AAAI Tutorial 2024

Aligning LLMs to Low-Resource Languages

Ahmet Üstün
Marzieh Fadaee

Nazar Beknazarov

Cohere For AI

Toloka

Welcome!

Cohere For AI - *Fundamental research that explores the unknown*

Ahmet Üstün
Marzieh Fadaee

Toloka AI - *Data & Model development company*

Nazar Beknazarov

Today.

Why should we work on multilingual LLM models?

Why LLMs struggle with non-English, specially low-resource languages?

What is instruction fine-tuning?

How we can gather a multilingual dataset for Low-Resource Languages?

How can we make models multilingual?

Why have some languages been left behind in technological progress?

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

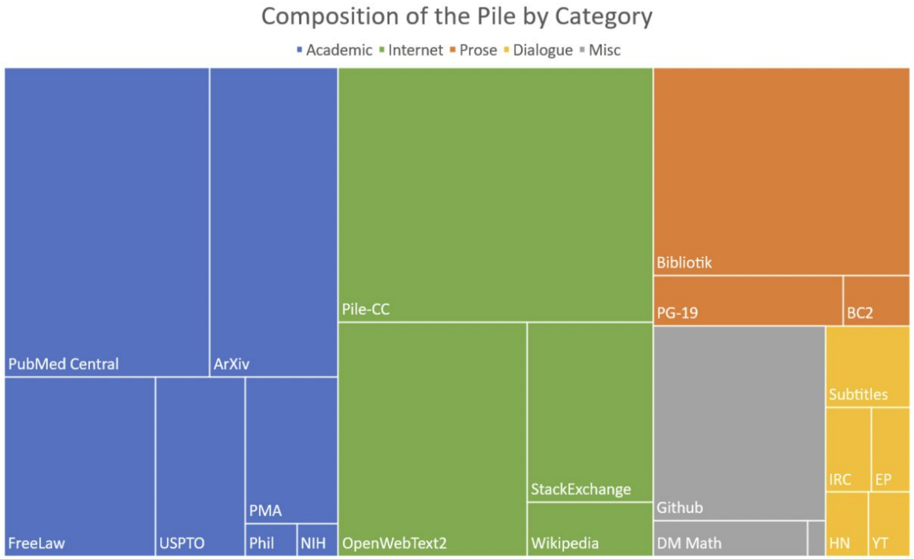


Figure 1: Treemap of Pile components by effective size.

Cohere For AI

Toloka

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

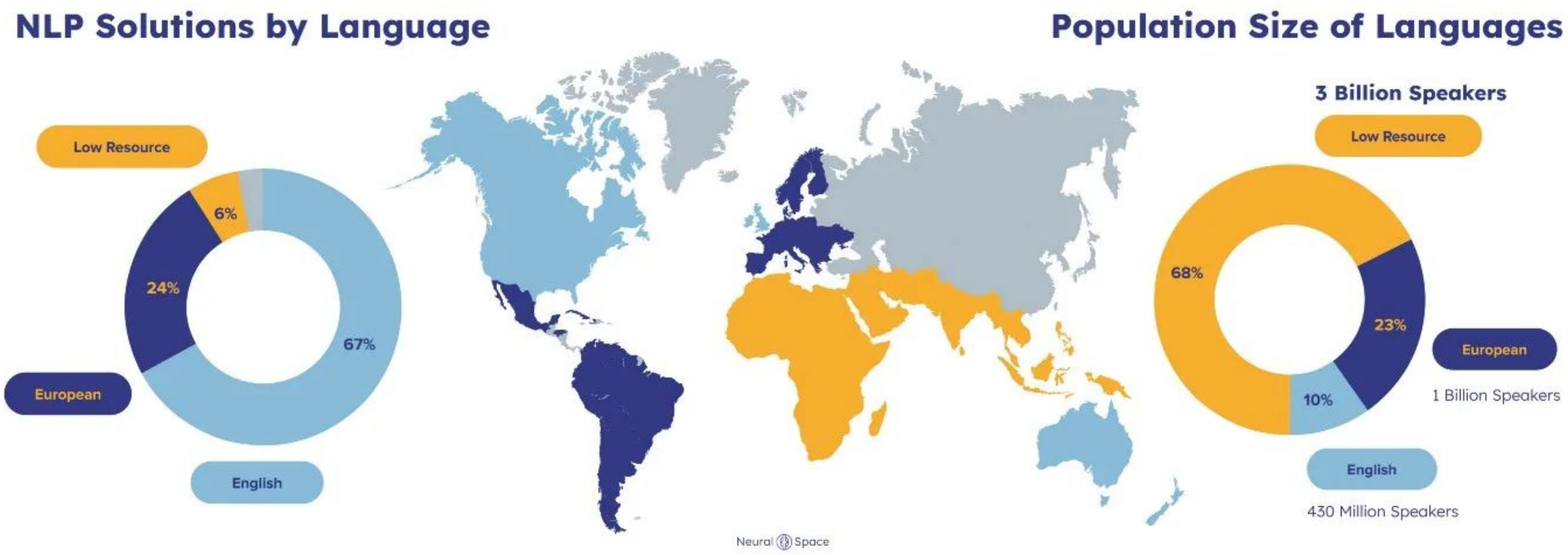
Number of papers in top NLP venues referencing language per 1 million speakers. [Van Etch et al. 2022]

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

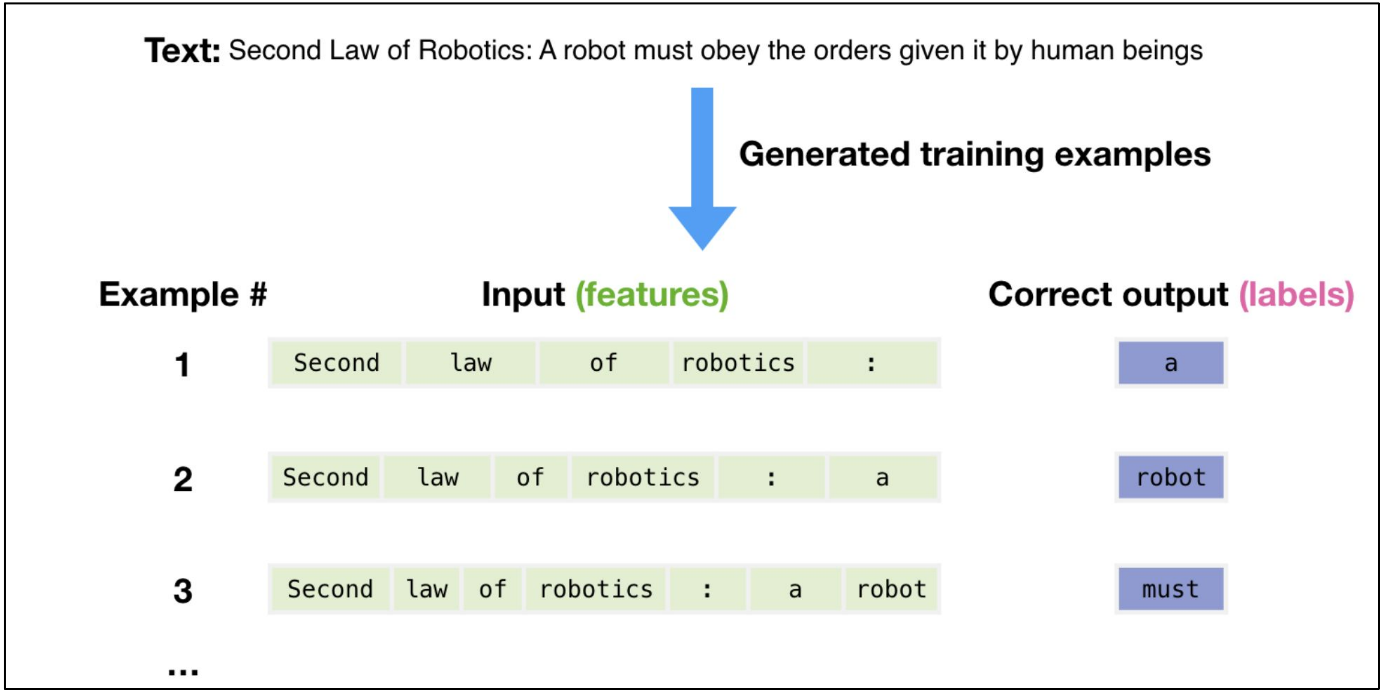
Open source multilingual state of art Large Language Models (LLM) are pre-trained a smaller subset of available languages.

Significant data inequality



Pretraining on larger and larger datasets in an unsupervised fashion.

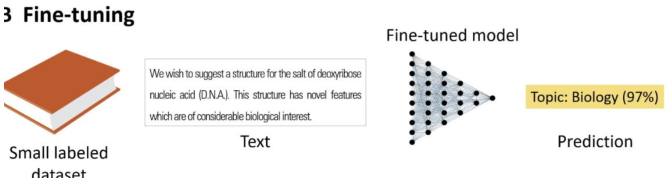
Step 1: Unsupervised pre-training of a transformer model on a massive web crawled dataset (i.e. train on the internet).



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

Cohere For AI

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



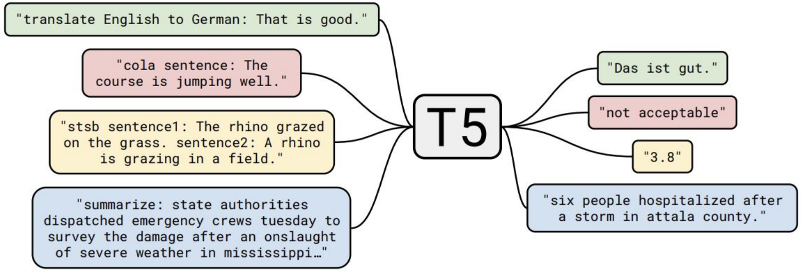


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across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

Finetuning on a single task → Finetuning on many different tasks

Cohere For AI

Toloka

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

Model for sentiment analysis

Model for topic categorization

Model for toxicity mitigation

Single model trained on multi-task finetuning

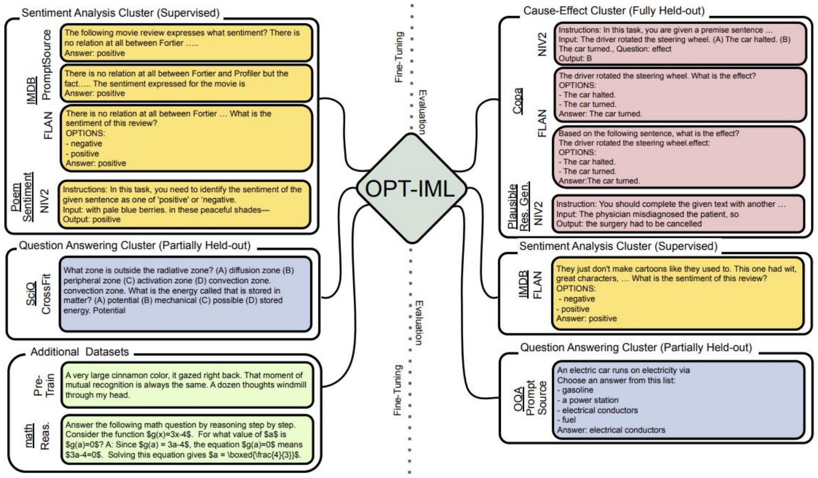
It turns out two ingredients have been particularly important at leading to breakthroughs in performance on zero and few shot tasks:

