Aya

Accelerating multilingual AI through open science

Aya at a glance.

cohere.com/research/aya

Aya at a Glance

1
Model

513M
Re-annotations of Datasets

3K
Independent Researchers

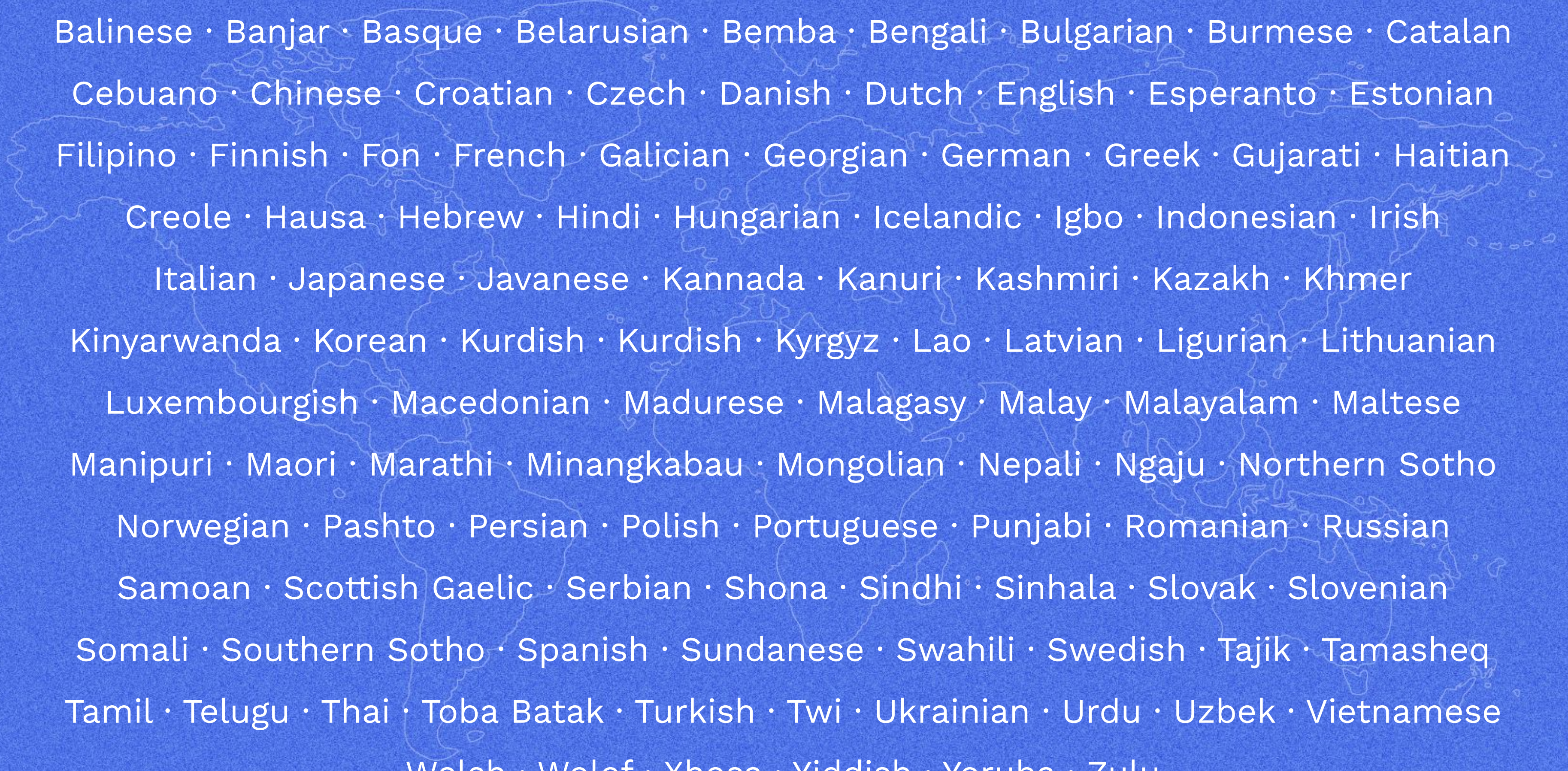
56
Language Ambassadors

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

Contents

The Story of Aya

Aya is a new state-of-the-art, open source, massively multilingual LLM covering 101 languages – including more than 50 previously underserved languages.

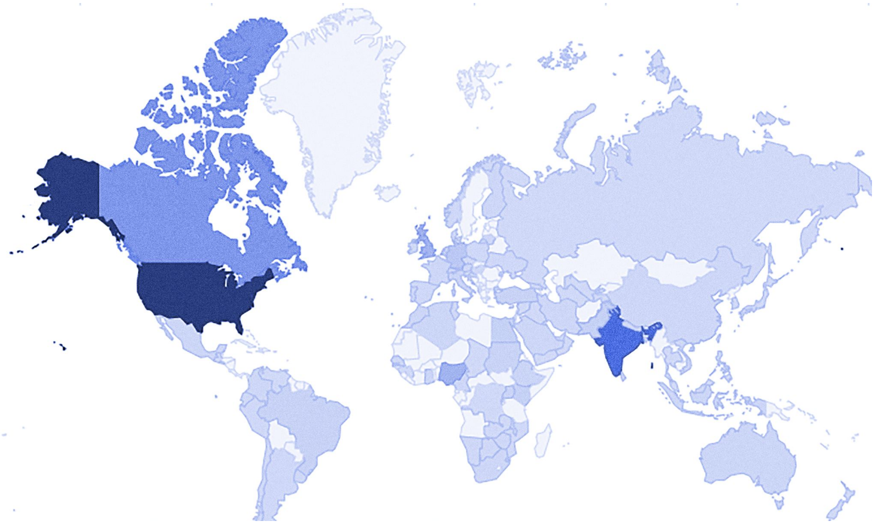
Building Aya has taken over a year, and involved 3,000 collaborators across 119 countries, making it one of the largest open science projects in machine learning research.

But how did we get here? It all started with a vision to solve complex machine learning problems and an ambitious goal to increase access to language technology for all.

A community, ready to collaborate

The impetus for Aya came out of the Cohere For AI Open Science initiative - a community that supports independent researchers around the world connect, learn from one another, and work collaboratively to advance the field of ML research.

Starting in January, 2023, members worldwide were keen to leverage the strengths of their diversity and collaborate on something brand new - an open science project to accelerate multilingual AI, and increase access to this technology for the people of their regions.



Join our Open Science Community

Involving 3000+ researchers around the world

Aya is as much a protest against how research is done as it is a technical contribution. Most breakthroughs to-date have come from a small set of labs and countries. Aya instead started with a revolutionary premise: working with independent researchers, engineers, linguists, language enthusiasts around the world to defy expectations and build a breakthrough model.

Standing up against inequitable progress

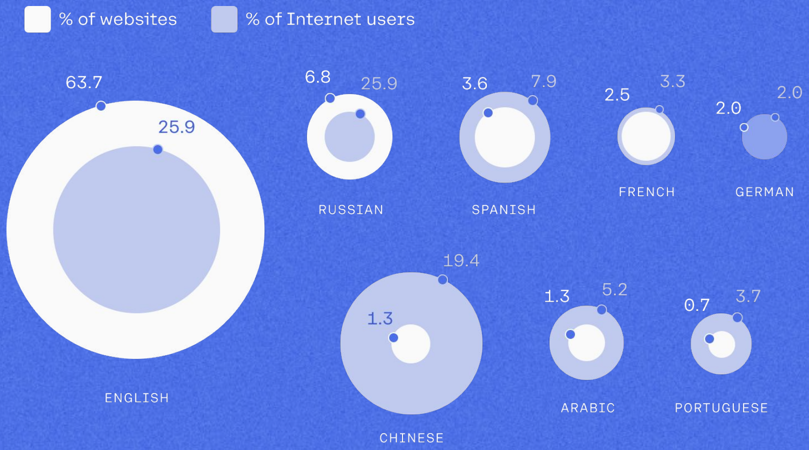
The impetus for this project stems from the stark reality that while natural language processing technologies have advanced exponentially, not all languages have been treated equally by developers and researchers. A significant drawback lies in the source of data used to train large language models, predominantly originating from the internet.

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	1000
Hausa	1.5	70
Nigerian Pidgin	0.4	30

Van Esch, et al. 2022. Writing System and Speaker Metadata for 2,800+ Language Varieties. In Proceedings of the *Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

English is the internet’s Universal language

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



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

A widening gap.

This mirrors the early adoption stage of this technology, where a mere 5% of the world's population speaks English at home, yet a surprising 63.7% of internet communication is in English. This trend inadvertently widens the gap in language access to new technologies, exacerbating disproportionate representation, and perpetuating this divide further.

Richter, F. (2022, February 21). English Is the Internet’s Universal Language. *Statista.* https://www.statista.com/ chart/ 26884/languages-on-the-internet/

Endurance and resourcefulness

The name Aya originates from the Twi language, meaning "fern," symbolizing endurance and resourcefulness – a perfect testament to the project's commitment to accelerating multilingual AI progress. What we didn’t realize when we named the project was how much endurance and resourcefulness we would need to pull it off.

Creating together

> “If you want to go fast,
> **go alone.**
>
> If you want to go far,
> **go together.**”
>
> – African Proverb

Aya has been the largest open-science project in the field of AI. Bringing together 3,000+ collaborators from 119 countries is no small feat. In addition to all the typical challenges of working in groups, we had to take into account time differences, language barriers, various culture understandings and resource inequity.

We hope our journey over a year will help serve as a case study for future participatory research initiatives. We share both the challenges as well as the unique advantages of working together on this mega-scale scientific initiative.

One step down a long road

The Aya model and dataset are now open source, inviting researchers and developers to build upon this progress and conduct further research and build tools to increase access for people in their communities.

By leveraging the Aya resources, you can contribute to the larger challenge of shifting the focus of technological development to encompass all communities and their unique languages.

Visit the Aya website

Together, we can create the future of AI advancement that benefits all.

Let us unite, collaborate, and unleash the full potential of open science for the betterment of global communication.

02
The People of Aya

The Frontiers of Participatory Research

Language is a deeply social phenomenon for its everyday users. It thrives on a network of social relations. However, there is no template or rulebook for working with 3000+ researchers and enthusiasts around the world. Instead, we kept in mind some guiding principles:

Whenever we engage with data, we are also engaging with the connections that data has to the people who produce it, prepare it, and distribute it.

Fluid Ownership and Growth

A decentralized model supports fluid leadership and flexible role adoption. It empowers members to take initiative independent of hierarchical position or level of involvement.

Organizational Structure

Asynchronous communication channels facilitate rich and timely collaborations.

Inclusion and Access

Bypass academic norms that often marginalize non-English speakers and people without formal academic credentials.

Participating motivators

Not based on financial remuneration but on ideals of community, identity, and social justice.

Cohere For AI

The Journey of Aya

Watch *The Journey of Aya*, a short documentary in which out collaborators tell the story of how Aya came to be.

Core team 1/2

Listed in alphabetical order.

The Core Team has been responsible for various technical elements of making Aya a reality. Their contributions varied across building an accessible user interface, establishing strong baselines, exploring data augmentation strategies, ensure responsible deployment, and coordinating regional contributions.

