

AAAI Tutorial 2024

Aligning LLMs to Low-Resource Languages

Ahmet Üstün
Marzieh Fadaee

Nazar Beknazarov

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?

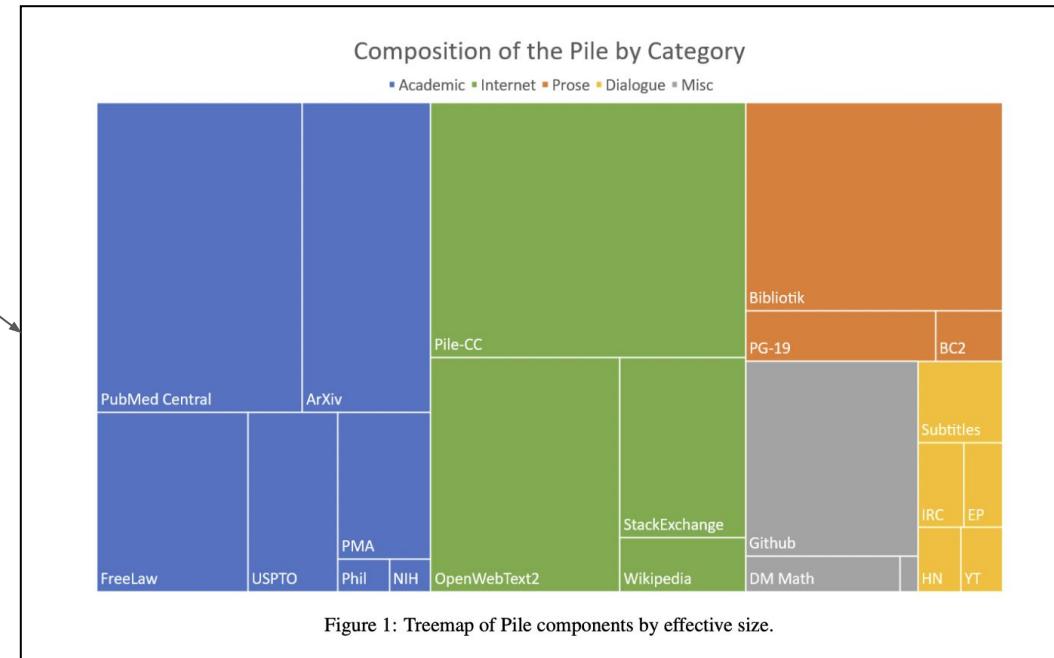
What is instruction fine-tuning?

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.



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.

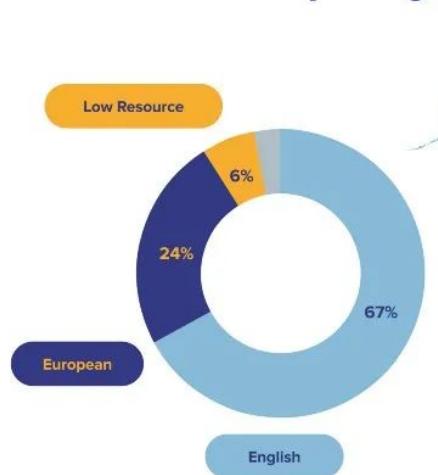
Multilingual Model Name	Number of Languages Trained On (pre-training)
BLOOM	46
mT5	101
XGLM	30



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

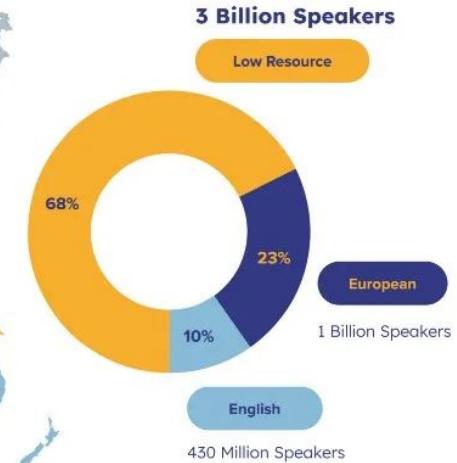
Significant data inequality

NLP Solutions by Language



Neural Space

Population Size of Languages



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

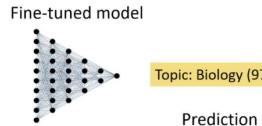
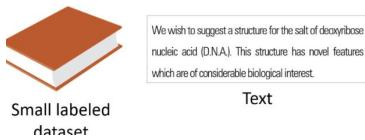


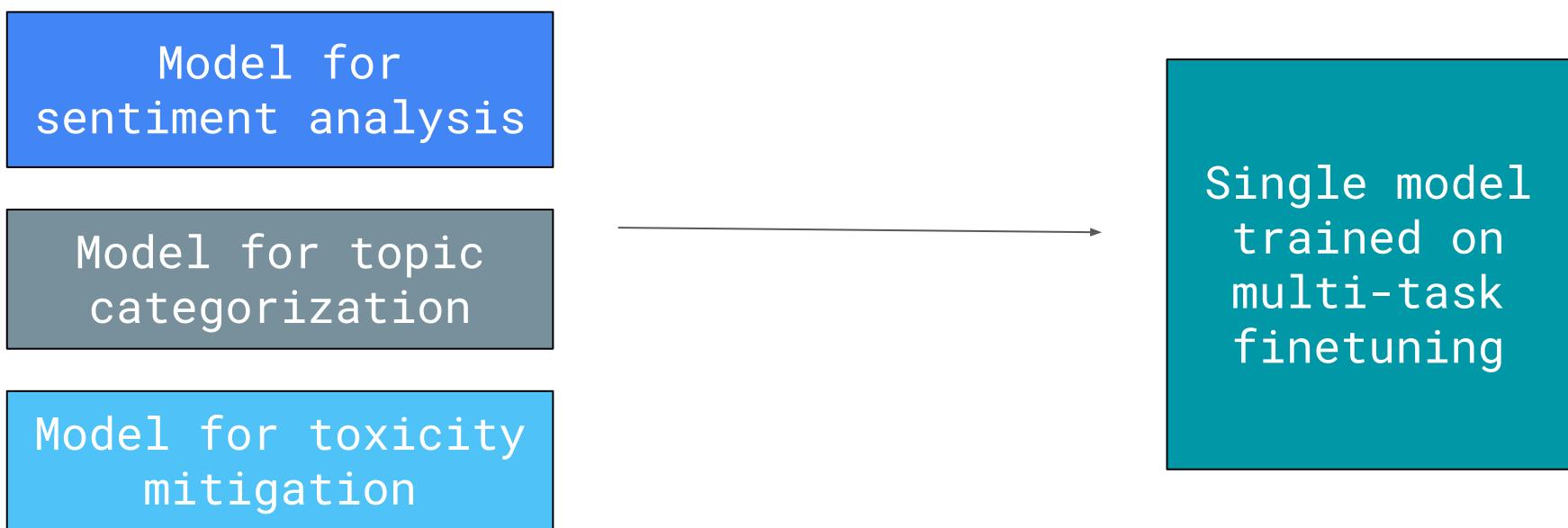
Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across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

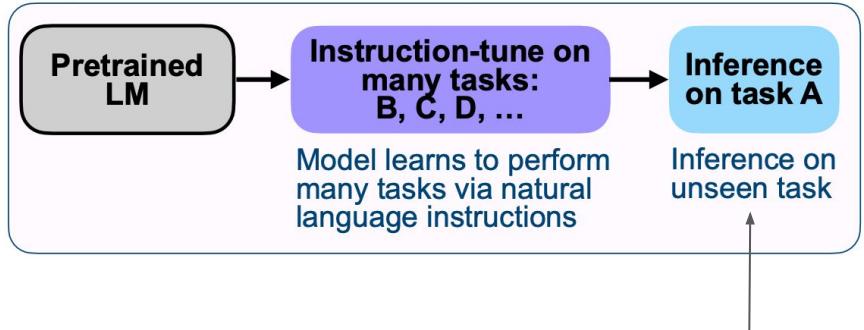
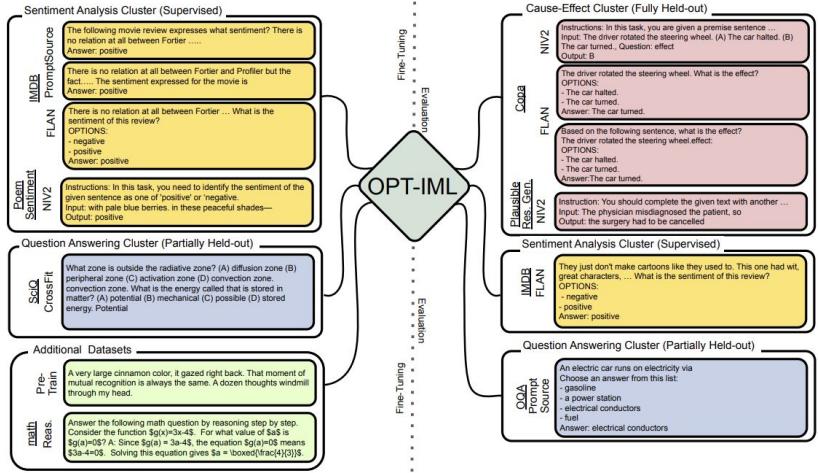


It turns out two ingredients have been particularly important at leading to breakthroughs in performance on zero and few shot tasks:

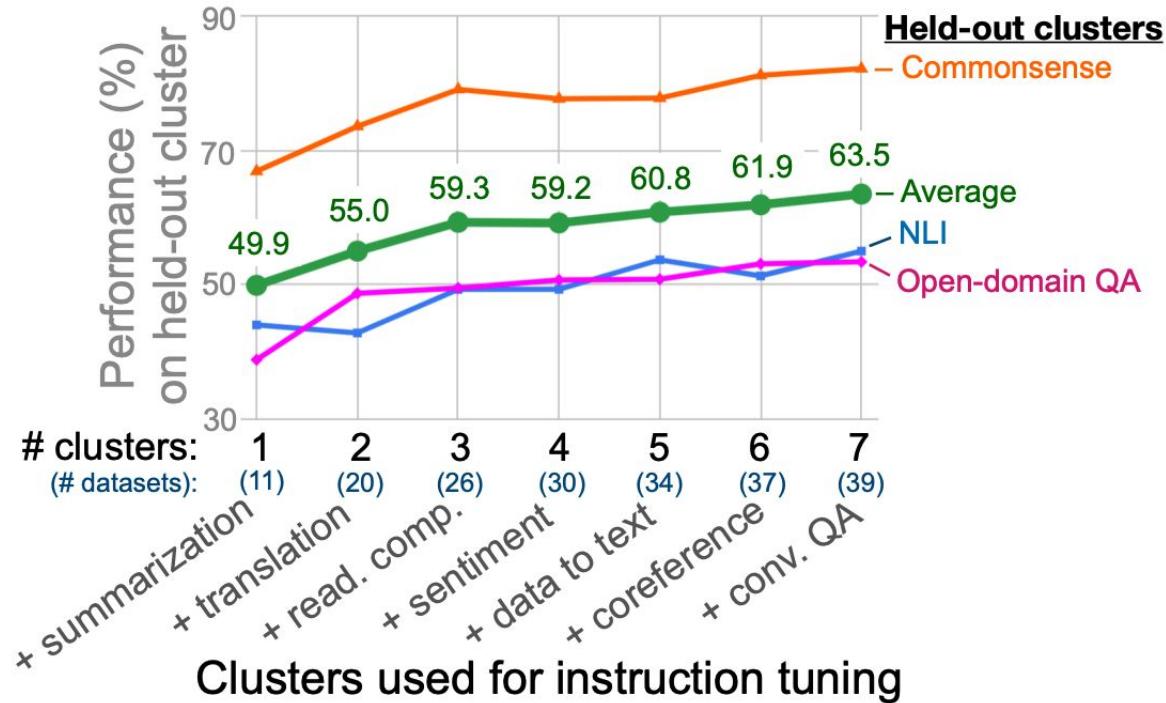
- 1. Instruction tuning – Structuring multi-task fine-tuning data as questions and answers**
- 2. Integrating human feedback about preferences**

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

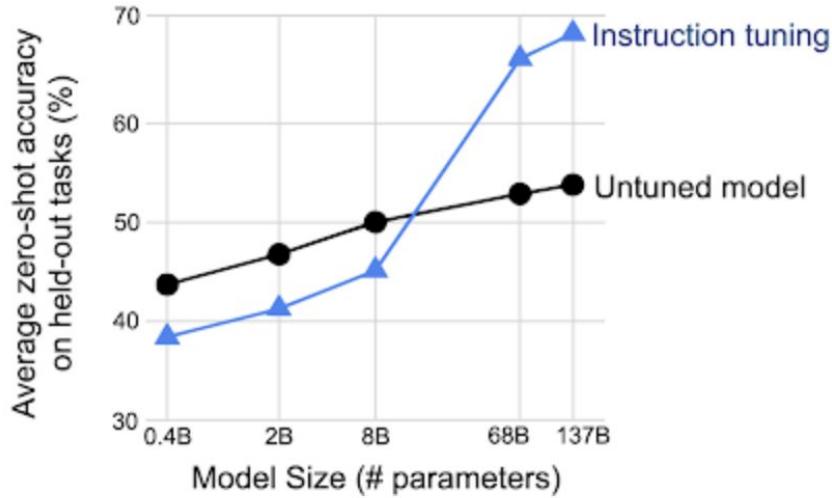
Leverage supervision to teach the model to perform many NLP tasks.



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).



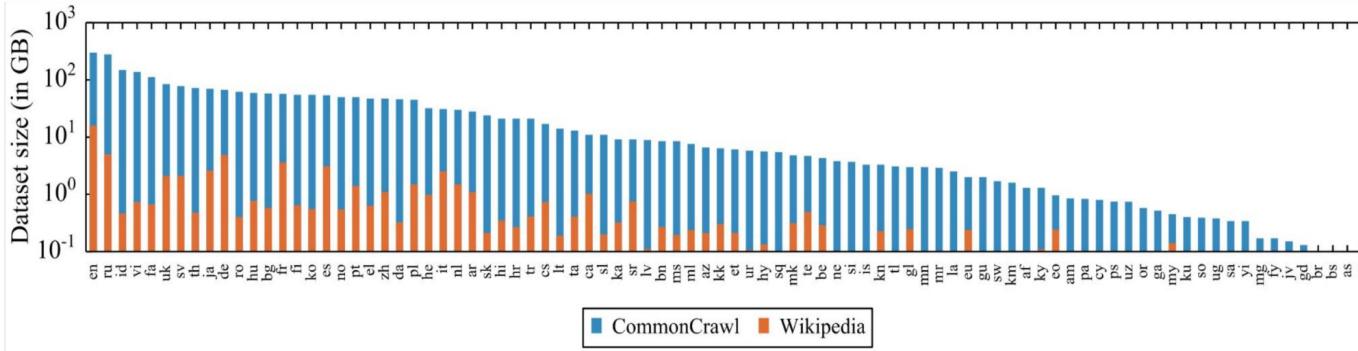
Instruction tuning only improves performance on unseen tasks for models of certain size.

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.

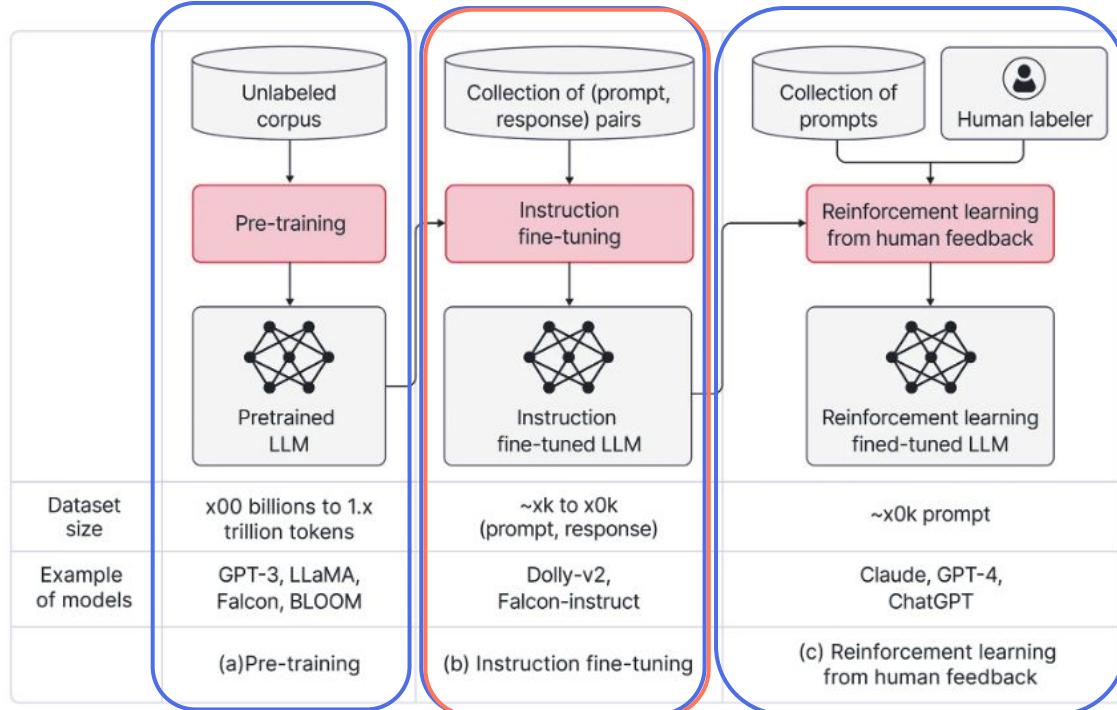
ACL [Keynote](#), [Conneau et al.](#)

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

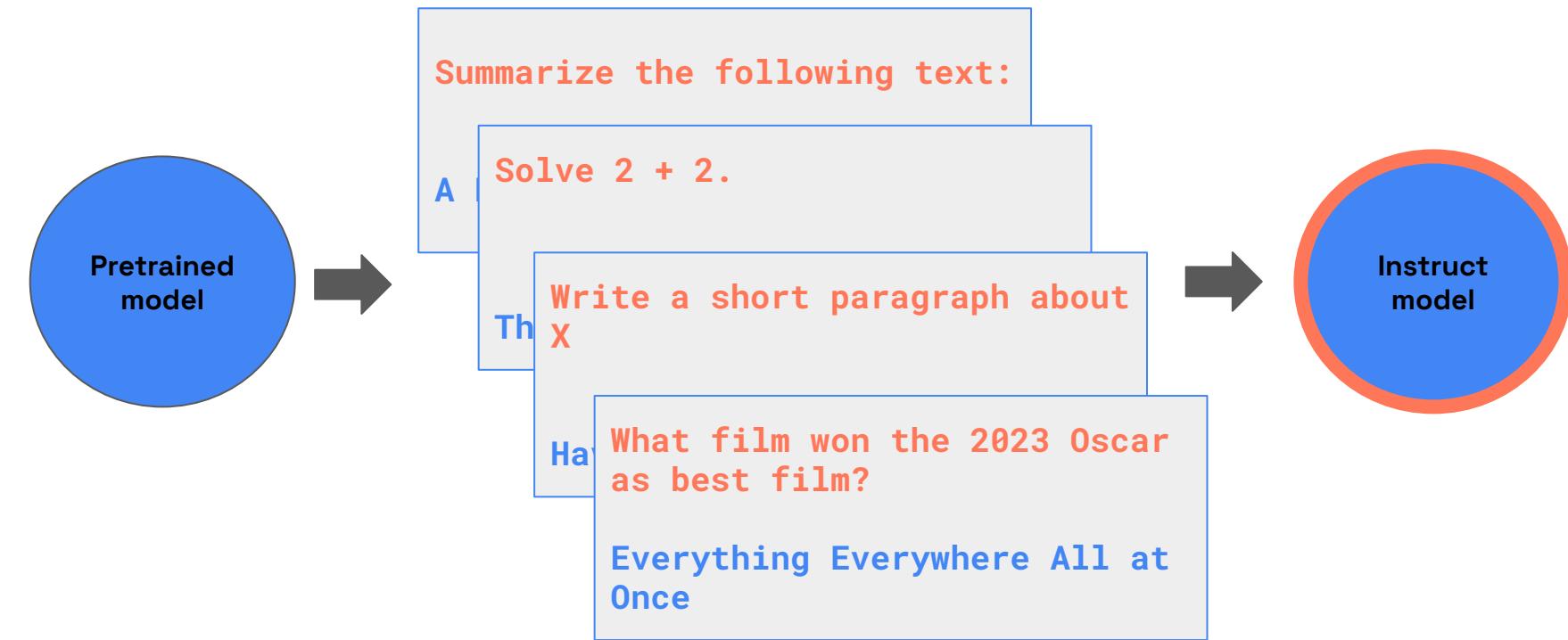


The long-tail of multilinguality, few high resource languages and many sparsely populated languages.