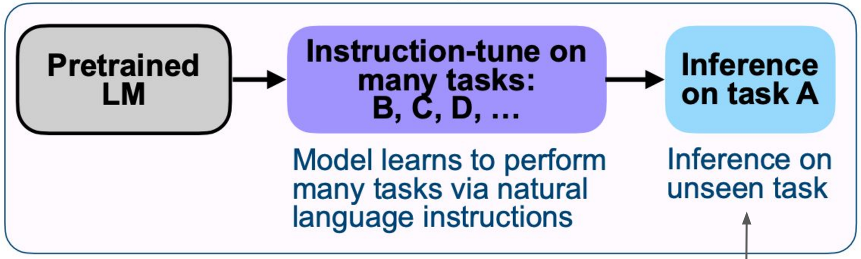
1. **Instruction tuning – Structuring multi-task fine-tuning data as questions and answers**

2. **Integrating human feedback about preferences**

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

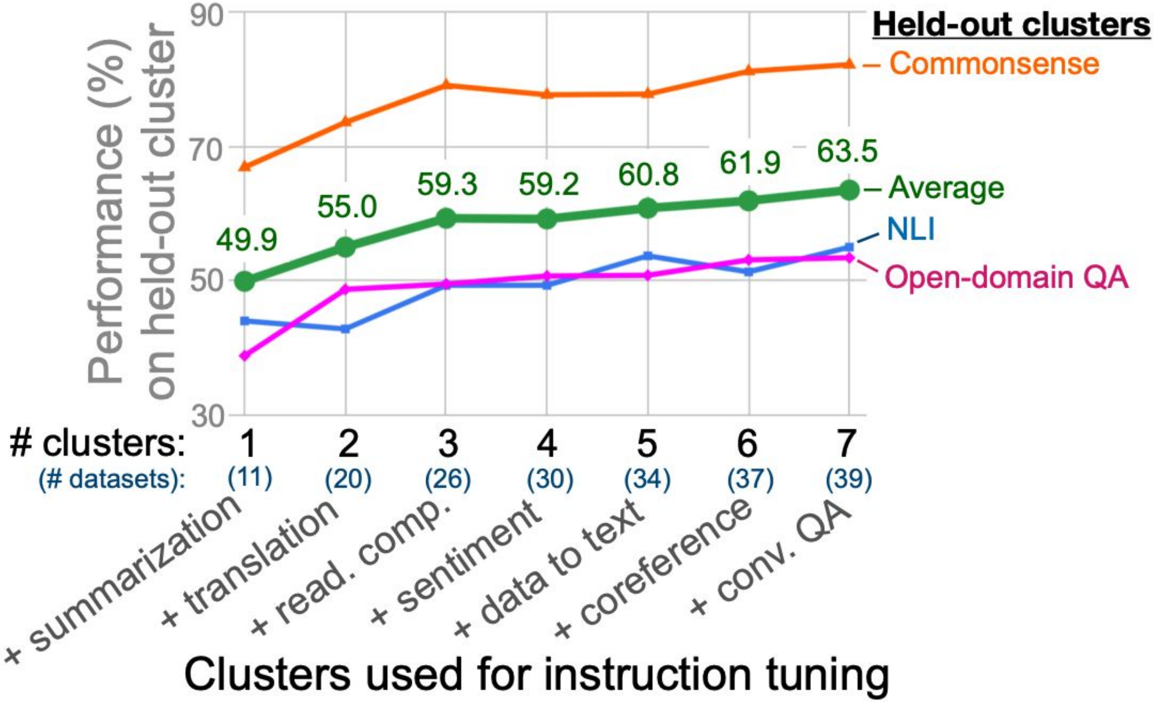
Leverage supervision to teach the model to perform many NLP tasks.





i.e., zero-shot!

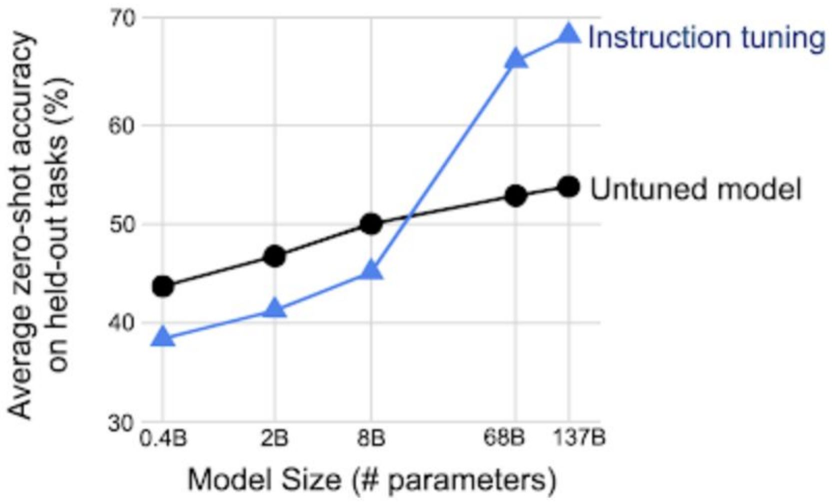
This combination – of multitask training and instruction style improves zero shot performance.



Cohere For AI

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

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



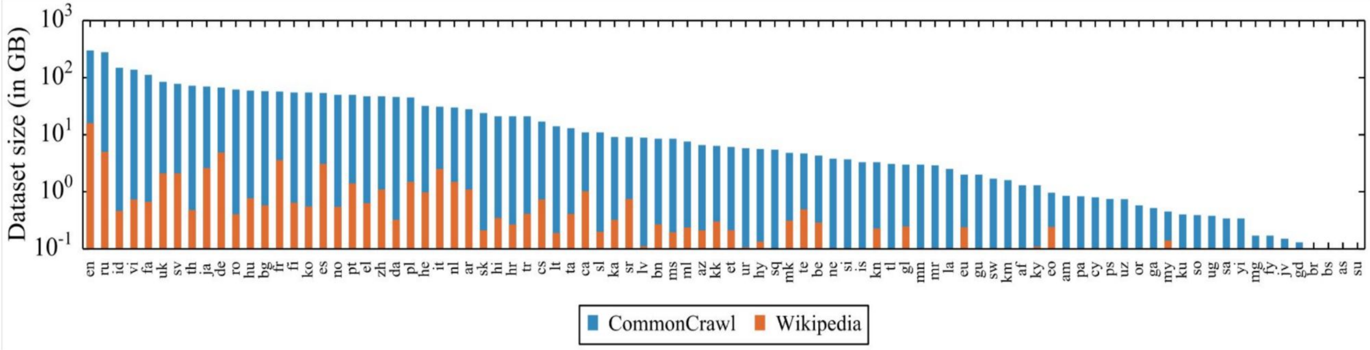
Instruction tuning only improves performance on unseen tasks for models of certain size.

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.

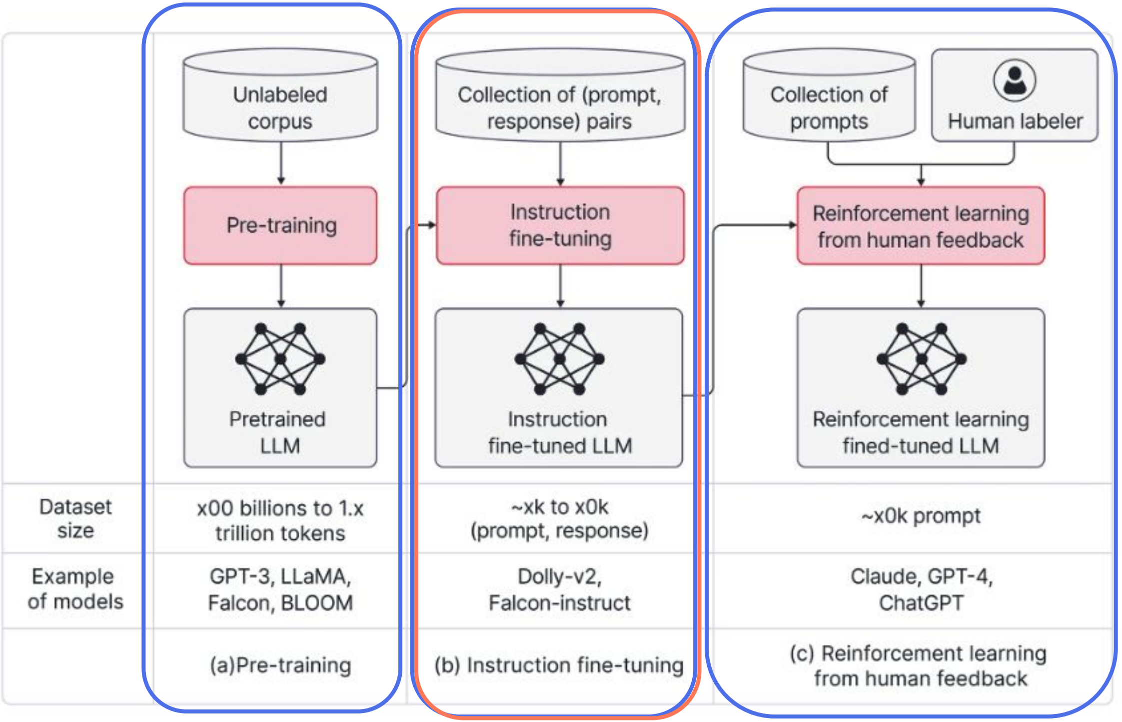
This makes instruction finetuning a particular promising research direction for multilingual, where there is a pronounced skew in the "haves" and "have nots."



