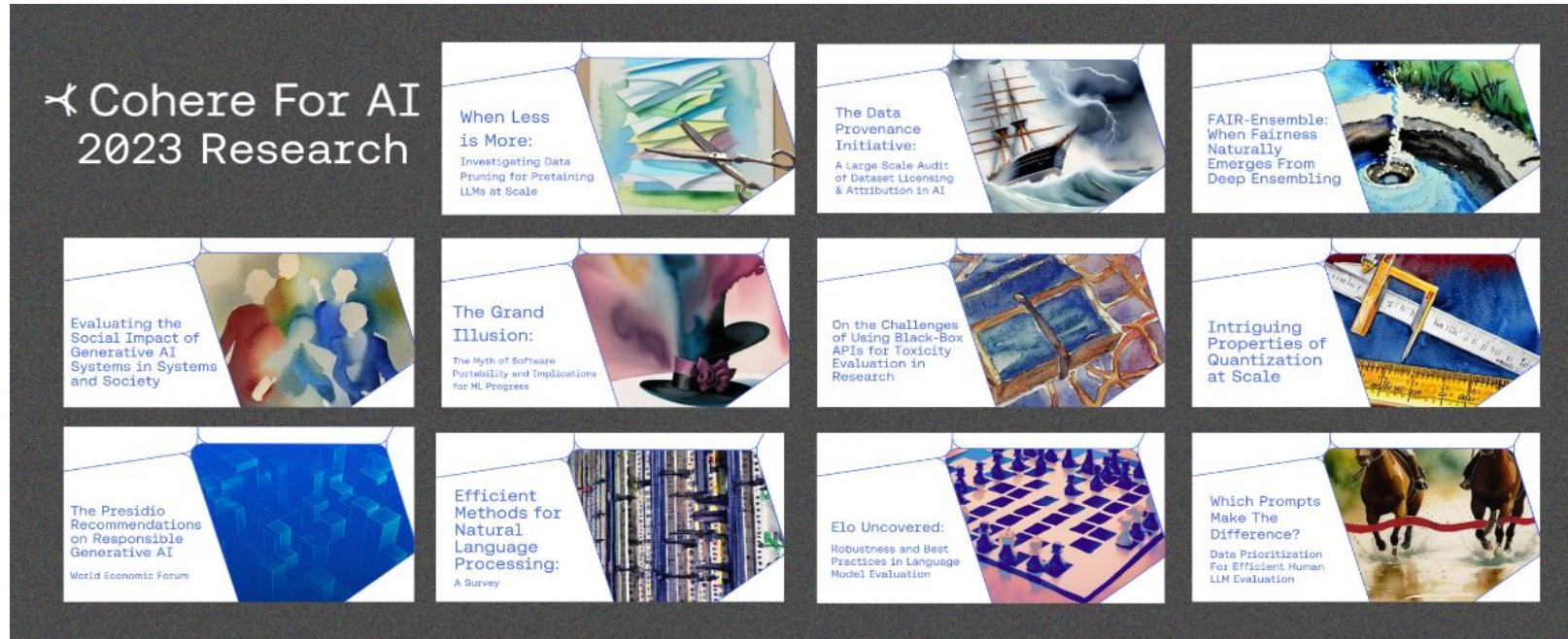
Cohere For AI is one of the first intentionally **hybrid research labs** -- with both a traditional industry lab and an open science lab.



The goal is to create new spaces where state of art research happens – and to empower more entry points into machine learning.

In 2023: our goal was to contribute state of art research in machine learning.

- Published 32 papers last year
- Including 7 papers published by our Research Scholars
- Collaborated across 40+ institutions and organizations



 Cohere For AI



Aya at a Glance

1 

Model

513M 

Re-annotations
of Datasets

3K 

Independent
Researchers

56 

Language
Ambassadors

119 

Countries

204K 

Original Human
Annotations

101 

Languages

31K 

Discord
Messages

Achinese · Afrikaans · Albanian · Amharic · Arabic · Arabic · Armenian · Azerbaijani
Balinese · Banjar · Basque · Belarusian · Bemba · Bengali · Bulgarian · Burmese · Catalan
Cebuano · Chinese · Croatian · Czech · Danish · Dutch · English · Esperanto · Estonian
Filipino · Finnish · Fon · French · Galician · Georgian · German · Greek · Gujarati · Haitian
Creole · Hausa · Hebrew · Hindi · Hungarian · Icelandic · Igbo · Indonesian · Irish
Italian · Japanese · Javanese · Kannada · Kanuri · Kashmiri · Kazakh · Khmer
Kinyarwanda · Korean · Kurdish · Kurdish · Kyrgyz · Lao · Latvian · Ligurian · Lithuanian
Luxembourgish · Macedonian · Madurese · Malagasy · Malay · Malayalam · Maltese
Manipuri · Maori · Marathi · Minangkabau · Mongolian · Nepali · Ngaju · Northern Sotho
Norwegian · Pashto · Persian · Polish · Portuguese · Punjabi · Romanian · Russian
Samoan · Scottish Gaelic · Serbian · Shona · Sindhi · Sinhala · Slovak · Slovenian
Somali · Southern Sotho · Spanish · Sundanese · Swahili · Swedish · Tajik · Tamasheq
Tamil · Telugu · Thai · Toba Batak · Turkish · Twi · Ukrainian · Urdu · Uzbek · Vietnamese
Welsh · Wolof · Xhosa · Yiddish · Yoruba · Zulu

Today, I'll talk about
instruction tuning under
severe constraints.

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

– Ludwig Wittgenstein

Why is multilingual
modelling so challenging?

Several key reasons it is challenging:

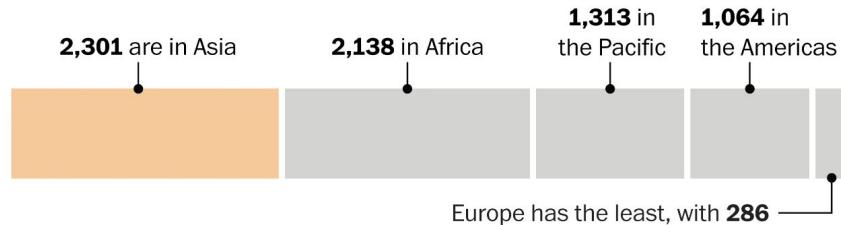
- Data scarcity
- Low quality data
- Access to compute (double low resource bind)
- Technical obstacles (weighting, tokenization)

Low resource languages can disproportionately benefit from the effectiveness of instruction finetuning.

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

● Critically endangered ● Seriously endangered ● Endangered



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

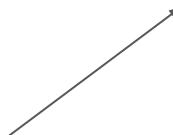
Languages are not treated equally by researchers. Some languages have received disproportionate attention and focus in NLP.

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.
[\[Van Etch et al. 2022\]](#)

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

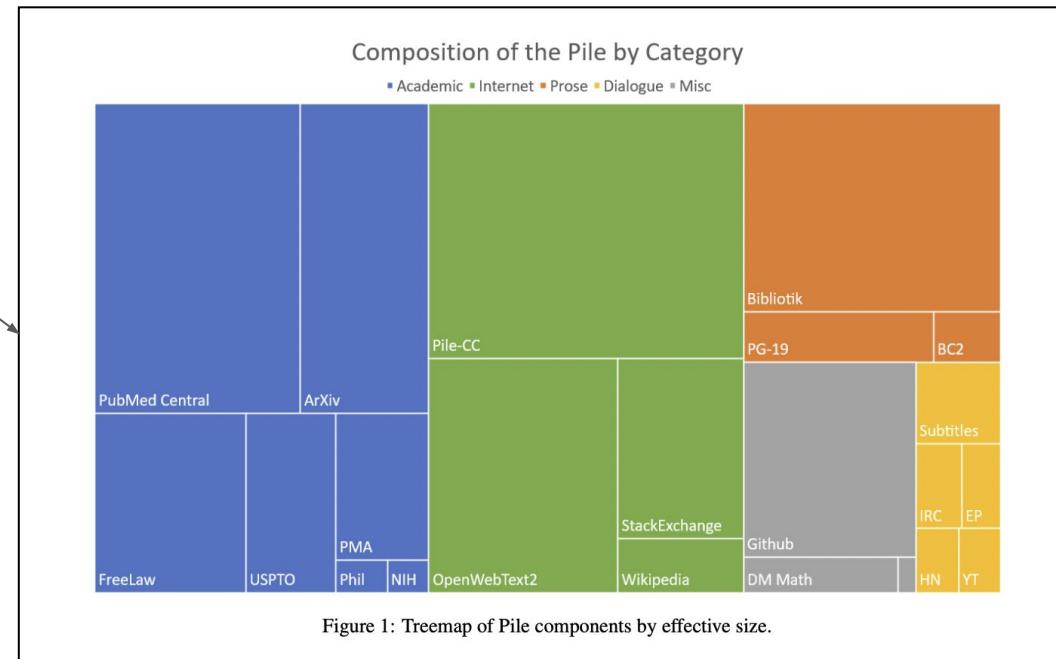
Multilingual Model Name	Number of Languages Trained On (pre-training)
BLOOM	46
mT5	101
XGLM	30



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.

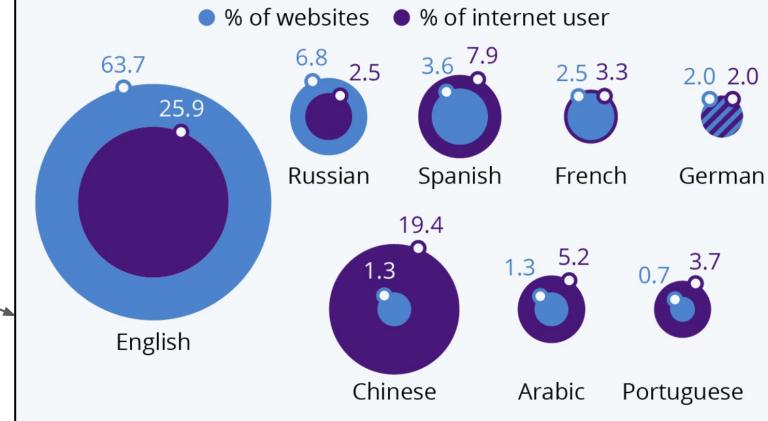


The composition of languages on the internet reflects the composition of early users.

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*

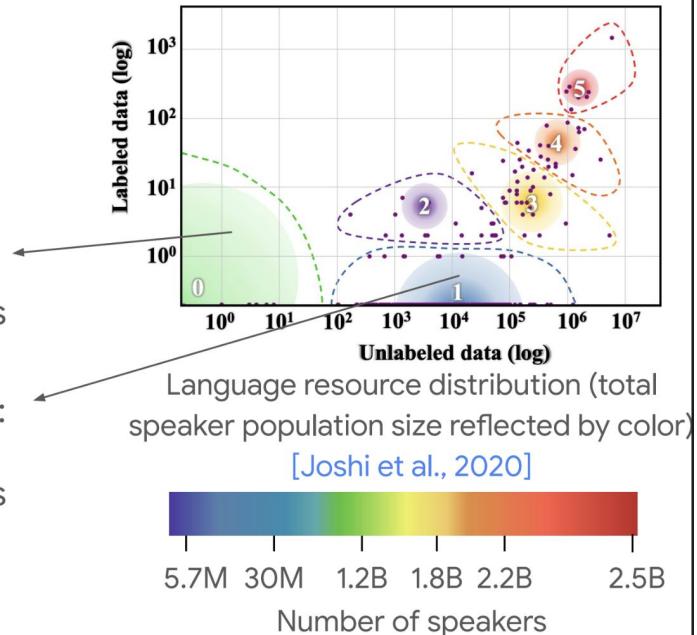


Under-resourced languages often have limited data available....

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



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

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

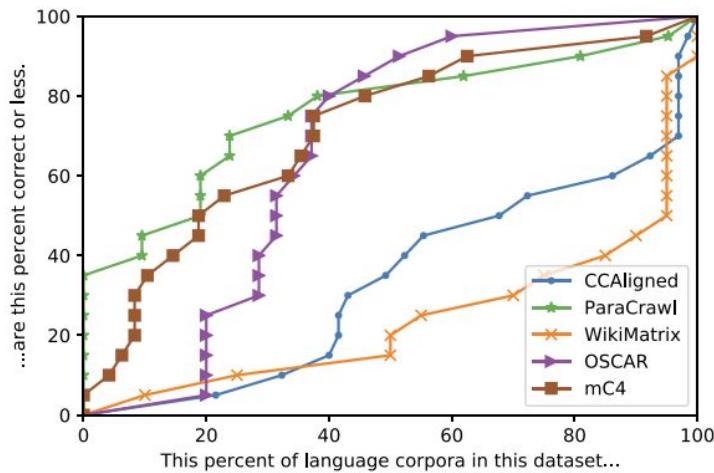
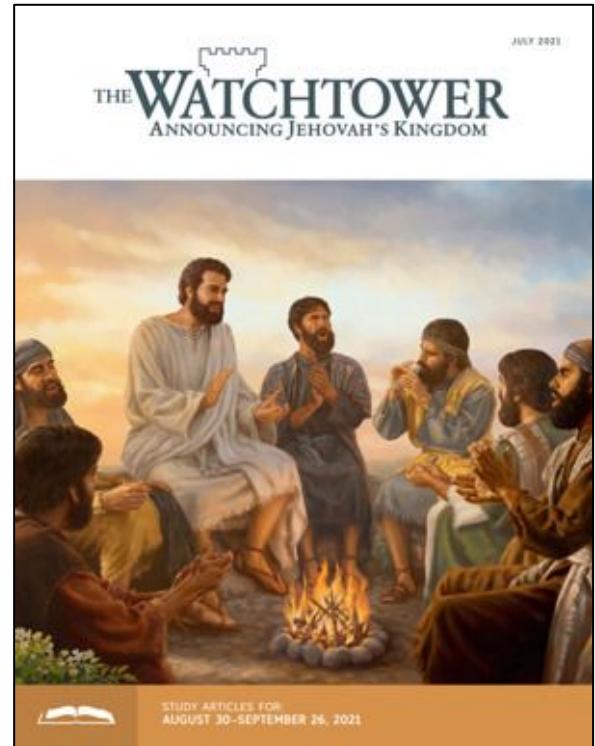


Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

... 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.



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

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

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

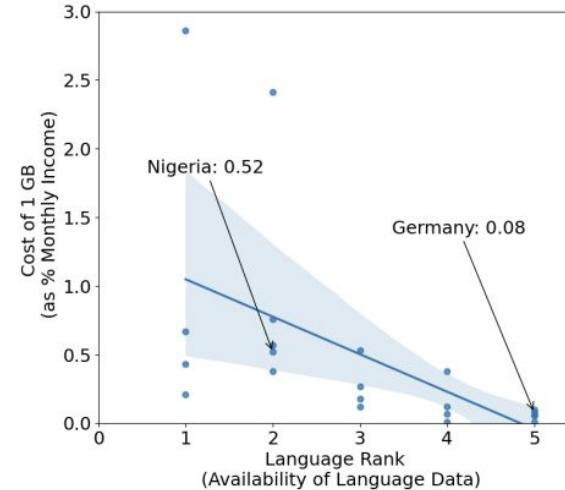
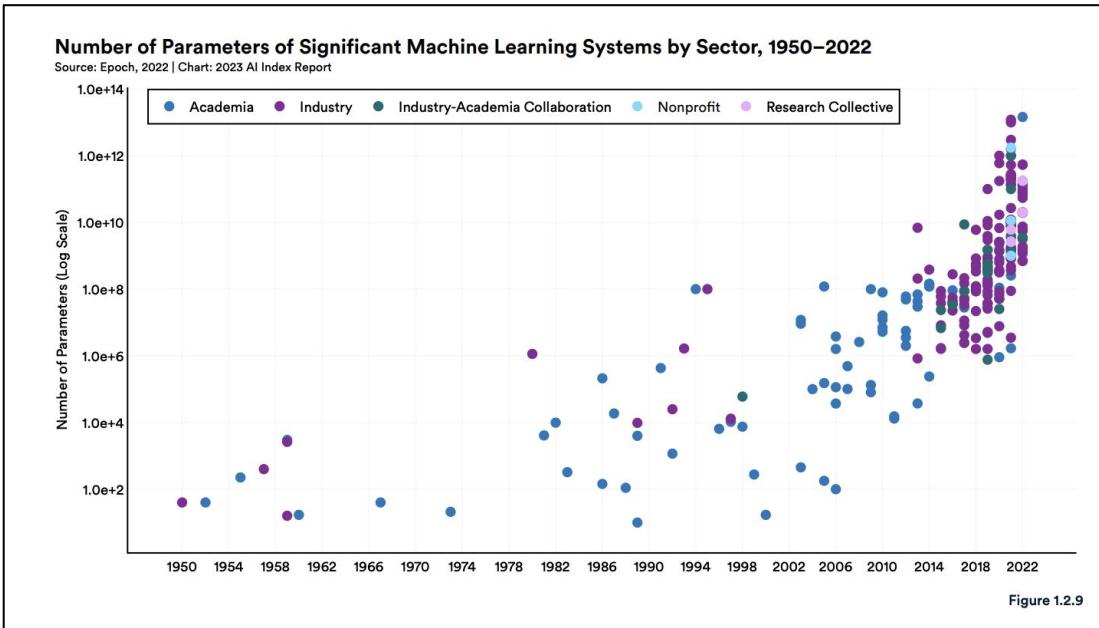


Figure 1: Cost of mobile data by country per language rank according to the taxonomy by [Joshi et al. \(2020\)](#).

Compute trends also amplify the disparity in who participates. Academia phased out, industry dominates.



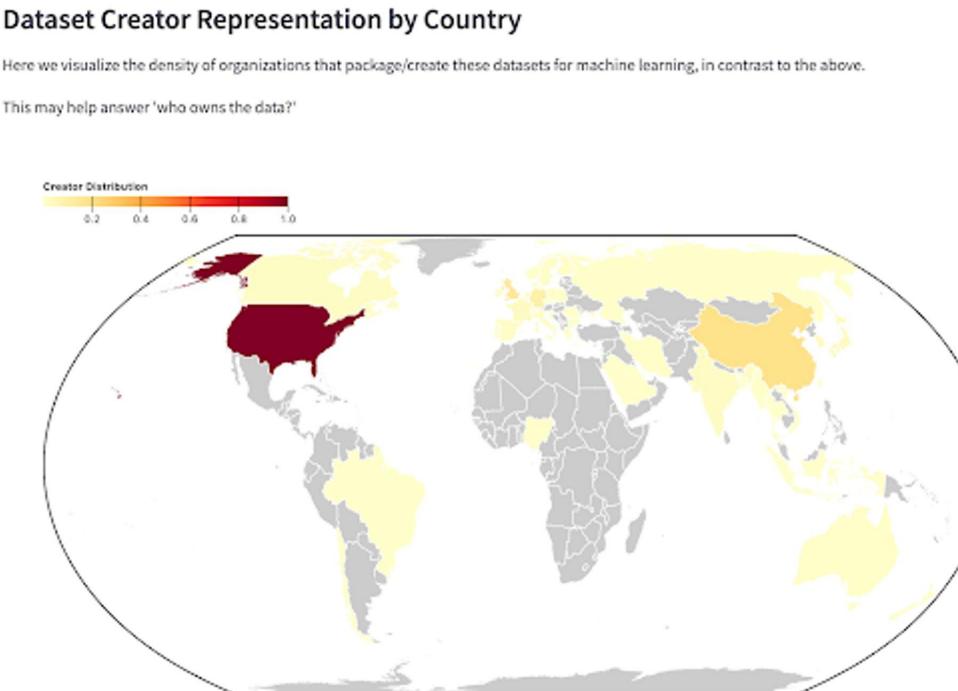
We are in a “bigger is better” race in machine learning.