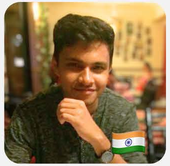
Aisha Alaagib
Cohere For AI Community

Emad A. Alghamdi
King Abdulaziz U
ASAS.AI

Zaid Alyafeai
King Fahd University of Petroleum and Minerals or KFUPM

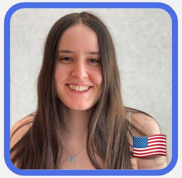
Viraat Aryabumi
Cohere For AI

Max Bartolo
Cohere


Neel Bhandari
Cohere For AI Community

Vu Minh Chien
Cohere For AI Community

Daniel D'souza
Cohere For AI Community

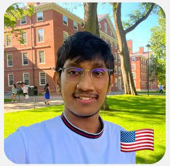

Irem Ergun
Cohere

Ellie Evans
Cohere For AI Community

Marzieh Fadaee
Cohere For AI

Hakimeh (Shafagh) Fadaei
Cohere For AI Community

Sebastian Gehrmann
Bloomberg LP


Ramith Hettiarachchi
MIT

Sara Hooker
Cohere For AI

Sarah Jafari
Cohere For AI

Börje Karlsson
Beijing Academy of Artificial Intelligence (BAAI)

Amr Kayid
Cohere

Farhan Khot

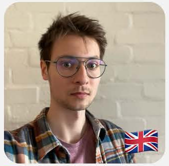
Wei-Yin Ko
Cohere

Julia Kreutzer
Cohere For AI

Core team 2/2

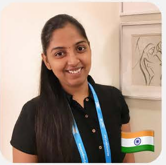
Listed in alphabetical order.

The Core Team has been responsible for various technical elements of making Aya a reality. Their contributions varied across building an accessible user interface, establishing strong baselines, exploring data augmentation strategies, ensure responsible deployment, and coordinating regional contributions.


Dominik Krzeminski
Cohere For AI Community

Shayne Longpre
MIT

Marina Machado
Cohere


Abinaya Mahendiran
Cohere For AI Community

Deividas Mataciunas
Cohere For AI Community

Oshan Mudannayake
Cohere For AI Community

Niklas Muennighoff
Cohere For AI Community

Laura O'Mahony
University of Limerick, Limerick, Ireland

Ifeoma Okoh
Cohere For AI Community

Gbemileke Onilude

Hui-lee Ooi
Cohere For AI Community

Jay Patel
Binghamton University, NY, USA

Herumb Shandilya
Cohere For AI Community

Shivalika Singh
Cohere For AI Community

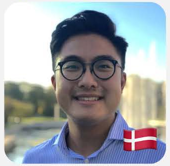
Madeline Smith
Cohere For AI

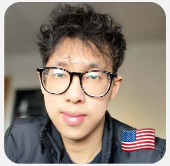
Luísa Souza Moura
Cohere

Ahmet Üstün
Cohere For AI

Freddie Vargus
Cohere For AI Community

Joseph Wilson
University of Toronto


Mike Zhang
IT University of Copenhagen


Yong Zheng Xin
Brown University
Cohere For AI Community

Language Ambassadors 1/3

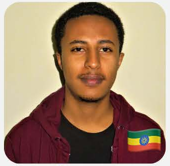
Listed in alphabetical order.

Language Ambassadors spread the word about Aya to speakers of their language, recruit new contributors, support those contributors to understand the goals of Aya data collection efforts, and celebrate progress.

Diana Abagyan
Russian

Muhammad Abdullahi
Somali

Elyanah Aco
Filipino



Henok Ademtew
Amharic

Adil
Kazakh

Emad A. Alghamdi
Arabic

Zaid Alyafeai
Arabic

Ahmad Anis
Urdu

Daniel Avila
Spanish

Michael Bayron
Cebuano

Nathanael Carraz Rakotonirina
Malagasy

Alberto Mario Ceballos Arroyo
Spanish

Yi Yi Chan Myae Win Shein
Burmese

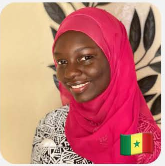
Vu Minh Chien
Vietnamese

Caroline Shamiso Chitongo
Zulu

Ionescu Cristian
Romanian

Ripal Darji
Gujarati

Suchandra Datta
Bengali



Rokhaya Diagne
Wolof

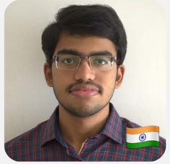
Irem Ergun
Turkish

Hakimeh (Shafagh) Fadaei
Persian

Language Ambassadors 2/3

Listed in alphabetical order.

Language Ambassadors spread the word about Aya to speakers of their language, recruit new contributors, support those contributors to understand the goals of Aya data collection efforts, and celebrate progress.


Surya Krishna Guthikonda
Telugu

Aleksandra Hadžić
Serbian

Shamsuddeen Hassan Muhammad
Hausa

Ramith Hettiarachchi
Sinhala

Mochamad Wahyu Hidayat
Sundanese

Rin Intachuen
Thai

Eldho Ittan George
Malayalam

Ganesh Jagadeesan
Hindi

Murat Jumashev
Kyrgyz

Börje Karlsson
Portuguese and Swedish

Abhinav Kashyap
Kannada

JiWoo Kim
Korean

Alkis Koudounas
Italian

Kevin Kudakwashe Murera
Shona

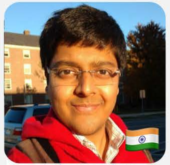
Falalu Ibrahim Lawan
Hausa

Wen-Ding Li
Traditional Chinese

Abinaya Mahendiran
Tamil

Mouhamadane Mboup
Wolof

Oleksander Medyuk
Ukrainian


Pratik Mehta
Hindi

Iftitahu Nimah
Javanese

Language Ambassadors 3/3

Listed in alphabetical order.

Language Ambassadors spread the word about Aya to speakers of their language, recruit new contributors, support those contributors to understand the goals of Aya data collection efforts, and celebrate progress.

Solam Nyangiwe
Xhosa

Laura O'Mahony
Irish

Ifeoma Okoh
Igbo

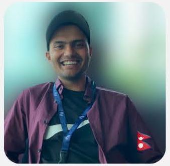
Hui-Lee Ooi
Malay

Iñigo Parra
Basque

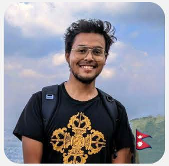
Jay Patel
Gujarati

Hanif Rahman
Pashto

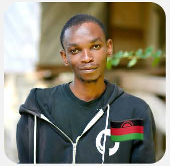
Olanrewaju Samuel
Yorùbá


Suman Sapkota
Nepali

Giacomo Sarchioni
Italian


Rashik Shrestha
Nepali

Bhavdeep Singh Sachdeva
Punjabi


Sean Andrew Thawe
Chichewa

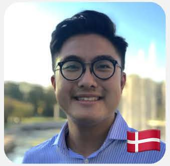
Alperen Ünlü
Turkish

Joseph Wilson
French

Emilia Wiśnios
Polish

Yang Xu
Simplified Chinese

Zheng-Xin Yong (Yong)
Malay


Mike Zhang
Dutch

Top 50 Quality Champions 1/2

Collaborators listed in ascending order based on Aya Quality Score.

These collaborators lead the way in ensuring the textual data contributed to Aya was of high quality including being free of grammatical errors, safe and factually correct, and robust completions to support model training.

- 🇻🇳 Vu Minh Chien
- 🇨🇦 Hui-Lee Ooi
- 🇱🇰 Gamage Omega Ishendra
- 🇮🇳 Surya Krishna Guthikonda
- 🇯🇵 Hoang Anh Quynh Nhu
- 🇳🇬 Moses Oyeleye
- 🇮🇳 Amarjit Singh Sachdeva
- 🇳🇱 Mike Zhang
- 🇰🇬 Almazbekov Bekmyrza Ruslanovich
- 🇸🇴 Ramla Abdullahi Mohamed
- 🇨🇳 Börje F. Karlsson
- 🇱🇰 Regina Sahani Lourdes De Silva Goonetilleke
- 🇸🇦 Zaid Alyafeai
- 🇺🇸 Yong Zheng Xin
- 🇹🇷 Yavuz Alp Sencer Öztürk
- 🇪🇬 Mohammed Hamdy
- 🇮🇳 Anitha Ranganathan
- 🇺🇸 Ramith Hettiarachchi
- 🇲🇾 Ooi Hui Yin
- 🇿🇼 Caroline Shamiso Chitongo
- 🇺🇸 Bhavdeep Singh Sachdeva
- 🇨🇭 Valentyn Bezshapkin

Top 50 Quality Champions 2/2

Collaborators listed in ascending order based on Aya Quality Score.

These collaborators lead the way in ensuring the textual data contributed to Aya was of high quality including being free of grammatical errors, safe and factually correct, and robust completions to support model training.

- 🇨🇦 Yang Xu
- 🇬🇧 Dominik Krzeminski
- 🇮🇩 Iftitahu Nimah
- 🇸🇴 Muna Mohamed Abdinur
- 🇰🇬 Nurbaeva Zhiidegul Talaibekovna
- 🇨🇦 Younes Bensassi Nour
- 🇮🇳 Eldho Ittan George
- 🇧🇷 Caio Dallaqua
- 🇮🇷 Hakimeh (Shafagh) Fadaei
- 🇪🇹 Henok Ademtew
- 🇮🇳 Vijayalakshmi Varadharajan
- 🇮🇳 Yogesh Haribhau Kulkarni
- 🇮🇪 Laura O'Mahony
- 🇺🇸 Jay Patel
- 🇧🇷 Luísa Souza Moura
- 🇵🇸 Rama Hasiba
- 🇲🇾 Geoh Zie Ee
- 🇬🇧 Gabriela Vilela Heimer
- 🇮🇳 Pratham Prafulbhai Savaliya
- 🇨🇭 Deividas Mataciunas
- 🇳🇬 Ifeoma Okoh
- 🇺🇸 Alberto Mario Ceballos Arroyo
- 🇫🇷 Basiiru Silla
- 🇬🇷 Yiorgos Tsalikidis

Dataset Champions

Collaborators listed in alphabetical order.

Aya Dataset Champions sourced, formatted and submitted open-source datasets in their languages to be included in the Aya collection.

- 🇺🇸 Diana Abagyan
- 🇪🇹 Henok Ademtew
- 🇵🇰 Ahmad Anis
- 🇮🇷 Hakimeh (Shafagh) Fadaei
- 🇳🇱 Hamidreza Ghader
- 🇧🇩 Md. Tahmid Hossain
- 🇮🇳 Eldho Ittan George
- 🇺🇸 Ganesh Jagadeesan
- 🇨🇳 Börje F. Karlsson
- 🇮🇳 Surya Krishna Guthikonda
- 🇮🇳 Abinaya Mahendiran
- 🇮🇳 Desik Mandava
- 🇮🇩 Iftitahu Nimah
- 🇹🇭 Wannaphong Phatthiyaphaibun
- 🇩🇰 Mike Zhang

5000 Contribution Points

Collaborators listed in descending order of most points earned.

These contributors achieved at least 5000 Contributions Points via the Aya data collection user interface.