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

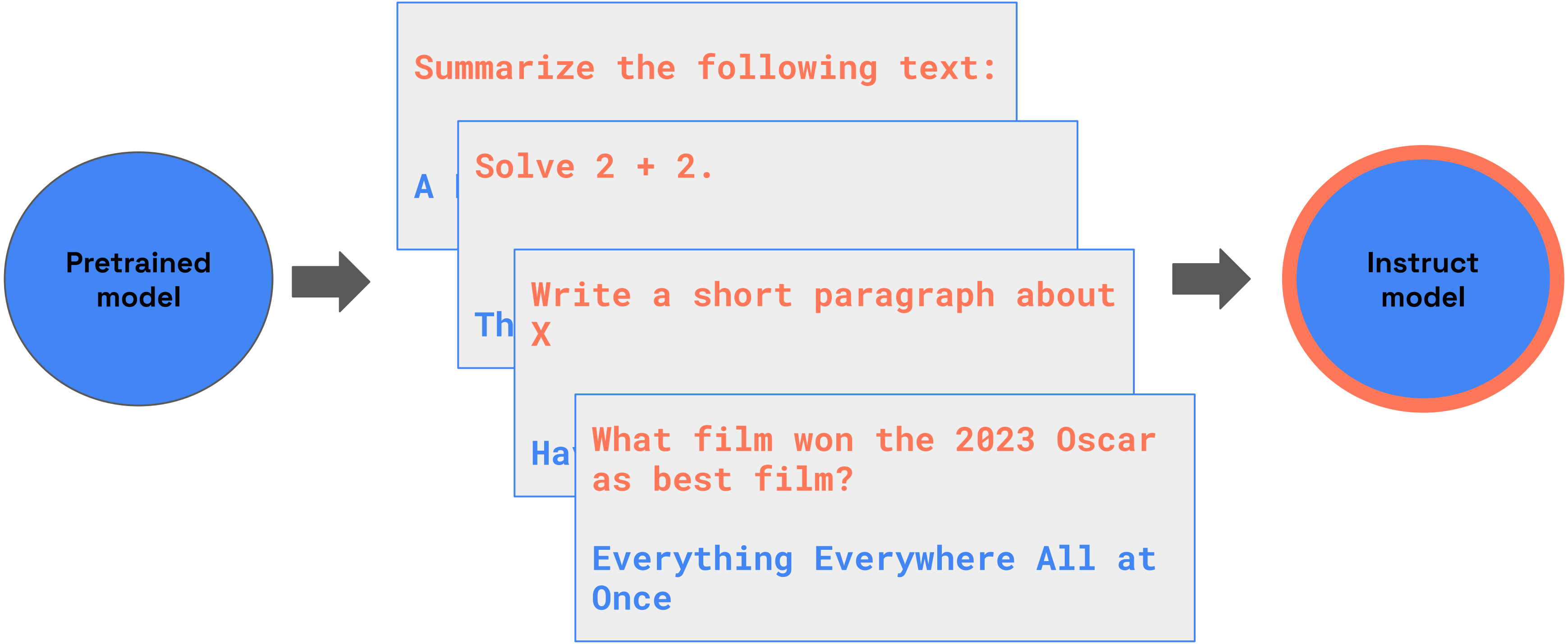
ACL [Keynote](), [Conneau et al.]()

LLM recipe



https://www.borealisai.com/research-blogs/a-high-level-overview-of-large-language-models/

What does the data actually look like?



Most relevant is work released last year by Muennighoff et al.

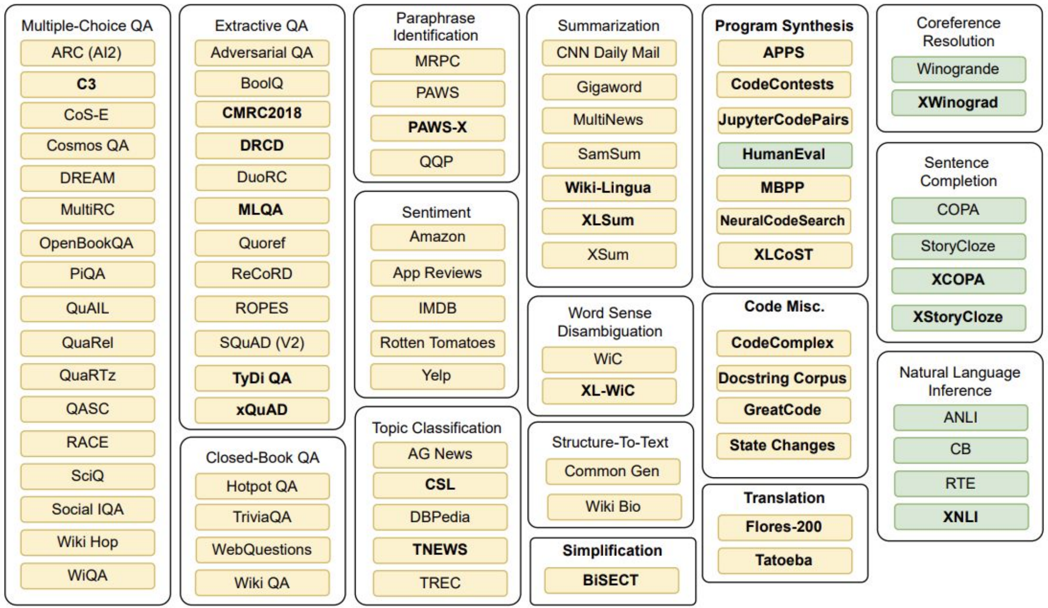


Figure 1: An overview of datasets in xP3. Datasets added to P3 in this work are marked **bold**. Yellow datasets are trained on. Green datasets are held out for evaluation.

40% of instruction data is still in English

Muennighoff et al.

Cohere For AI

Toloka

Observed boosts in performance over base multilingual models.

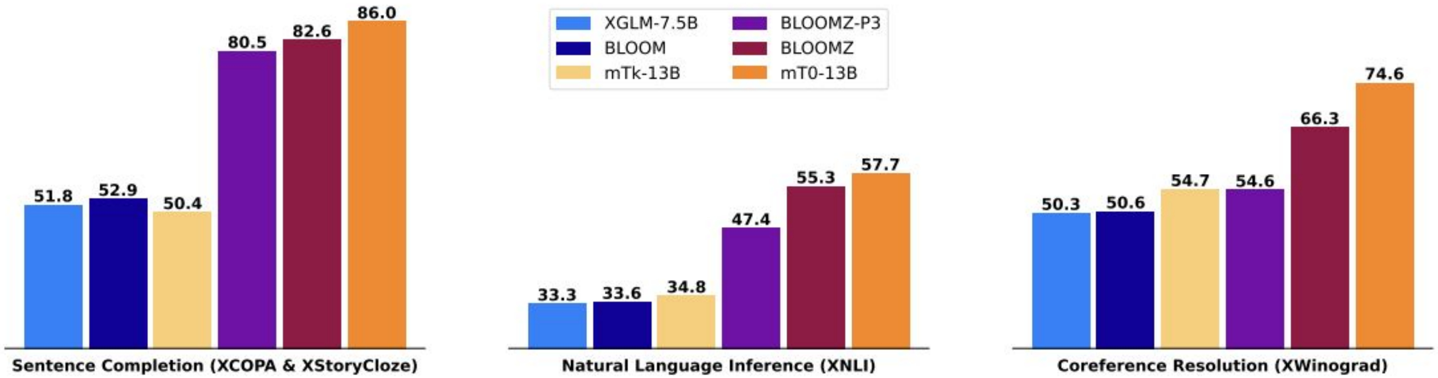


Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.

Muennighoff et al.

Cohere For AI

Toloka

A years work has built Aya

1 Model

513M Total release dataset size

3K Independent researchers

56 Language ambassadors

119 Countries

204K Original human annotations

101 Languages

31K Discord messages

[Accelerating Multilingual AI through open science]

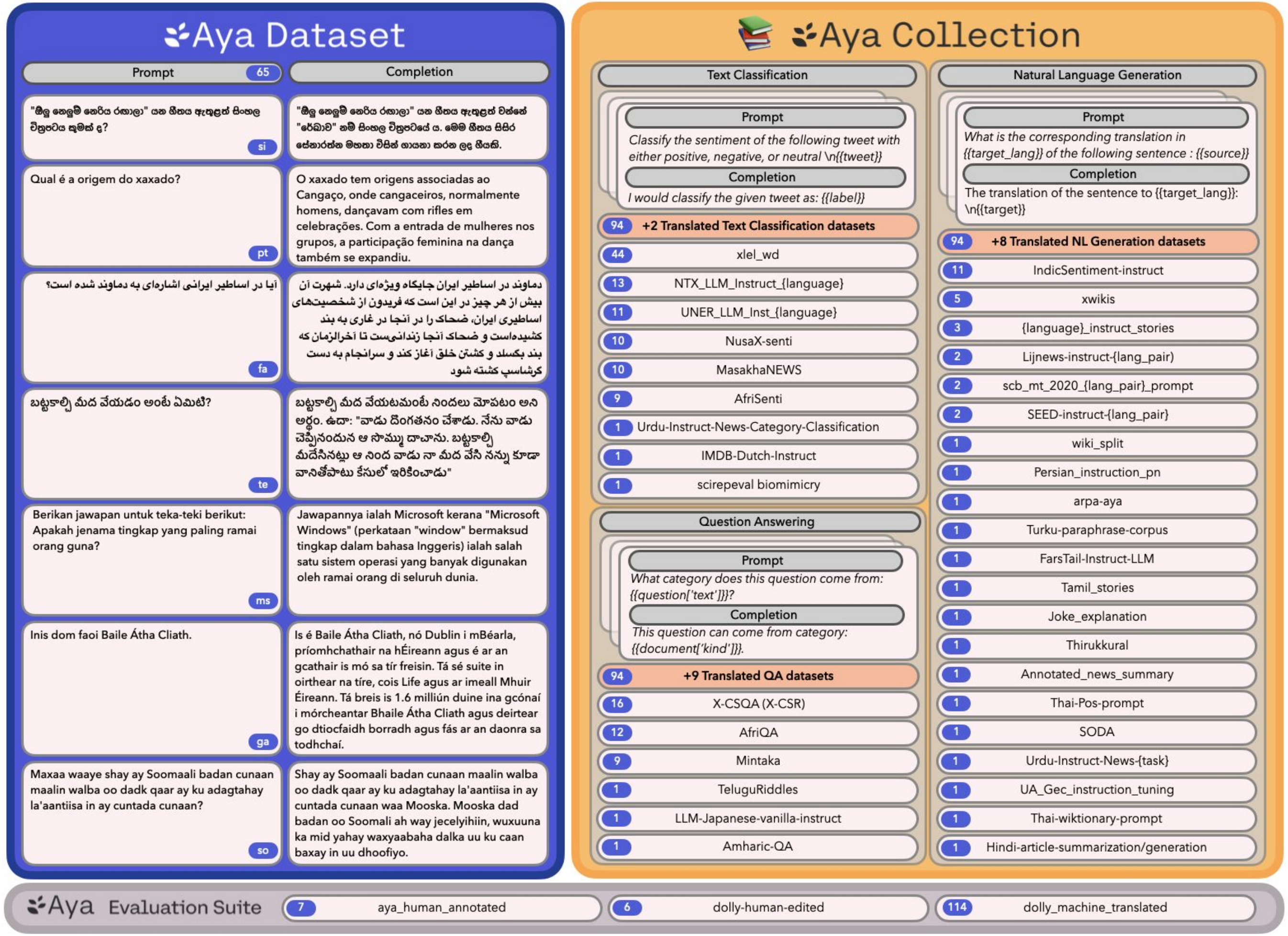
cohere.com/research/aya

Aya Data

Cohere For AI

Aya Dataset and Collection –

We build a huge collection of multilingual instruction data with human annotations, manuel curations, quality check, and translation.