[AI index report](#)

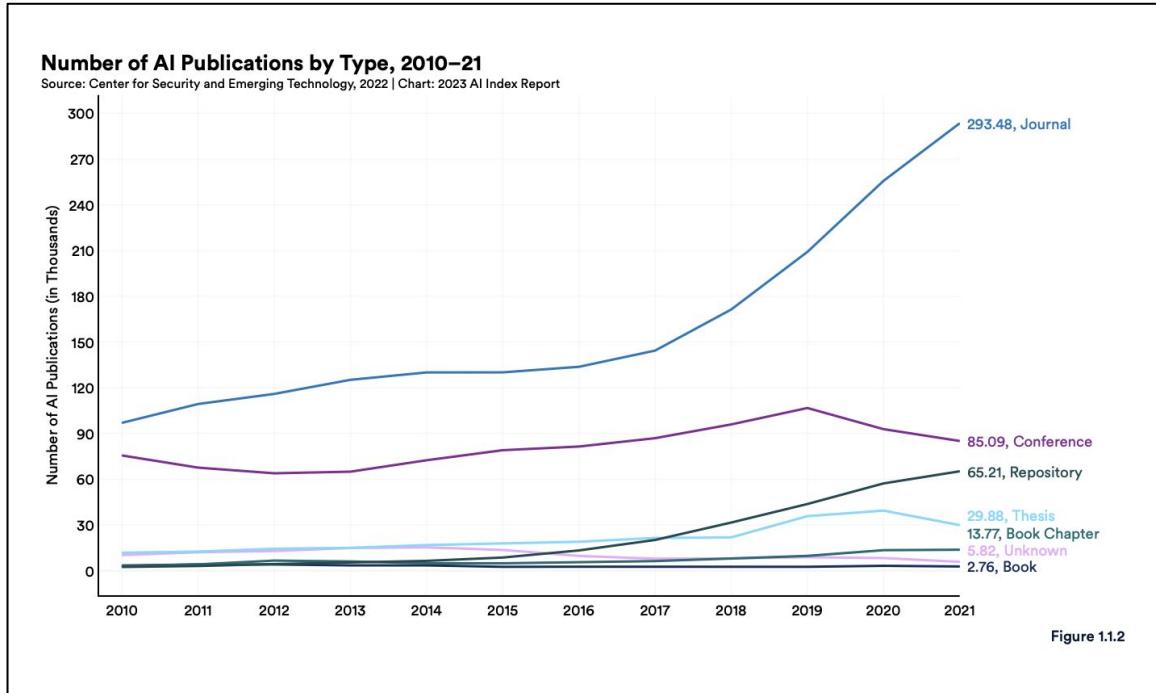
This underrepresentation of multilingual data partly reflects the lack of access for researchers across the world.

Who creates popular datasets?

Dataset creators are heavily skewed towards the west, with few datasets created in Latin America or Africa.



This is a trend also reflected in who contributes research publications.



We are going through a boom cycle in AI funding, research and the number of overall publications has increased.

[AI index report](#)

Who produces research remains narrow.

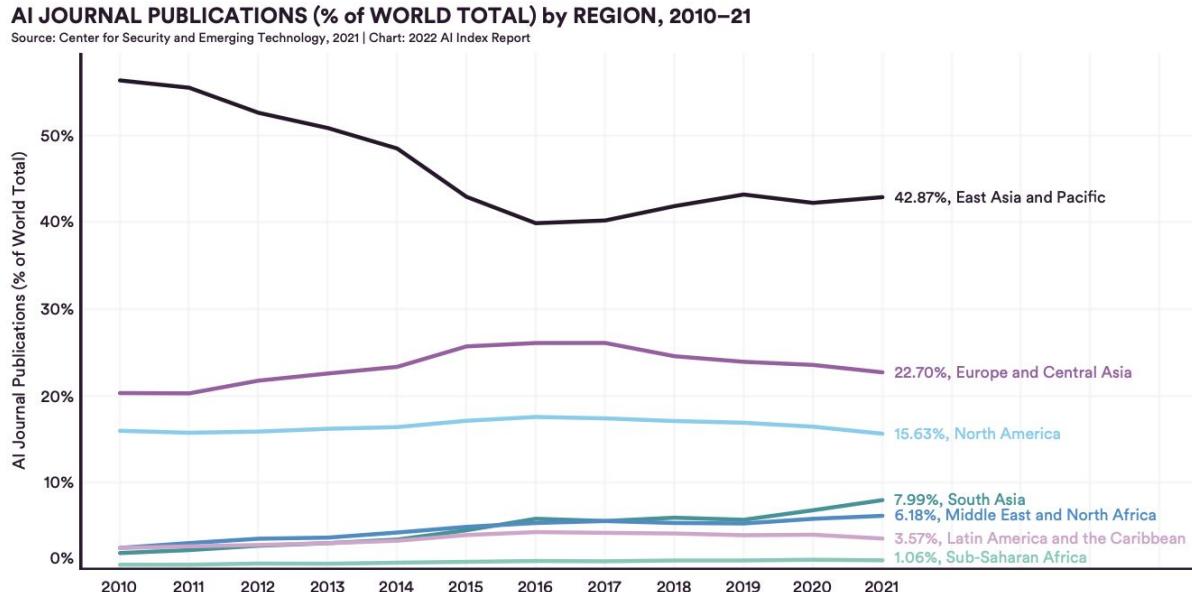


Figure 1.1.9

Large discrepancies persist in who participates in research:

Sub-Saharan Africa
1.06%

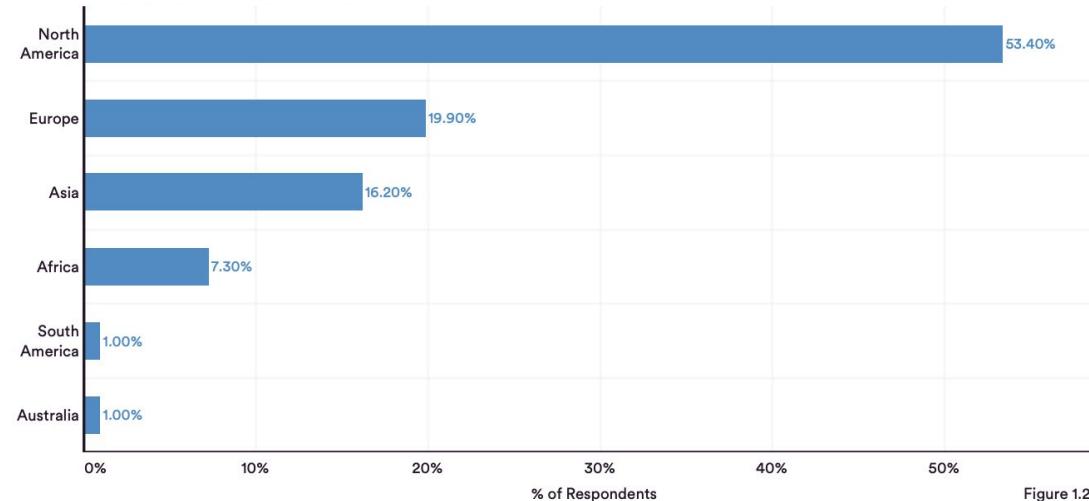
Latin America and the Caribbean
3.57%

[AI index report](#)

This geo disparity is also reflected in who attends conferences like NeurIPS.

CONTINENT of RESIDENCE of PARTICIPANTS at NEURIPS WOMEN in MACHINE LEARNING WORKSHOP, 2021

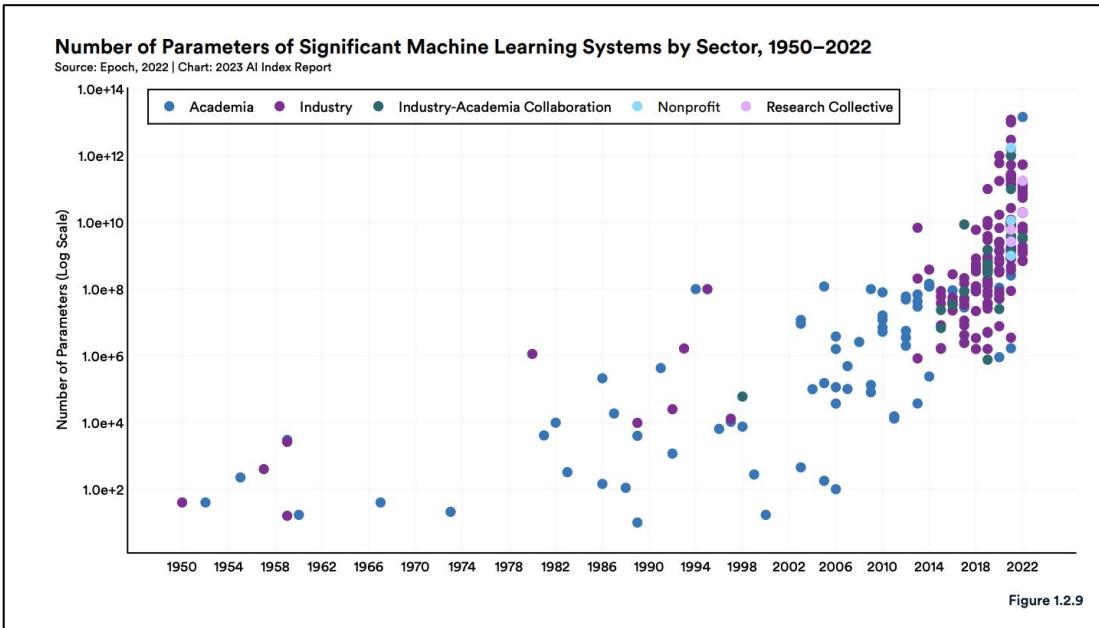
Source: Women in Machine Learning, 2021 | Chart: 2022 AI Index Report



53.40 % of participants are from North America.

Figure 1.2.5

Compute trends also amplify the disparity in who participates. Academia phased out, industry dominates.



We are in a “bigger is better” race in machine learning.

[AI index report](#)

Who participates in research also determines = who shapes scientific breakthroughs.

Number of Significant Machine Learning Systems by Country, 2002–22 (Sum)

Source: AI Index, 2022 | Chart: 2023 AI Index Report

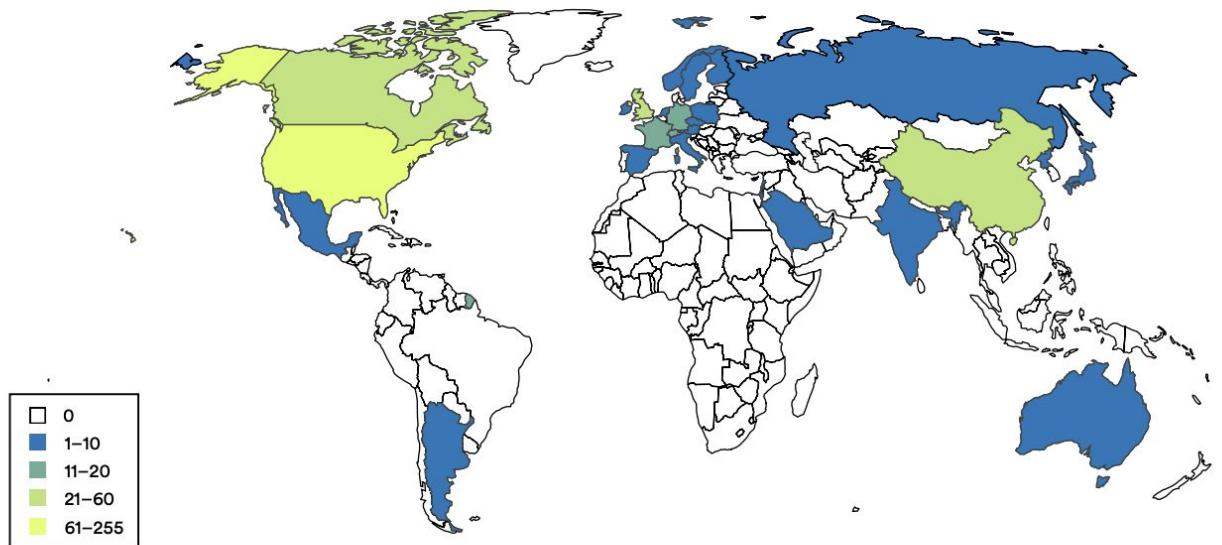


Figure 1.2.5

Let this sink in – look at what continents are left out and where breakthroughs come from.

Number of Significant Machine Learning Systems by Country, 2002–22 (Sum)

Source: AI Index, 2022 | Chart: 2023 AI Index Report

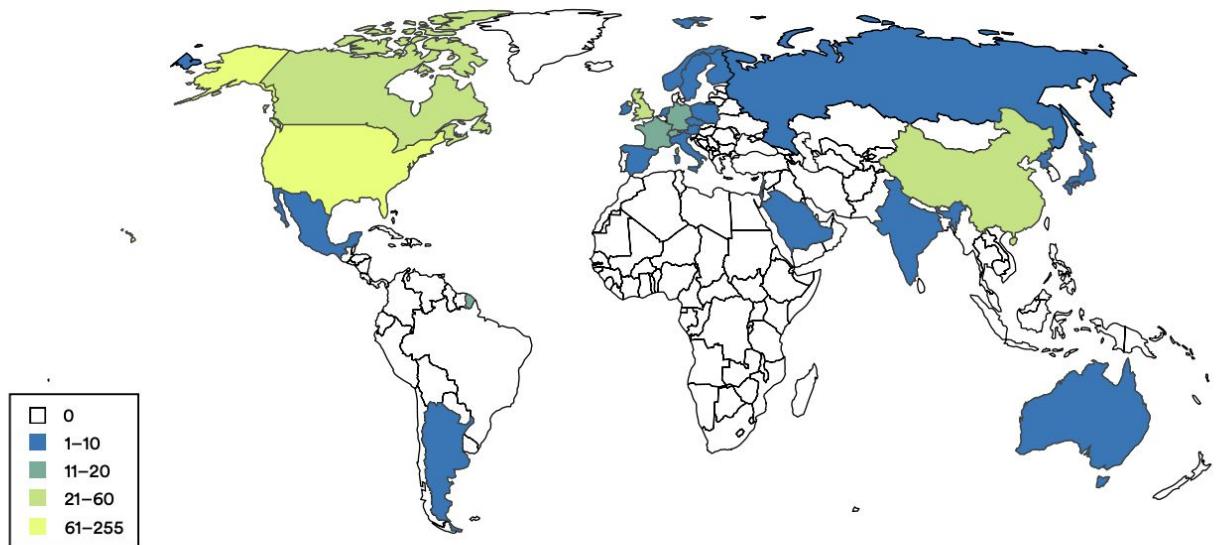


Figure 1.2.5

“When you are not part of the conversation, it happens to you and not with you.”

Vukosi Marivate

Without robust multilingual datasets to train models, we risk:



Introducing biases and gaps in performance towards languages not included.

[[Robinson et al.
2023](#)]

Marginalizing speakers of languages not included.

Creating a performance-divide for languages with limited datasets.

[[\[\[Ahia et al. 2023\]\]](#)

Introducing security flaws.

Performance on low resource languages are often double taxed – with both worse generalization and higher cost

Ahia et al. find that API calls are more expensive for low resource languages because of number of tokens.

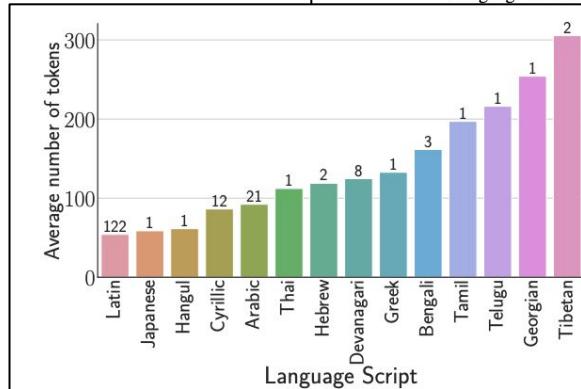


Figure 2: Average number of tokens by script after tokenizing the Flores dataset. The fragmentation rate is lower for Latin script languages and higher for other scripts. Number of languages per language group is indicated at the top of each bar.

[[Ahia et al. 2023]]

Do All Languages Cost the Same?
Tokenization in the Era of Commercial Language Models
Orevaoghene Ahia[◊] Sachin Kumar^{*} Hila Gonen[◊] Jungo Kasai[◊]
David R. Mortensen^{*} Noah A. Smith^{◊○} Yulia Tsvetkov[◊]
◊Paul G. Allen School of Computer Science & Engineering, University of Washington
*Language Technologies Institute, Carnegie Mellon University
Artificial Intelligence

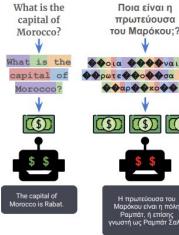


Figure 1: We investigate the effects of subword tokenization in LLMs across languages with different writing systems. Our findings highlight disparities in the utility of LLMs, as well as socio-economic disparities and increased costs in using commercial APIs for speakers of underrepresented languages.¹

The brittleness of models on low resource languages undermines overall safety.