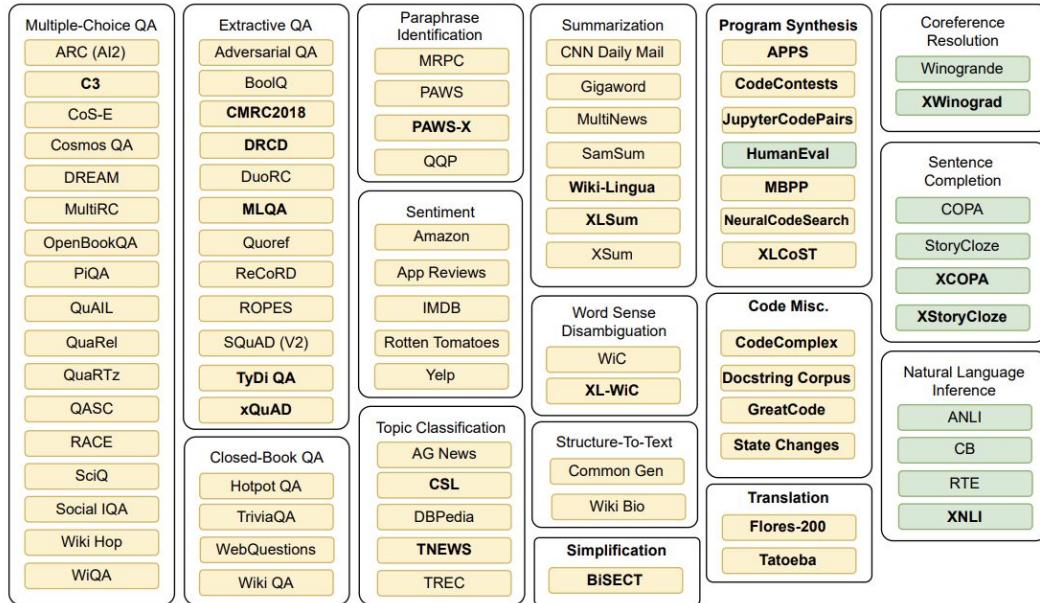
LLM recipe



What does the data actually look like?



Most relevant is work released by Muennighoff et al.



40% of instruction
data is still in
English

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.

Observed boosts in performance over base multilingual models.

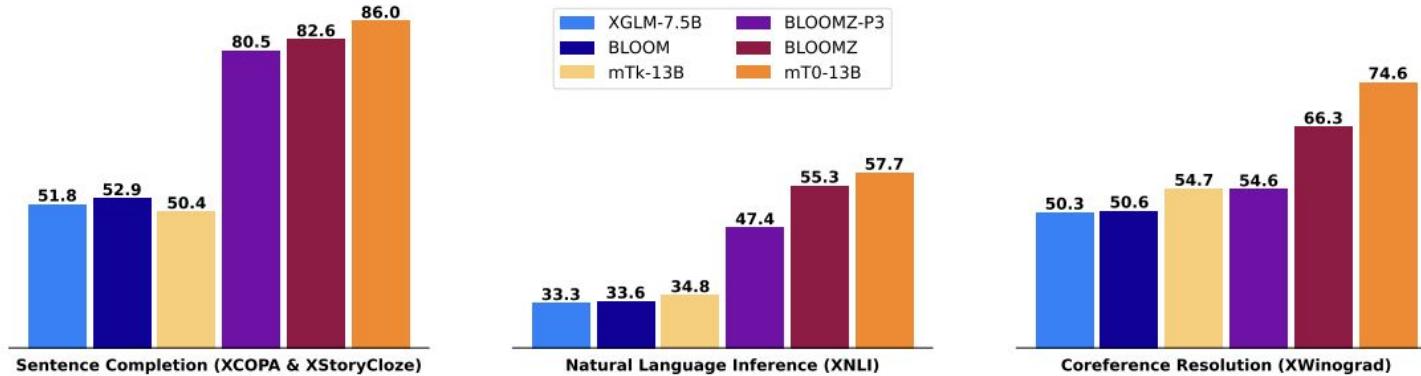


Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.



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

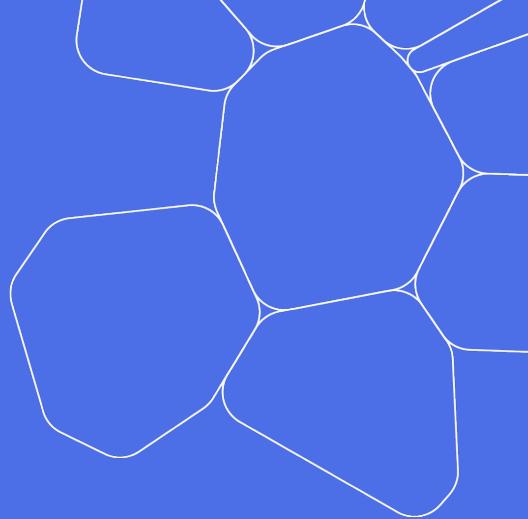
101

Languages

31K

Discord
messages

Aya Datasets



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

The image displays the Aya Dataset and Collection interface, featuring two main sections: Aya Dataset (blue background) and Aya Collection (orange background).

Aya Dataset:

- Prompt:** Aya Dataset 65
 - "କ୍ଷେତ୍ର ପାଇଁ କରିବ ପାଇଁ" ଏବଂ କିମ୍ବା ଦୟାଜଳତ ମିଳାଇ ପାଇଁ କାହାର ଅଭିଭାବ କିମ୍ବା?
 - Qual é a origem do xadrez?
 - ایا در اساطیر ایرانی اشاره‌ای به دعاوند شده است؟
 - ବୁଝାଲ୍ପି ମୁଦ୍ରାର ଅନ୍ତର୍ଭାବ ଏମଣି?
 - Berikan jawapan untuk teka-teki berikut: Apakah jenama tingkap yang paling ramai orang guna?
 - Inis dom faoi Baile Átha Cliath.
 - Maxaa waaye shay ar Soomaali badan cunaan maalin walba oo dadk qaar ay ku adtaghayta la'aantisa in ay cundaan cunaan waa Mooska. Mooska dad badan oo Soomali ah way jeelyihin, wuxuuna ka mid yahay waxayabba dalka uu ku caan bayax in uu dhoofiyo.
- Completion:**
 - "କ୍ଷେତ୍ର ପାଇଁ କରିବ ପାଇଁ" ଏବଂ କିମ୍ବା ଦୟାଜଳତ ମିଳାଇ "ପାଇଁ" କାହିଁ କିମ୍ବା ସ୍ଵରୂପରେ ଏବଂ ଏହି କିମ୍ବା ଏକାର୍ଥରେ କାହାର ମିଳାଇ ଅବଶ୍ୟକ ଏବଂ କିମ୍ବା.
 - O xadrez tem origens associadas ao Cangaço, onde cangaceiros, normalmente homens, dançavam com rifles em celebrações. Com a entrada de mulheres nos grupos, a participação feminina na dança também se expandiu.
 - دعاوند در اساطیر ایران چیکاو پیویمه‌ای دارد. شهوت آن بیش از هر چیز دیگر است که فردیون از شخصیت‌های اساطیری ایران، حسکا را از غاری به بند کنند و هدایت و خصماً آنها را تا آخرالیام که پند پکسل و کشتن خلخ اغزار کند و سرانجام به سنت کوشاسب کشته شوند.
 - ବୁଝାଲ୍ପି ମୁଦ୍ରାର ଅନ୍ତର୍ଭାବ ଏମଣି?
 - Jawapannya ialah Microsoft kerana "Microsoft Windows" (perkataan "window" bermaksud tingkap dalam bahasa Inggeris) ialah salah satu sistem operasi yang banyak digunakan oleh ramai orang di seluruh dunia.
 - Is Baile Átha Cliath, nó Dublin i mBéalair, príomhchathair na hÉireann agus é ar an gcaithair is mó sa tir freisin. Tá sé suite in oirthear na tire, cois Life agus ar meall Mhuir Éireann. Tá breis is 1.6 milíon duine ina gcónaí i mórcheantar Bhaile Átha Cliath agus deirtear go dtíofaíodh borradh agus fás ar an daonra sa todhchaí.
 - Shay ar Soomaali badan cunaan maalin walba oo dadk qaar ay ku adtaghayta la'aantisa in ay cundaan cunaan waa Mooska. Mooska dad badan oo Soomali ah way jeelyihin, wuxuuna ka mid yahay waxayabba dalka uu ku caan bayax in uu dhoofiyo.