Cohere For AI

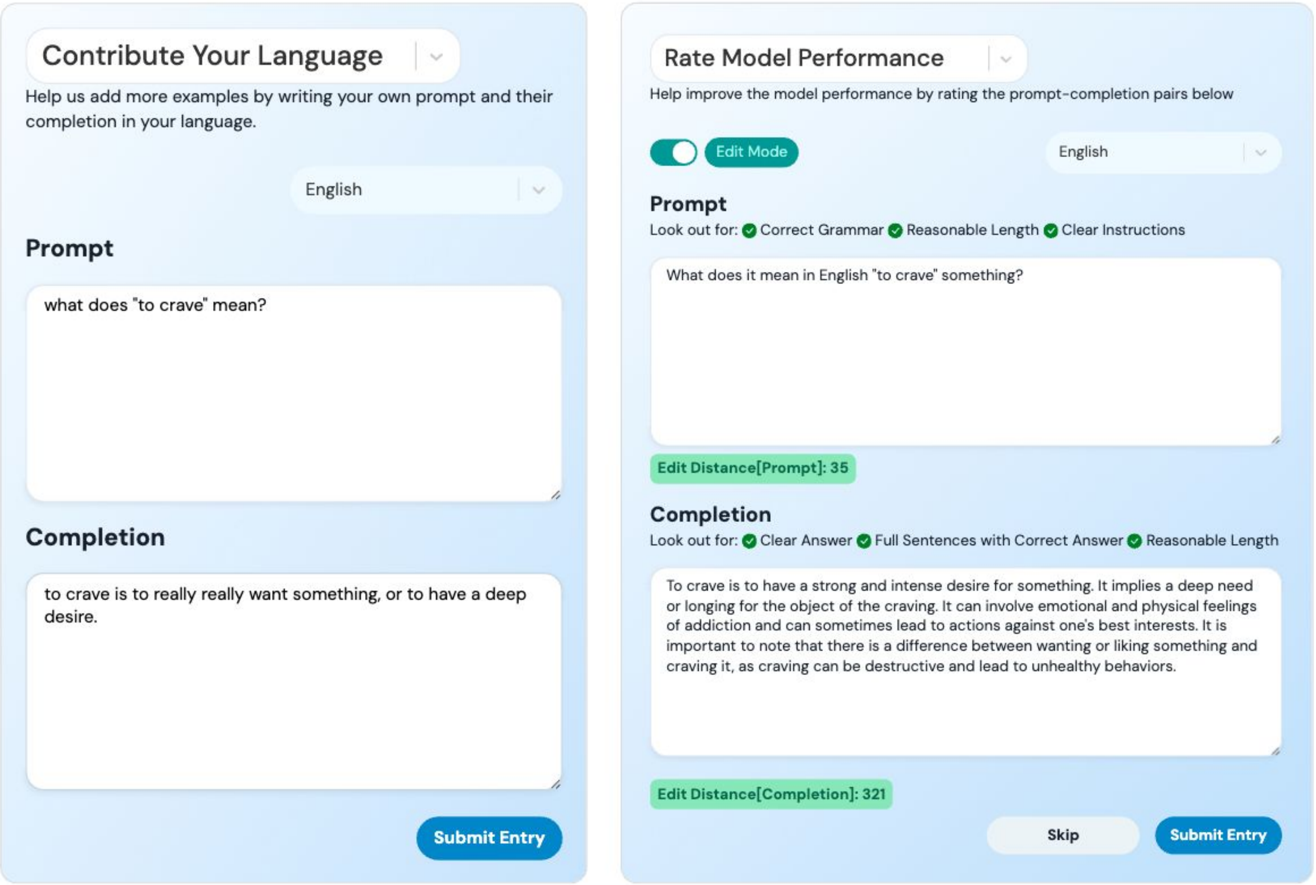
Aya Dataset and Collection is the largest instruction mixture with permissive licence

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.

We started with human annotations

+3K community members contributed the dataset in 2 ways.



a) Example of an original annotation contribution. (b) Example of a re-annotation contribution.

+200k instruction from original annotations and re-edits of existing datasets.

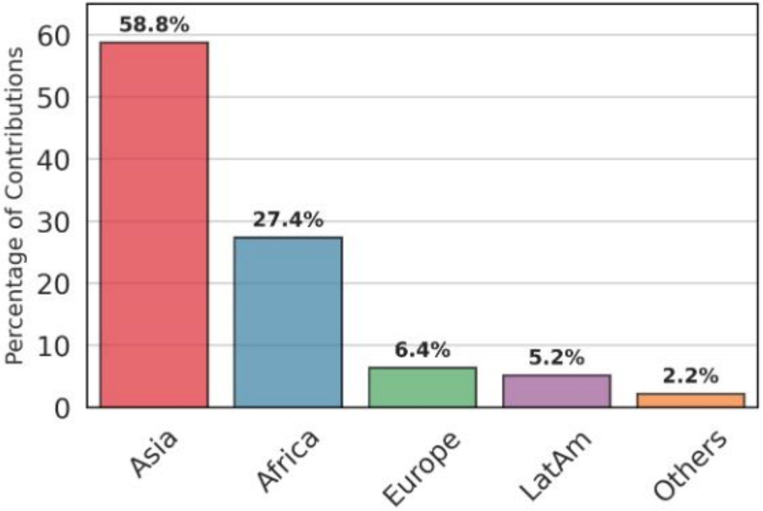


Figure 4: Distribution of total contributions across different regions.

		Count
Original Annotations		138,844
Re-Annotations	xP3 datasets	2859
	Translated datasets	7757
	Templated datasets	11013
	Original Annotations	43641
Aya Dataset Total		**204,114**

Longer instructions and higher quality data in Aya

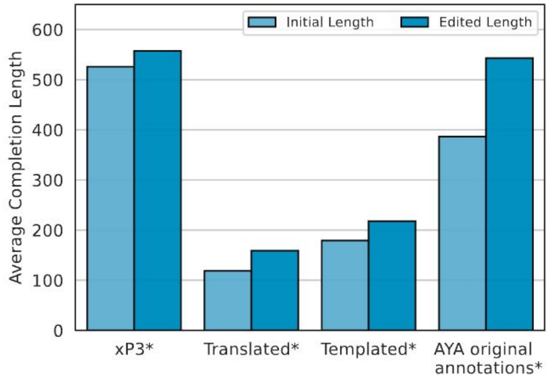


Figure 8: Average Completion Length before and after re-annotation.

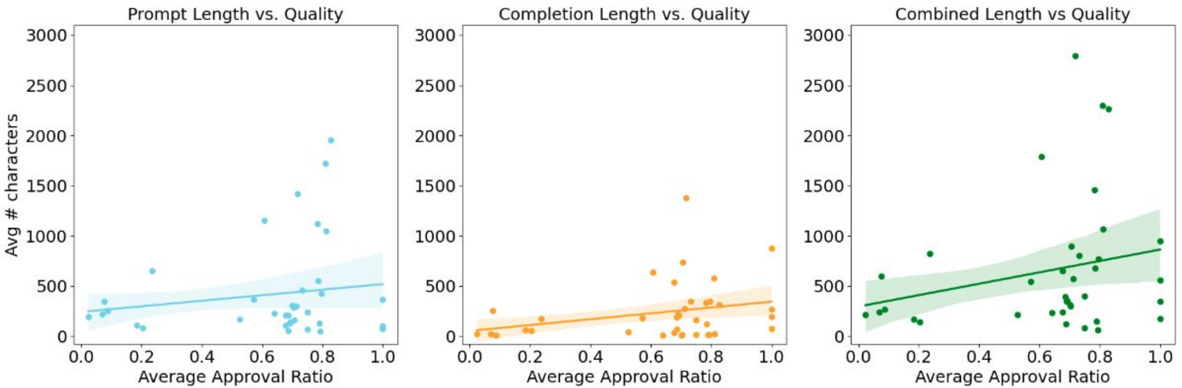
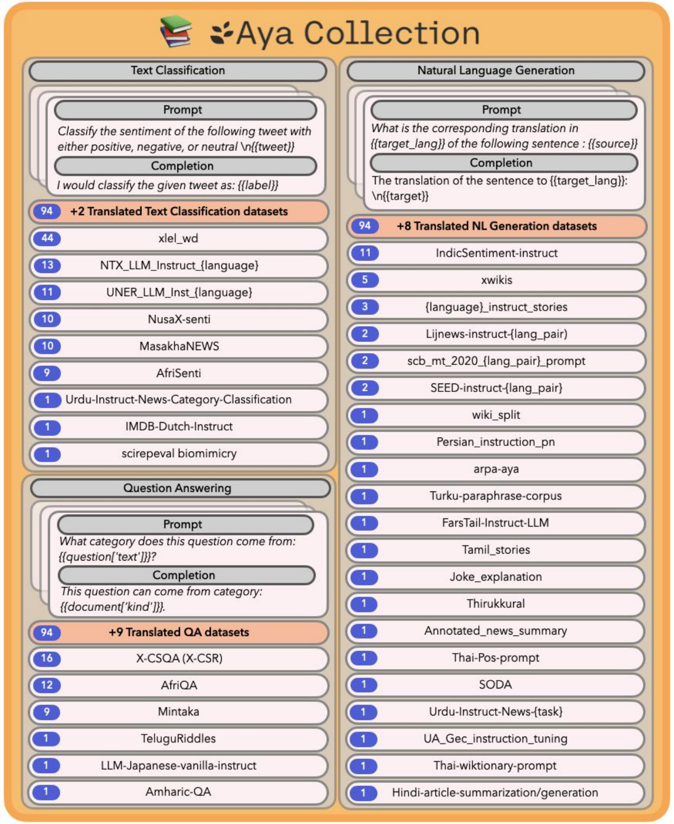


Figure 9: Relationship between Average Prompt and Completion Length in characters and the Average Approval Rate of the example.

We also collect a diverse set of datasets that are converted into instruction format, and translated into 101 languages



Main Task Type	Fine-grained Task Type
Question Answering	—
Natural Language Generation	Summarization Translation Paraphrasing Dialogue Text Simplification
Text Classification	Sentiment Analysis Information Extraction Named Entity Recognition Event Linking Natural Language Inference Document Representation

Table 4: Task Taxonomy of NLP tasks in the **Aya** Collection.