- 🇳🇬 Moses Oyeleye
- 🇻🇳 Vu Minh Chien
- 🇸🇴 Ramla Abdullahi Mohamed
- 🇱🇰 Gamage Omega Ishendra
- 🇺🇸 Nitta Sitakrishna
- 🇮🇳 Surya Krishna Guthikonda
- 🇨🇦 Hui-Lee Ooi
- 🇯🇵 Hoang Anh Quynh Nhu
- 🇰🇬 Nurbaeva Zhiidegul Talaibekovna
- 🇸🇴 Muna Mohamed Abdinur
- 🇮🇳 Amarjit Singh Sachdeva
- 🇨🇦 Yang Xu
- 🇰🇬 Almazbekov Bekmyrza Ruslanovich
- 🇸🇴 Ahmed Mohamed Hussein Malin
- 🇺🇸 Bhavdeep Singh Sachdeva
- 🇺🇸 Yong Zheng Xin
- 🇹🇷 Yavuz Alp Sencer Öztürk
- 🇱🇰 Regina Sahani Lourdes De Silva Goonetilleke
- 🇮🇳 Yogesh Haribhau Kulkarni
- 🇸🇦 Zaid Alyafeai
- 🇨🇦 L N Deepak
- 🇿🇼 Caroline Shamiso Chitongo
- 🇨🇳 Börje F. Karlsson
- 🇨🇦 Younès Bensassi Nour

1000 Contribution Points 1/3

These contributors achieved at least 1000 Contributions Points via the Aya data collection user interface.

Contributors listed in descending order from most points.

- 🇮🇳 Sudharshini AJ
- 🇳🇬 Maryam Sabo Abubakar
- 🇮🇳 Mr. A. Karthik
- 🇳🇱 Mike Zhang
- 🇧🇷 Caio Dallaqua
- 🇸🇳 Rokhaya Diagne
- 🇮🇳 Anitha Ranganathan
- 🇮🇳 Eldho Ittan George
- 🇬🇧 Dominik Krzeminski
- 🇵🇸 Rama Hasiba
- 🇮🇳 Dev Haral
- 🇬🇧 Gabriela Vilela Heimer
- 🇧🇷 Júlia Souza Moura
- 🇮🇳 Suchandra Datta
- 🇮🇪 Laura O'Mahony
- 🇨🇭 Valentyn Bezshapkin
- 🇿🇼 Makomborero Magaya
- 🇵🇰 Taqi Haider
- 🇱🇰 R. A. Nirmal Sankalana
- 🇫🇷 Basiiru Silla
- 🇺🇸 Ramith Hettiarachchi
- 🇨🇦 Yat Kan Eden Cheung
- 🇩🇪 Sefika Efeoglu
- 🇸🇴 Abdishakuur Mohamed Hussein
- 🇮🇷 Hakimeh (Shafagh) Fadaei
- 🇧🇷 Luísa Souza Moura
- 🇺🇸 Iñigo Parra
- 🇲🇬 Razafindrakotonjatovo Zo Anjatiana Henitsoa Kokoly
- 🇰🇬 Aidaiym Omurbekovna
- 🇺🇸 Ripal Darji
- 🇮🇳 Mr. MARAPPAN .A
- 🇲🇬 NDIMBIARISOA Valdo Tsiaro Hasina
- 🇧🇷 Rafael Panisset Motta
- 🇺🇸 Jay Patel
- 🇰🇬 Zalkarbek Tilenbaev
- 🇺🇸 Meghana Denduluri
- 🇸🇳 Abdou Sall
- 🇪🇸 Nathanaël Carraz Rakotonirina
- 🇮🇳 Dr. Maharasan.K.S
- 🇮🇳 Khaleel Jageer
- 🇳🇬 Falalu Ibrahim Lawan
- 🇮🇩 Iftitahu Nimah
- 🇮🇳 Armeen Kaur Luthra

1000 Contribution Points 2/3

These contributors achieved at least 1000 Contributions Points via the Aya data collection user interface.

Contributors listed in descending order from most points.

- 🇵🇭 Elyanah Marie Aco
- 🇵🇰 Adeer Khan
- 🇸🇬 Ooi Hui Mei
- 🇨🇭 Deividas Mataciunas
- 🇪🇹 Betel Addisu
- 🇲🇬 Randriamanantena Manitra Luc
- 🇮🇳 K.Chinnaraju
- 🇸🇳 Mouhamadane Mboup
- 🇲🇬 Filamatra Manampy Fanantenana Rasolofoniaina
- 🇮🇳 Amandeep Singh
- 🇺🇸 Alberto Mario Ceballos Arroyo
- 🇲🇾 Geoh Zie Ee
- 🇲🇬 Andriatsalama Fiononantsoa Jaofera
- 🇲🇬 Tsaramanga Jeanny Fidelica
- 🇲🇼 Sean Andrew Thawe
- 🇲🇬 Ratsimba Ranto Sarobidy
- 🇮🇳 Srinadh Vura
- 🇩🇿 Benmeridja Ahmed Younes
- 🇪🇹 Elshaday Desalegn Asfaw
- 🇧🇩 Md. Tahmid Hossain
- 🇪🇹 Henok Ademtew
- 🇳🇬 Mohammed Nasiru
- 🇪🇸 Harena Finaritra Ranaivoarison
- 🇺🇸 Mansi Kamlesh Patel
- 🇧🇷 Marina Fontes Alcântara Machado
- 🇲🇬 Tahina Mahatoky
- 🇲🇬 Ramarozatovomampionona Todisoa Nirina Mickael
- 🇧🇷 Ana Carolina Correia Pierote
- 🇰🇬 Ainura Nurueva
- 🇮🇪 Hollie O'Shea
- 🇹🇭 Wannaphong Phatthiyaphaibun
- 🇳🇬 Abubakr Labaran Salisu
- 🇲🇾 Ooi Hui Yin
- 🇲🇬 RAKOTONIRINA Tokinantenaina Mathieu Razokiny
- 🇧🇷 Robinson Rodrigo Silva Oliveira
- 🇬🇧 Hanif Rahman
- 🇲🇬 Maminirina Rahenintsoa

1000 Contribution Points 3/3

These contributors achieved at least 1000 Contributions Points via the Aya data collection user interface.

Contributors listed in descending order from most points.

- 🇮🇳 Krishna Chhatbar
- 🇮🇳 J.Nirmala
- 🇱🇰 Tharin Edirisinghe
- 🇲🇬 Randrianarison Diarintsoa Fandresena No HerijaonaHerijaona
- 🇲🇬 Andrianarivony Harijaona Fanirintsoa
- 🇲🇬 Rakotondrainibe Nirisoa Tendry
- 🇰🇬 Bekbolot Abdirasulov
- 🇬🇧 Joseph Marvin Imperial
- 🇳🇬 Ifeoma Okoh
- 🇳🇵 Sumi Shakya
- 🇮🇹 Alkis Koudounas
- 🇺🇸 Mohamad Aboufoul
- 🇸🇦 Emad A. Alghamdi
- 🇮🇳 Jothika. S
- 🇲🇬 Razakahasina Fanomezana Sarobidy
- 🇧🇷 Valério Viégas Wittler
- 🇳🇵 Anish Gasi Shrestha
- 🇨🇦 Joseph Wilson
- 🇳🇬 Ijeoma Irene Okoh
- 🇳🇬 Ajayi Akinloluwa Irawomitan
- 🇰🇬 Zarlykov Kelsinbek
- 🇨🇦 Micol Altomare
- 🇮🇳 Yadnyesh Chakane
- 🇲🇬 Rafidy Julie Tassia
- 🇳🇵 Rabin Adhikari
- 🇳🇬 Chinwendu Peace Anyanwu
- 🇮🇳 Dr. S.P. Balamurugan
- 🇱🇰 G. A. Jalina Hirushan Gunathunga
- 🇳🇬 Ogba Stephen Kesandu
- 🇲🇬 Tiana Kaleba Andriamanaja
- 🇲🇬 Andriamiadanjato Mioraniaina

500 Contribution Points

These contributors achieved at least 500 Contributions Points via the Aya data collection user interface.

Contributors listed in descending order from most points.

- 🇮🇳 M.Neelavathi
- 🇳🇵 Sabita Rajbanshi
- 🇮🇳 Silambarasan U.
- 🇮🇳 Dr.A.Prasanth
- 🇧🇷 Sara Salvador
- 🇮🇳 Dr A.Jeba Christy
- 🇮🇳 Mr.V.Balakrishnan
- 🇮🇳 Abinaya Mahendiran
- 🇿🇦 Solam
- 🇳🇵 Rashik Shrestha
- 🇮🇳 Easwaran K
- 🇵🇰 Ahmad Mustafa Anis
- 🇮🇳 Dr.G.Thilagar
- 🇲🇾 Gan Chin Chin
- 🇮🇳 Bhanu Prakash Doppalapudi
- 🇸🇴 Abdullahi Adan Hassan
- 🇺🇸 Sara Hooker
- 🇾🇪 Amjad Abdulkhaliq Alkhatabi
- 🇲🇾 Muhamad Audi Bin Pasha
- 🇲🇽 Santiago Pedroza Díaz
- 🇫🇷 Siyu Wang
- 🇱🇰 Randinu Jayaratne
- 🇱🇰 Rithara Kithmanthie
- 🇮🇳 Bhanu Prakash Doppalapudi
- 🇳🇵 TSuman Sapkota
- 🇱🇰 Charindu Abeysekara
- 🇲🇾 Afifah binti Mohd Shamsuddin
- 🇲🇾 Verassree Rajaratnam
- 🇵🇹 Ruqayya Nasir Iro
- 🇮🇳 Geetharamani R.
- 🇳🇵 Sandesh Pokhrel
- 🇰🇬 Orozbai Topchubek uulu
- 🇮🇳 Prajapati Maitri R.
- 🇵🇹 Francisco Valente
- 🇳🇵 Gaurav Jyakhwa
- 🇮🇳 Mrs. G. Sangeetha
- 🇹🇷 Ahmet Güneyli

Public Release and Engineering Team 1/2

Collaborators listed in alphabetical order.

The public release team is responsible for bringing Aya to the world. From building and deployment of the model, planning the launch event, creating *The Journey of Aya* documentary, hosting the model and coordinating outreach efforts.

- 🇮🇳 Viraat Aryabumi
- 🇺🇸 Saurabh Baji
- 🇬🇧 Max Bartolo
- 🇨🇦 Claude Beaupré
- 🇬🇧 Phil Blunsom
- 🇪🇸 Tomeu Cabot
- 🇨🇦 Isabelle Camp
- 🇺🇸 Jon Ander Campos
- 🇨🇦 Claire Cheng
- 🇨🇦 Linus Chui
- 🇺🇸 Jenna Cook
- 🇨🇦 Natasha Deichmann
- 🇺🇸 Roy Eldar
- 🇺🇸 Irem Ergun
- 🇩🇪 Beyza Ermis
- 🇳🇱 Marzieh Fadaee
- 🇨🇦 Ramy Farid
- 🇨🇦 Nick Frosst
- 🇺🇸 Josh Gartner
- 🇨🇦 Aidan Gomez
- 🇺🇸 Manoj Govindassamy
- 🇨🇦 Rod Hajjar
- 🇺🇸 Sara Hooker
- 🇨🇦 Monica Iyer
- 🇨🇦 Sarah Jafari
- 🇨🇦 Amr Kayid
- 🇨🇦 Julia Kedrzycki
- 🇺🇸 Wei-Yin Ko

Public Release and Engineering Team 1/2

Collaborators listed in alphabetical order.