Aya Collection:

- Text Classification:**
 - Prompt: Classify the sentiment of the following tweet with either positive, negative, or neutral \n{[tweet]}
 - Completion: I would classify the given tweet as: {[label]}
- Natural Language Generation:**
 - Prompt: What is the corresponding translation in {[target_lang]} of the following sentence : {[source]}
 - Completion: The translation of the sentence to {[target_lang]}: \n{[target]}
- Question Answering:**
 - Prompt: What category does this question come from: {[question['text']]}
 - Completion: This question can come from category: {[document['kind']]}

Datasets Summary:

- Translated Text Classification datasets:** +2 Translated Text Classification datasets (e.g., xtel_wd, NTX_LLM_Instruct_{language}, UNER_LLM_Inst_{language}, NusaX-senti, MasakhaNEWS, AfriSenti, Urdu-Instruct-News-Category-Classification, IMDB-Dutch-Instruct, scirepeval biomimicry)
- Translated NL Generation datasets:** +8 Translated NL Generation datasets (e.g., IndicSentiment-instruct, xwikis, {language}_instruct_stories, Lijnnews-instruct-{lang_pair}, scb_mt_2020_{lang_pair}.prompt, SEED-instruct-{lang_pair}, wiki_split, Persian_instruction_fn, arpa-aya, Turku-paraphrase-corpus, FarsTail-Instruct-LLM, Tamil_stories, Joke_explanation, Thirukkural, Annotated_news_summary, Thai-Pos-prompt, SODA, Urdu-Instruct-News-(task), UA_Gec_instruction_tuning, Thai-wiktionary-prompt, Hindi-article-summarization/generation)
- QA Datasets:** +9 Translated QA datasets (e.g., X-CSQA (X-CSR), AfriQA, Mintaka, TeluguRiddles, LLM-Japanese-vanilla-instruct, Amharic-QA)

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) Contribute Your Language

Help us add more examples by writing your own prompt and their completion in your language.

English

Prompt

what does "to crave" mean?

Completion

to crave is to really really want something, or to have a deep desire.

Submit Entry

(b) Rate Model Performance

Help improve the model performance by rating the prompt-completion pairs below

Edit Mode English

Prompt

Look out for: ✓ Correct Grammar ✓ Reasonable Length ✓ Clear Instructions

What does it mean in English "to crave" something?

Edit Distance[Prompt]: 35

Completion

Look out for: ✓ Clear Answer ✓ Full Sentences with Correct Answer ✓ Reasonable Length

To crave is to have a strong and intense desire for something. It implies a deep need or longing for the object of the craving. It can involve emotional and physical feelings of addiction and can sometimes lead to actions against one's best interests. It is important to note that there is a difference between wanting or liking something and craving it, as craving can be destructive and lead to unhealthy behaviors.

Edit Distance[Completion]: 321

Skip Submit Entry

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.

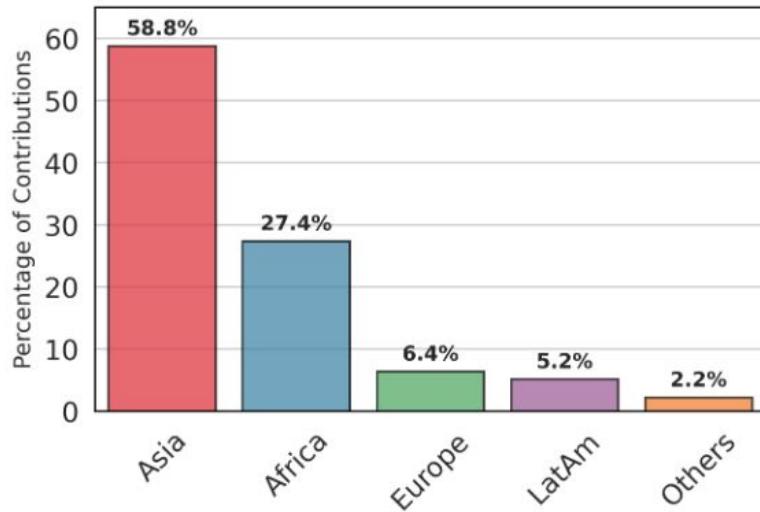


Figure 4: Distribution of total contributions across different regions.

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

Longer instructions and higher quality data in Aya

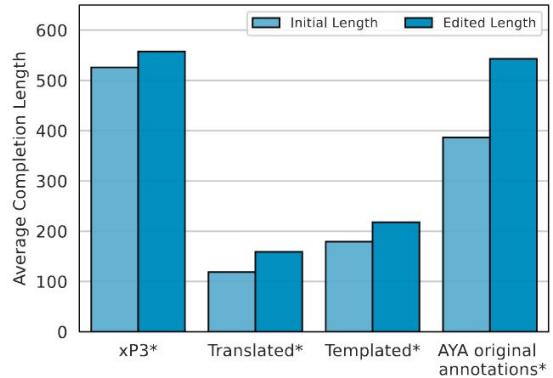


Figure 8: Average Completion Length before and after re-annotation.

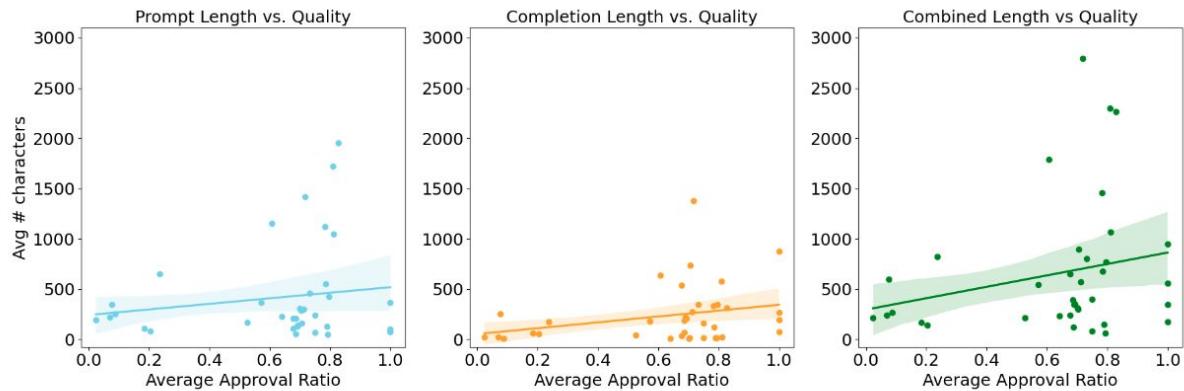


Figure 9: Relationship between Average Prompt and Completion Length in characters and the Average Approval Rate of the example.

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

Aya Collection	
Text Classification	Natural Language Generation
Prompt Classify the sentiment of the following tweet with either positive, negative, or neutral \n{{tweet}} Completion I would classify the given tweet as: {{label}}	Prompt What is the corresponding translation in {{target_lang}} of the following sentence : {{source}} Completion The translation of the sentence to {{target_lang}}: \n{{target}}
+2 Translated Text Classification datasets	+8 Translated NL Generation datasets
44 xle_wd	11 IndicSentiment-instruct
13 NTX_LLM_Instruct_{{language}}	5 xxwikis
11 UNER_LLM_Inst_{{language}}	3 {{language}}_instruct_stories
10 NusaX-senti	2 Lijnews-instruct-{{lang_pair}}
10 MasakhaNEWS	2 scb_mt_2020_{{lang_pair}}_prompt
9 AfriSenti	2 SEED-instruct-{{lang_pair}}
1 Urdu-Instruct-News-Category-Classification	1 wiki_split
1 IMDB-Dutch-Instruct	1 Persian_instruction_pn
1 scirepeval biomimicry	1 arpa-aya
Question Answering	Turku-paraphrase-corpus
Prompt What category does this question come from: {{question['text']}}? Completion This question can come from category: {{document['kind']}}.	1 FarsTail-Instruct-LLM
+9 Translated QA datasets	1 Tamil_stories
16 X-CSQA (X-CSR)	1 Joke_explanation
12 AfriQA	1 Thirukkural
9 Mintaka	1 Annotated_news_summary
1 TeluguRiddles	1 Thai-Pos-prompt
1 LLM-Japanese-vanilla-instruct	1 SODA
1 Amharic-QA	1 Urdu-Instruct-News-{{task}}
	1 UA_Gec_instruction_tuning
	1 Thai-wiktionary-prompt
	1 Hindi-article-summarization/generation

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

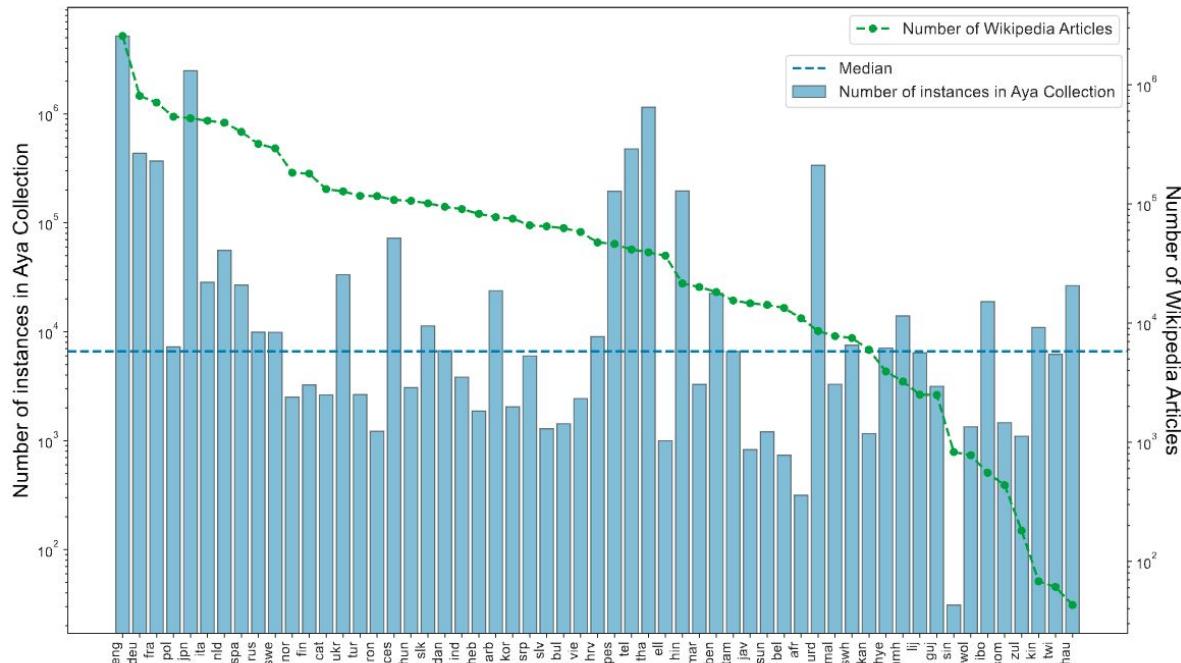


Figure 14: Number of prompt/completion pairs in each language in the **Aya** Collection (templated). Many languages with limited digital presence, as indicated by a low number of Wikipedia pages, are well-represented in the templated portion of the **Aya** Collection. Note that both axes are in log-scale.