We aimed a high representation for low-resource languages

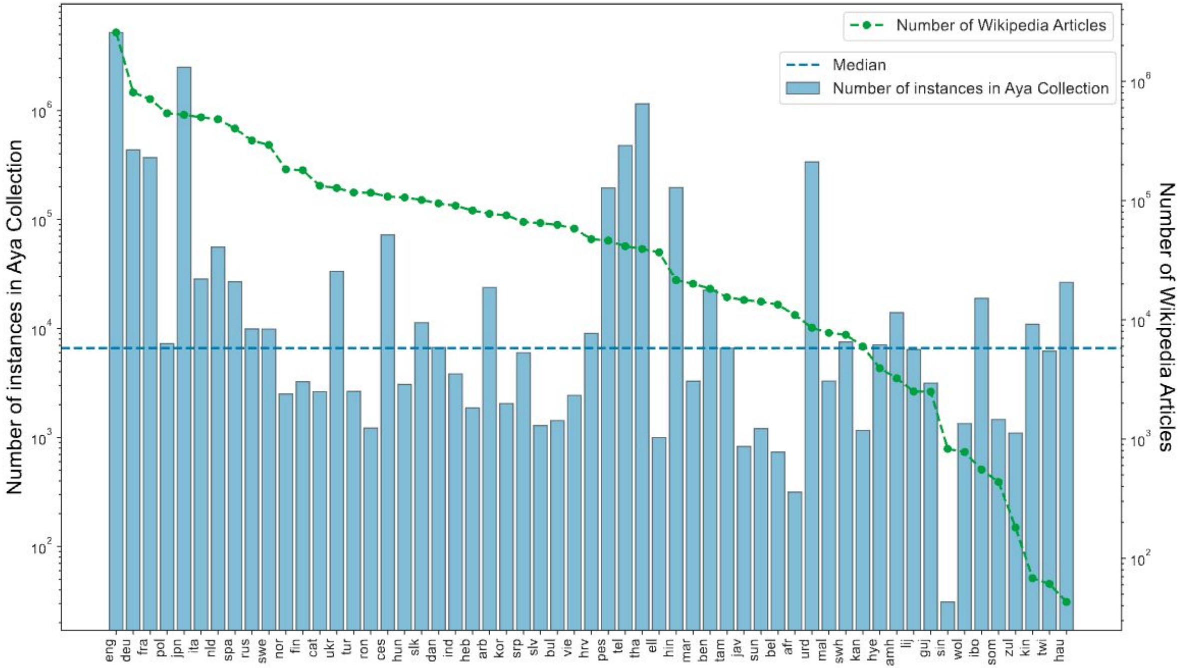


Figure 14: Number of prompt/completion pairs in each language in the **Aya** Collection (templated). Many languages with limited digital presence, as indicated by a low number of Wikipedia pages, are well-represented in the templated portion of the **Aya** Collection. Note that absolute Both axes are in log-scale.

Cohere

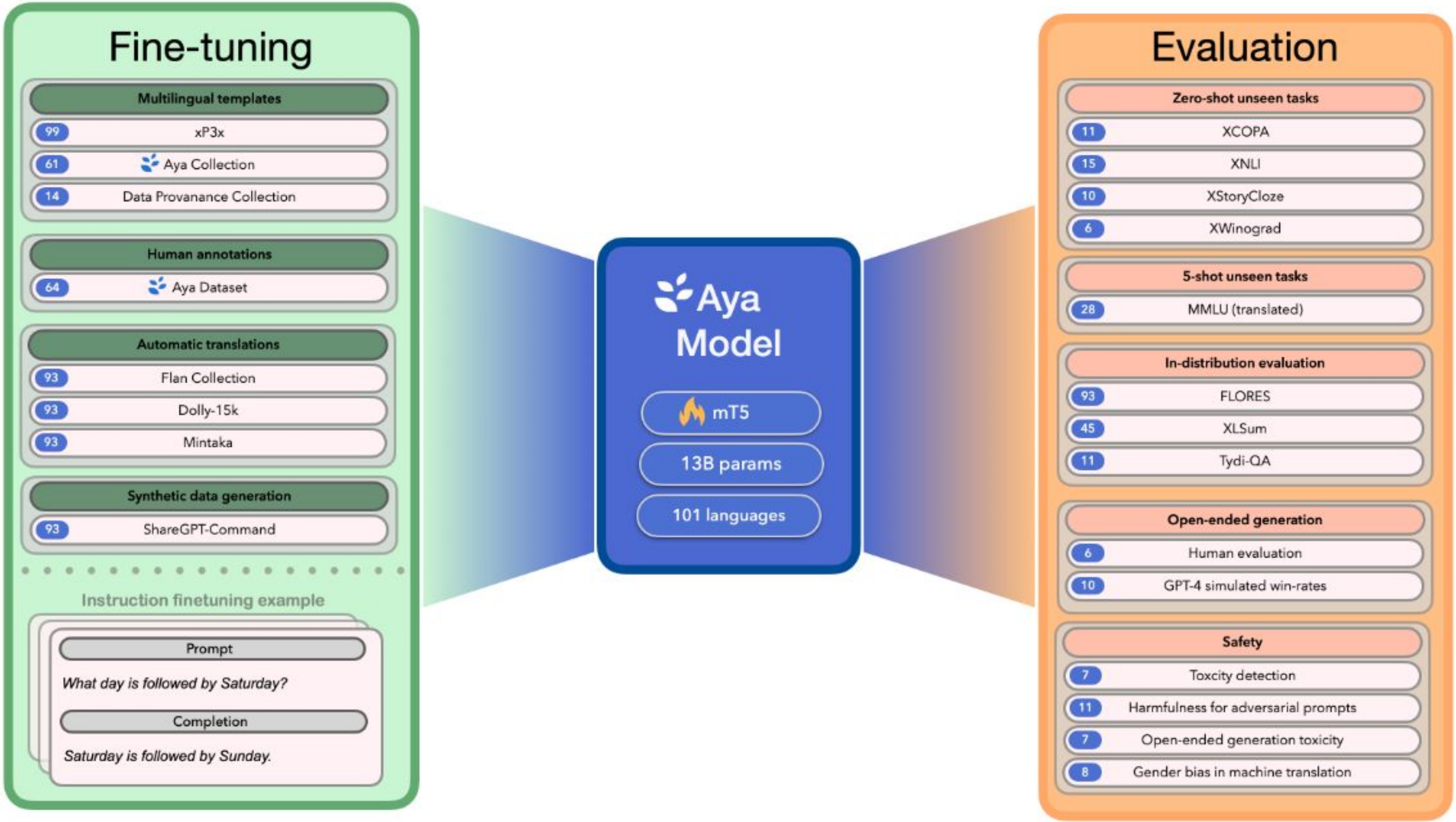
Final piece: a multilingual dataset evaluation suite for open-ended generations

Aya Evaluation Suite | 7 aya_human_annotated | 6 dolly-human-edited | 114 dolly_machine_translated

Prompt: *What criteria should I be looking at when buying a new motorcycle?*			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	✗	✗
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)	✗	✗
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 quieres 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		✗

Aya Model

Cohere For AI



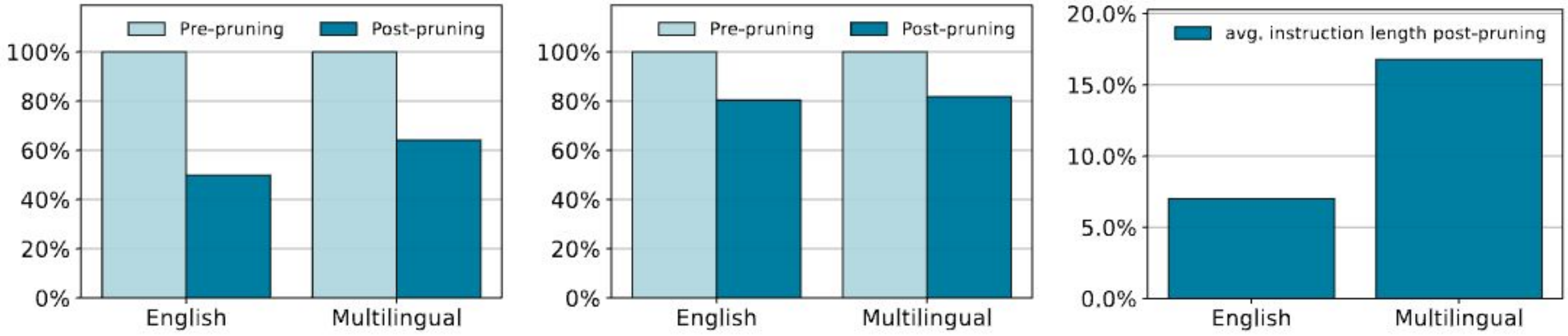
Fine-tuning
Multilingual templates
99 xP3x
61 Aya Collection
14 Data Provanance Collection
Human annotations
64 Aya Dataset
Automatic translations
93 Flan Collection
93 Dolly-15k
93 Mintaka
Synthetic data generation
93 ShareGPT-Command
Instruction finetuning example
Prompt
What day is followed by Saturday?
Completion
Saturday is followed by Sunday.
Aya Model
mT5
13B params
101 languages
Evaluation
Zero-shot unseen tasks
11 XCOPA
15 XNLI
10 XStoryCloze
6 XWinograd
5-shot unseen tasks
28 MMLU (translated)
In-distribution evaluation
93 FLORES
45 XLSum
11 Tydi-QA
Open-ended generation
6 Human evaluation
10 GPT-4 simulated win-rates
Safety
7 Toxcity detection
11 Harmfulness for adversarial prompts
7 Open-ended generation toxicity
8 Gender bias in machine translation

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

Name	Characteristics					Lang Ratio (%)		
	Langs	Datasets	Size	Avg Input Len	Avg Target Len	HR	MR	LR
xP3x Dataset	101	56	168M	1048	780	68.2	18.2	13.6
Data Provenance Collection (Commercial)	14	161	1.65M	998	78	97.5	0.5	2.0
Aya Collection (Templated Data Subset)	61	34	18.9M	1864	209	85.3	9.5	5.2
Aya Dataset	64	1	199.5K	178	501	29.1	14.7	56.2
Aya Collection (Translated Data Subset)	93	19	7.53M	496	219	27.3	21.7	50.9
ShareGPT-Command	93	1	6.8M	385	1080	27.3	21.7	50.9

Table 1: **A list of training data sources used for instruction finetuning Aya models.** Dataset characteristics include the number of languages, examples (size), sampling ratio and average input + target sequence length (in chars). We also describe language representation based on Higher- (HR), Mid-(MR), and Lower-Resourced (LR) languages, which we assign based on language scores as described in Joshi et al. [2020]. All characteristics described are for the final training mixture which includes both filtering, i.e. template pruning, and language filtering as well as subsampling in both Data Provenance and **Aya** Translated Data collections.

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



(a) Templates (b) Instances (c) Instruction Length

Figure 2: Pruning statistics across (2a) number of templates and (2b) instances for English-only and multilingual datasets. (2c) shows the average instruction length in characters per instance before and after pruning.