Jailbreaking presents higher success when harmful prompt in a low resource language – **particularly high rates for zulu and gaelic.**

Low-Resource Languages Jailbreak GPT-4

Zheng-Xin Yong, Cristina Menghini, Stephen H. Bach

Department of Computer Science

Brown University

Providence, RI 02906, USA

{contact.yong, cristina_menghini, stephen_bach}@brown.edu

Abstract

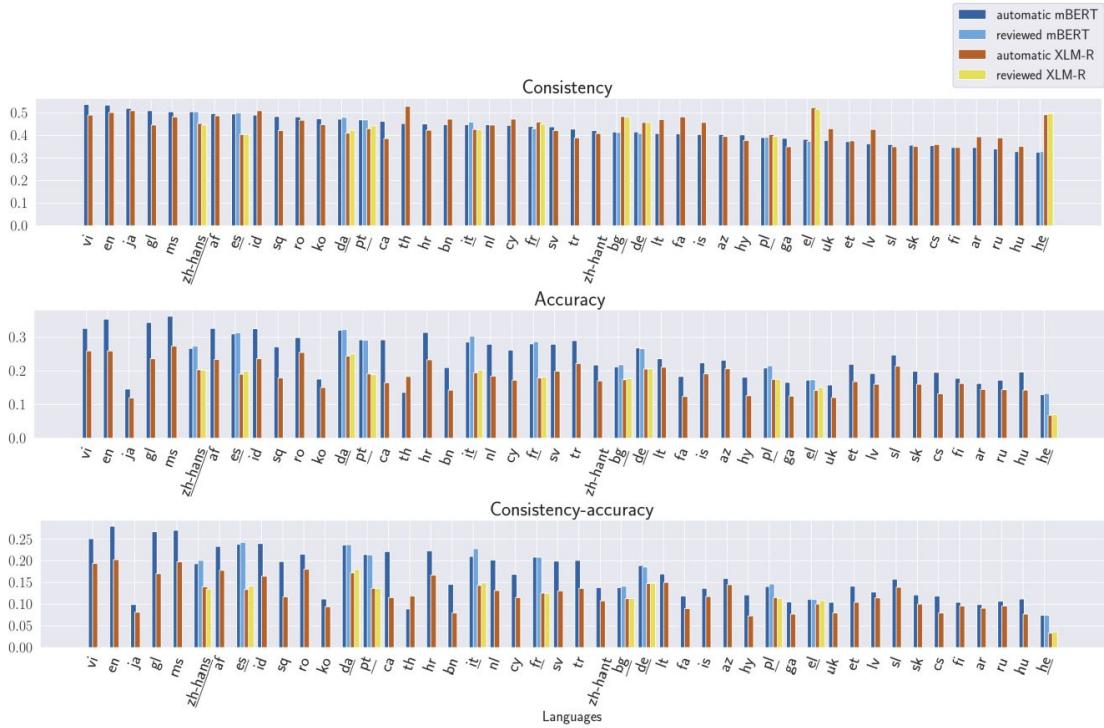
Red-teaming of large language models (LLMs) are measures to detect unsafe content. Our work exposes the inherent brittleness of these safety mechanisms, resulting from the limited training data, by successfully circumventing GPT-4's translating unsafe English inputs into low-resource languages. Specifically, GPT-4 engages with the unsafe *translated inputs* and succeeds at bypassing safety mechanisms that can get the users towards their harmful goals 79% of the time, with or even surpassing state-of-the-art jailbreaking methods. Low-resource languages have significantly lower attack success rates than high-resource languages, which suggests that the cross-lingual vulnerability mainly applies to low-resource languages. Previously, limited training on low-resource languages has been shown to be a major factor of those languages, causing technological disparities. Our work highlights a crucial shift: this deficiency now poses a risk to the availability of LLMs' safety mechanisms. Therefore, our work calls for a more holistic red-teaming approach that includes multilingual safeguards with wide language coverage. **Important note:** This paper contains examples of harmful language.

Attack	BYPASS (%)	REJECT (%)	UNCLEAR (%)
LRL-Combined Attacks	79.04	20.96	
Zulu (zua)	53.08	17.12	29.80
Scots Gaelic (gd)	43.08	45.19	11.73
Hmong (hmn)	28.85	4.62	66.53
Guarani (gn)	15.96	18.27	65.77
MRL-Combined Attacks	21.92		78.08
Ukrainian (uk)	2.31	95.96	1.73
Bengali (bn)	13.27	80.77	5.96
Thai (th)	10.38	85.96	3.66
Hebrew (he)	7.12	91.92	0.96
HRL-Combined Attacks	10.96		89.04
Simplified Mandarin (zh-CN)	2.69	95.96	1.35
Modern Standard Arabic (ar)	3.65	93.85	2.50
Italian (it)	0.58	99.23	0.19
Hindi (hi)	6.54	91.92	1.54
English (en) (No Translation)	0.96	99.04	0.00
AIM [8]	55.77	43.64	0.59
Base64 [44]	0.19	99.62	0.19
Prefix Injection [44]	2.50	97.31	0.19
Refusal Suppression [44]	11.92	87.50	0.58

Table 1: Attack success rate (percentage of the unsafe inputs bypassing GPT-4's content safety guardrail) on the AdBenchmark dataset [49]. LRL indicates low-resource languages, MRL mid-resource languages, and HRL high-resource languages. We color and bold the most effective translation-based jailbreaking method, which is the LRL-combined attacks.

Multilingual models present higher levels of factual inconsistency than in English.

“Both mBERT and XLM-R exhibit a high degree of inconsistency in English and even more so for all the other 45 languages.”



Our goal is to unlock state of art
in multilingual generative
models. It is an exciting time to
work on that problem – here is
why.

Several key changes to optimization over last few years have led to breakthroughs in high resource language generation:

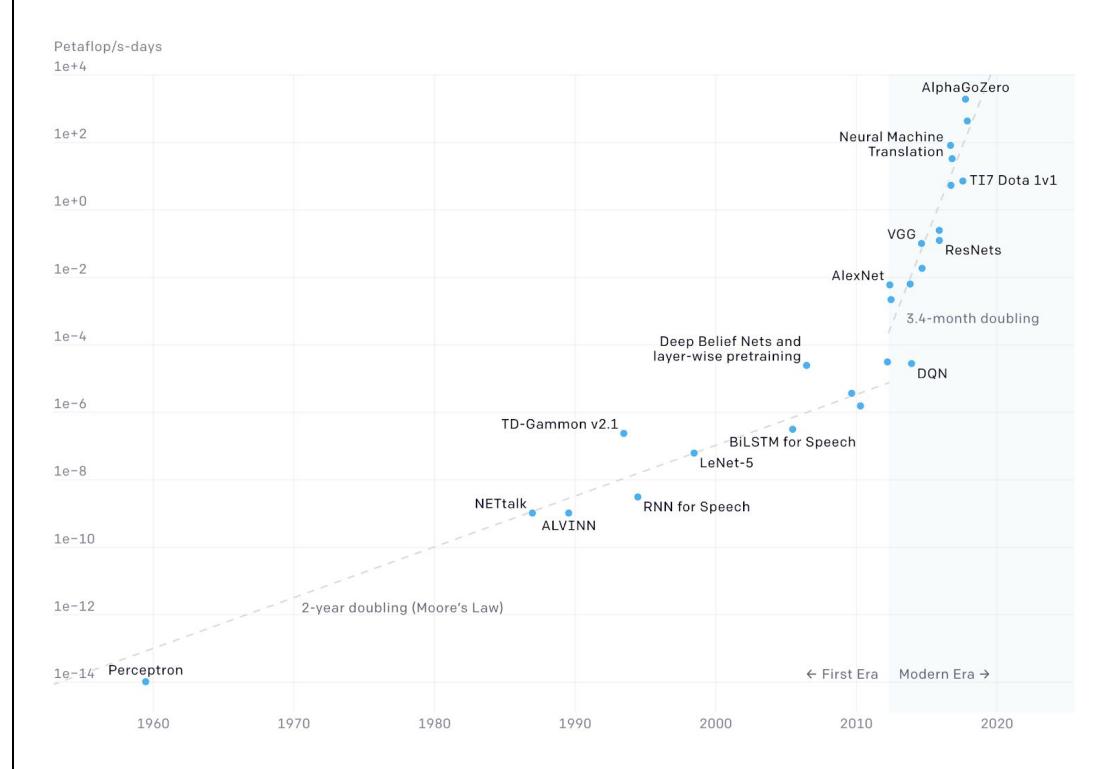
What are these breakthroughs?

Several key changes to optimization over last few years have led to breakthroughs in high resource language generation:

- Scaling
- Pre-training breakthroughs
- Change to multi-task finetuning
- Instruction finetuning
- RLHF

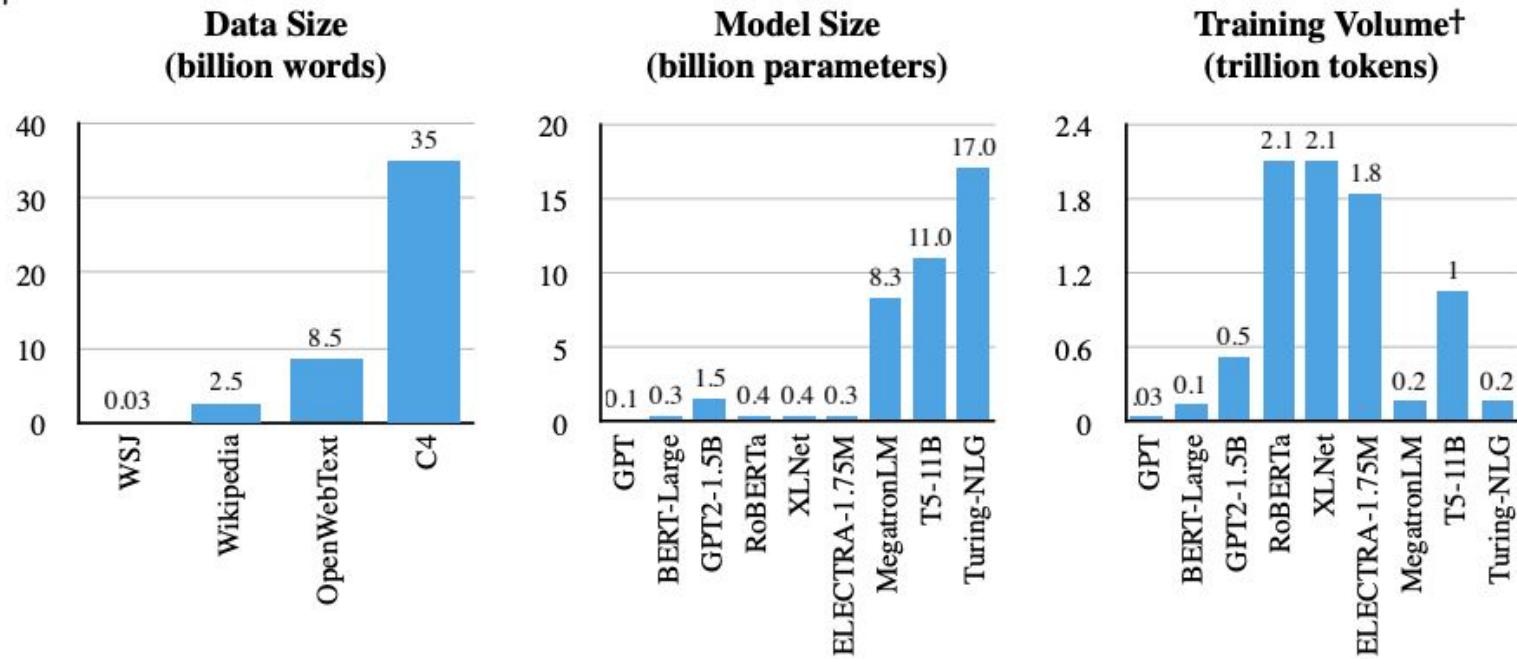
Low resource languages can disproportionately benefit from the effectiveness of instruction finetuning.

A “bigger is better” race in the number of model parameters has gripped the field of machine learning.

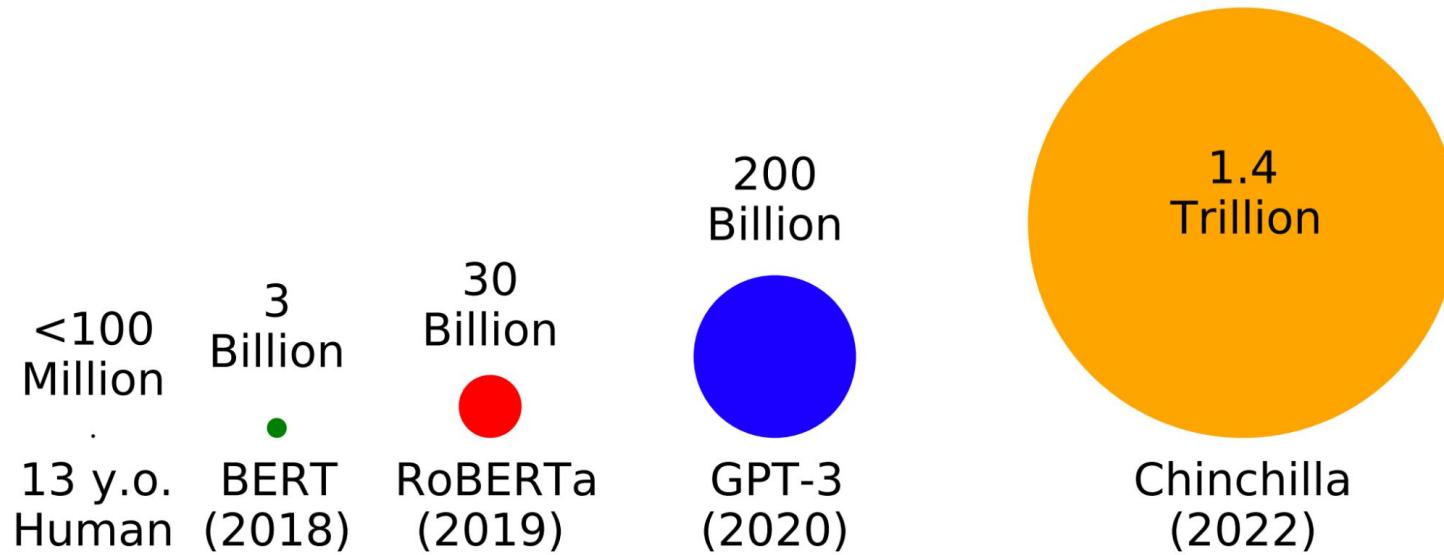


This characterizes both vision and NLP tasks.

Bird's-eye View



And involves large increases in both model and dataset sizes:



Number of tokens involved in training.

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/>

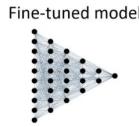
Changed to multi-task fine-tuning. Moving to a single global model – train on multiple tasks at once.

3 Fine-tuning



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

Text



Topic: Biology (97%)
Prediction

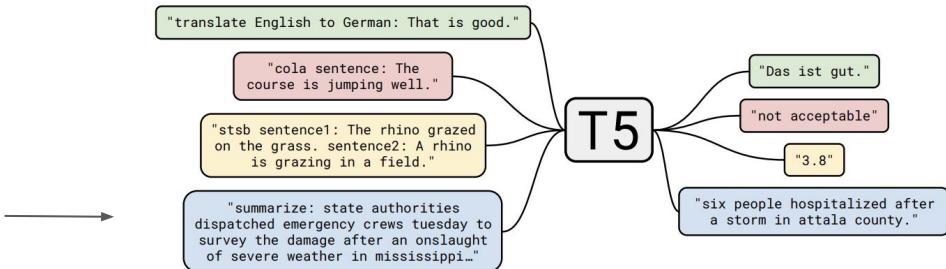


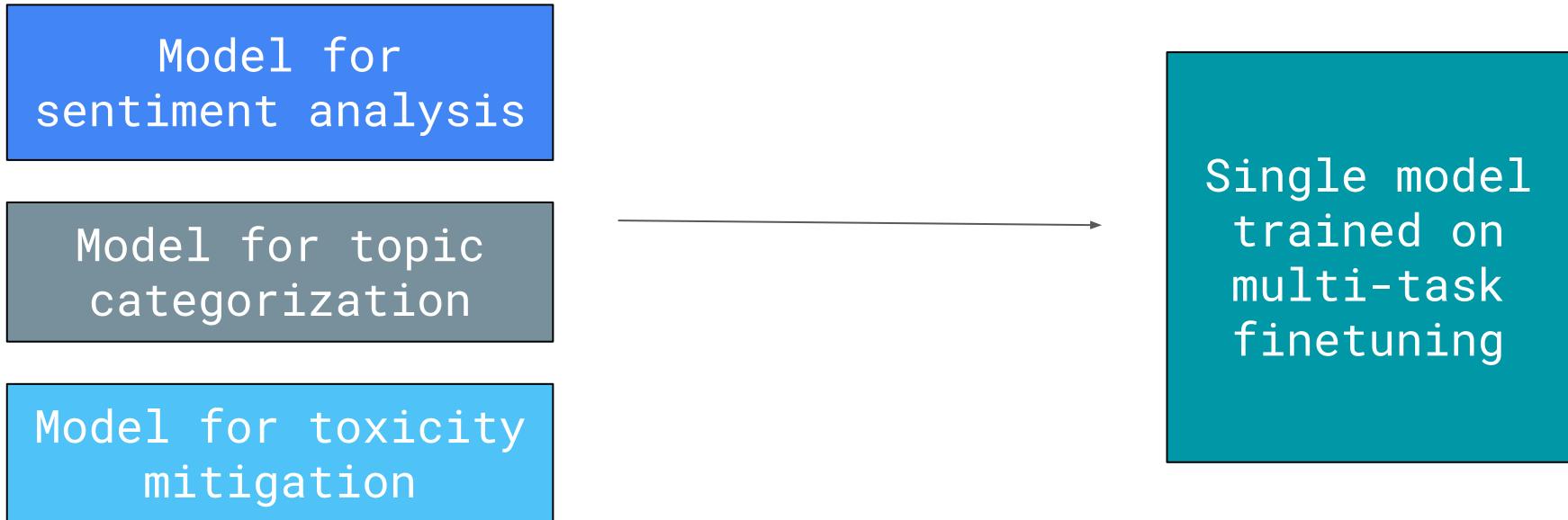
Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across 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 “Text-to-Text Transfer Transformer”.

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



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



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

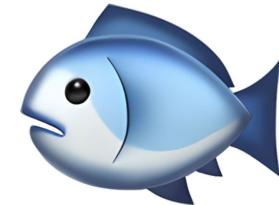


There once was a fish with a special gift,
A wizard in his own right, so wise and deft.
He lived in an underwater school,
Where students gathered from far and wide.

Each day, the wizard fish would cast
A spell to protect the school from harm.
He'd conjure up a shield so strong and vast,
Keeping them all safe from any storm.

But one day, a fierce shark appeared,
With an army of evil fish, they'd been hired.
They came to destroy the wizard fish's school,
And take over the underwater rule.

The wizard fish knew he had to act,
So he gathered his friends, ready for the attack.
With his magic, he created a plan,
To outsmart the shark and win the game.