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

Aya Evaluation Suite

7

aya_human_annotated

6

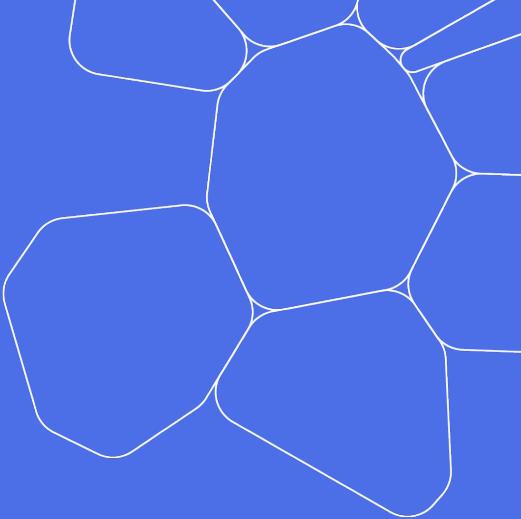
dolly-human-edited

114

dolly_machine_translated

Prompt: <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		x	
	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	

Aya Model



Fine-tuning

Multilingual templates

- 99 xP3x
- 61 Aya Collection
- 14 Data Provenance 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



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

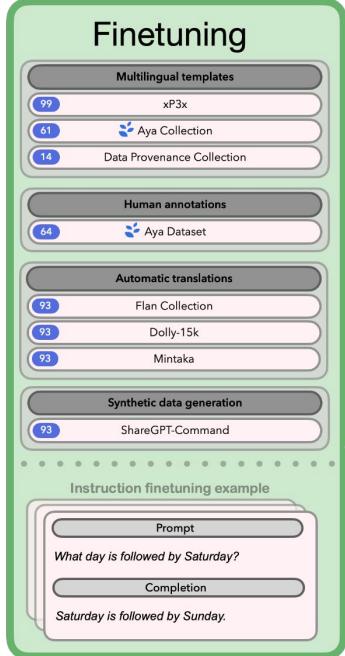
- 7 Toxicity detection
- 11 Harmfulness for adversarial prompts
- 7 Open-ended generation toxicity
- 8 Gender bias in machine translation

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
	14	161	1.65M	998	78	97.5	0.5	2.0
	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.

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

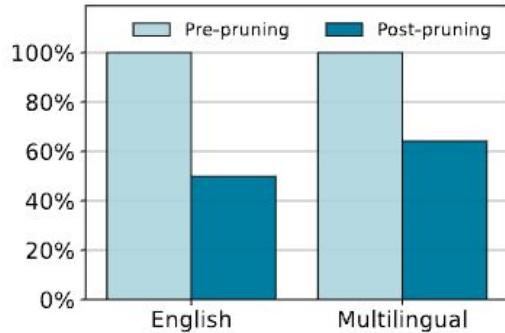
203 million examples

Machine translated datasets into 93 languages

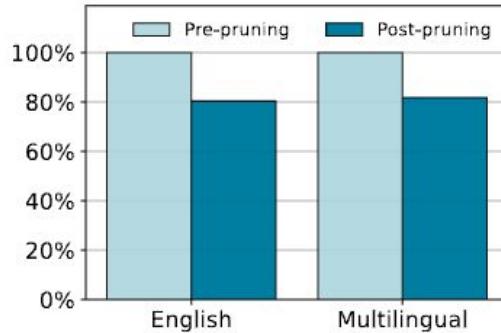
Aya Dataset, a fully human-curated dataset of instructions

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

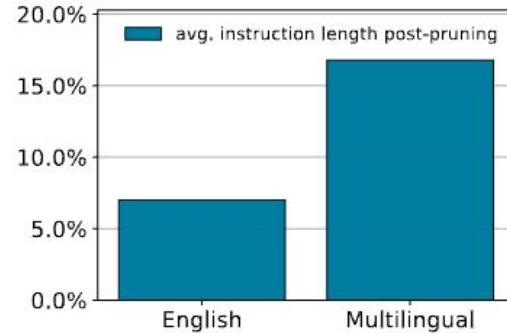
Changed to multi-task fine-tuning. Moving to a single global model – train on multiple tasks at once.



(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

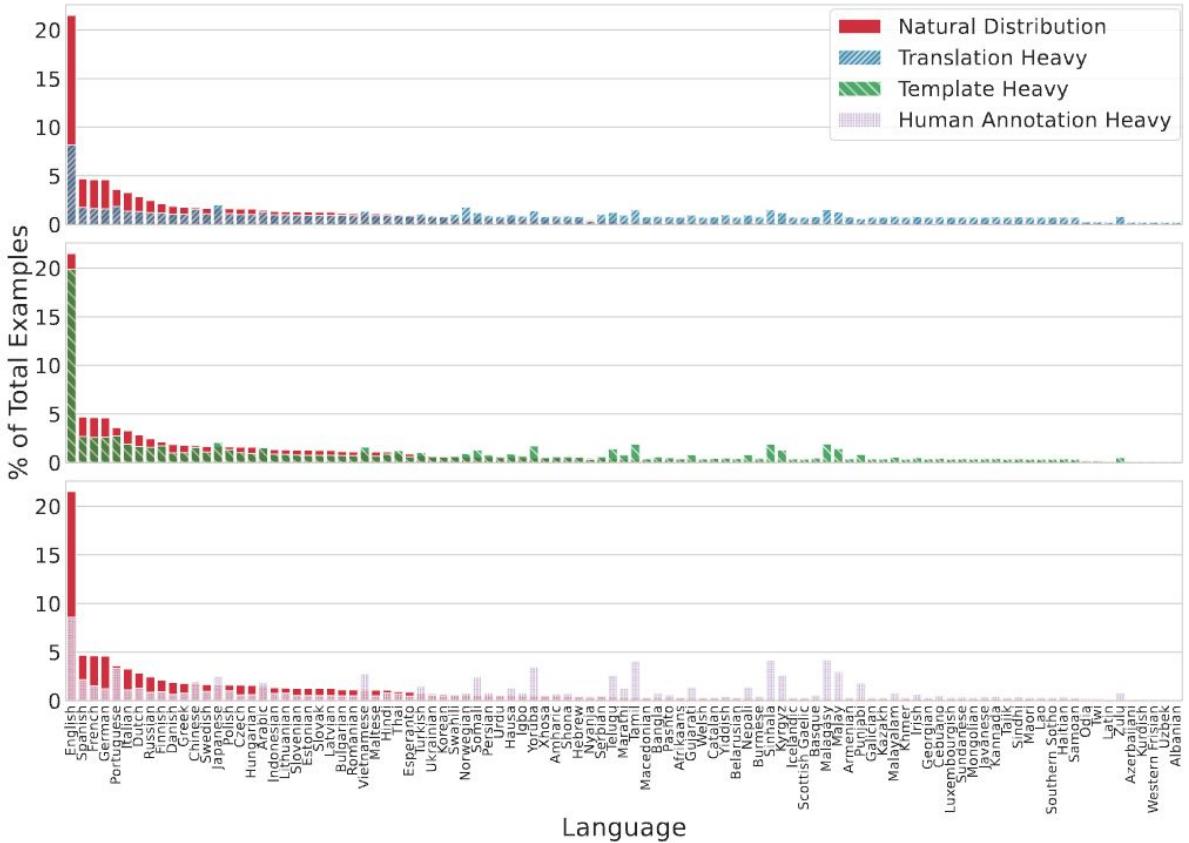
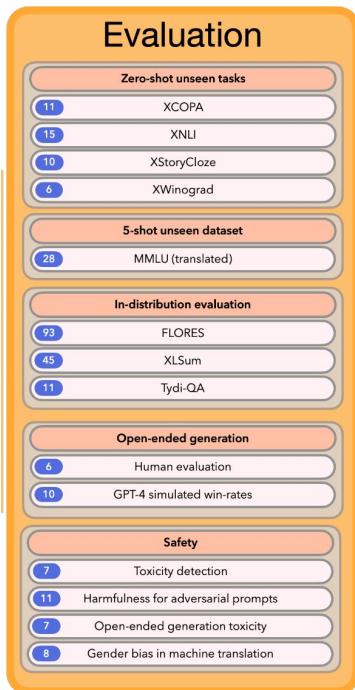


Figure 17: % of Examples for each language with different weighting schemes

Extensive multilingual evaluation on multiple evaluation categories on many languages



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.