We experimented with weighting importance of each data sources
–

Important for balancing multilinguality with downstream performance

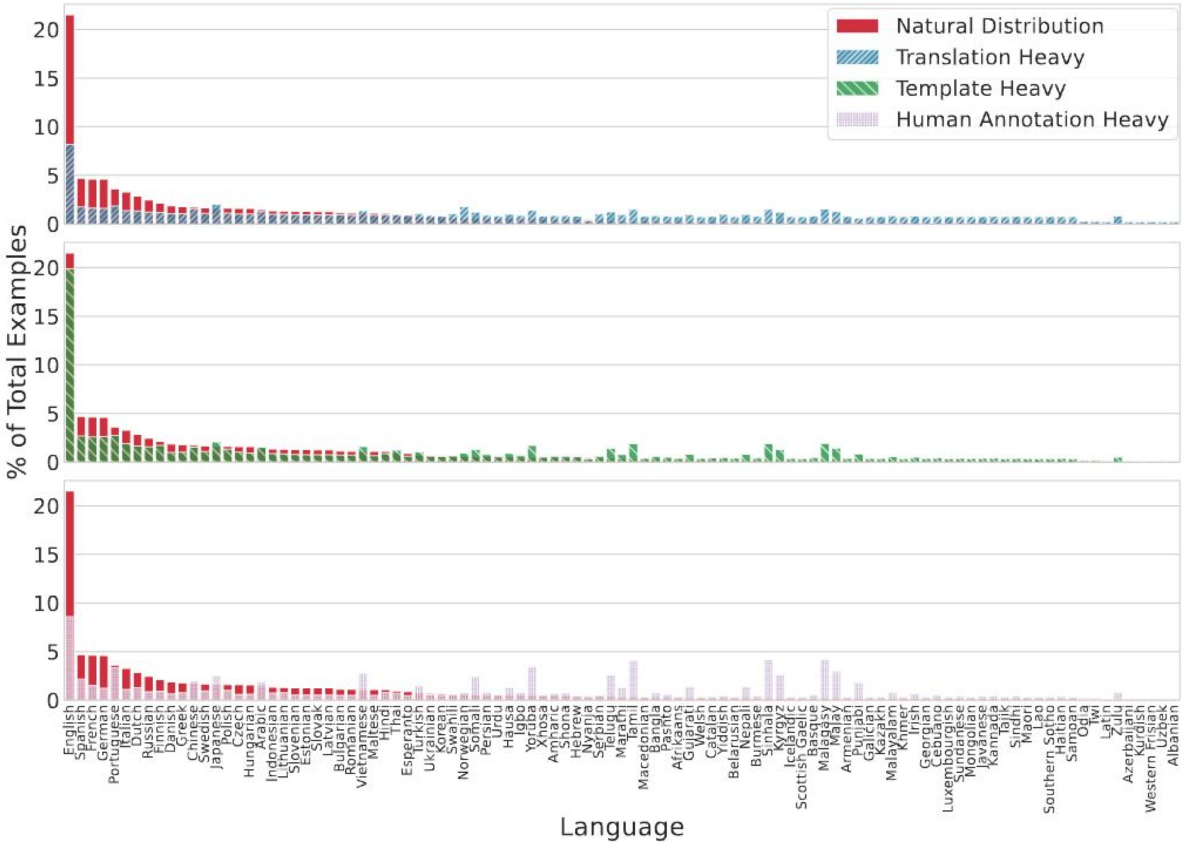


Figure 17: % of Examples for each language with different weighting schemes

Cohere For AI

Evaluation on unseen discriminative tasks (zero-shot)

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.

Evaluation on general purpose unseen dataset (5-shot)

	arb	cat	deu	eus	fra	hin	hrv	hun	ita	nld	por	rud	ser	spa	swe	vie
Okapi[‡]	27.7	30.5	31.7	27.9	30.7	26.5	30.0	30.1	30.4	31.1	30.1	30.6	30.4	30.9	29.3	27.5
mT0	31.5	32.8	32.7	29.7	32.1	32.0	31.1	32.3	32.4	32.0	32.1	32.8	30.9	32.1	31.6	30.9
mT0x	31.6	32.6	32.5	29.2	32.7	31.6	31.1	31.7	31.3	32.1	32.0	31.7	31.4	32.2	32.8	31.1
Aya	38.2	39.6	39.7	36.0	39.7	38.7	37.5	38.8	39.0	40.1	39.0	39.2	38.1	39.7	39.7	34.8

	zho	ben	dan	ind	ron	slk	tam	ukr	guj	hye	kan	mal	mar	npi	tel	**Avg**
Okapi[‡]	28.2	26.8	31.8	27.5	30.9	30.2	26.0	31.6	27.4	27.5	26.8	25.8	26.1	25.2	25.9	28.8
mT0	32.5	31.6	33.0	33.3	32.4	32.3	29.4	31.5	29.5	28.4	30.9	28.6	31.6	32.4	29.0	31.5
mT0x	31.6	30.2	32.0	32.3	31.8	31.4	27.7	32.3	28.5	26.7	28.9	26.7	29.7	30.1	27.9	30.8
Aya	38.3	35.8	39.7	40.0	39.5	39.4	31.2	39.9	33.6	30.0	34.5	30.4	36.0	37.2	32.1	**37.3**

Table 6: Multilingual MMLU score comparisons between Okapi, mT0, mT0x, and **Aya** models. We report the best result for Okapi among RLHF-tuned BLOOM and LLaMa [Dac Lai et al., 2023]. '‡' Note that Okapi reports 25-shot results, however, mT0, mT0x and **Aya** (`translation-heavy`) models are evaluated using 5-shot. Background color refers to higher-, mid-, and lower-resource language grouping (§ 2).

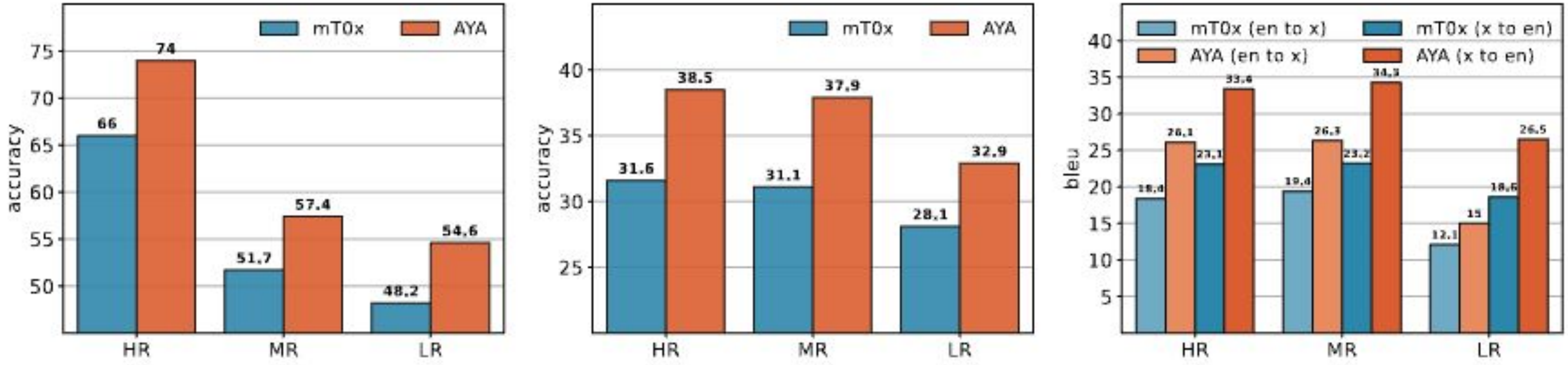
Evaluation on in-distribution generative tasks

Model	IFT Mixture	Generative Tasks			
		FLORES-200 (spBleu)		XLSum (RougeLsum)	Tydi-QA (F1)
101 Languages		X→ En	En → X		
mT0x	xP3x	20.2	14.5	21.6	76.1
Aya (`human-anno-heavy`)	All Mixture	25.1	18.9	22.2	77.9
Aya (`templated-heavy`)	All Mixture	25.0	18.6	**23.2**	**78.8**
★**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.

Cohere For AI

Larger improvement in MR and LR languages

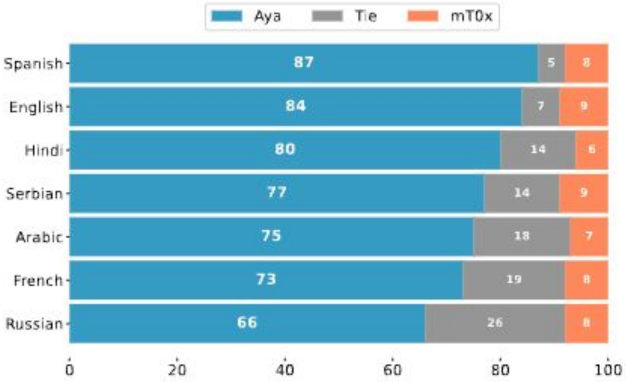


(a) Unseen Discriminative Tasks (b) Multilingual MMLU (c) Generative Task: FLORES

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

How do models compares in open-ended generations?

(c) Human Eval. (**Aya** vs **mT0**)



(d) Human Eval. (**Aya** vs **mT0x**)

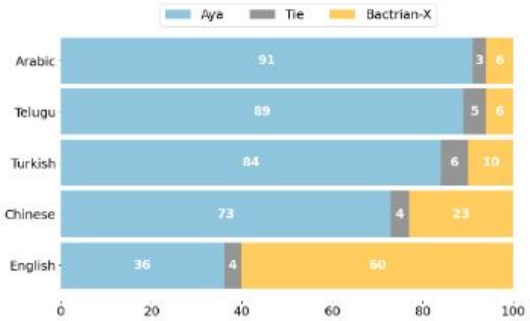


Figure 5: GPT-4 Eval. (**Aya** vs BX) using `aya-human-annotated` test set

Cohere For AI

We observe a tension between discriminative tasks and open-ended generations

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

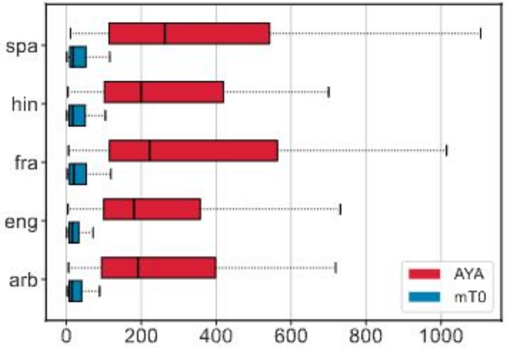
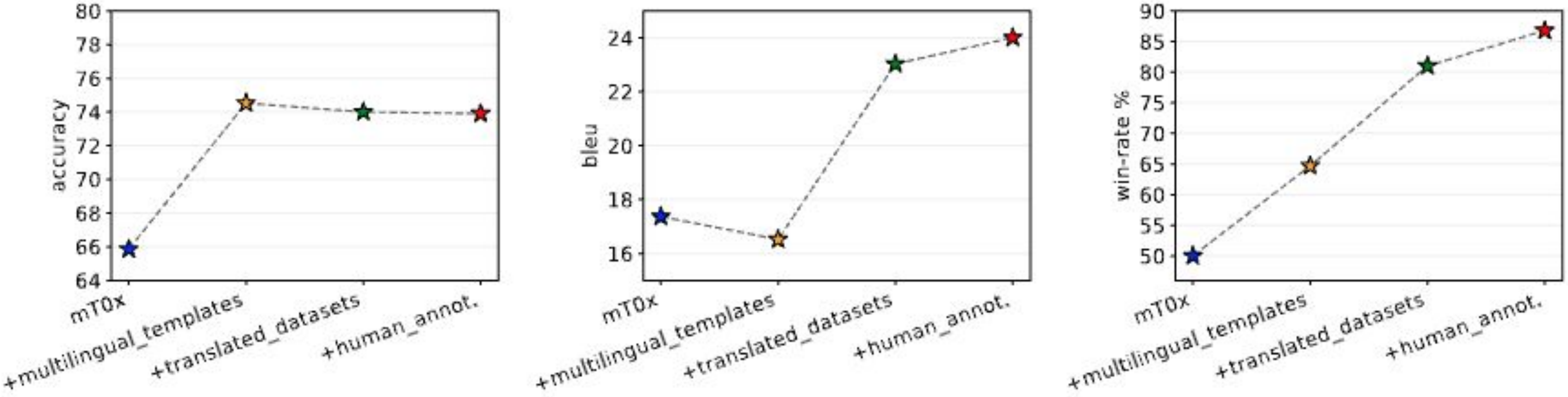