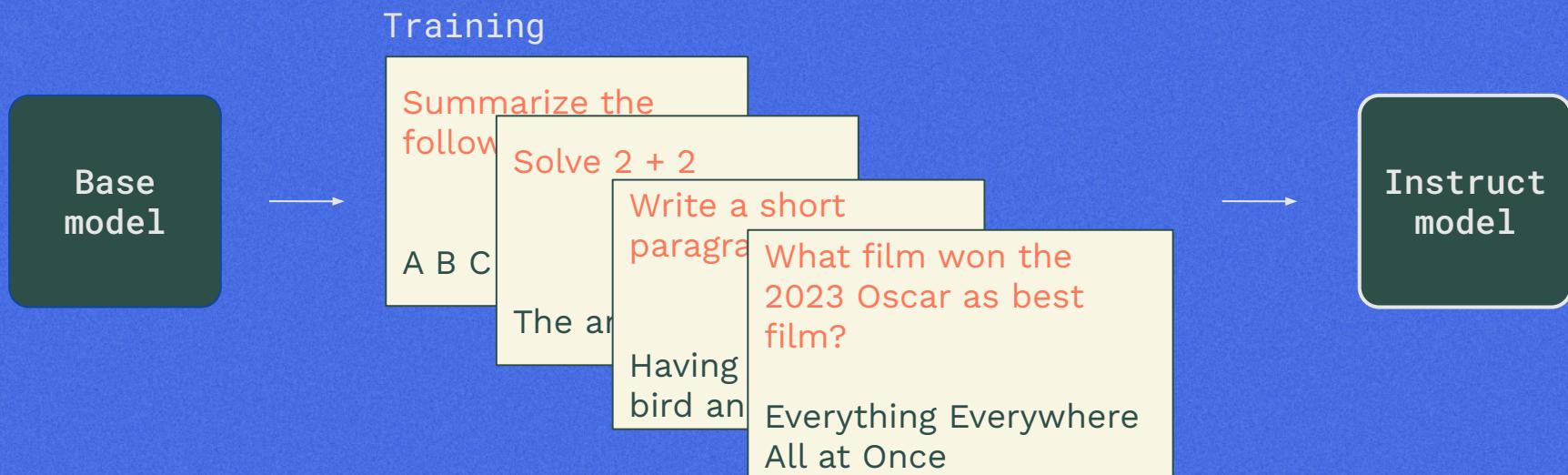


It turns out two ingredients have been particularly important at leading to breakthroughs in performance on zero and few shot tasks:

- 1. Instruction tuning – Structuring multi-task fine-tuning data as questions and answers**
- 2. Integrating human feedback about preferences**

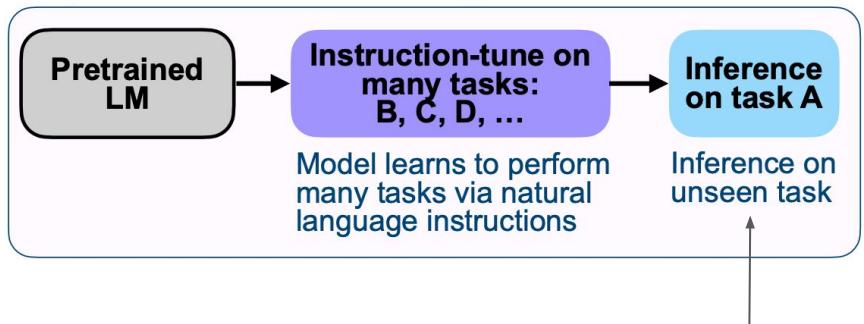
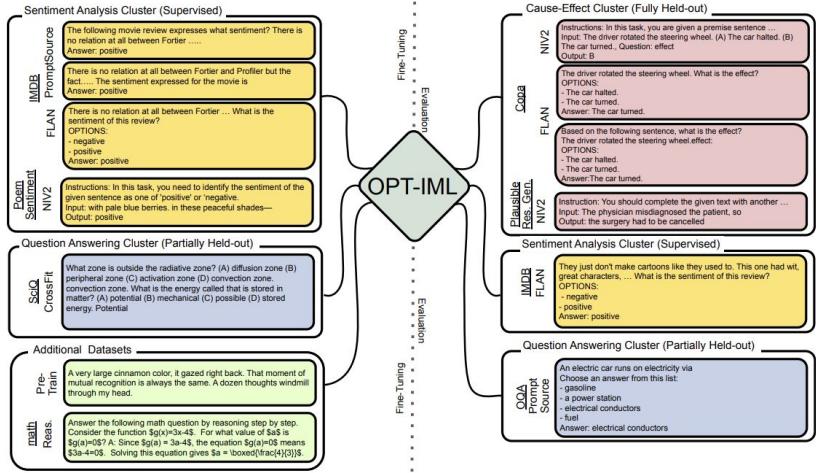
What Is Instruction Fine-Tuning?

Instruction Fine-Tuning (IFT) is a form of model training that enables models to better understand and act upon instructions. It is based on the idea that we can use everyday language to ask a model to perform a task and in return the model generates an accurate response in natural language.

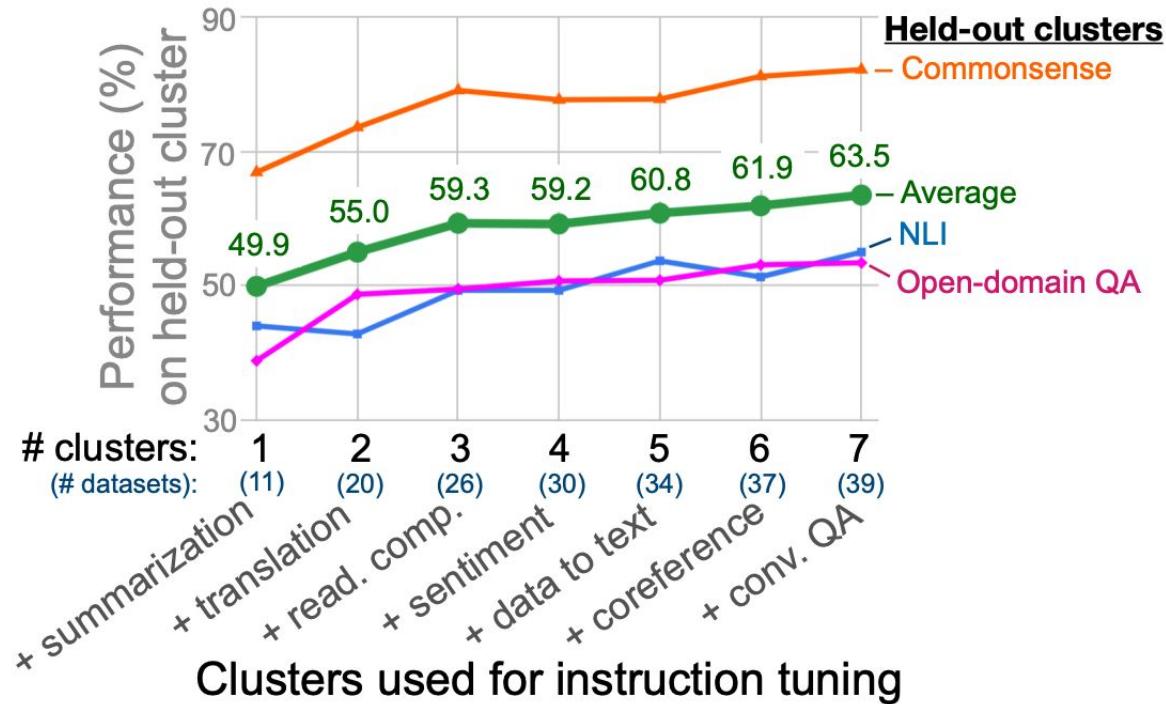


Instruction Tuning – Finetuning a LLM on a collection of tasks described by instructions to improve performance on unseen tasks.

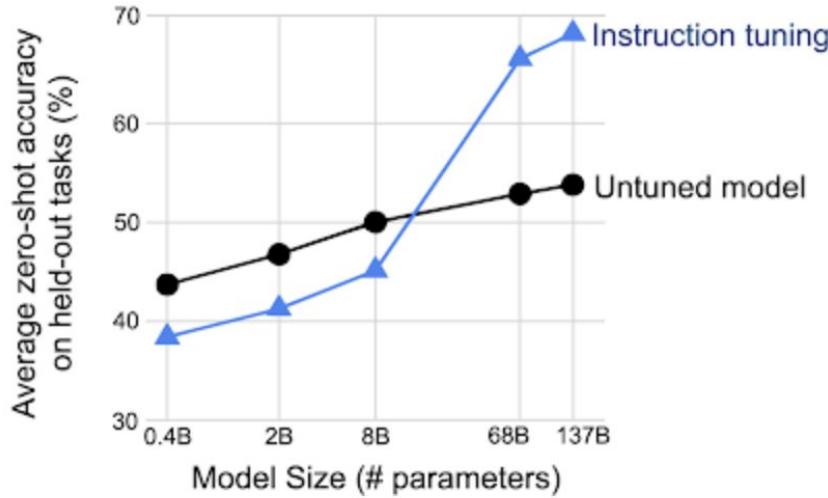
Leverage supervision to teach the model to perform many NLP tasks.



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



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



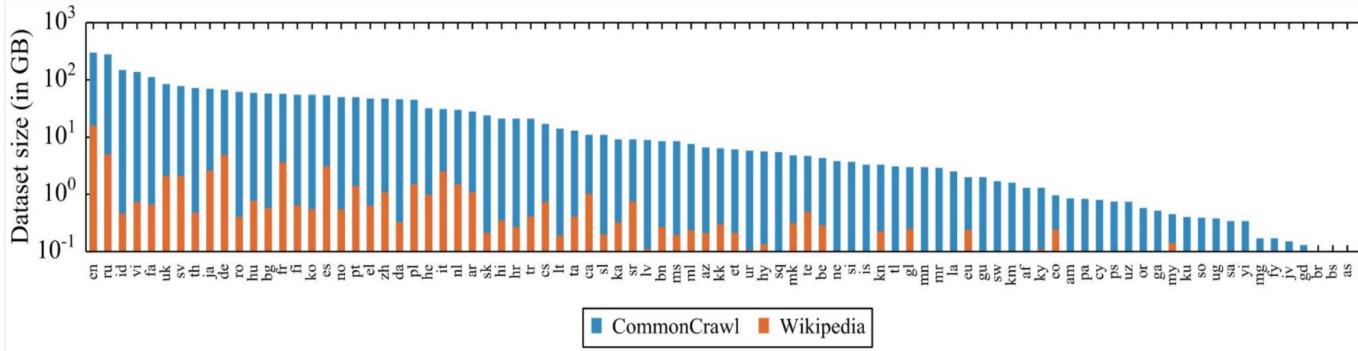
Instruction tuning only improves performance on unseen tasks for models of certain size.

Zero shot performance 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 (which typically impacts low resource languages more [Oreva et al. 2021](#)) – would be great if a model generalized to a task out of the box.

ACL [Keynote](#), [Conneau et al.](#)

This makes instruction finetuning a particular promising research direction for multilingual, where there is a pronounced skew in the “haves” and “have nots.”

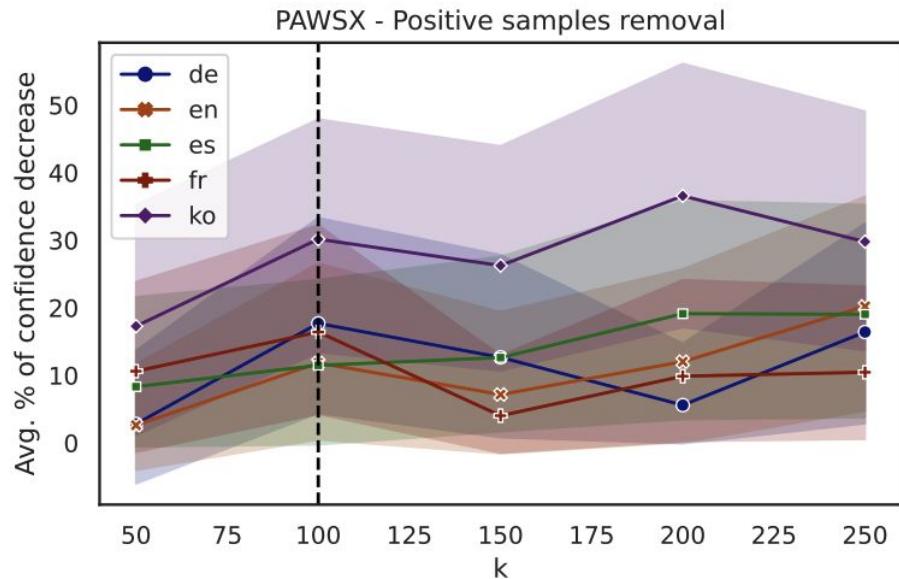


The long-tail of multilinguality, few high resource languages and many sparsely populated languages.

Choenni et al. also observe that multi-task finetuning benefits multilingual tasks in-distribution performance.

Cross-lingual sharing increases as finetuning progresses.

Languages can support one another by playing both reinforcing as well as complementary roles.



This was the starting point for
our year long open science AYA
multilingual project kicked off.

The Journey of



14 month movement which resulted in state of art dataset and model.

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 Patel⁴,
Deividas Mataciunas¹, Laura O'Mahony⁵, Mike Zhang⁶, Ramith Hettiarachchi⁷,
Joseph Wilson⁸, Marina Machado³, Luisa Souza Moura³, Dominik Krzemiński¹,
Hakimeh Fadaei¹, Irem Ergün³, Ifeoma Okoh¹, Aisha Alaagib¹,
Oshan Mudannayake¹, Zaid Alyafeai⁹, Vu Minh Chien¹, Sebastian Ruder³,
Surya Guthikonda¹, Emad A. Alghamdi¹⁰, Sebastian Gehrmann¹¹,
Niklas Muennighoff¹, Max Bartolo³, Julia Kreutzer¹², Ahmet Üstün¹²,
Marzieh Fadaee¹², and Sara Hooker¹²

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

Read our research, [Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning](#).

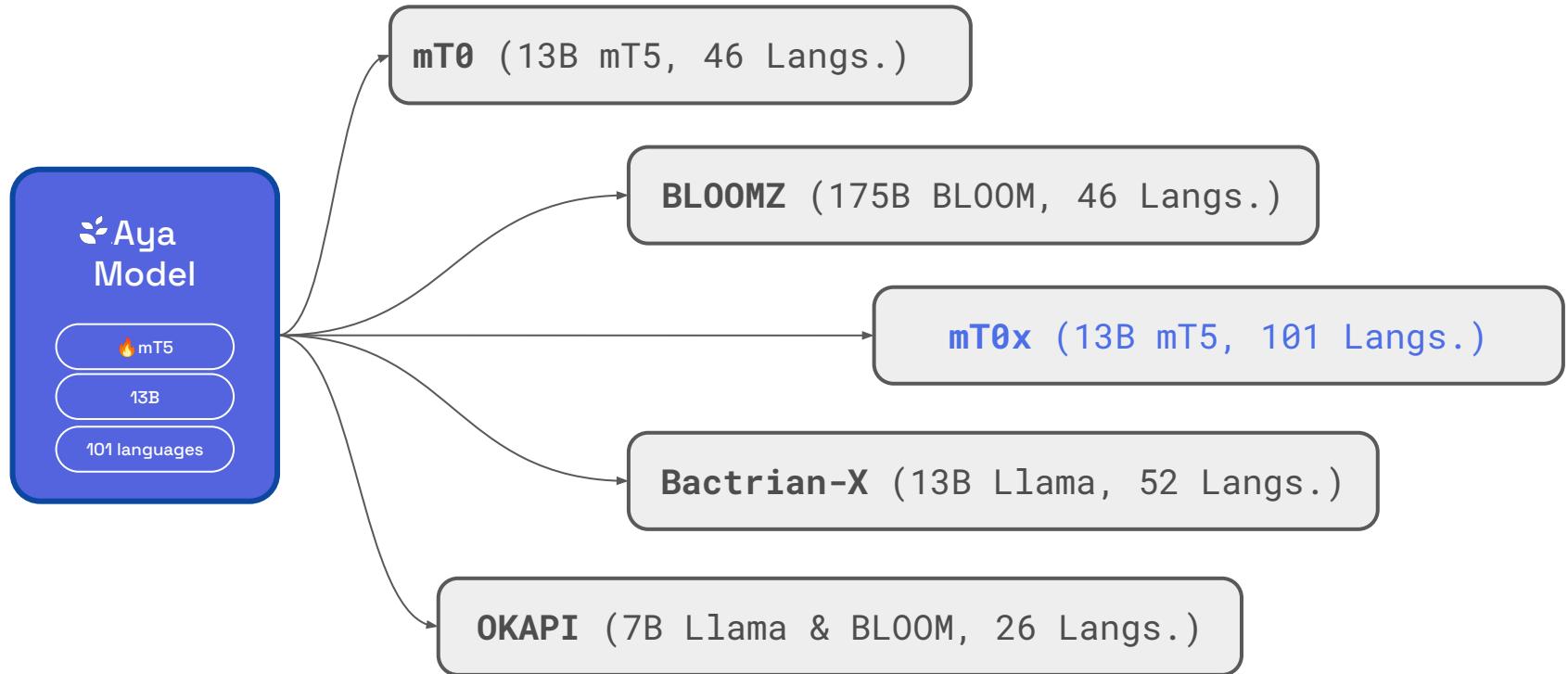
Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model

Ahmet Üstün^{♦1}, Viraat Aryabumi^{♦1}, Zheng-Xin Yong^{♦2,4},
Wei-Yin Ko^{♦3}, Daniel D'souza^{♦4}, Gbemileke Onilude⁵,
Neel Bhandari⁴, Shivalika Singh⁴, Hui-Lee Ooi⁴, Amr Kayid³,
Freddie Vargus⁴, Shayne Longpre⁶, Niklas Muennighoff⁴,
Marzieh Fadaee¹, Julia Kreutzer¹, and Sara Hooker¹

¹Cohere For AI, ²Brown University, ³Cohere, ⁴Cohere For AI Community, ⁵Carnegie Mellon University, ⁶MIT

Read our research, [Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model](#).

Expanding multilingual to double languages. Comparison with state-of-the-art baselines



Release largest multilingual data collection to-date.

Dataset

65 languages

Human-written instances
from fluent native speakers

204K instances

https://hf.co/datasets/CohereForAI/aya_dataset

Collection

115 languages

Templating and Translating
existing datasets

513M instances

https://hf.co/datasets/CohereForAI/aya_collection

Evaluation

101 languages

Mixture of human-curated,
postedits, and translations

23K instances

https://hf.co/datasets/CohereForAI/aya_evaluation_suite

This was the starting point for
our year long open science AYA
multilingual project kicked off.

03

Aya Dataset & Collection



3000 collaborators building a cross-institutional dataset.



A research initiative spanning 119 countries.

02 The People of Aya

Core team 1/2

Listed in alphabetical order.



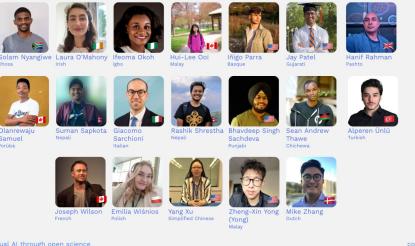
Accelerating multilingual AI through open science

cohere.com/research/aya

02 The People of Aya

Language Ambassadors 3/3

Listed in alphabetical order.



Accelerating multilingual AI through open science

cohere.com/research/aya

02 The People of Aya

Core team 2/2

Listed in alphabetical order.



19

02 The People of Aya

Language Ambassadors 1/3

Listed in alphabetical order.



Accelerating multilingual AI through open science

cohere.com/research/aya

02 The People of Aya

Language Ambassadors 2/3

Listed in alphabetical order.

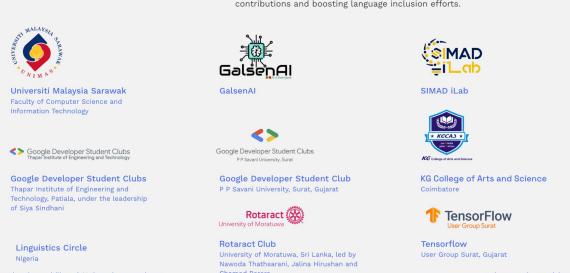


Accelerating multilingual AI through open science

cohere.com/research/aya

02 The People of Aya

Partner Organizations



20

21

Aya Dataset and Collection –
We build a huge collection of multilingual instruction data with **human annotations**, manual curations, quality check, and translation.