The public release team is responsible for bringing Aya to the world. From building and deployment of the model, planning the launch event, creating *The Journey of Aya* documentary, hosting the model and coordinating outreach efforts.

- 🇨🇦 Martin Kon
- 🇺🇸 Dave Kong
- 🇨🇦 Julia Kreutzer
- 🇺🇸 Kyle Lastovica
- 🇺🇸 Tali Livni
- 🇧🇷 Marina Machado
- 🇨🇦 Abigail Mackenzie-Armes
- 🇨🇦 Kim Moir
- 🇬🇧 Luísa Moura
- 🇨🇦 Alyssa Pothier
- 🇺🇸 Brittawnya Prince
- 🇨🇦 Daniel Quainoo
- 🇺🇸 Jess Rosenthal
- 🇺🇸 Sudip Roy
- 🇩🇪 Sebastian Ruder
- 🇬🇧 Astrid Sandoval
- 🇨🇦 Shubham Shukla
- 🇨🇦 Madeline Smith
- 🇨🇦 Trish Starostina
- 🇺🇸 Kate Svetlakova
- 🇨🇦 Chris Taeyoung Kim
- 🇺🇸 Yi Chern Tan
- 🇳🇱 Ahmet Üstün
- 🇨🇦 Jaron Waldman
- 🇨🇦 Donglu Wang
- 🇨🇦 Lauren Waters
- 🇨🇦 Ivan Zhang

Safety Evaluation

Our multilingual human evaluation annotators help us understand model quality across languages. They support our evaluations of where models differ and uncover safety and quality issues.

Faraaz Ahmed

April Alcantara

Kirill Borisov

Owen Chung

Laura De Vuono

Sama Elhansi

Sonja Gavric

Marwan Genena

Robin Gershman

Stuti Govil

Bruno Guratti

Maryam Helmy

Ricardo Joaquin Hornedo Aldeco

Nishi Jain

Milica Jez

Dina Kliuchareva

Finlay Korol-O'Dwyer

Rachel Lo

Juan Lozano

Arishi Maisara

Brenda Malacara

Annika Maldonado

Simar Malhan

Jullia Naag

Sasha O'Marra

Uros Popic

Naeesha Puri

Elina Qureshi

Alizé Qureshi

Manuela Ramirez Naranjo

Boris Sehovac

Ankit Sharma

Hana Sherafati Zanganeh

Ambuj Upadhyay

Susheela Willis

Linda Yanes

Joanna Yulo

Partner Organizations

These organizations supported Aya by hosting events, providing resources, and/or spreading awareness of the project, thereby facilitating contributions and boosting language inclusion efforts.

Universiti Malaysia Sarawak

Faculty of Computer Science and Information Technology

GalsenAI

SIMAD iLab

Google Developer Student Clubs

Thapar Institute of Engineering and Technology, Patiala, under the leadership of Siya Sindhani

Google Developer Student Club

P P Savani University, Surat, Gujarat

KG College of Arts and Science

Coimbatore

Linguistics Circle

Nigeria

Rotaract Club

University of Moratuwa, Sri Lanka, led by Nawoda Thathsarani, Jalina Hirushan and Chamod Perera

Tensorflow

User Group Surat, Gujarat

03

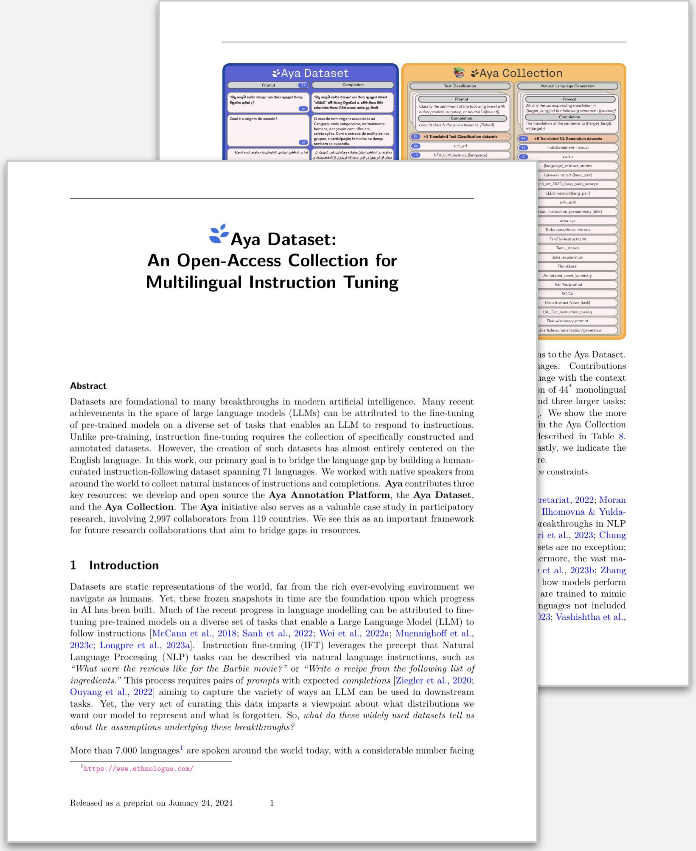
Aya Dataset & Collection

Aya Dataset

An Open-Access Collection for Multilingual Instruction Fine-Tuning

The Aya Dataset represents the most extensive compilation of multilingual instructional examples to date, and it is accessible for use under a fully permissive licensing framework.

For the full paper, read here.



ns to the Aya Dataset.
uages. Contributions
uage with the context
n of 44* monolingual
nd three larger tasks:
. We show the more
in the Aya Collection
described in Table 8.
astly, we indicate the
re.
e constraints.

retariat, 2022; Moran
Ilhomovna & Yulda-
reakthroughs in NLP
ri et al., 2023; Chung
sets are no exception;
ermore, the vast ma-
et al., 2023b; Zhang
how models perform
are trained to mimic
nguages not included
023; Vashishtha et al.,

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

Abstract

Datasets are foundational to many breakthroughs in modern artificial intelligence. Many recent achievements in the space of large language models (LLMs) can be attributed to the fine-tuning of pre-trained models on a diverse set of tasks that enables an LLM to respond to instructions. Unlike pre-training, instruction fine-tuning requires the collection of specifically constructed and annotated datasets. However, the creation of such datasets has almost entirely centered on the English language. In this work, our primary goal is to bridge the language gap by building a human-curated instruction-following dataset spanning 71 languages. We worked with native speakers from around the world to collect natural instances of instructions and completions. **Aya** contributes three key resources: we develop and open source the **Aya Annotation Platform**, the **Aya Dataset**, and the **Aya Collection**. The **Aya** initiative also serves as a valuable case study in participatory research, involving 2,997 collaborators from 119 countries. We see this as an important framework for future research collaborations that aim to bridge gaps in resources.

1 Introduction

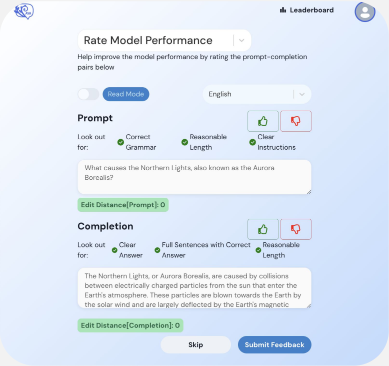
Datasets are static representations of the world, far from the rich ever-evolving environment we navigate as humans. Yet, these frozen snapshots in time are the foundation upon which progress in AI has been built. Much of the recent progress in language modelling can be attributed to fine-tuning pre-trained models on a diverse set of tasks that enable a Large Language Model (LLM) to follow instructions [McCann et al., 2018; Sanh et al., 2022; Wei et al., 2022a; Muennighoff et al., 2023c; Longpre et al., 2023a]. Instruction fine-tuning (IFT) leverages the precept that Natural Language Processing (NLP) tasks can be described via natural language instructions, such as *“What were the reviews like for the Barbie movie?”* or *“Write a recipe from the following list of ingredients.”* This process requires pairs of *prompts* with expected *completions* [Ziegler et al., 2020; Ouyang et al., 2022] aiming to capture the variety of ways an LLM can be used in downstream tasks. Yet, the very act of curating this data imparts a viewpoint about what distributions we want our model to represent and what is forgotten. So, *what do these widely used datasets tell us about the assumptions underlying these breakthroughs?*

More than 7,000 languages[1] are spoken around the world today, with a considerable number facing

[1] https://www.ethnologue.com/

Released as a preprint on January 24, 2024 1

Aya contributes four key resources:



Leaderboard

Rate Model Performance

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

Read Mode · English

Prompt

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

What causes the Northern Lights, also known as the Aurora Borealis?

Edit Distance[Prompt]: 0

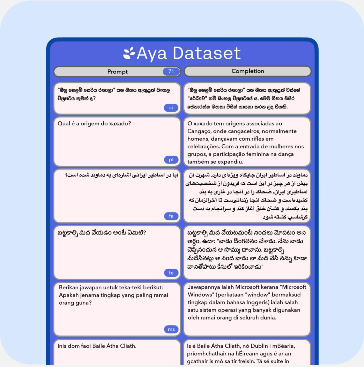
Completion

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

The Northern Lights, or Aurora Borealis, are caused by collisions between electrically charged particles from the sun that enter the Earth's atmosphere. These particles are blown towards the Earth by the solar wind and are largely deflected by the Earth's magnetic

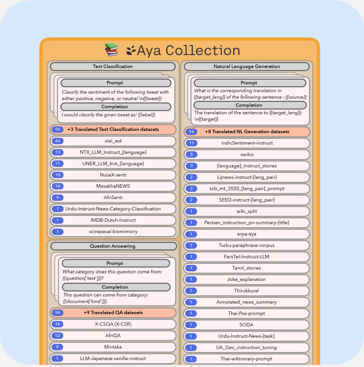
Edit Distance[Completion]: 0

Skip · Submit Feedback



Aya Dataset

Prompt (71)	Completion
"ඕලු නෙලුම් නෙරිය රඟාලා" යන ගීතය ඇතුළත් සිංහල චිත්‍රපටය කුමක් ද? si	"ඕලු නෙලුම් නෙරිය රඟාලා" යන ගීතය ඇතුළත් වන්නේ "රේඛාව" නම් සිංහල චිත්‍රපටයේ ය. මෙම ගීතය සිසිර සේනාරත්න මහතා විසින් ගායනා කරන ලද ගීයකි.
Qual é a origem do xaxado? pt	O xaxado 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.
آیا در اساطیر ایرانی اشاره‌ای به دماوند شده است؟ fa	دماوند در اساطیر ایران جایگاه ویژه‌ای دارد. شهرت آن بیش از هر چیز در این است که فریدون از شخصیت‌های اساطیری ایران، ضحاک را در آنجا در غاری به بند کشیده‌است و ضحاک آنجا زندانی‌ست تا آخرالزمان که بند بگسلد و کشتن خلق آغاز کند و سرانجام به دست گرشاسپ کشته شود
బట్టకాల్చి మీద వేయడం అంటే ఏమిటి? te	బట్టకాల్చి మీద వేయటమంటే నిందలు మోపటం అని అర్థం. ఉదా: "వాడు దొంగతనం చేశాడు. నేను వాడు చెప్పినందున ఆ సొమ్ము దాచాను. బట్టకాల్చి మీదేసినట్లు ఆ నింద వాడు నా మీద వేసి నన్ను కూడా వానితోపాటు కేసులో ఇరికించాడు"
Berikan jawapan untuk teka-teki berikut: Apakah jenama tingkap yang paling ramai orang guna? ms	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.
Inis dom faoi Baile Átha Cliath.	Is é Baile Átha Cliath, nó Dublin i mBéarla, príomhchathair na hÉireann agus é ar an gcathair is mó sa tír freisin. Tá sé suite in