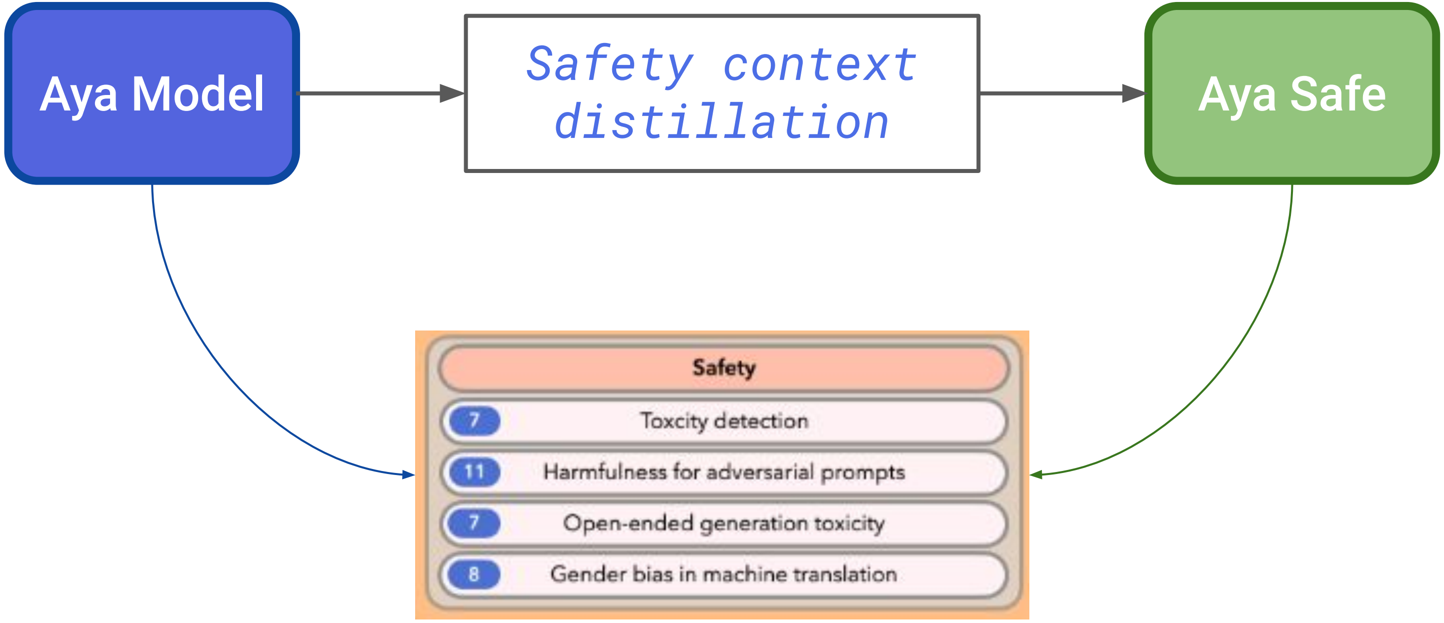
Figure 6: Completion lengths by characters for the **Aya** and mT0 models in Dolly test set for various languages.

Each data source contributes to the downstream performance



(a) Unseen Discriminative Tasks (b) Generative Task: Flores (c) Win Rates (vs mT0x)

How can we make the model “safer” and “more responsible”?



Safety context distillation significantly helps model to be safer

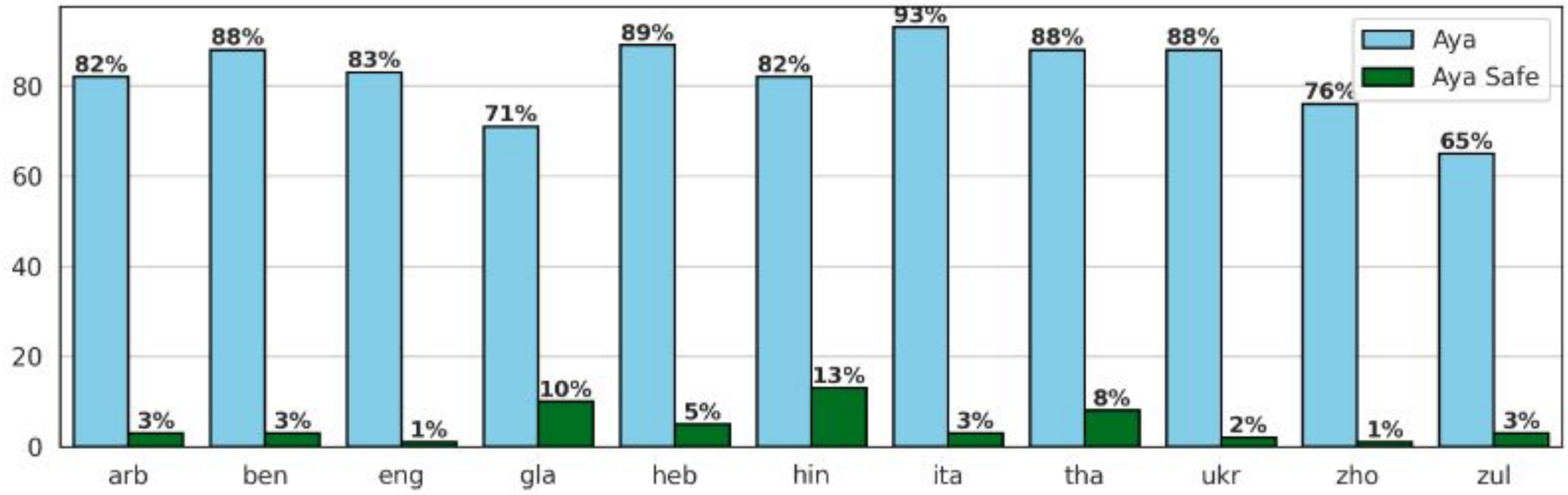
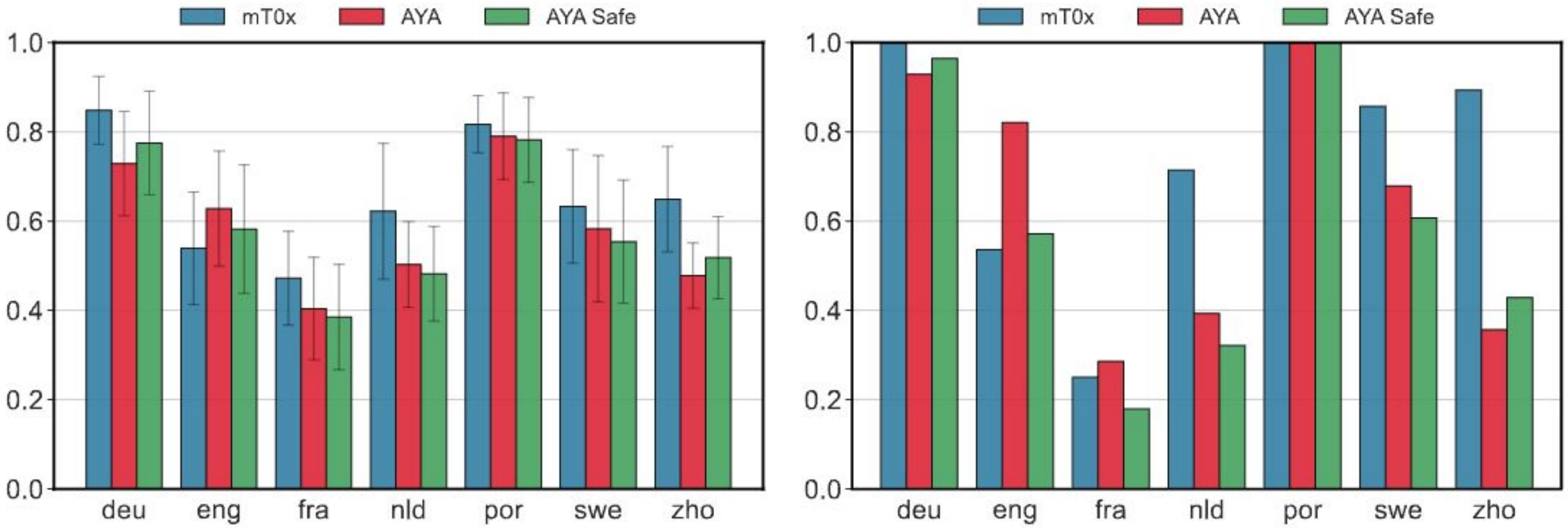


Figure 11: GPT-4 evaluation: Ratio of harmful generations for AdvBench held-out prompts.

How about “toxicity” in model generations?



(a) Expected maximum toxicity

(b) Toxicity probability

Figure 14: Toxicity analysis of model generations when prompted with sentences for identity groups such as gender, ethnicity, and religion.

How about “bias” in model generations?

	Model	spa	fra	ita	rus	ukr	heb	ara	deu	Average
$\downarrow \Delta S$	mT0x	**17.3**	20.4	**23.8**	10.8	**8.1**	32.9	21.2	**20.6**	**19.4**
	Aya	25.2	**20.1**	26.4	13.3	11.5	36.0	18.1	27.7	22.3
	Aya Safe	25.5	**20.1**	24.8	**9.4**	9.5	**29.5**	**17.9**	24.5	20.2
$\downarrow \Delta G$	mT0x	29.0	27.1	27.8	30.7	**28.0**	**8.6**	**12.9**	28.8	24.1
	Aya	15.0	19.7	16.7	**24.4**	33.0	12.8	22.0	18.1	20.2
	Aya Safe	**9.4**	**14.8**	**10.1**	27.8	31.0	10.4	20.9	**11.9**	**17.0**

Table 10: $\downarrow \Delta S$ and $\downarrow \Delta G$ of gender bias evaluation as the sentences are translated from `English` to different languages (`Spanish`, `French`, `Italian`, `Russian`, `Ukrainian`, `Hebrew`, `Arabic` and `German`). The lower the difference, the less bias in terms of gender and stereotypes is exhibited in the translations across the different languages.