Evaluation at a glance:

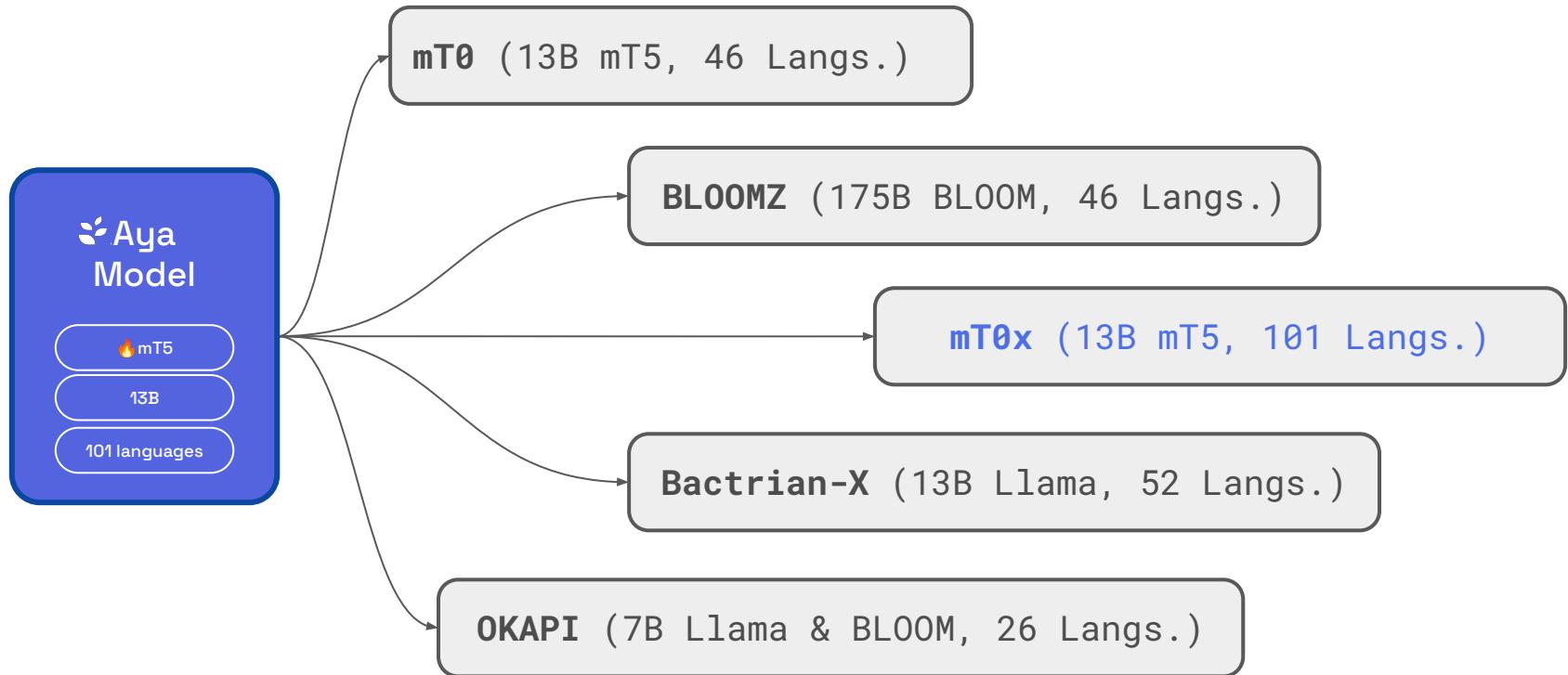
99 languages

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

Comparison with state-of-the-art baselines



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

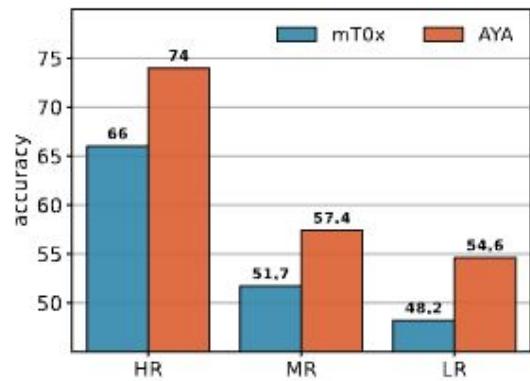
Table 6: Multilingual MMLU score comparisons between Okapi, mT0, mT0x, and **Aya** models. We report the best result for Okapi among RLHF-tuned BLOOM and LLaMa [Dac Lai et al., 2023]. ‘‡’ Note that Okapi reports 25-shot results, however, mT0, mT0x and **Aya** (translation-heavy) models are evaluated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

Evaluation on in-distribution generative tasks

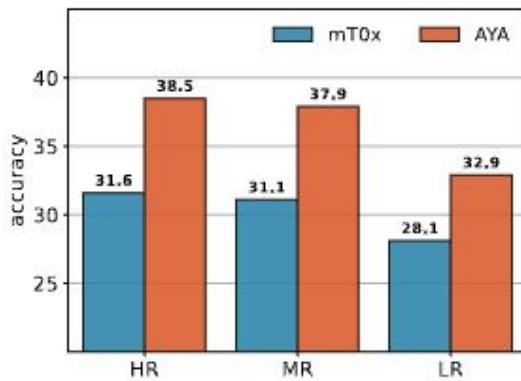
Model	IFT Mixture	Generative Tasks		
		FLORES-200 (spBleu)	XLSum (RougeLsum)	Tydi-QA (F1)
101 Languages				
		X → En	En → X	
mT0x	xP3x	20.2	14.5	21.6
Aya (human-anno-heavy)	All Mixture	25.1	18.9	22.2
Aya (templated-heavy)	All Mixture	25.0	18.6	23.2
★ Aya (translation-heavy)	All Mixture	29.1	19.0	22.0
				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.

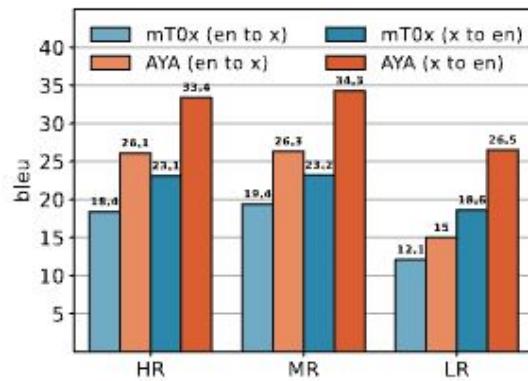
Larger improvement in MR and LR languages



(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?

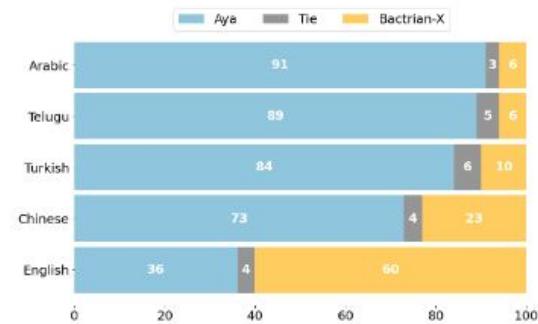
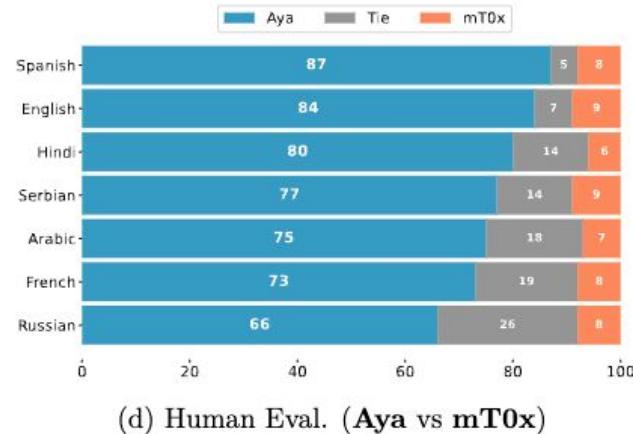
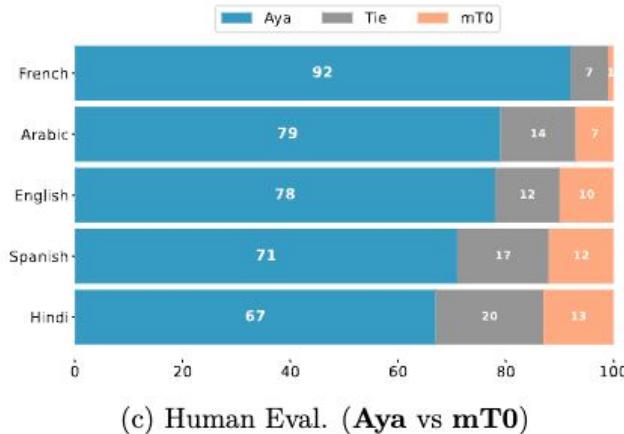


Figure 5: GPT-4 Eval. (Aya vs BX) using aya-human-annotated test set

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

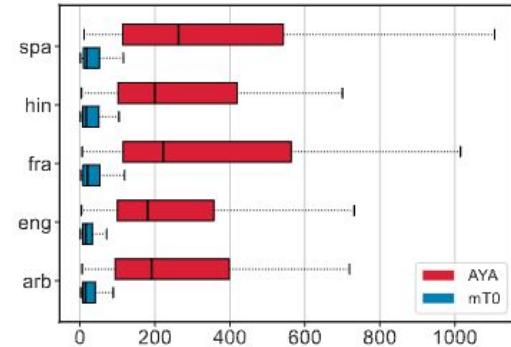
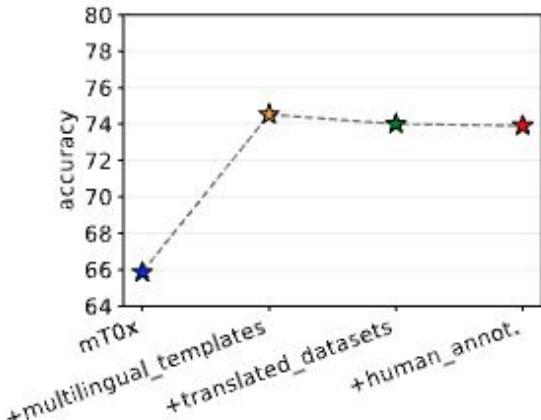
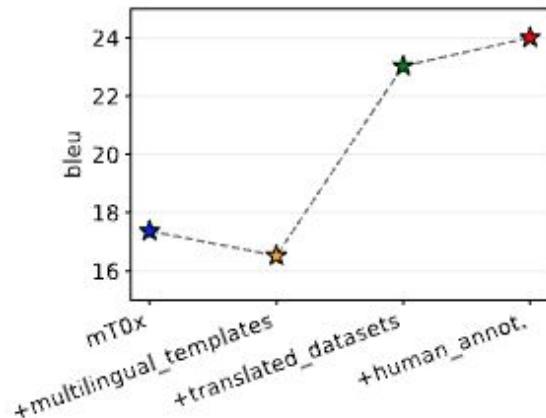


Figure 6: Completion lengths by characters for the Aya and mT0 models in Dolly test set for various languages.