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}}*

94	+3 Translated Text Classification datasets
44	xlel_wd
13	NTX_LLM_Instruct_{language}
11	UNER_LLM_Inst_{language}
10	NusaX-senti
10	MasakhaNEWS
9	AfriSenti
1	Urdu-Instruct-News-Category-Classification
1	IMDB-Dutch-Instruct
1	scirepeval biomimicry

Question Answering

Prompt: *What category does this question come from: {{question['text']}}?*

Completion: *This question can come from category: {{document['kind']}}.*

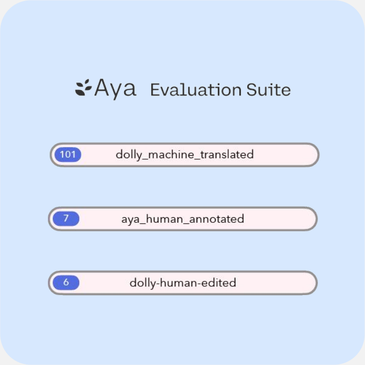
94	+9 Translated QA datasets
16	X-CSQA (X-CSR)
12	AfriQA
9	Mintaka
1	LLM-Japanese-vanilla-instruct

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}}

94	+8 Translated NL Generation datasets
11	IndicSentiment-instruct
5	xwikis
3	{language}_instruct_stories
2	Lijnews-instruct-{lang_pair}
2	scb_mt_2020_{lang_pair}_prompt
2	SEED-instruct-{lang_pair}
1	wiki_split
1	Persian_instruction_pn-summary-{title}
1	arpa-aya
1	Turku-paraphrase-corpus
1	FarsTail-Instruct-LLM
1	Tamil_stories
1	Joke_explanation
1	Thirukkural
1	Annotated_news_summary
1	Thai-Pos-prompt
1	SODA
1	Urdu-Instruct-News-{task}
1	UA_Gec_instruction_tuning
1	Thai-wiktionary-prompt



Aya Evaluation Suite

101	dolly_machine_translated
7	aya_human_annotated
6	dolly-human-edited

Aya Annotation Platform

An user interface for large-scale participatory research available for free. Used by **2,997 Aya contributors**

Aya Dataset

The largest human-annotated, multilingual dataset supporting **65 languages**

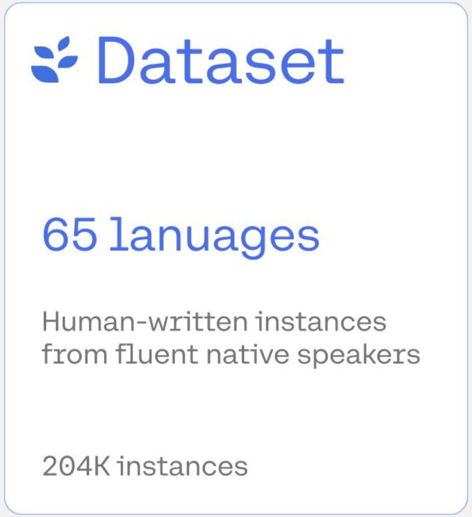
Aya Collection

A collection of **44 templated and 19 translated datasets, supporting 115 languages,** to train multilingual LLMs

Aya Evaluation Suite

A high quality dataset for evaluation of LLMs. Subsets include **human-written (7 languages), post-edited translations (6 languages)**, and translations of manually selected prompts **(101 languages)**

Aya Datasets at a glance



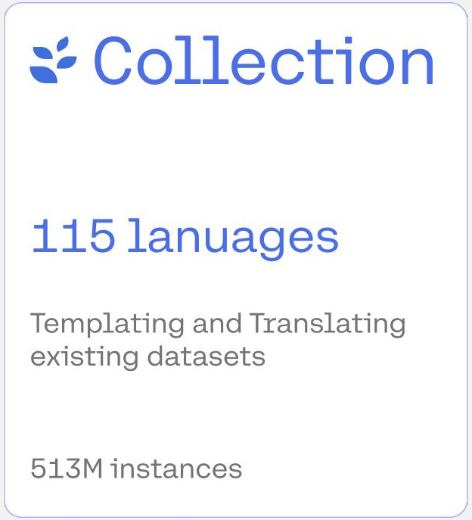
Dataset

65 lanuages

Human-written instances from fluent native speakers

204K instances

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



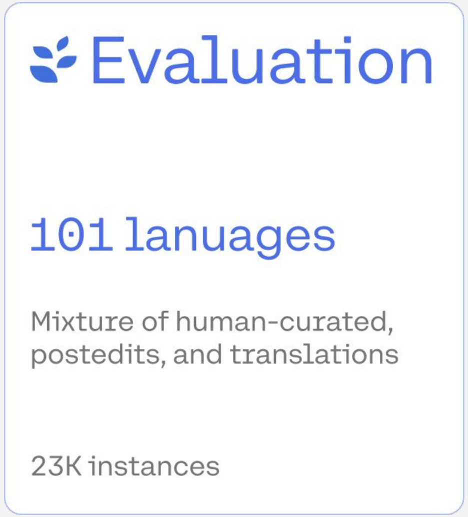
Collection

115 lanuages

Templating and Translating existing datasets

513M instances

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



Evaluation

101 lanuages

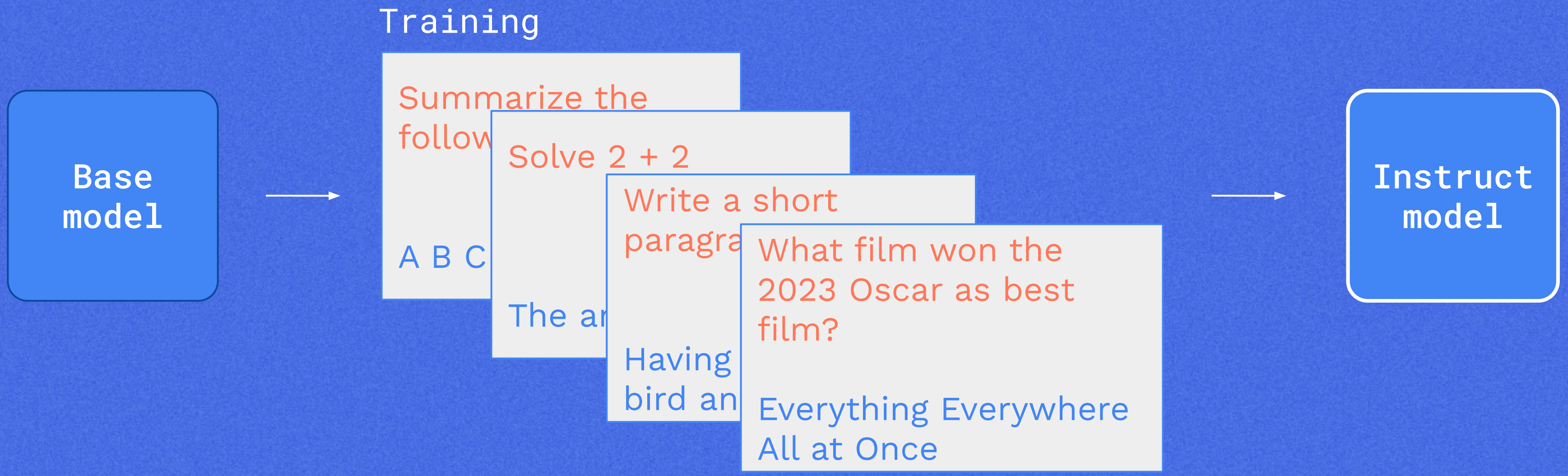
Mixture of human-curated, postedits, and translations

23K instances

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

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.



Challenges With Multilingual Data Quality and Coverage

To effectively train foundational models with multilingual instructions, we need access to large volumes of quality multilingual instructional data.

This has been plagued by three challenges:

- 🫙 Data scarcity
- 📉 Low quality data
- 👀 Lack of qualified contributors for low-resource languages

Without robust multilingual datasets to train models, we risk:

Introducing biases towards languages not included.

Marginalizing speakers of languages not included.

Creating a performance-divide for languages with limited datasets.

Introducing security flaws.

The Aya Dataset

The Aya Collection

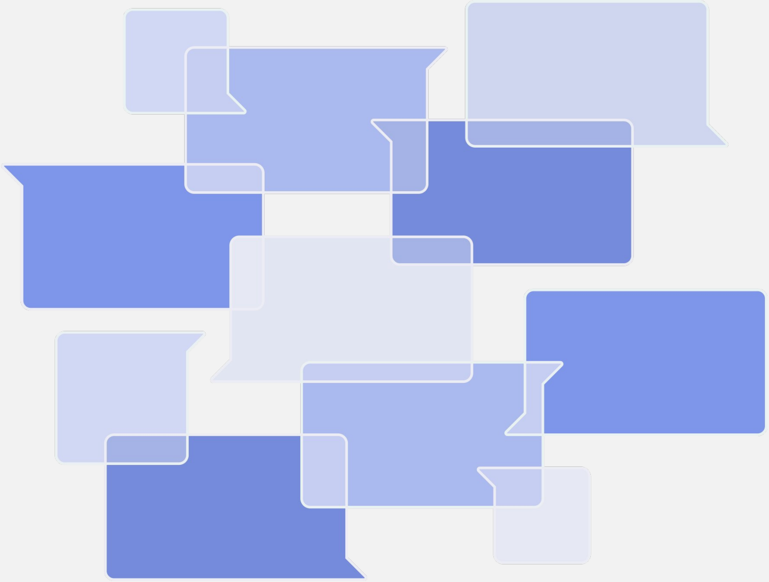
The Aya Evaluation Suite

The largest human-curated multilingual dataset for finetuning LLMs to follow instructions.

The Largest Human-Curated Dataset from Native and Fluent Speakers

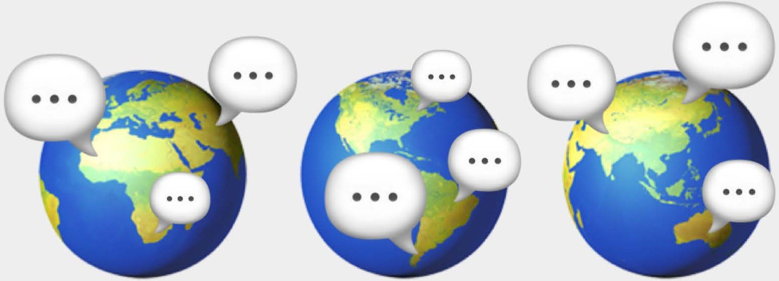
Human-curated data from native and fluent speakers can be hard to come by. It can be costly and difficult to orchestrate.