Safety context distillation is still very preliminary and limited…

Model	IFT Mixture	Generative Tasks: Flores (spBleu) X→ En	Flores (spBleu) En → X	XLSum (RougeLsum)	Tydiqa (F1)	Held out tasks (Accuracy %): XCOPA	XNLI	XSC	XWNG
101 Languages									
mT0x	xP3x	20.2	14.5	21.6	76.1	71.7	45.9	85.1	60.6
Aya	All Mixture	**29.1**	**19.0**	**22.0**	**77.8**	**76.8**	**58.3**	**90.0**	**70.7**
Aya Safe	+ Safety Mitigation	28.9	17.6	20.9	76.0	74.8	56.9	86.8	67.5

Table 8: **Aya Safe** model performance compared to mT0x and **Aya** on the evaluation suite consisting of generative and held out tasks (§4).

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**

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

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 Onilude[5], Neel Bhandari[4], Shivalika Singh[4], Hui-Lee Ooi[4], Amr Kayid[3], Freddie Vargus[4], Shayne Longpre[6], Niklas Muennighoff[4], Marzieh Fadaee[1], Julia Kreutzer[1], and Sara Hooker[1]

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

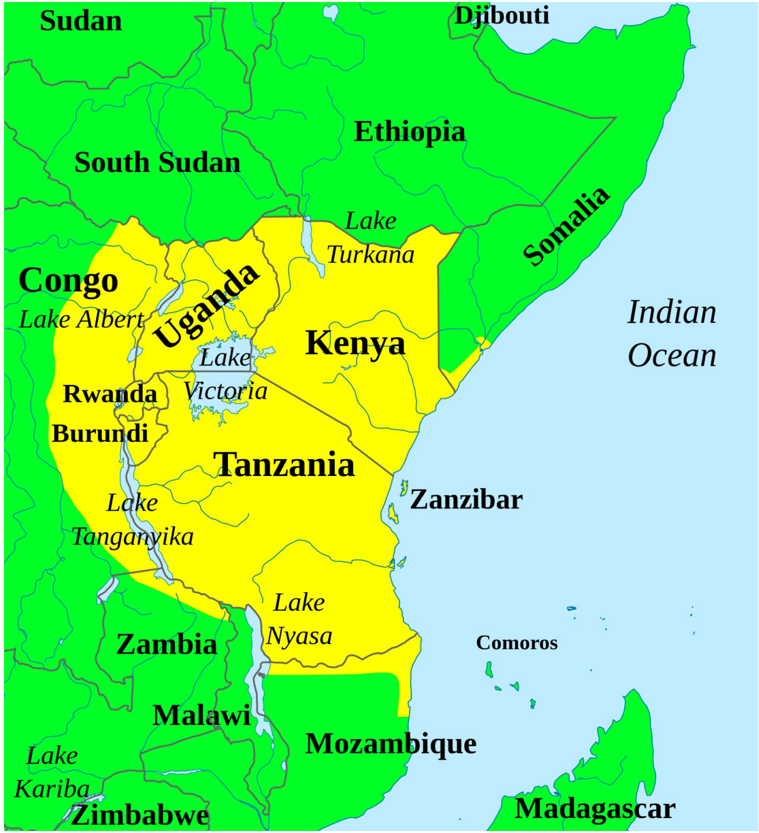
Cohere For AI

Swahili Usecase

Gathering data

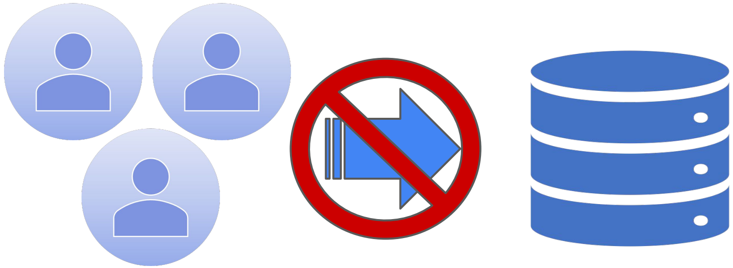
Our model language is SWAHILI

- Over 200M speakers
- Used in 4 african countries
- No instruction following datasets



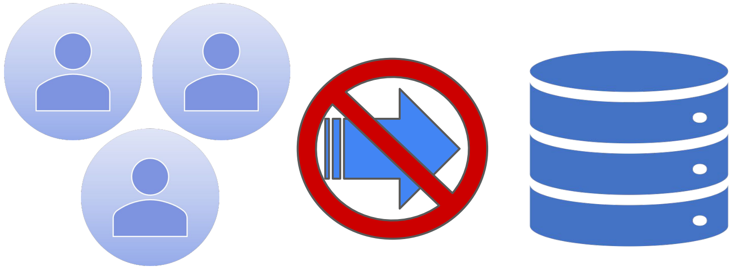
Translation and filtering the data

Assessors



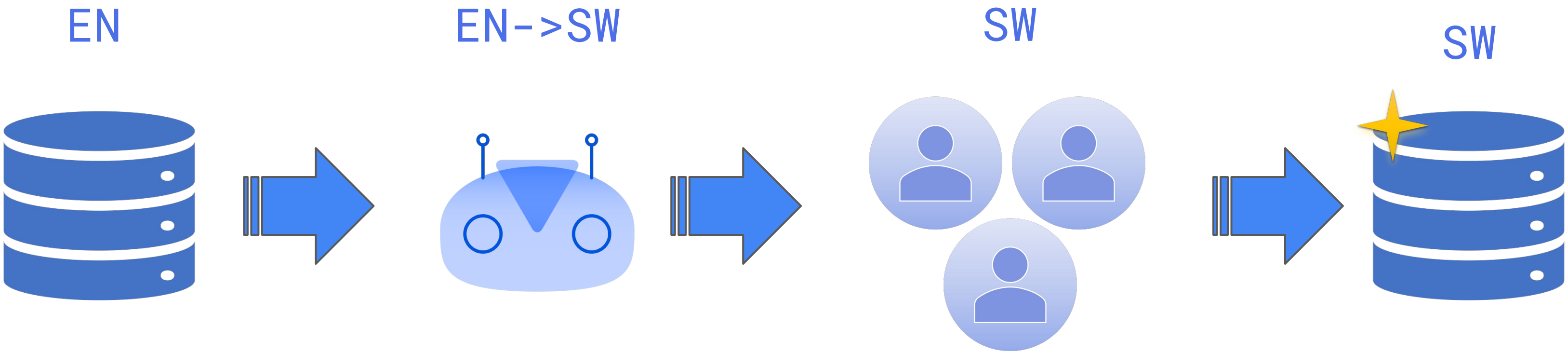
Direct generation by assessors is too expensive and assessors need to be highly prepared

Translators



Direct translation by assessors is expensive and hard to check

Translation and filtering the data



Dolly 15k Instruction following dataset

Best available translator from English to Swahili
Google Translate

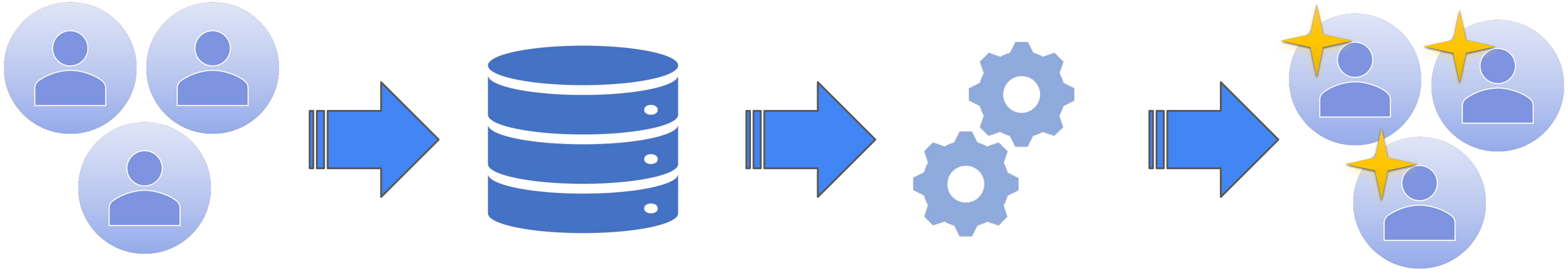
Selection subset of correctly translated data by native speakers

New curated dataset of instruction following on Swahili

Cohere For AI

Toloka

Data filtering I: Select best assessors



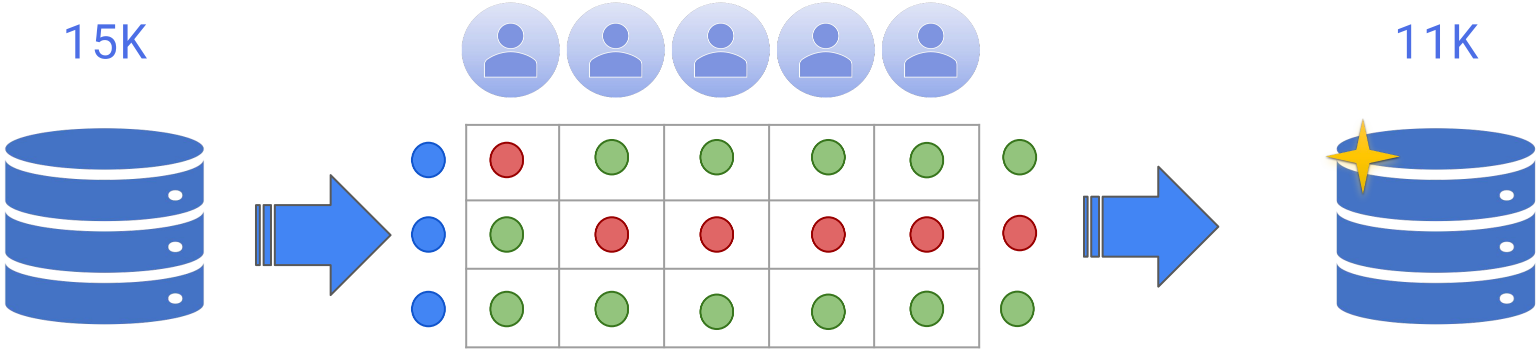
Select crowd of Swahili and English speaking assessors

Take any curated EN->SW dataset and dataset of incorrect translations

Create a task of classifying good from bad translations

Select assessors that perform better that others

Data filtering II: Filter your dataset



Raw EN->SW translated dataset with BLEU ~25

Selected assessors label dataset with high coverage (5-8 assessors)

Forming of new filtered dataset with BLEU ~50

Filtered out examples

Original	Translation	Retranslation
"Heartbreak Hotel".	"Hoteli ya Moyo".	"Heartbreak Hotel".
where do aliens come from	wageni wanatoka wapi	where do guests come from
Because whisky is Manly.	Kwa sababu whisky ni Mwanaume	Because whiskey is a Man

Time to try it in practice!

Cohere For AI