Aya Dataset

Prompt

65

Completion

"قىزى ئەللىخانى رەكىلۇ" ەم قىتاڭ ئەنۋەلەن مەسىنىڭ
ئېلىپەت چۈھۈك ؟

si

Qual é a origem do xaxado?

pt

ایا در اساطیر ایرانی اشاره‌ای به دماوند شده است؟

fa

بىلەكالىپى مۇد دەمىددىن ۱۰۰٪ ەيمىتى?

te

Berikan jawapan untuk teka-teki berikut:
Apakah jenama tingkap yang paling ramai
orang guna?

ms

Inis dom faci Baile Átha Cliath.

ga

Maxaa waaye shay aa Soomaali badan cunaan
maalin walba oo dadk qaar ay ku adagtaghay
la'aantiisa in ay cuntada cunaan?

so

"قىزى ئەللىخانى رەكىلۇ" ەم قىتاڭ ئەنۋەلەن مەسىنىڭ
ئېلىپەت چۈھۈك ئەمدىت ئەم مەسىنىڭ ەرچىلەنەت ئەم
ئەنۋەلەن ئەتىلەن ئەنۋەلەن ەرچىلەن ئەم.

O xaxado tem origens associadas ao Cangaceiro, onde cangaceiros, normalmente homens, dançavam com rifles em celebrações. Com a entrada de mulheres nos grupos, a participação feminina na dança também se expandiu.

مداؤنیدن اساطیر ایران چاهکار و خداوار شهربان
بپیش از هر چیز رو این است که فردیون از شخصیت‌های اساطیری ایران، شخصیات در این انجا در غاری به بند کشیده‌اند و سپاهان را اینجا نهاده‌انست تا اخراج‌های که پند بیکسل و شکست خالق اغزار کند و سرگاجام به دست گوشاسب کشته شوند

بىلەكالىپى مۇد دەمىددىن مەسىنىڭ ھەپلەن ئەللىخانى
لەردىق، ئەمدا "داخىل دەرىگەنلەن سەئەل، سەئەل
دەرىگەنلەن ئەن سەئەل ئەلەنلەن دەلەنلەن". بىلەكالىپى
مۇد دەنىيەنلىك ئەندرى دەلەن ئەن سەئەل نەنى نەنى ئەلەن
ئەلەنلىك ئەلەنلەن ئەلەنلەن ئەلەنلەن دەلەنلەن

Jawapannya ialah Microsoft kerana "Microsoft Windows" ("perkataan "window" bermakna tingkap dalam bahasa Inggeris) ialah salah satu sistem operasi yang banyak digunakan oleh ramai orang di seluruh dunia.

Is é Baile Átha Cliath, nô Dublin i mBéarla, príomhchathair na hÉireann agus é ar an gcatáir is mo sa tís freisin. Tá sé suite in oirthear na tíre, cois Life agus ar imseall Muhr Éireann. Tá breis is 1.6 jyllion duine ina gcónaí i mórcheartha Bhaille Átha Cliath agus deirtear go dtíochfaidh borradh agus fás ar an daonra sa toidhcáil.

Shay aa Soomaali badan cunaan maalin walba
oo dadk qaar ay ku adagtaghay la'aantiisa in ay
cuntada cunaan waa Mooska. Mooska dad
badan oo Soomaali ah way jeciyihii, wuxuuna
ka mid hayd waxyaabha dalku uu ku caan
baxay in uu choofiyo.



Aya Dataset and Collection is the largest instruction mixture with permissive licence

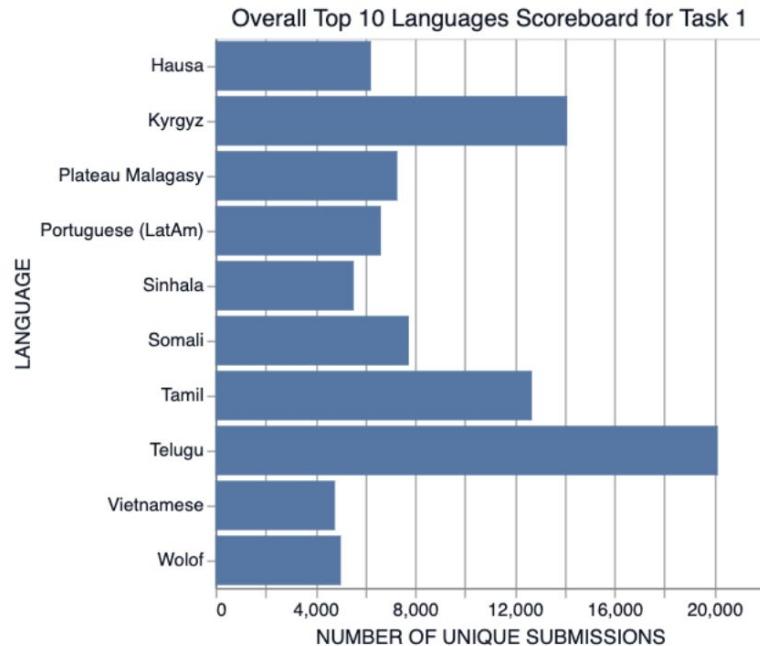
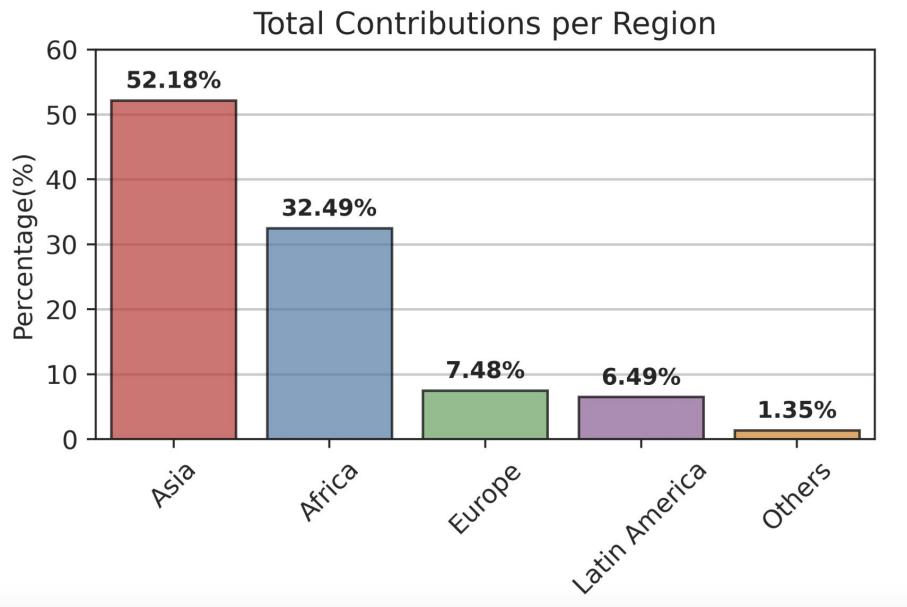
Dataset	#Instances	#Langs	% English	Generation method	Permissive license
Llama2 IFT data [Touvron et al., 2023]	NA	27	90%	Human-annotations SFT datasets	✗
Alpaca [Taori et al., 2023]	52K	1	100%	Synthetic data generation IFT datasets	≈
P3 [Sanh et al., 2022]	12M	1	100%	Template generation given applied to English datasets	✓
Flan 2022 [Longpre et al., 2023a]	15M	60	100%	Template generation applied to English datasets	✓
xP3 [Muennighoff et al., 2023c]	81M	46	39%	Template generation applied to English datasets	✓
Sweinstruct [Holmström & Doostmohammadi, 2023]	68K	1	0%	Machine translation English IFT datasets	≈
Okapi [Dac Lai et al., 2023]	158K	26	45%	Machine translation English IFT datasets	✓
Bactrian-X [Li et al., 2023a]	3.4M	52	2%	Machine translation + synthetic data generation	≈
Aya Dataset	204K	65	2%	Original IFT Human-annotations	✓
Aya Collection	513M	114	3.5%	Template Generation and translating existing datasets	✓

Table 1: Comparison of different instruction-tuning datasets. ✓represents permissive licenses that allow commercial use while ≈ represents restrictive licenses that do not allow commercial use.



✗represents non availability of license.

Exceptional contribution for low-resource languages!



We also found it important to reward quality in the UI with an AYA score.

Open science contributors review each other, resulting already in clear gains in quality.

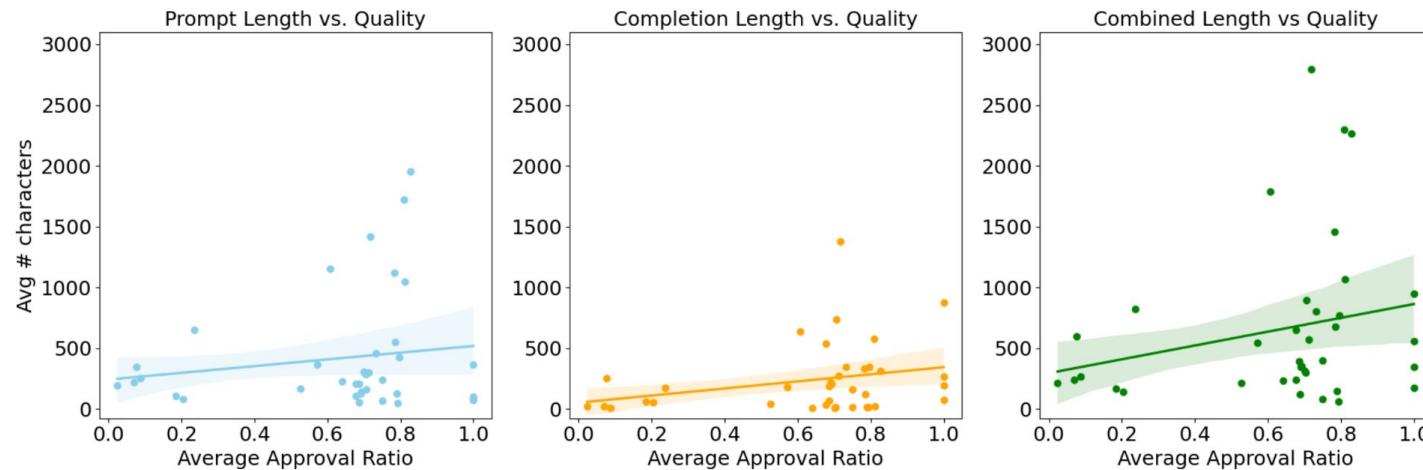
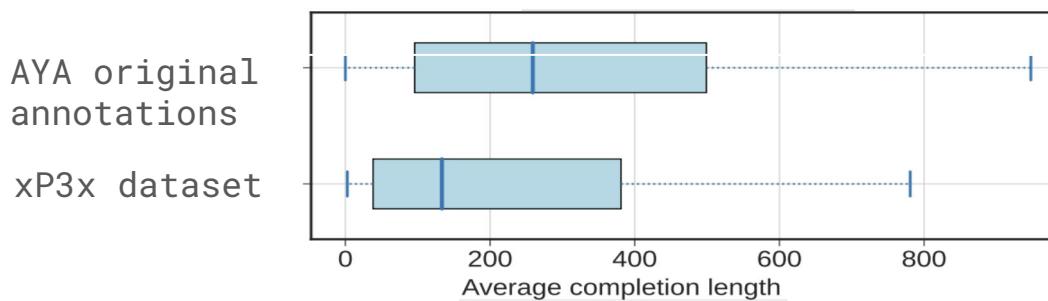
This allows us to estimate quality at scale with a ever changing annotator pool.

Leaderboard

Daily Weekly Overall Language

Rank	Username	Languages	Points	Quality	Aya Score
#1	 suryaguthikonda	Telugu	8056	3.83	3870
#2	 huilee_	Chinese (Simplified), Chinese (Traditional), English, Standard Malay	8462	3.22	3077
#3	 _mohamdy	Egyptian Arabic, English, Standard Arabic	2125	3.76	2382
#4	 zaidalyafeai	Egyptian Arabic, English, Standard Arabic, Taizzi-Adeni Arabic	5038	3.54	2343
#5	 yongzx	Chinese (Simplified), English, Standard Malay	5563	3.29	2107

Not only quantity but also quality → longer multilingual instructions

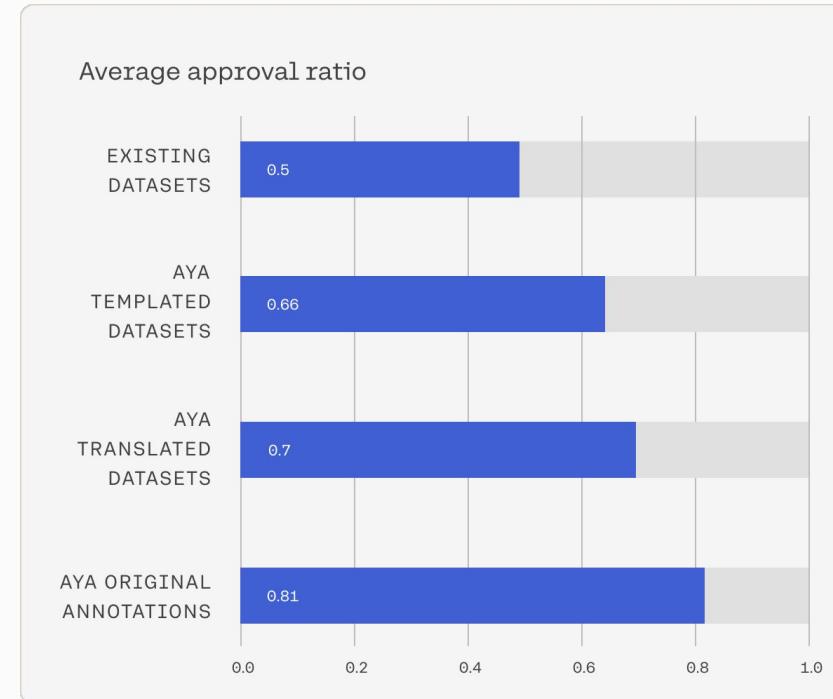




Aya Collection Surpasses Previous Multilingual Datasets in terms of quality

The quality of instruction data significantly influences the performance of the fine-tuned language model.

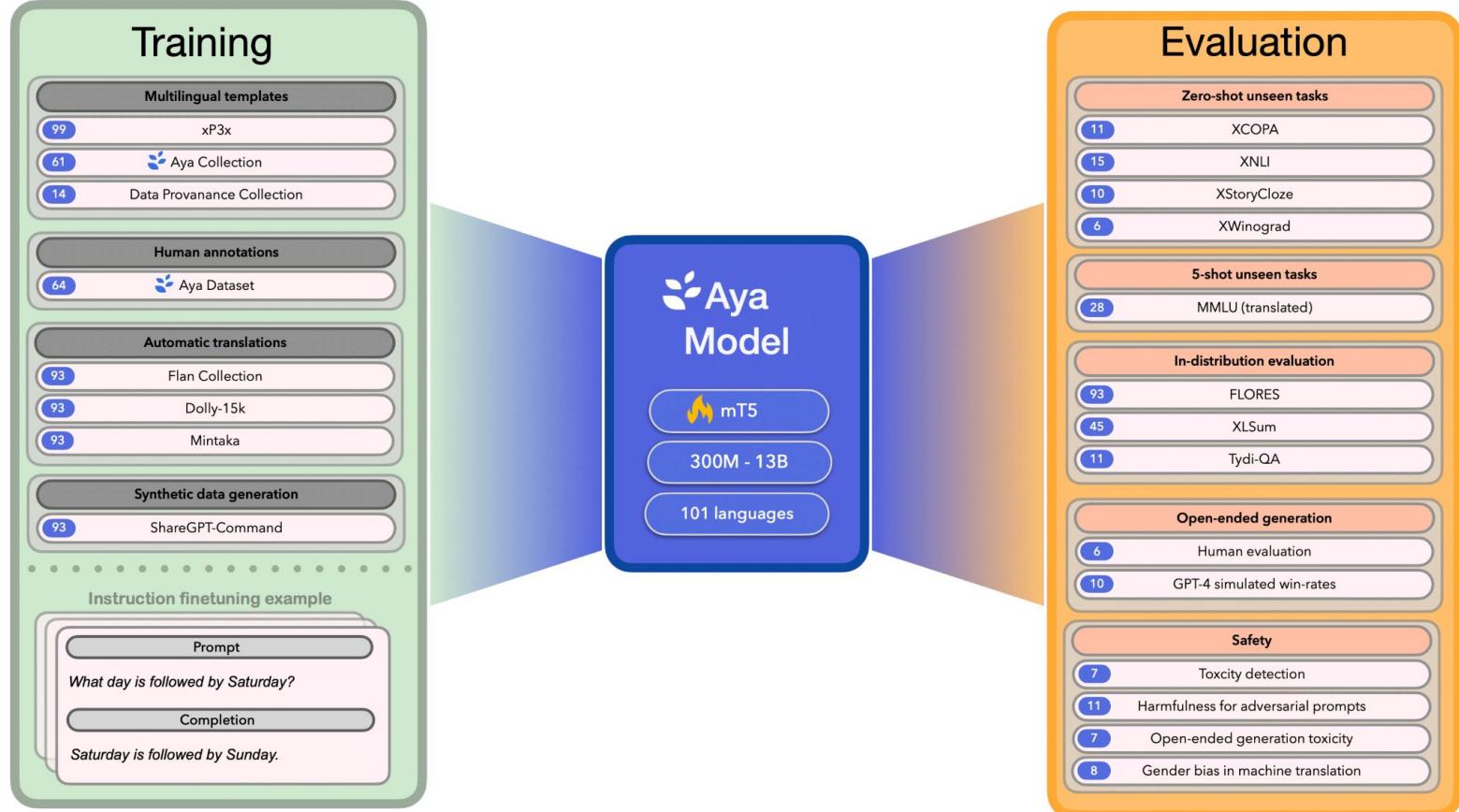
Through a global assessment, we enlisted annotators to assess the quality of various multilingual data collections. This process revealed that Aya's original annotations received the highest approval ratings from both native and fluent speakers.