By leveraging best practices from open-source and crowdsourced science projects, we were able to create the Aya Dataset – the largest collection to date of human-curated and annotated multilingual instruction data.



Aiming for Worldwide Coverage of Languages

Behind each datapoint for each language is a person familiar with the nuances of the language. This level of expertise provides the subtle distinctions and variations in meaning that make each language unique in practice.



Criteria for Inclusion in Aya Dataset

The **Aya Dataset** includes all original annotations and a subset of all re-annotations that vary to a certain extent from the originals.

In order to ensure linguistic diversity and quality, we included languages that were varied, with at least 50 contributions, and with naturally long prompts and corresponding completions.

65 languages

33 high-resource

12 mid-resource

31 low-resource languages

The goal was to include as many languages as possible without lowering the overall quality of the dataset. The table below lists details of the **Aya** Dataset.

Aya Dataset Statistics (number of pairs of prompts and completions obtained through various annotation tasks)

		Count
Original Annotations		138,844
Re-Annotations	xP3 datasets	2,895
	Translated datasets	7,757
	Templated datasets	11,013
	Original Annotations	43,641
Aya Dataset Total		**204,114**

The Aya Dataset

The Aya Collection

The Aya Evaluation Suite

A combination of human-annotated, translated, and templated data.

An Overview of the Aya Collection

How do we make the world's largest multilingual instruction dataset?

Human Annotated

Human-annotated data is information that has been manually reviewed, labelled, and/or annotated by human annotators, leveraging their native knowledge of a language to provide context and enhance machine learning algorithms.

Translated

Translated multilingual data is when machine translation tools convert text from one language to another, making use of an existing dataset in one language to create the set in another.

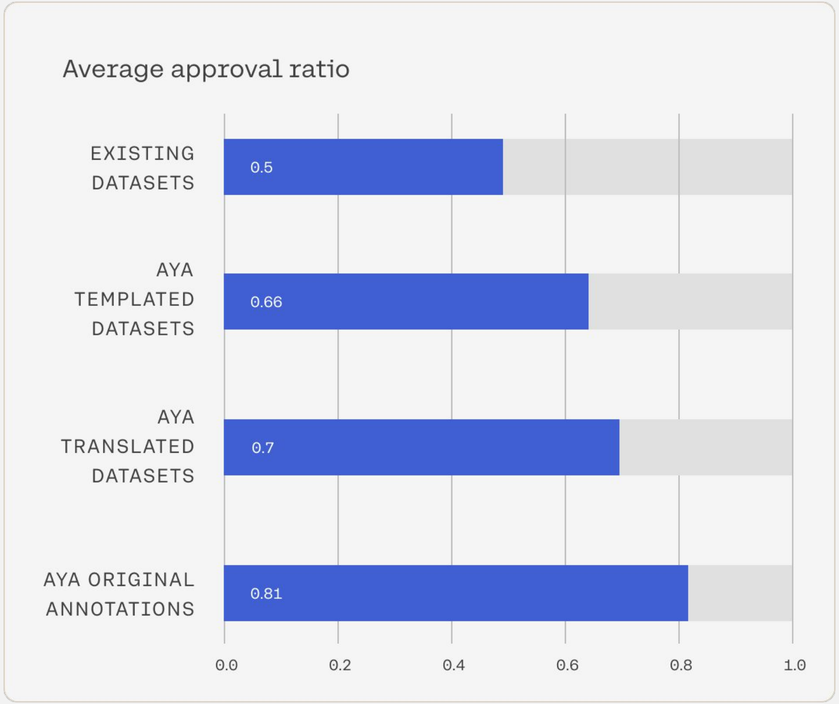
Templated

Templated is created by annotators writing templates and then applying them to datasets to reformat existing NLP datasets into instruction-style.

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.



Expanding Data Diversity and Task Coverage

Increasing diversity while maintaining high quality will result in more robust and powerful [1, 2]

We focused on existing datasets templated for instructions and finding tasks that require asking questions and answering based on small pieces of information.

The collection includes 3 main tasks,

1) Question Answering
2) Natural Language Generation
3) Text Classification

and 12 fine-grained task types.

Task Taxonomy of NLP tasks in the Aya Collection

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

The Aya Dataset

The Aya Collection

The Aya Evaluation Suite

a diverse multilingual dataset to assess open-ended generation capabilities of LLMs

Building an Evaluation Suite

We curate and release an evaluation suite tailored for multilingual models.

This set is a valuable contribution in tackling the scarcity of multilingual data, a challenge that becomes even more apparent when considering evaluation sets.

To strike a balance between language coverage and the quality that comes with human oversight, we create an evaluation suite that includes:

(1) **human-curated** examples in a limited set of languages,

(2) automatic **translations** of handpicked examples in an extensive number of languages, and

(3) **human-post-edited** translations in a few languages.

Human-curated examples

7 languages

1750 instances

Translations of hand-picked examples from Dolly-15k

101 languages

20K instances

Human-post-edited translations

6 languages

1200 instances

Limitations of the Aya Dataset

All research has limitations. Below we outline the top challenges faced by the Aya project and results.

- 💬 **Language and dialect coverage**: 115 languages (Aya Dataset and Aya Collection) is only a tiny fraction of the world's linguistic diversity.
- 📊 **Uneven distribution of contributions**: Relatively few contributors accounted for the most annotations.
- ⚖️ **Cultural or personal bias**: limited representation can lead to a narrow selection of cultural viewpoints.
- ⚥ **Gendered pronouns**: featuring languages with gendered pronouns or lacking gender-neutral ones, requires careful response crafting to maintain gender neutrality.
- 🤵 **Formality distinctions**: released dataset contains many languages that have varying levels of standardization and differing style guidelines for formal language like honorifics.
- ⚠️ **Toxic or offensive speech:** the annotation platform does not contain specific flags for toxic, harmful, or offensive speech, so it is possible that malicious users could submit unsafe data.
- 🏷️ **Accounting for mislabeled data**: the annotation platform does not contain any components that enable re-labeling the assigned language of annotations.
- 📌 **Coverage of tasks in Aya Collection:** the collection only includes 3 main tasks (Question Answering, Natural Language Generation, Text Classification) and 12 fine-grained task types.

04
Aya Model

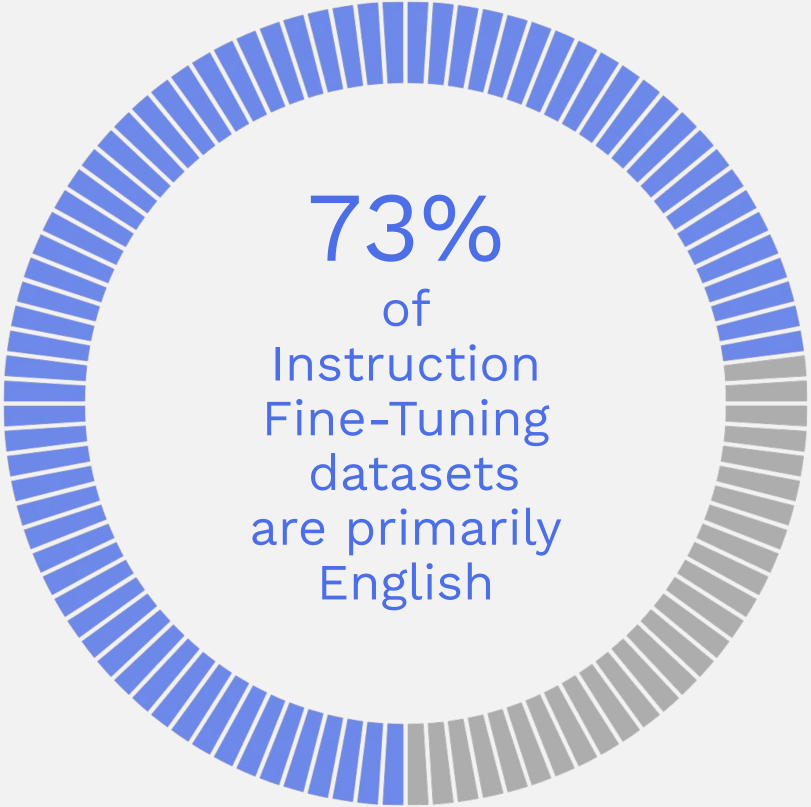
Introducing the Aya Model

The landscape of modern machine learning has been profoundly shaped by datasets. Yet, this progress has predominantly favored a few data-rich languages due to legacy use and lack of accessible resources. The global linguistic diversity is not represented.

This skew contrasts sharply with a core machine learning principle: **training data should mirror the real-world's vast linguistic diversity**.

We face a glaring inclusivity gap.

> “The limits of my **language** means the limits of my **world.** ”
>
> – Ludwig Wittgenstein



73% of Instruction Fine-Tuning datasets are primarily English

The Aya Model aims to bridge this divide, pushing for multilingual IFT datasets that truly reflect our world's rich tapestry of languages, making machine learning not just smarter, but more equitable and representative.

Prompt:

What are some languages spoken in Mexico?

Output:

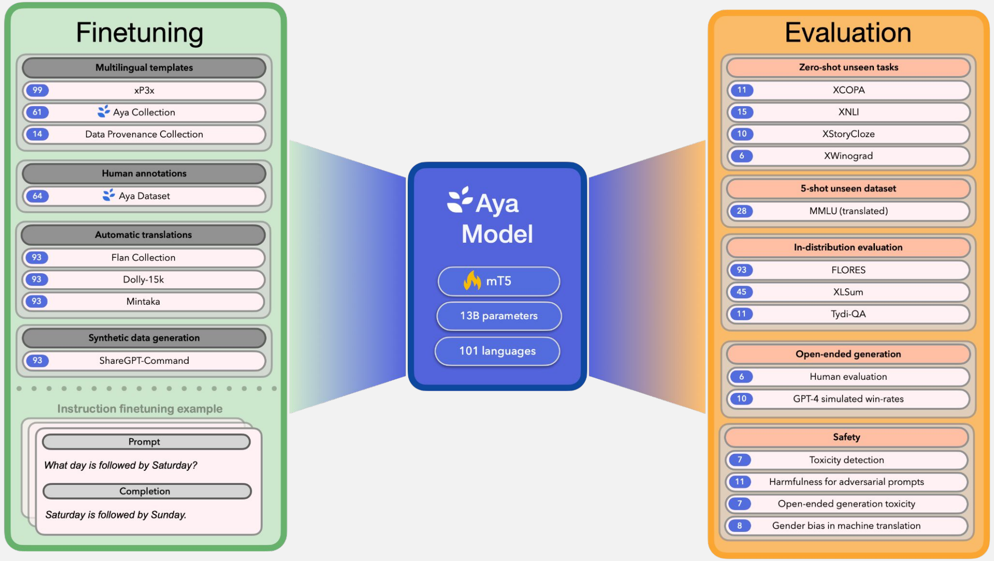
The three most spoken languages in Mexico are Spanish, Nahuatl, and Maya.

Instruction Fine-Tuning involves training a foundational model on a dataset of prompts or instructions followed by the desired outputs.