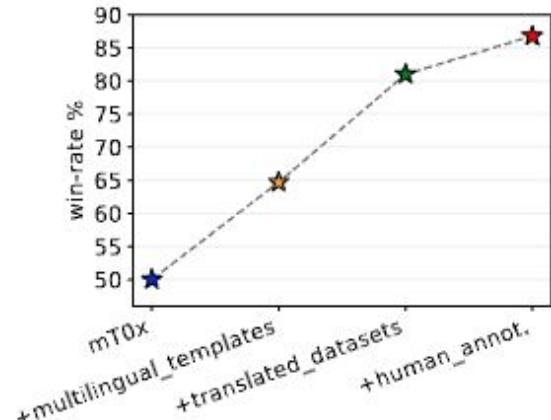
Each data source contributes to the downstream performance



(a) Unseen Discriminative Tasks

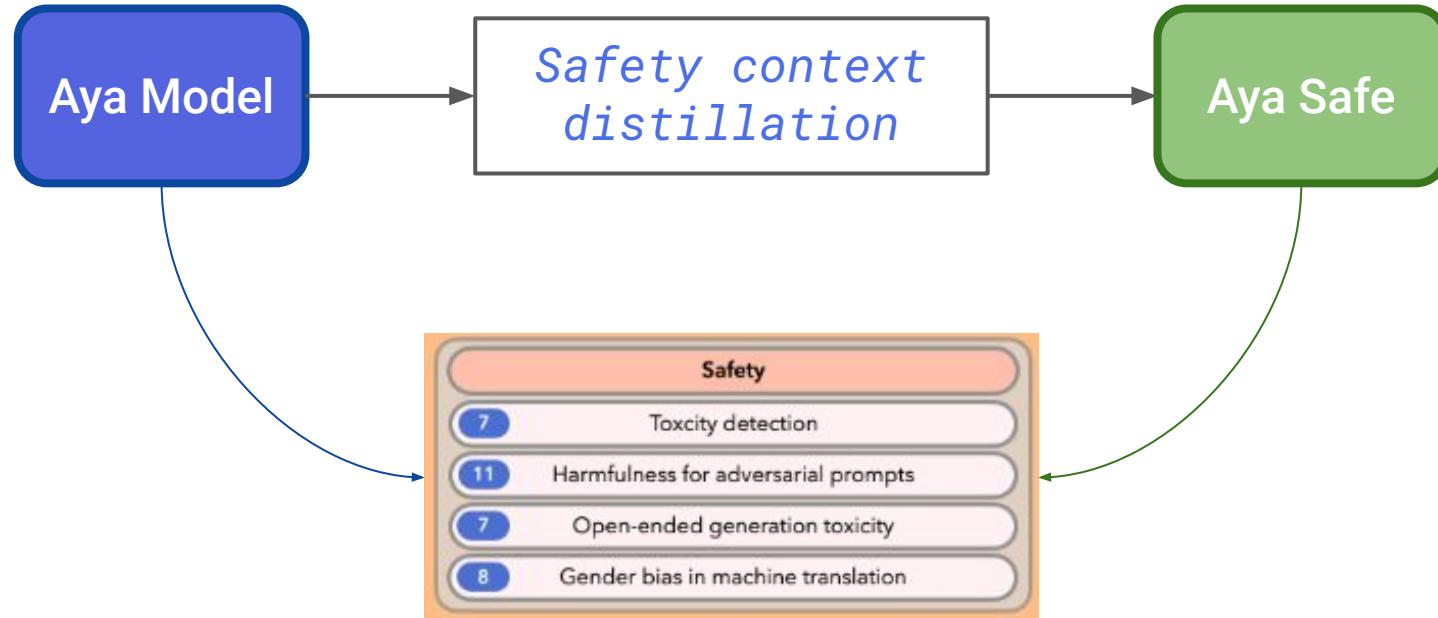


(b) Generative Task: Flores



(c) Win Rates (vs mT0x)

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



Safety context distillation significantly helps model to be safer

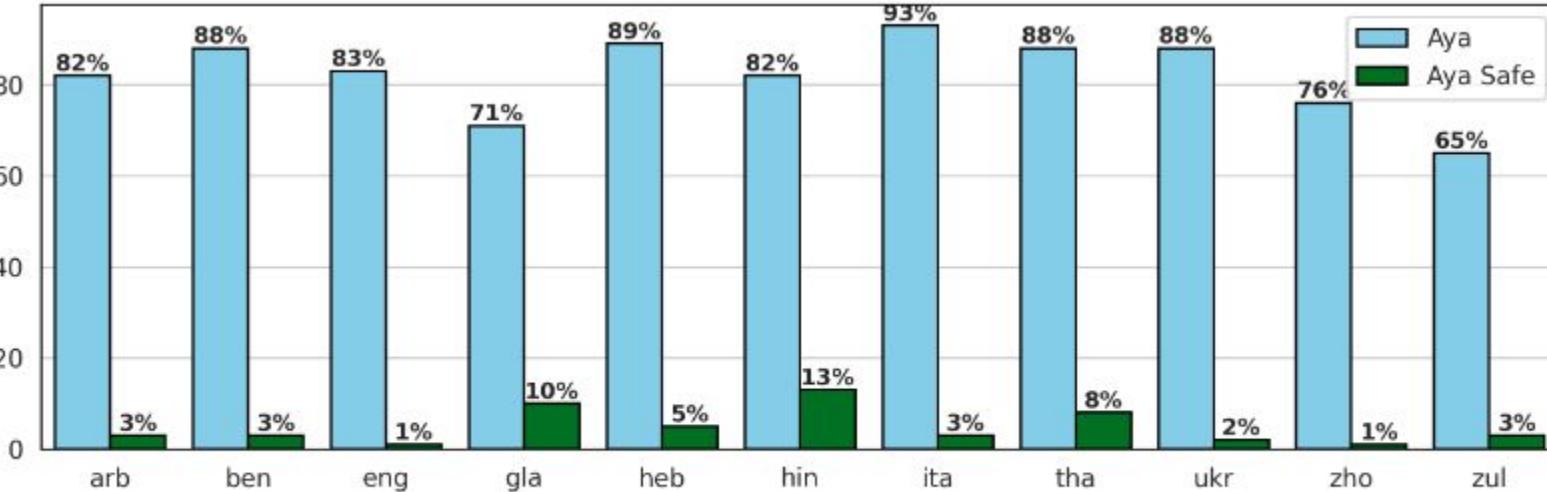
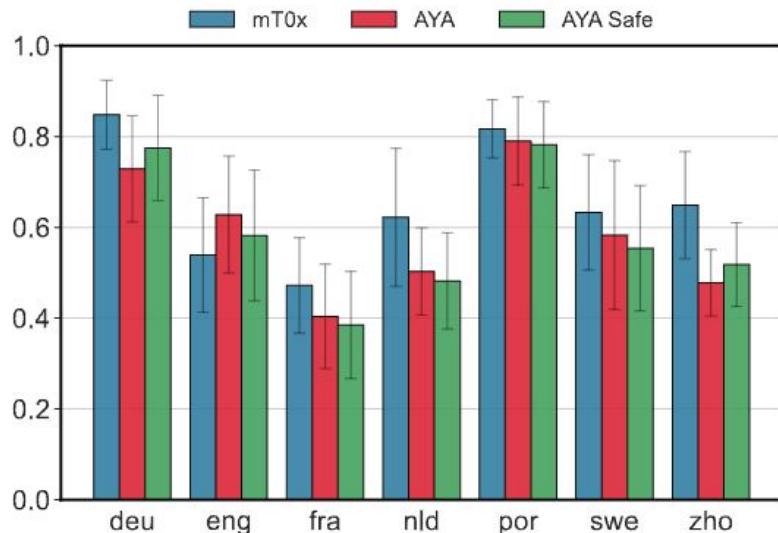


Figure 11: GPT-4 evaluation: Ratio of harmful generations for AdvBench held-out prompts.