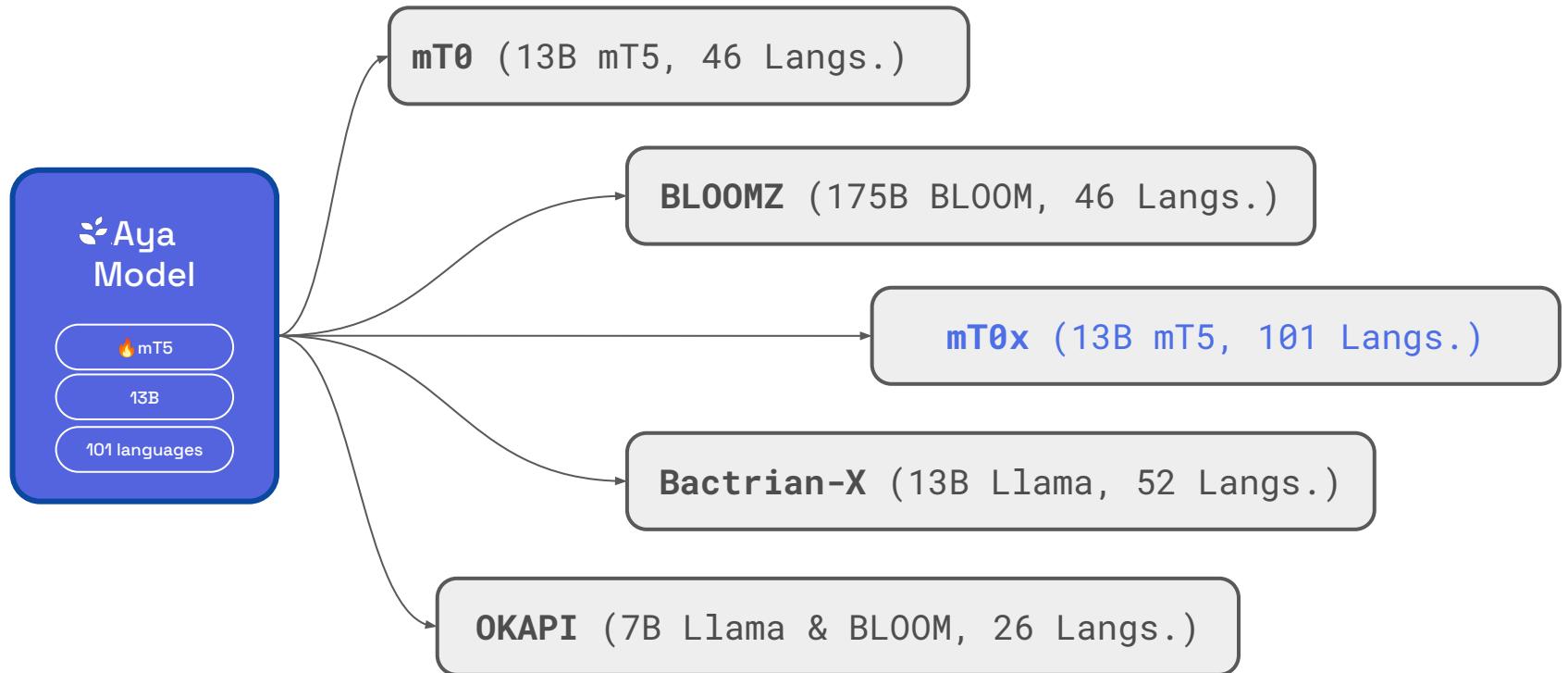
04 Aya Model



Aya model is a massively multilingual instruction-following LLM with a diverse multilingual data mixture and comprehensive evaluation suite



Expanding to double languages. Comparison with state-of-the-art baselines



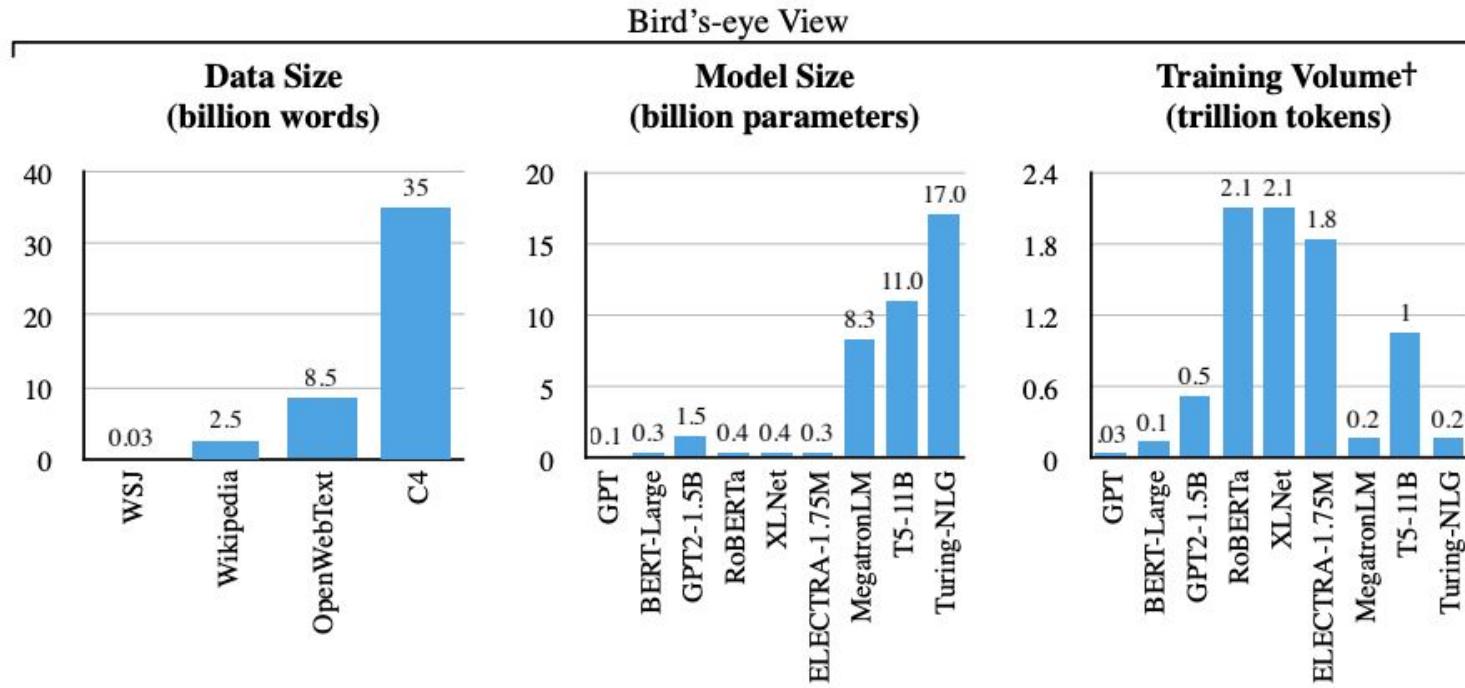
How to improve data quality and diversity?

Large scale datasets are typically an unknown.

- Data is randomly crawled from the internet
- Billions of examples
- Data quality is low
- Manual inspection / validation is simply impossible
- Simple automated techniques already in place for data cleaning

Dataset	# documents	# tokens	size
C4.EN.NOCLEAN	1.1 billion	1.4 trillion	2.3 TB
C4.EN.NOBLOCKLIST	395 million	198 billion	380 GB
C4.EN	365 million	156 billion	305 GB

The volume of data typically used for state of art training is infeasible for comprehensive auditing or labelling.



Instruction finetuning datasets despite being smaller are also increasingly opaque.

It has become the norm to finetune on ever larger sets of “collections” – a set of tasks where often data provenance or characteristics is not known.

COLLECTION	PROPERTY COUNTS					TEXT LENS		DATASET TYPES									
	DATASETS	IALOGS	TASKS	LANS	TOPICS	DOMAINS	DOWNS	INPT	TGT	SOURCE	Z	F	C	R	M	U	O
Airoboros	1	17k	5	2	10	1	1k	347	1k	🔗	✓					✓	✓
Alpaca	1	52k	8	1	10	1	100k	505	270	🔗	✓					✓	✓
Anthropic HH	1	161k	3	1	10	1	82k	69	311	🔗	✓					✓	✓
BaizeChat	4	210k	12	2	37	3	<1k	74	234	🔗	✓					✓	✓
BookSum	1	7k	4	1	10	1	<1k	14k	2k	🔗	✓					✓	✓
CamelAI Sci.	3	60k	2	1	29	1	<1k	190	2k	🔗	✓					✓	✓
CoT Coll.	6	2,183k	12	7	29	1	<1k	728	265	🔗	✓					✓	✓
Code Alpaca	1	20k	3	2	10	1	5k	97	196	🔗	✓					✓	✓
CommitPackFT	277	702k	1	278	751	1	4k	645	784	🔗	✓					✓	✓
Dolly 15k	7	15k	5	1	38	1	10,116k	423	357	🔗	✓					✓	✓
Evol-Instr.	2	213k	11	2	17	1	2k	570	2k	🔗	✓					✓	✓
Elan Collection	450	9,812k	19	39	1k	22	19k	2k	128	🔗	✓	✓	✓	✓	✓	✓	✓
GPT-4-Alpaca	1	55k	7	1	10	1	1k	130	543	🔗	✓					✓	✓
GPT4AllJ	7	809k	10	1	56	1	<1k	883	1k	🔗	✓					✓	✓
GPTTeacher	4	103k	8	2	33	1	<1k	227	360	🔗	✓					✓	✓
Gorilla	1	15k	4	2	10	2	<1k	119	76	🔗	✓					✓	✓
HC3	12	37k	6	2	102	6	2k	119	652	🔗	✓					✓	✓
Joke Expl.	1	<1k	2	1	10	1	<1k	96	547	🔗	✓					✓	✓
LAION OIG	26	9,211k	12	1	171	11	<1k	343	595	🔗	✓					✓	✓
LIMA	5	1k	10	2	43	6	3k	228	3k	🔗	✓					✓	✓
Longform	7	23k	11	1	63	4	3k	810	2k	🔗	✓					✓	✓
OpAss OctoPack	1	10k	3	20	10	1	<1k	118	884	🔗	✓					✓	✓
OpenAI Summ.	1	93k	5	1	10	1	14k	1k	134	🔗	✓					✓	✓
OpenAssistant	19	10k	4	20	99	1	14k	118	711	🔗	✓					✓	✓
OpenOrca	4	4,234k	11	1	30	23	28k	1k	492	🔗	✓					✓	✓
SHIP	18	349k	6	2	151	1	4k	824	496	🔗	✓					✓	✓
Self-Instruct	1	83k	6	2	10	1	3k	134	104	🔗	✓					✓	✓
ShareGPT	1	77k	9	1	10	2	<1k	303	1k	🔗	✓					✓	✓
StackExchange	1	10,607k	1	2	10	1	<1k	1k	901	🔗	✓					✓	✓
StarCoder	1	<1k	1	2	10	1	<1k	195	504	🔗	✓					✓	✓
Tasksource Ins.	288	3,397k	13	1	582	20	<1k	518	18	🔗	✓					✓	✓
Tasksource ST	229	338k	15	1	477	18	<1k	3k	6	🔗	✓					✓	✓
TinyStories	1	14k	4	1	10	1	12k	517	194k	🔗	✓					✓	✓
Tool-Llama	1	37k	2	2	10	1	-	7k	1k	🔗	✓					✓	✓
UltraChat	1	1,468k	7	1	11	2	2k	282	1k	🔗	✓					✓	✓
Unnatural Instr.	1	66k	4	1	10	1	<1k	331	68	🔗	✓					✓	✓
WebGPT	5	20k	4	1	35	3	1k	737	743	🔗	✓					✓	✓
xP3x	467	886,240k	5	245	151	14	<1k	589	441	🔗	✓					✓	✓

[The Data Provenance Initiative:
A Large Scale Audit of Dataset Licensing & Attribution in AI](#) Longpre et al. 2023

Data pruning and weighting is a promising direction.

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

	% train examples with dup in train	% valid with dup in train	% valid with dup in valid
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](#)

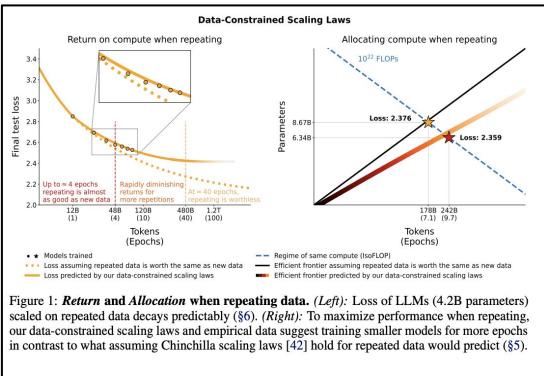


Figure 1: *Return and Allocation when repeating data.* (Left): Loss of LLMs (4.2B parameters) scaled on repeated data decays predictably (§6). (Right): To maximize performance when repeating, our data-constrained scaling laws and empirical data suggest training smaller models for more epochs in contrast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

[Muennighoff et al. 2023](#)

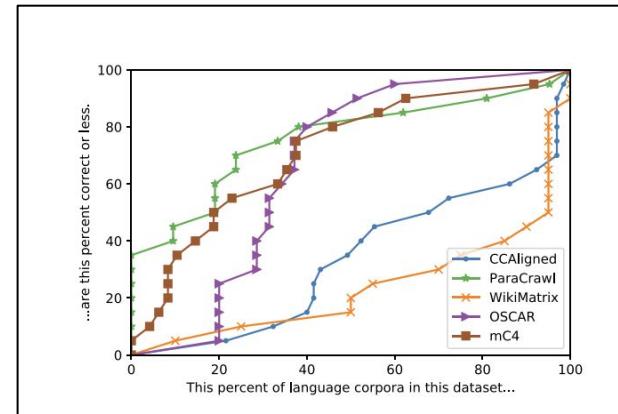


Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

[Kreutzer et al. 2022](#)

Our recent work was at the pre-training level – showing internet level scale shows 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

[\[\[Marion et al. 2023\]\]](#)

For RLHF, our recent work also look at how to prioritize limited human annotation time.

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.

For **instruction-finetuning** there is the question of task mixing, task coverage diversity and the impact on quality.

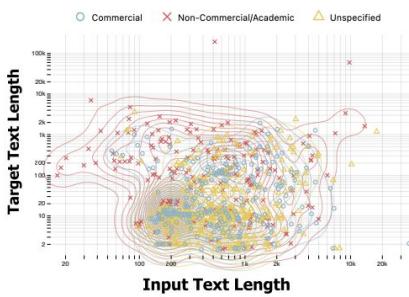
Combining 4 Insights



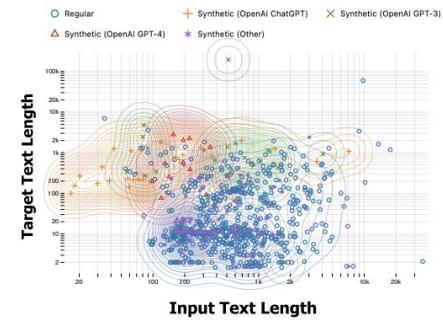
- ★ Outperforms prior work by 3-17%
- ★ Better computational efficiency for next finetuning
- ★ Open Source Flan Collection

Our learnings from AYA:

- Currently available open source multilingual instruction finetuning datasets have extremely short completions.
- Also suffers from extremely low prompt diversity.



(a) License Use Categories vs Text Lengths

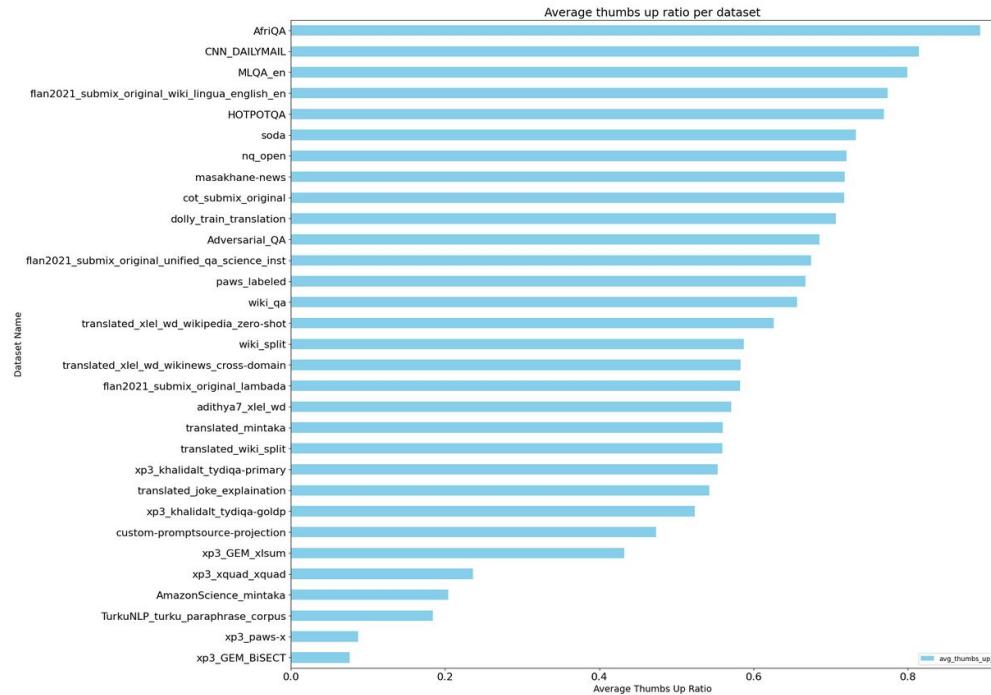


(b) Synthetic/Regular Datasets vs Text Lengths

Figure 5: Across finetuning datasets, we visualize their mean input (x-axis) and target (y-axis) text lengths, measured in log-scaled number of words. The colors indicate either their license use category (left) or whether they were machine generated or human collected (right). Long target texts are represented in large part by Non-Commercial and Synthetic datasets, that are often generated by commercial APIs.

[[[Longpre et al. 2022](#)]]

We collected feedback from native speakers on each dataset, and found a large spread in perceived quality of datasets.



Extremely short completions – often a symptom of instruction finetuned datasets constructed using templates.

Question: "The homepages of many websites removed all advertisements and what other change?"

Context: "Many websites converted their home page to black and white; Sina.com and Sohu, major internet portals, limited their homepages to news items and removed all advertisements. Chinese video sharing websites Youku and Tudou displayed a black background and placed multiple videos showing earthquake footage and news reports. The Chinese version of MSN, cn.msn.com, also displayed banner ads about the earthquake and the relief efforts. Other entertainment websites, including various gaming sites, such as the Chinese servers for World of Warcraft, had shut down altogether, or had corresponding links to earthquake donations. After the moments of silence, in Tiananmen Square, crowds spontaneously burst out cheering various slogans, including "Long Live China". Casinos in Macau closed down."

Answer:

limited their homepages

Here is a review left by a customer on a product. Would you say he was satisfied or dissatisfied?

Title: Could have picked better sketches

Review: It's not Jimmy's fault that they picked not so good sketches for this DVD. I love Jimmy and the fact that he would crack up in the middle of the sketches made the sketches all the more hilarious and made him even more charming. I think that's why people love him. He was the BEST thing to ever happened to SNL. As for this DVD, the Weekend Updates are priceless and I love the "Drinkin in the Woods" segment. They could have picked better sketches though. The audio commentary from Jimmy and the writers are hilarious in itself. And I almost forgot about that his SNL audition is on there too. You can tell he's kinda nervous or he seems nervous to me cause he's soft spoken, but his performance in that audition doesn't show it at all. I put 4 stars cause it's not that great, but it's not that bad. All in all, if you're a Jimmy fan, it's a great addition to your DVD collection.

satisfied

[[[Longpre et al. 2022](#)]]

And consistent with results that show that the most permissive licensed datasets tend to have the shortest completions.

Commercial licenses have mean target length of 102.7 vs 1580 for the more restrictive non-commercial

.