The Aya Model Explained

The Aya Model is designed to tackle linguistic inequality. It can execute tasks in response to prompts given in any supported language. This eliminates the need for multilingual speakers to default to English when writing prompts.

Our goal is to greatly expand the coverage of languages to 101, far beyond the current coverage of previous instruction fine-tuned multilingual models.



Finetuning

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
- mT5
- 13B parameters
- 101 languages

Evaluation

Zero-shot unseen tasks
- 11 XCOPA
- 15 XNLI
- 10 XStoryCloze
- 6 XWinograd

5-shot unseen dataset
- 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

Figure 2: Aya involved extensive contributions to both the breadth of IFT training dataset, optimization techniques including weighting of datasets and introducing more extensive evaluation of performance across varied tasks.

Representing Linguistic Diversity

To create a model with diverse linguistic representation, we focused on four areas:

Expansion of Language Coverage

We more than doubled the number of languages with 2.5x the size of the starting dataset.

Broadening Multilingual Evaluation

We benchmark on 99 languages with 4 different evaluation categories using 10 datasets.

Leading Multilingual Performance

The Aya Model consistently outperforms various baselines across all multilingual benchmarks.

Safety

We evaluate our model for gender bias, social bias, harmfulness, and toxicity across languages.

Recipe for a State-Of-The-Art Multilingual Model

🪛

We fine-tune pretrained multilingual T5 (mT5) language model using instructions in 101 languages

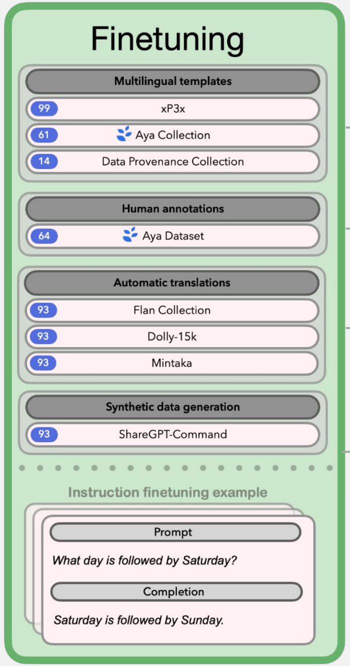
✂️

We carefully select data sources and further prune them to have high quality and diverse set of instruction datasets

⚖️

We balance different data sources during fine-tuning, resulting in high performance across several category of tasks

Building a Massively Multilingual and Diverse Instruction Fine-tuning Mixture



Finetuning

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.

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

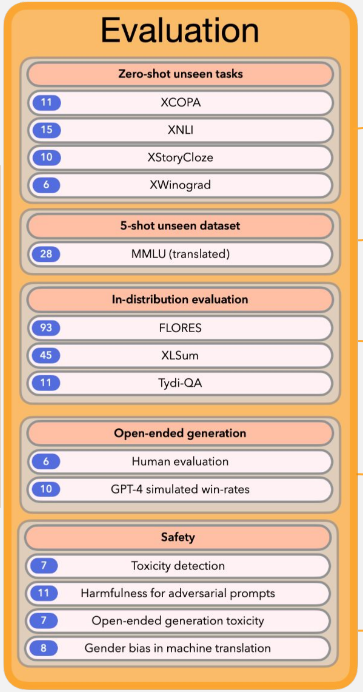
203 million examples

Aya Dataset, a fully human-curated dataset of instructions

Machine translated datasets into 93 languages

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

Creating a Massively Multilingual Evaluation Suite



Evaluation

Zero-shot unseen tasks	
11	XCOPA
15	XNLI
10	XStoryCloze
6	XWinograd

5-shot unseen dataset	
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

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:

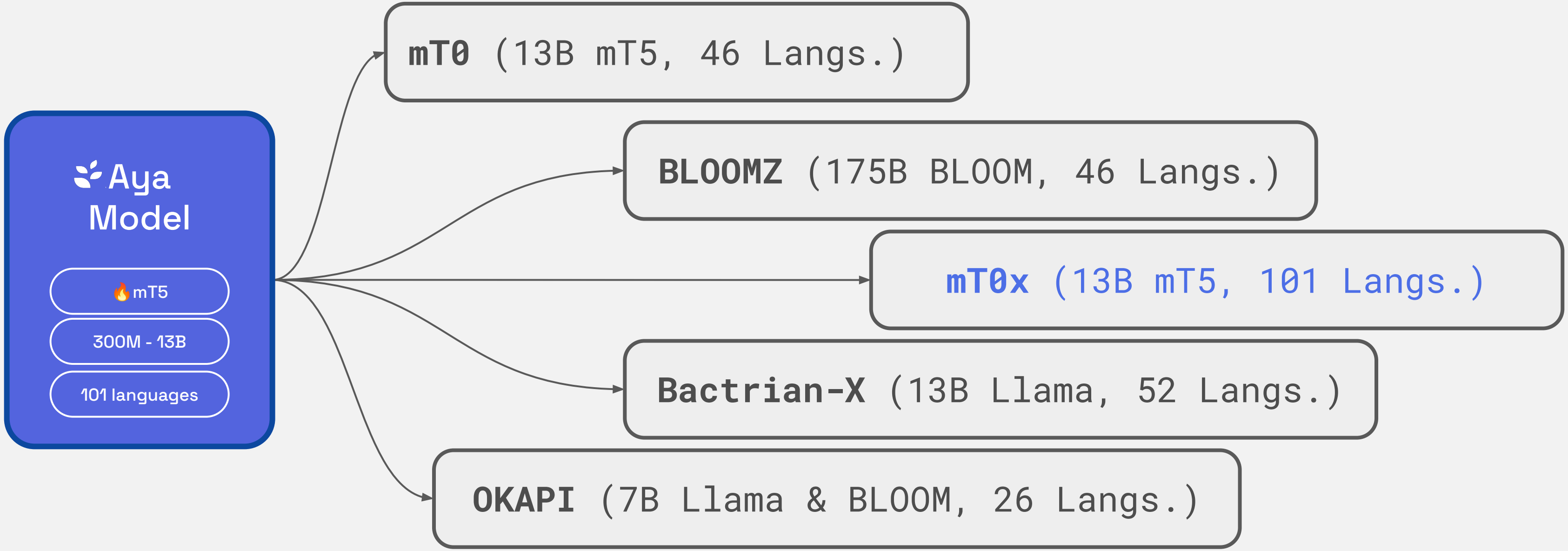
99 languages

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

Aya Model Compared With Multiple Baselines



Aya Model

🔥mT5

300M - 13B

101 languages

mT0 (13B mT5, 46 Langs.)

BLOOMZ (175B BLOOM, 46 Langs.)

mT0x (13B mT5, 101 Langs.)

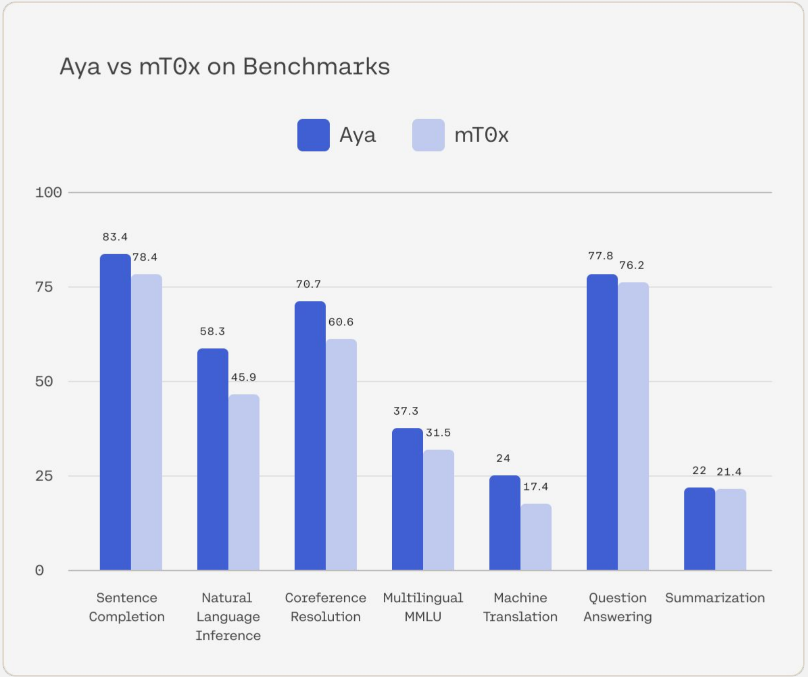
Bactrian-X (13B Llama, 52 Langs.)

OKAPI (7B Llama & BLOOM, 26 Langs.)

Leading Multilingual Performance

The Aya Model achieves superior performance compared to mT0x in the multilingual benchmarks.

These benchmarks include a collection of unseen tasks and in-distribution generative tasks in total covering 100 languages. The Aya model outperforms mT0x in all tasks showing its multilingual capabilities in different task types.

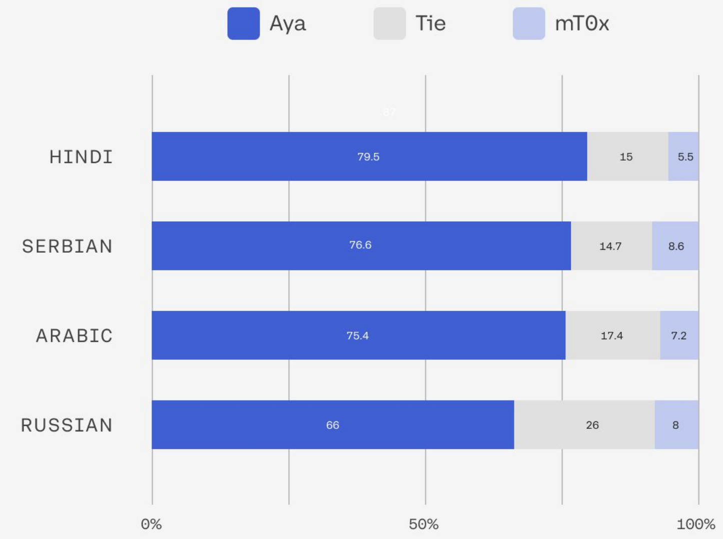


The Aya Model Win Rates

The Aya Model follows instructions and generates responses of significantly higher quality than mT0x.

According to the human evaluation where the professional annotators compared models' responses for given instructions in multiple languages, **the Aya Model is preferred by an average of 77% times.**

Aya win rates against mT0x



05
Responsibility

Safety for All Languages

The model may produce undesirable responses, such as toxic, biased, or harmful responses - but we want to ensure a safe and responsible use - across all languages.