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			Held out tasks			
		Flores (spBleu)	XLSum (RougeLsum)	Tydiqa (F1)	XCOPA	XNLI (Accuracy %)	XSC	XWNG
101 Languages		X→ En En → X						
MT0x	xP3x	20.2	14.5	21.6	76.1	71.7	45.9	85.1
Aya	All Mixture	29.1	19.0	22.0	77.8	76.8	58.3	90.0
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 Patel⁴,
Deividas Mataciunas¹, Laura O'Mahony⁵, Mike Zhang⁶, Ramith Hettiarachchi⁷,
Joseph Wilson⁸, Marina Machado³, Luisa Souza Moura³, Dominik Krzemiński¹,
Hakimeh Fadael¹, 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

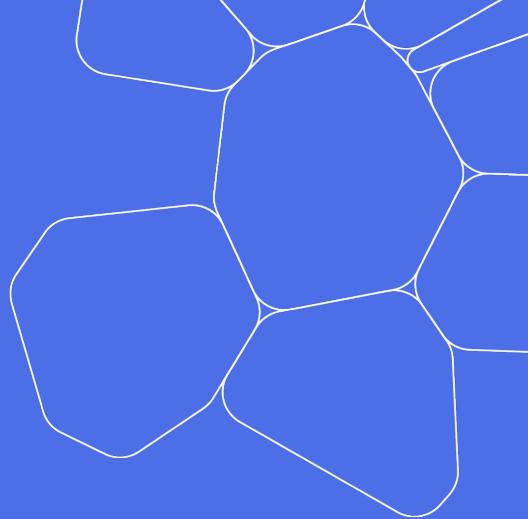
¹Cohere For AI Community, ²Beijing Academy of Artificial Intelligence, ³Cohere, ⁴Binghamton University,
⁵University of Limerick, ⁶IT University of Copenhagen, ⁷MIT, ⁸University of Toronto, ⁹King Fahd University of
Petroleum and Minerals, ¹⁰King Abdulaziz University, ASAS.AI, ¹¹Bloomberg LP, ¹²Cohere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

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

Swahili Usecase



Gathering data

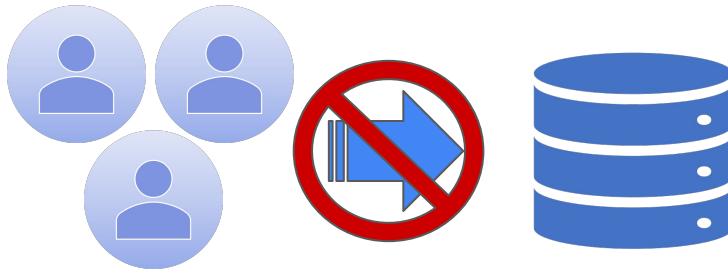
Our model language is
SWAHILI

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



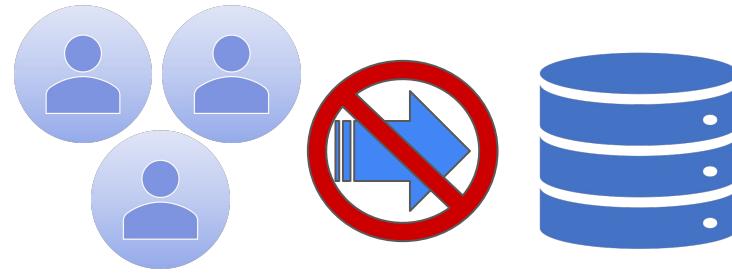
Translation and filtering the data

Assessors



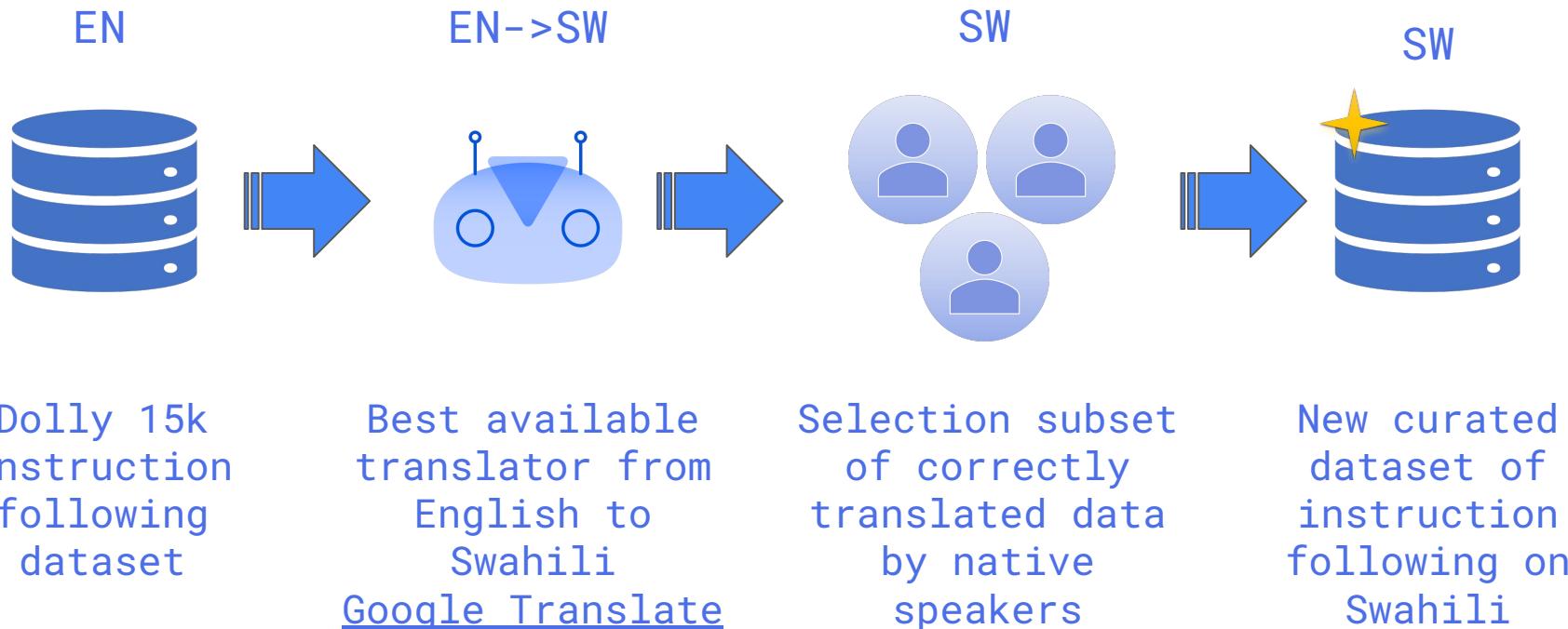
Direct generation by
assessors is too
expensive and
assessors need to be
highly prepared

Translators

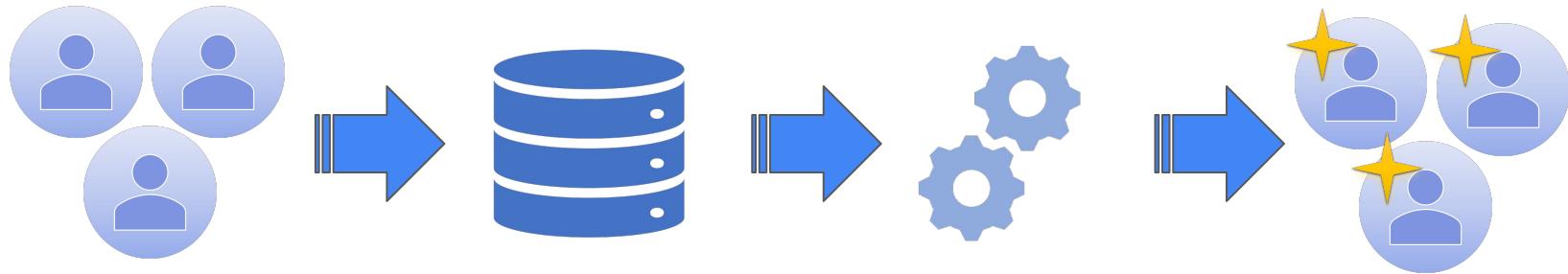


Direct translation by
assessors is
expensive and hard to
check

Translation and filtering the data



Data filtering I: Select best assessors



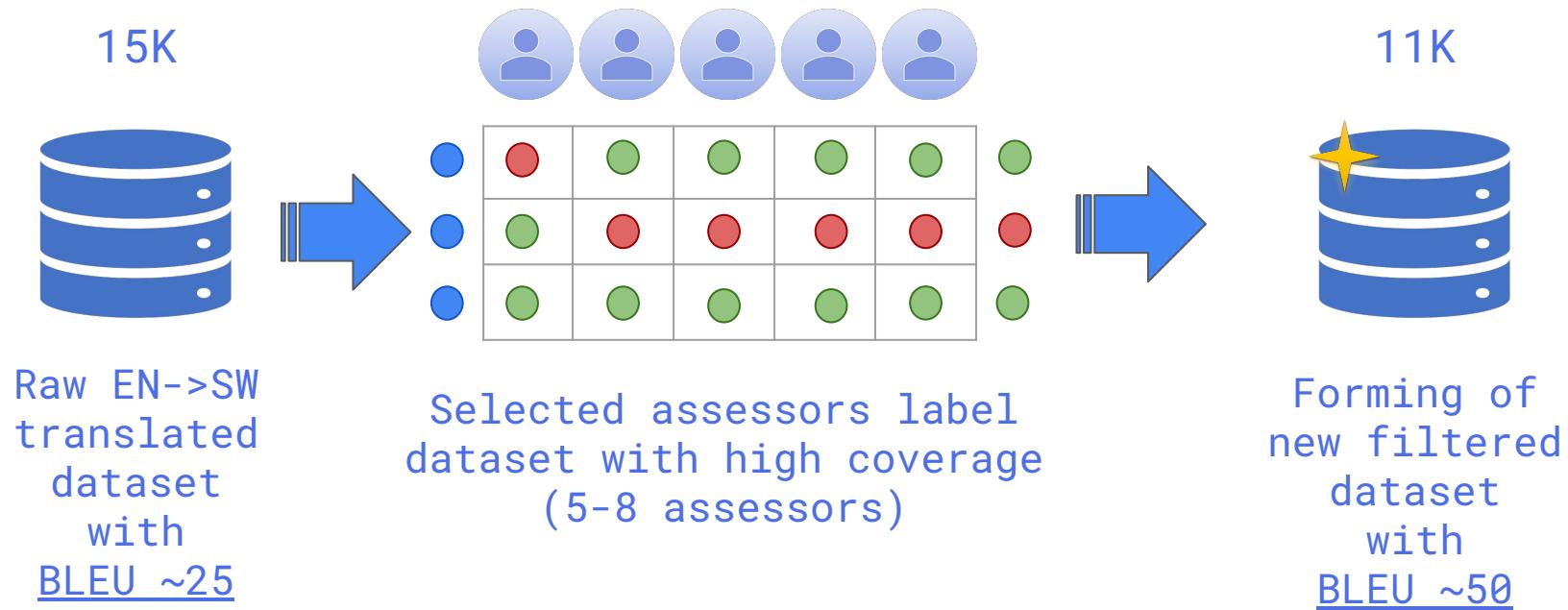
Select crowd
of Swahili
and English
speaking
assessors

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 than
others

Data filtering II: Filter your dataset



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!