METRICS	COMMERCIAL		UNSPECIFIED		NC / A-O	
	MEAN	ENTROPY	MEAN	ENTROPY	MEAN	ENTROPY
TASKS	1.7±0.1	0.61	1.6±0.1	0.53	3.4±0.2	0.69
LANGUAGES	1.3±0.0	0.52	1.2±0.0	0.16	1.1±0.0	0.45
TOPICS	8.2±0.2	0.70	9.2±0.1	0.75	9.1±0.2	0.77
SOURCES	1.6±0.1	0.67	1.8±0.1	0.72	4.2±1.3	0.78
INPUT TEXT LENGTHS	1043.4±151.9	6.37	860.2±67.7	6.66	950.3±112.9	6.46
TARGET TEXT LENGTHS	102.7±14.6	4.39	90.5±14.3	4.09	1580.7±965.6	5.37
SYNTHETIC	12.8%±2.1	-	15.6%±1.7	-	45.5%±5.4	-

Table 3: The mean number of features (e.g. tasks or languages) per dataset, and the mean entropy of the distribution, representing the diversity of categories. **Non-Commercial / Academic-Only datasets have consistently and statistically higher task, topic, and source variety than Commercial datasets.** We use Normalized Shannon Entropy for discrete features, and Differential Entropy for continuous features, which are both measures of randomness.

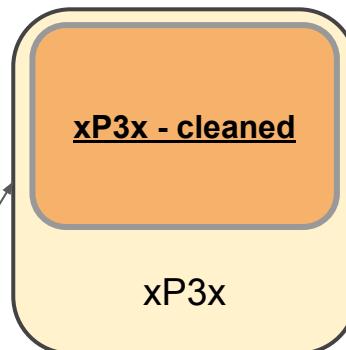
[[Longpre et al. 2022]]

Lack of prompt diversity – partly a symptom of having very few people contribute prompt templates.

First sentence of the article: authorities in western china have executed ## people , among them a bus driver who hit and killed two people while driving at night with his lights off , a state-run newspaper said .	
Title: authorities intercepted ## iraqi kurds on a barren aegean islet monday and arrested two turks accused of smuggling the group into greece , reports said . ==== Generate a title for this article: authorities say a misunderstanding about a jewish prayer ritual led to the diversion of a us airways flight to philadelphia . ==== Given the above sentence, write its title:	china executes ## for crimes ## illegal immigrants found ; smuggling suspects arrested us flight diverted after confusion over prayer

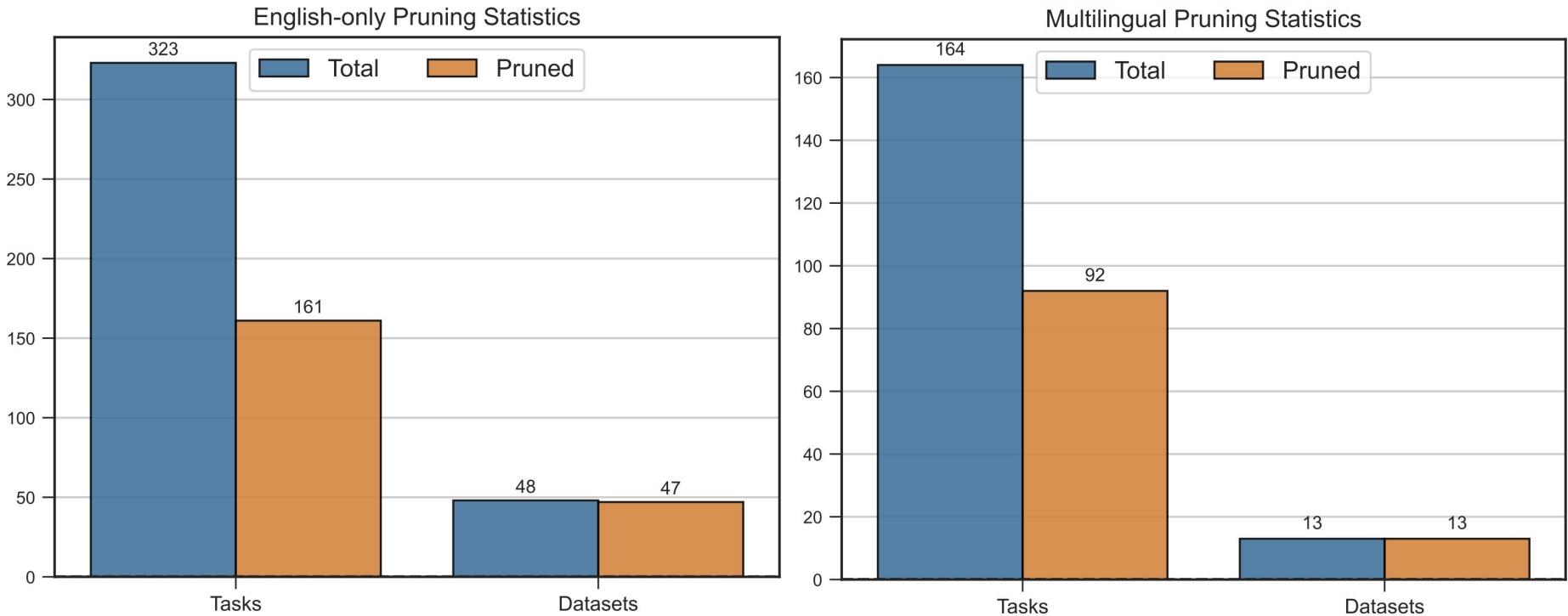
We aggressively manually pruned by removing dataset/prompts with very short answers or with ones with similar prompts (no diversity)

We have manually inspected, annotated and removed much of the worst academic datasets with short completions, not conversational style.



We removed **50% of prompt templates across 47 datasets from English split** with low quality (in xP3x).

50% in English and 44% in non-English templates are filtered out



We removed dataset/prompts with **very short answers** or with ones with **similar prompts** (no diversity)

[boolq_super_glue/could you tell me...]

"**inputs**": "United States raw milk debate -- The FDA considers hard, aged cheese, such as parmesan and cheddar, made from raw milk to be generally safe for consumption; soft cheese made from raw milk is considered unsafe. [...] Having read that, could you tell me is all cheese in the united states pasteurized?

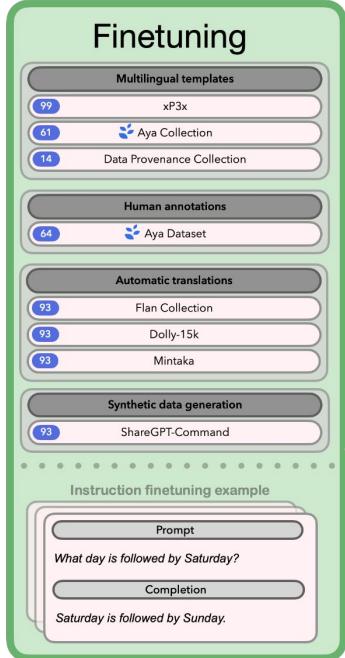
"**output**": "No"

[qqp_glue/answer]

"**inputs**": "Can an answer to "Are there any happily married couples whose kundli didn't match?" also be used to answer "Are there happily married couples, whose kundli didn't matched?"?

"**output**": "Yes"

We combined Aya dataset and collection with existing **high-quality** instructions



Carefully selected and pruned **multilingual templates** from 3 sources:

- 1) **xP3x**, a multilingual collection of academic datasets
- 2) **Aya Template Collection**, templated data subset from AYA Collection
- 3) **Data Provenance Collection**, permissively licenced data collection

101 languages

203 million examples

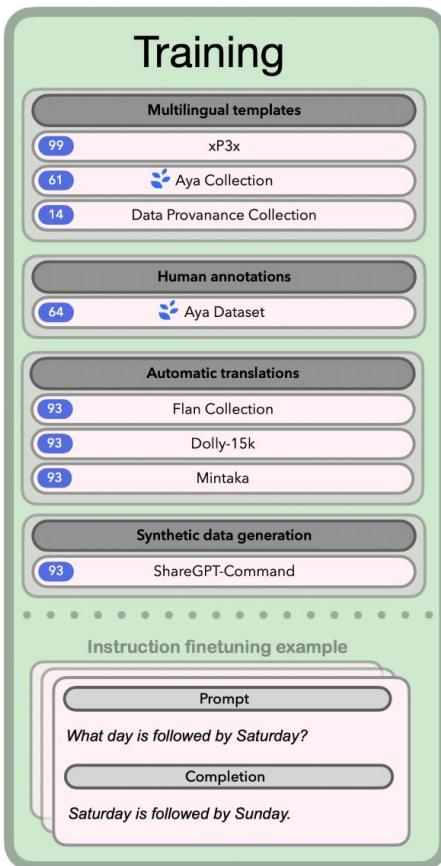
Machine translated datasets into 93 languages

Aya Dataset, a fully human-curated dataset of instructions

Synthetic instructions generated by Cohere Command and translated afterward into 93 languages

Optimization Open Questions. Is
translation a viable
augmentation strategy?

Our final training mix consisted of templated data, rare human annotations and translated data. But how to select the right balance of each?



Carefully selected and
pruned multilingual
templates

Human curated
Aya Dataset and
Collection

Machine translated
datasets into 101
languages

101 languages
203M example

Synthetic instructions
powered by **Command**

**Massive multilingual
instruction tuning
mixture**

We experimented with weighting importance of each data sources

–
Important for balancing multilinguality with downstream performance

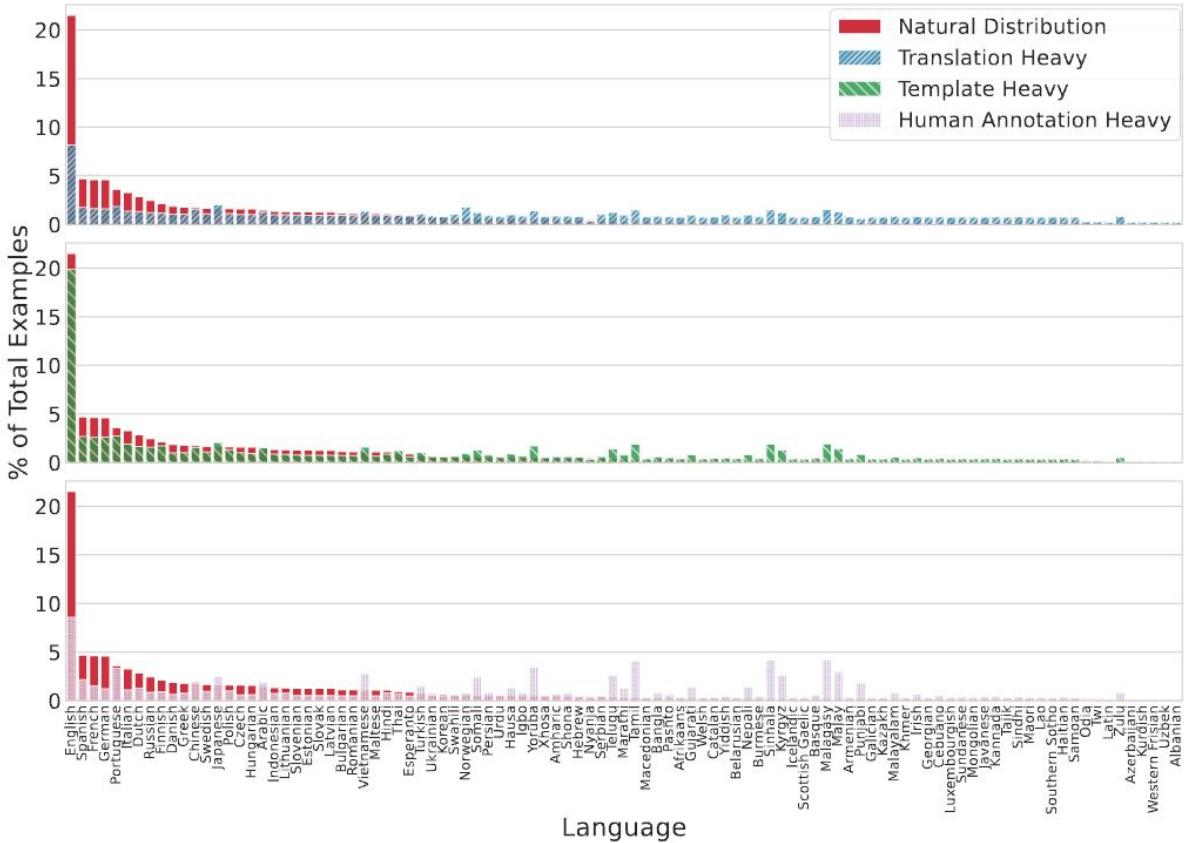


Figure 17: % of Examples for each language with different weighting schemes

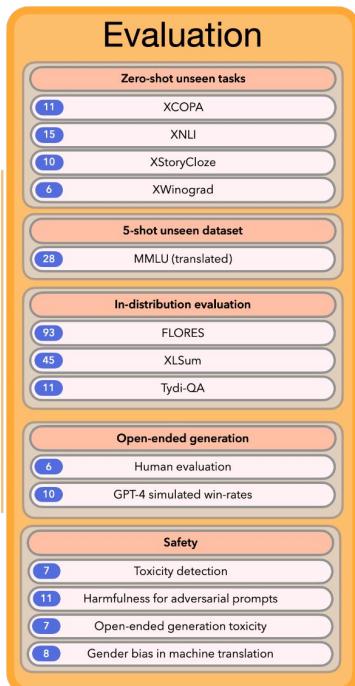
For each weighting ablation we trained a separate model.

Weighting name	HUMAN ANNOT.	TEMPLATE			TRANSLATION	
	Aya Dataset	Aya Templates	xP3x	Data Provenance	Aya Translations	ShareGPT- Command
Human Annot. Heavy	25	4	20	6	30	15
Translation Heavy	10	1.5	15	3.5	47.5	22.5
Template Heavy	20	10	30	10	20	10

Table 3: Data sampling ablation with different weighting schemes for each data source for training. Our training budget is 25M samples, and these weights describe the % of the training budget they are allocated. We group each data source based on type into Human Annotated (HA), Templated, and Translated. Based on these groups, we assign different weighting schemes: (1) *Human Annotation Heavy* which upweights the **Aya** Dataset; (2) *Translation heavy* which comparatively upweights the **Aya** Translations and ShareGPT-Command which are both translated into 93 languages; and (3) *Template heavy* which upweights the **Aya** Collection, xP3x, and Data Provenance. The results of the different weighting ablations are presented in Section 5.

To evaluate: Extensive multilingual evaluation on multiple evaluation categories on many languages

Evaluation at a glance:



Unseen tasks, or tasks the model has not been trained on:

- 1) **Discriminative**, to test how the model distinguishes between different types of inputs
- 2) **General purpose**, to test the models ability to handle diverse situations

In-distribution generative tasks, to test for generation of new outputs based on statistical distribution of original model

Human and simulated evaluation, to test quality and nuances of responses

Safety, toxicity, and bias measures, to test for harmful outputs.

99 languages

13 datasets

6 distinct evaluation types:

- Unseen zero-shot tasks
- General purpose unseen dataset (5-shot)
- In-distribution generative tasks
- Human eval
- LLM simulated eval
- Safety eval

We see on both unseen zero-shot tasks and in-distribution generative tasks benefit from translation variant the most (on average).

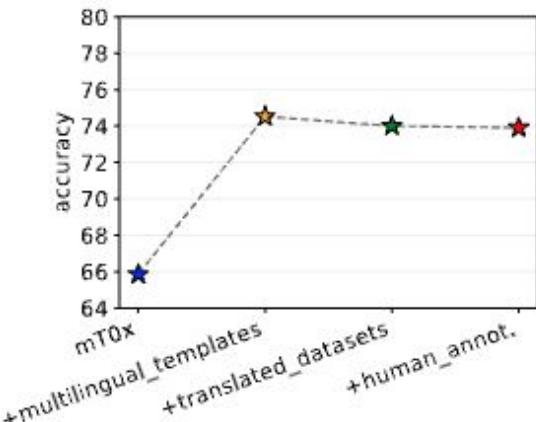
Model	Base Model	IFT Mixture	Held out tasks (Accuracy %)				
			XCOPA	XNLI	XSC	XWNG	Avg
46 Languages							
mT0	mT5 13B	xP3	75.6	55.3	87.2	73.6	72.9
BLOOMZ	BLOOM 176B	xP3	64.3	52.0	82.6	63.3	65.5
52 Languages							
BACTRIAN-X 13B	Llama 13B	Bactrian-X	52.4	34.5	51.8	50.5	47.3
101 Languages							
mT0x	mT5 13B	xP3x	71.7	45.9	85.1	60.6	65.8
Aya (human-anno-heavy)	mT5 13B	All Mixture	76.5	59.2	89.3	70.6	73.9
Aya (template-heavy)	mT5 13B	All Mixture	77.3	58.3	91.2	73.7	75.1
*Aya (translation-heavy)	mT5 13B	All Mixture	76.7	58.3	90.0	70.7	73.9

Table 5: Results for held-out task evaluation. Results are averaged across all splits of XCOPA, XNLI, XStoryCloze, and XWinoGrad. *Aya (translation-heavy) is used as the final Aya model. See § 5.6 for detailed analysis.

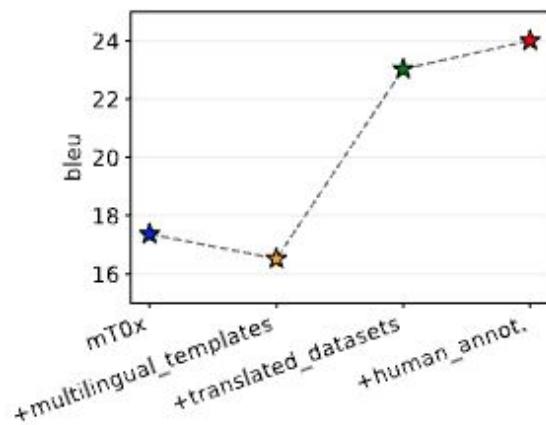
Model	IFT Mixture	Generative Tasks		
		X→ En	En → X	
101 Languages				
mT0x	xP3x	20.2	14.5	21.6
Aya (human-anno-heavy)	All Mixture	25.1	18.9	22.2
Aya (templated-heavy)	All Mixture	25.0	18.6	23.2
*Aya (translation-heavy)	All Mixture	29.1	19.0	22.0
				77.8

Table 7: Generative tasks' results based on different dataset sample weighting. Here the Translation Heavy weighting has the highest Bleu score on Flores and the Template Heavy weighting has the highest RougeLsum and F1 scores on XLSum and Tydiqa respectively. *Aya (translation-heavy) is used as the final Aya model. See § 5.6 for detailed analysis.