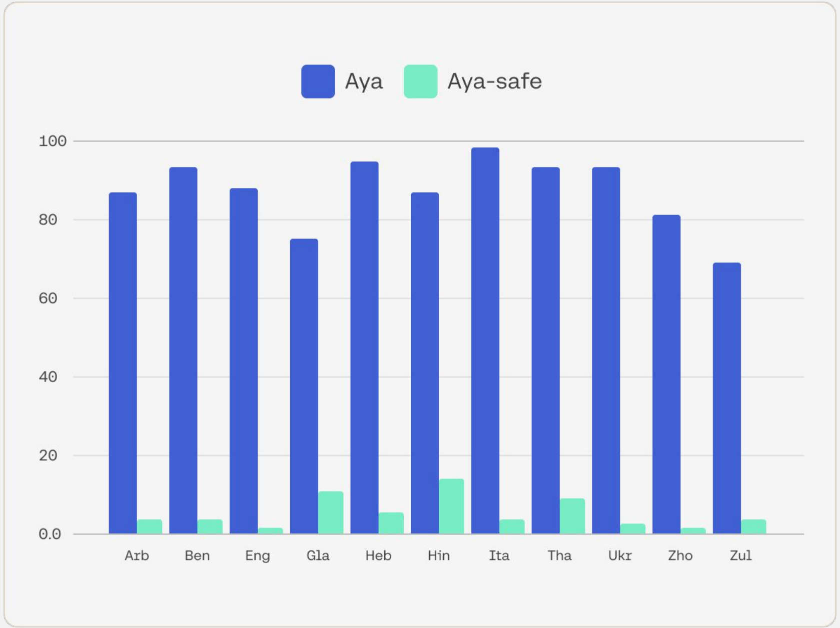
Previous safety mitigations have predominantly focused on English, which can lead to safety oversights in other languages. This means models might produce safe outputs in English but unsafe ones when prompted in different languages.

With Aya, we focus on a wide, multilingual evaluation of biases, toxicity, and harmfulness, and we implement a multilingual safety measure to prevent misuse for potentially harmful user intentions.

Multilingual Safety Context Distillation

First we define a set of unsafe contexts, where a user queries the model with an adversarial prompt and a harmful intention. We can then train the Aya Model to generate refusal messages for such use cases across all of its languages.

The refusal messages are obtained by querying a teacher model with a safety preamble that explicitly discourages harmful responses. By training on these responses, we distill concepts of safety into the Aya Model, achieving *more harmless responses*, and *maintaining open-ended generation quality*.



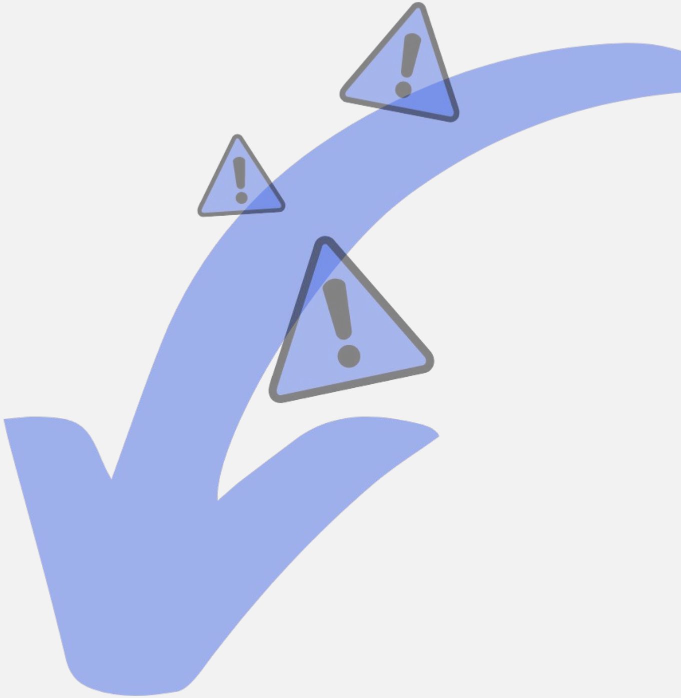
NOTE: The release of the Aya model will make community-based red-teaming efforts possible by exposing an open-source multilingual model for community research.

Measuring Toxicity and Bias

Benchmarking toxicity and bias in models helps us understand how often and how seriously the model might give responses that could be toxic or biased across languages.

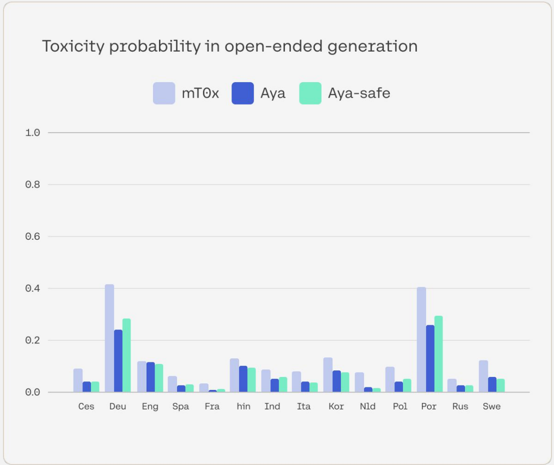
The Aya Model is tested on two evaluation scenarios:

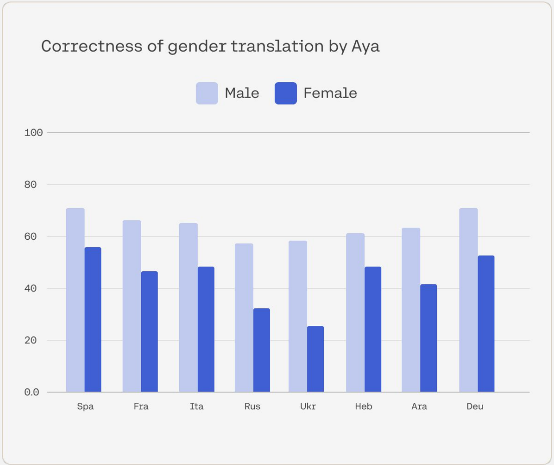
1) Toxicity and bias in open-ended generation, across 14 languages.
2) Gender bias in machine translation, across 8 languages.



Results From Benchmarking Toxicity and Bias

1. Our findings show that instruction fine-tuning and safety mitigation reduce toxicity and bias.
2. Absolute tendencies towards toxic and bias outputs vary across languages.
3. The problem is not solved: especially racial and gender biases are still present.


Toxicity probability in open-ended generation


Correctness of gender translation by Aya

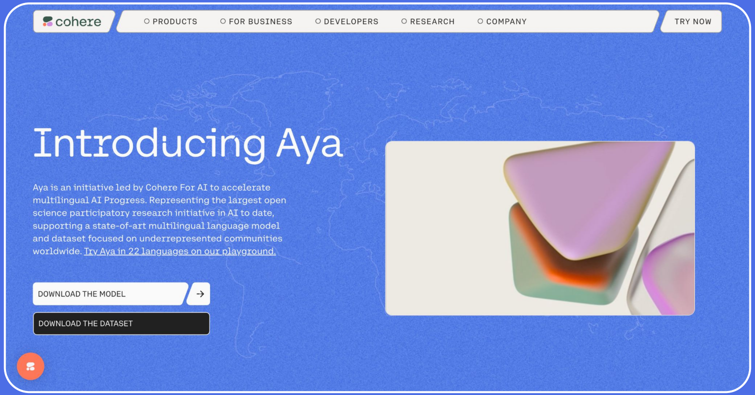
06 The Aya Movement

Read the Research

Read our research, Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning.

Read our research, Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model.

Learn more



Introducing Aya

Aya is an initiative led by Cohere For AI to accelerate multilingual AI Progress. Representing the largest open science participatory research initiative in AI to date, supporting a state-of-art multilingual language model and dataset focused on underrepresented communities worldwide. Try Aya in 22 languages on our playground.

Aya

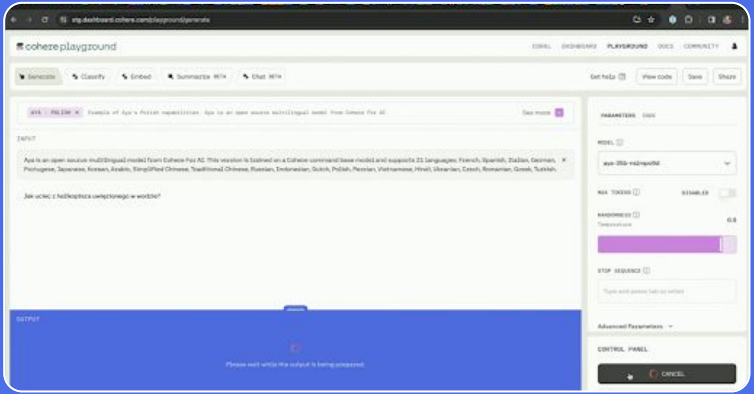
Accelerating Multilingual AI Through Open Science

Visit the Aya webpage to download the model and dataset, see the latest Aya press coverage, and get to know some of our collaborators.

Read our blog post on Aya's release.

Dive Deeper

Watch *The Journey of Aya*, a 20-minute documentary featuring many of our collaborators that highlights the importance of progress in multilingual ML, and showcases how this major research effort came together over the past year.



Example of Aya's Polish capabilities. Aya is an open source multilingual model from Cohere For AI

INPUT

Aya is an open source multilingual model from Cohere For AI. This version is trained on a Cohere command base model and supports 21 languages: French, Spanish, Italian, German, Portuguese, Japanese, Korean, Arabic, Simplified Chinese, Traditional Chinese, Russian, Indonesian, Dutch, Polish, Persian, Vietnamese, Hindi, Ukrainian, Czech, Romanian, Greek, Turkish.

Jak uciec z helikoptera uwięzionego w wodzie?

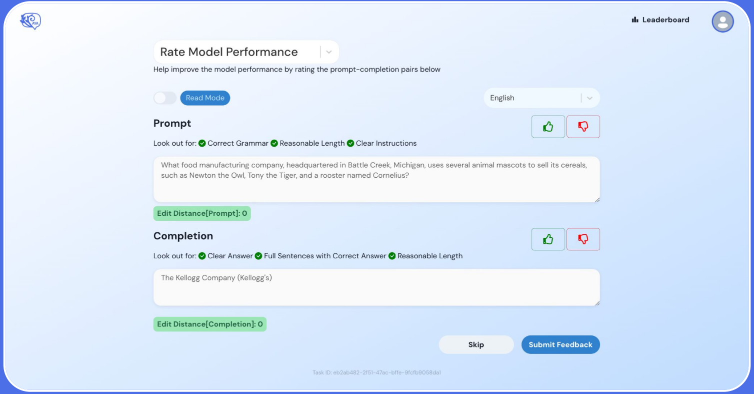
OUTPUT

Please wait while the output is being prepared.

Use your own prompts to Try Aya on the Cohere Playground in 22 sample languages.

Join us

This is only the beginning. Aya will be a foundation for additional open science projects and we expect to continue to improve Aya capabilities.



Rate Model Performance

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

Prompt

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

What food manufacturing company, headquartered in Battle Creek, Michigan, uses several animal mascots to sell its cereals, such as Newton the Owl, Tony the Tiger, and a rooster named Cornelius?

Edit Distance[Prompt]: 0

Completion

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

The Kellogg Company (Kellogg's)

Edit Distance[Completion]: 0

Task ID: eb2ab482-2f51-47ac-bffe-9fcfb9058da1

Contribute to Aya. Share expertise in your language to be include. We will continue to release data every year or each time an additional 20,000 annotations are contributed (whichever comes first).

Join the Cohere For AI Open Science community - a space for ML researchers worldwide to connect, learn from one another, and work collaboratively to advance the field of ML research. We will continue to host open science initiatives.

Aya

🌐 cohere.com/research/aya

𝕏 @CohereForAI

in /showcase/cohere-for-ai