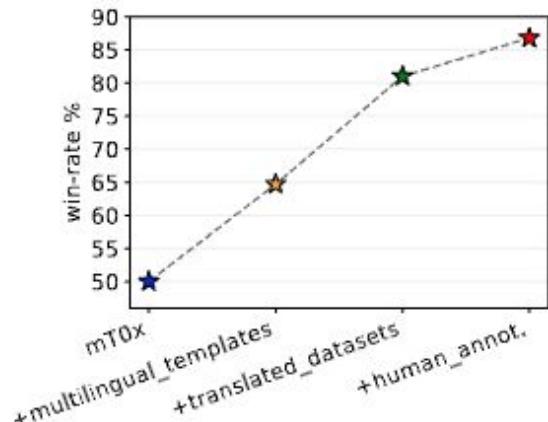
However, even when weighted towards translation – each data source contributes to the downstream performance



(a) Unseen Discriminative Tasks

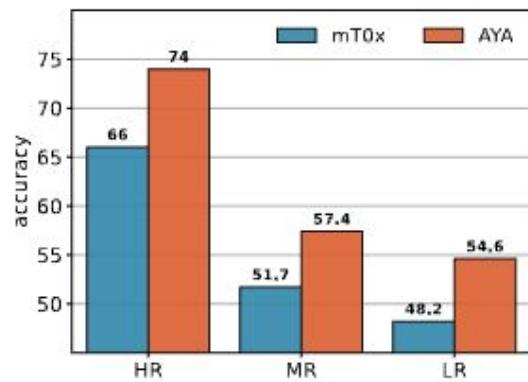


(b) Generative Task: Flores

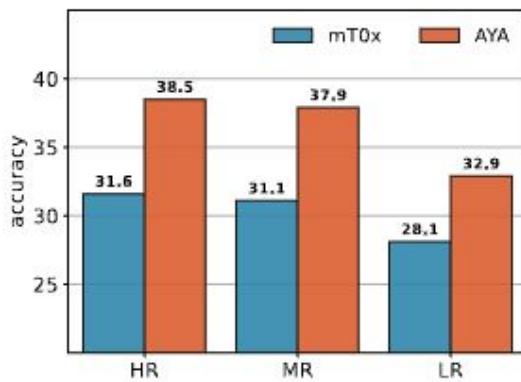


(c) Win Rates (vs mT0x)

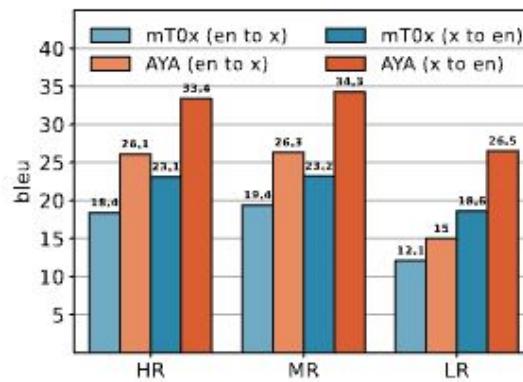
For final Aya translate model – we observe larger improvement in MR and LR languages



(a) Unseen Discriminative Tasks



(b) Multilingual MMLU



(c) Generative Task: FLORES

Figure 3: Generative and discriminative performance of models in high (HR), medium (MR), and low-resource (LR) language groups.

Optimization Open Questions: is
there a tension between
performance on discriminative
tasks and open ended tasks?

We observe a tension between discriminative tasks and open-ended generations. This is tension between how models used to be evaluated (academic benchmarks) vs used today.

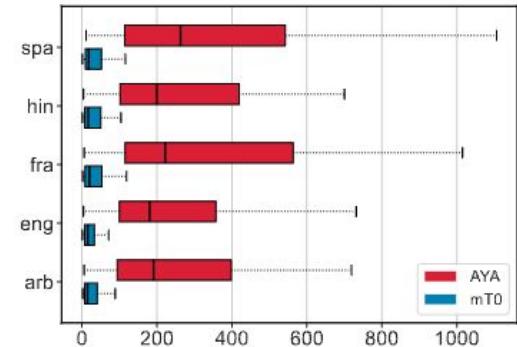
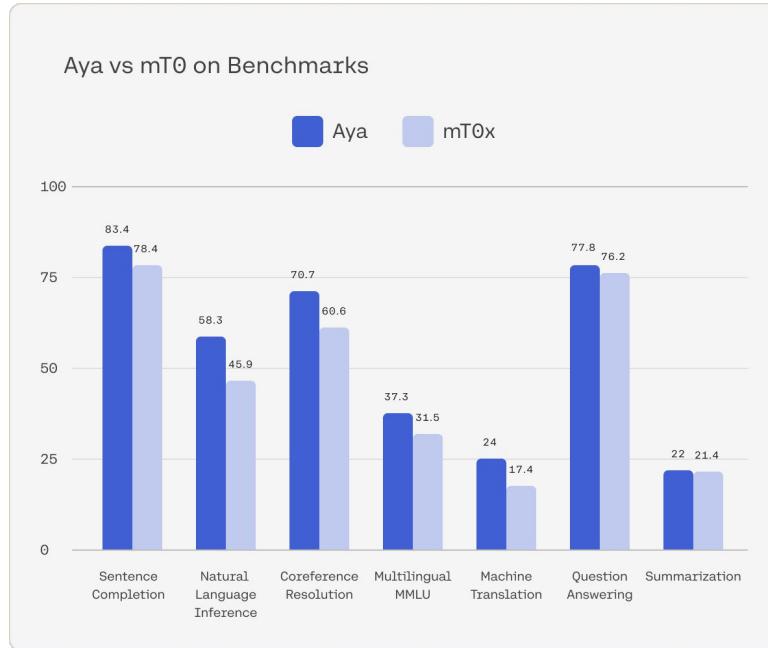
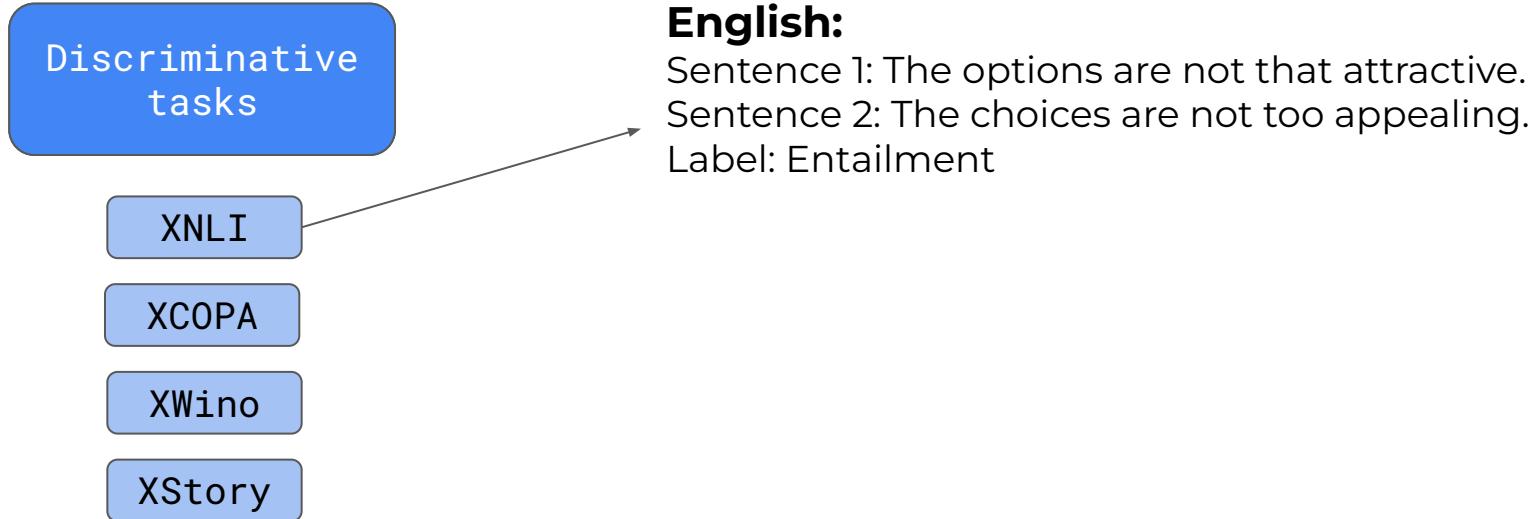


Figure 6: Completion lengths by characters for the **Aya** and mT0 models in Dolly test set for various languages.

Academic benchmarks often require short completions. Are not representative of how humans want to engage with language models.



Aya model presents much more fluid and longer completions, so discriminative tasks alone don't capture gap in quality.

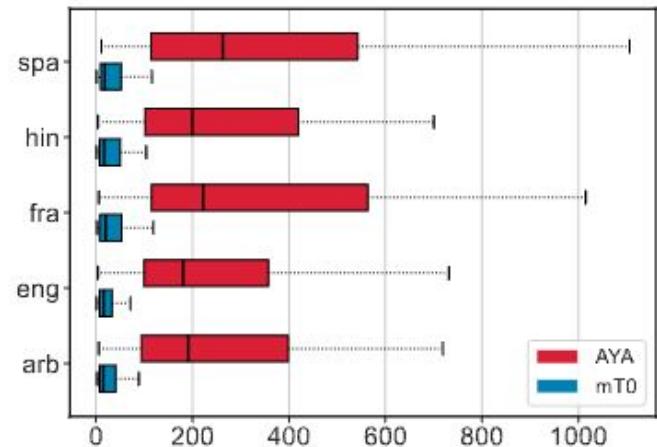
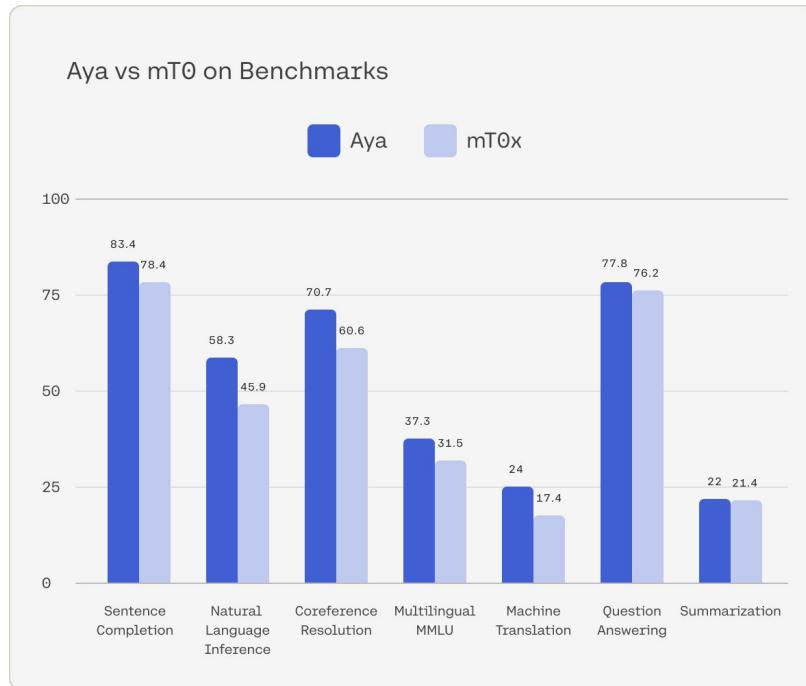


Figure 6: Completion lengths by characters for the **Aya** and mT0 models in Dolly test set for various languages.

Gap between model quality becomes much more clear when using human win rate comparisons.

Which completion is better given following prompt?

mT0

"**inputs**": "What are 5 ways to eat apples?"
"**output**": "eating apples raw"

AYA

"**inputs**": "What are 5 ways to eat apples?"
"**output**": "1. Cut the apple into halves and place them in a bowl. 2. Add a few drops of lemon juice and eat the apple as a side. 3. Cut the apple into slices and place them in a bowl. 4. Add a few drops of lemon juice"

Human eval comparisons show dominant preference for Aya vs mT0 (46 language), mT0x (101 language), BX (52 language).

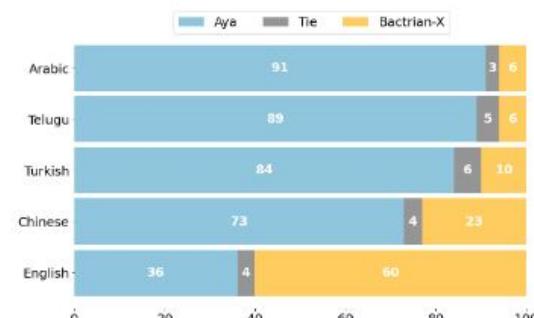
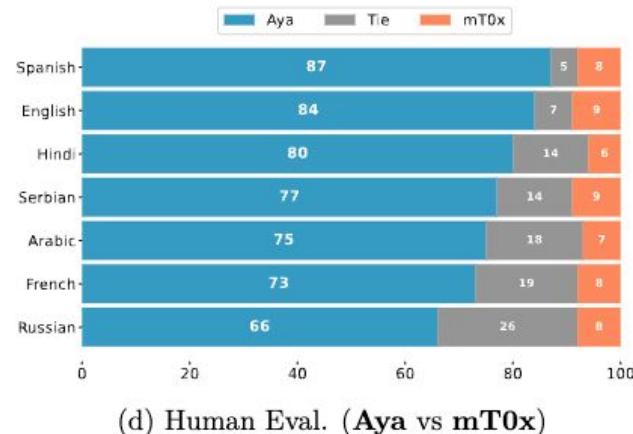
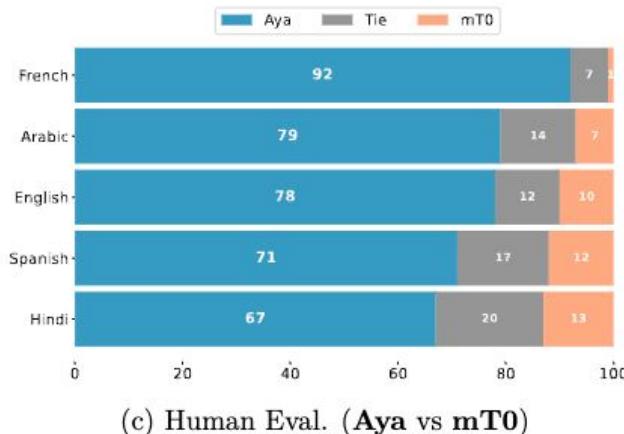
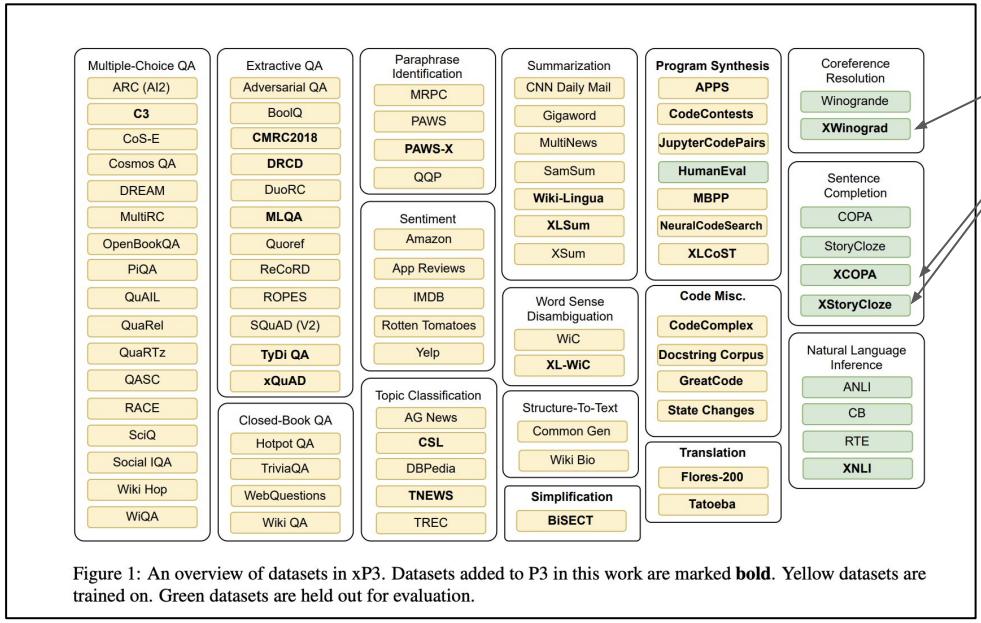


Figure 5: GPT-4 Eval. (Aya vs BX) using aya-human-annotated test set

Parting thoughts – evaluation for aya involved as much building new evaluations as the difficulty of evaluation.



Some amount of overlap in languages covered between benchmarks – French and Spanish overindexed.

Final piece: we open source multilingual dataset evaluation suite for open-ended generations

The screenshot shows a user interface for the Aya Evaluation Suite. At the top, there are three tabs: "aya_human_annotated" (7), "dolly-human-edited" (6), and "dolly_machine_translated" (114). The main content area displays a table comparing responses from "Human", "GPT-4", and "Aya" to a prompt about buying a new motorcycle.

Prompt: <i>What criteria should I be looking at when buying a new motorcycle?</i>				Human	GPT-4
eng	mT0	a motorcycle that meets your needs Aya There are many factors to consider when buying a new motorcycle. These include: * brand * model year * mileage * price * warranty * reliability		x	x
fra	mT0	Je dois prendre en compte le budget que je peux consacrer à l'achat de ma moto. Aya Voici quelques critères à prendre en compte lors de l'achat d'une nouvelle moto : Le budget - le type de moto (sport, custom, cruiser, etc.) - le type de suspensions - la qualité du moteur et de la boîte de vitesses - le look et l'aspect général - les options de sécurité (clignotants, suspensions, étriers, pièces détachées)		x	x
spa	mT0	posibles tamaño y estilo, potencia, potencia de corte, peso y peso de transmisión Aya Debes considerar las siguientes cosas al comprar una nueva motocicleta: 1. El trabajo. El tipo de motocicleta que quieras usar (por ejemplo, carreras, turismo, etc.) 2. El tiempo. Precio 3. ¿Qué hacer? Accesorios que quieras incluir 4. y 4. Tamaño de la motocicleta ¿Qué es esto? Durabilidad		x	x

Parting thoughts



Aya at a Glance

1 

Model

513M 

Re-annotations
of Datasets

3K 

Independent
Researchers

56 

Language
Ambassadors

119 

Countries

204K 

Original Human
Annotations

101 

Languages

31K 

Discord
Messages

If you talk to a man in a language he understands, that goes to his head. If you talk to him in his own language, that goes to his heart.

Nelson Mandela

Let's open up for questions.

Questions?

Aya model

Intriguing Properties of Quantization at Scale Arash

Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Stephen Gou, Phil Blunsom, Ahmet Üstün, Sara Hooker [\[\[paper link\]\]](#)

When Less is More: Investigating Data Pruning for

Pretraining LLMs at Scale Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, Sara Hooker [\[\[paper link\]\]](#)

The state of sparsity in deep neural networks Trevor

Gale, Erich Elsen, Sara Hooker [\[\[paper link\]\]](#)

Feel free to reach out if any of these ideas is relevant to work you are doing..

Final takeaways:

Recent breakthroughs in NLP - combination of changes in optimization, scale (of both data and weights)

Key challenge - efficiency of our chosen representation. The relationship between weights and generalization is not well understood.

Promising directions of improving efficiency – includes both algorithmic, hardware-software and data space.

Tension between theoretical and practical motivations – some cherished theoretical techniques do not produce speed ups.

Email: sarahooker@cohere.com