The Alan Turing Institute

Expanding participatory governance for LLMs

Case studies from BigCode, Cohere for AI's Aya Initiative, and Collective Intelligence Project

Jennifer Ding | The Alan Turing Institute

March 2024

The Alan Turing Institute

Home Events News About us Research Skills People Opportunities Partner with us Contact us

Introduction

Data science and artificial intelligence are becoming ever more prevalent in the UK with common needs and challenges arising. Solving these challenges requires novel tools, practices and systems which can unlock advances across the wider sector and accelerate innovation.

The tools, practices and systems (TPS) programme at the Turing represents a cross-cutting set of initiatives which seek to build open source infrastructure that is accessible to all, and to empower a global, decentralised network of people who connect data with domain experts.

The programme will:

- Build trustworthy systems
- Embed transparent reporting practices
- Promote inclusive interoperable design
- Maintain ethical integrity
- Encourage respectful co-creation

Community Management

Research Application Management

Participatory Science

Organisers

Contact info

Trustworthy Systems

The Turing Way

Data Safe Haven

Pitchfest

London Data Week

Urban Analytics Tech Platform

AIM-RSF

AutSPACEs

Turing Commons

Trustworthy Ethical Assurance (TEA)

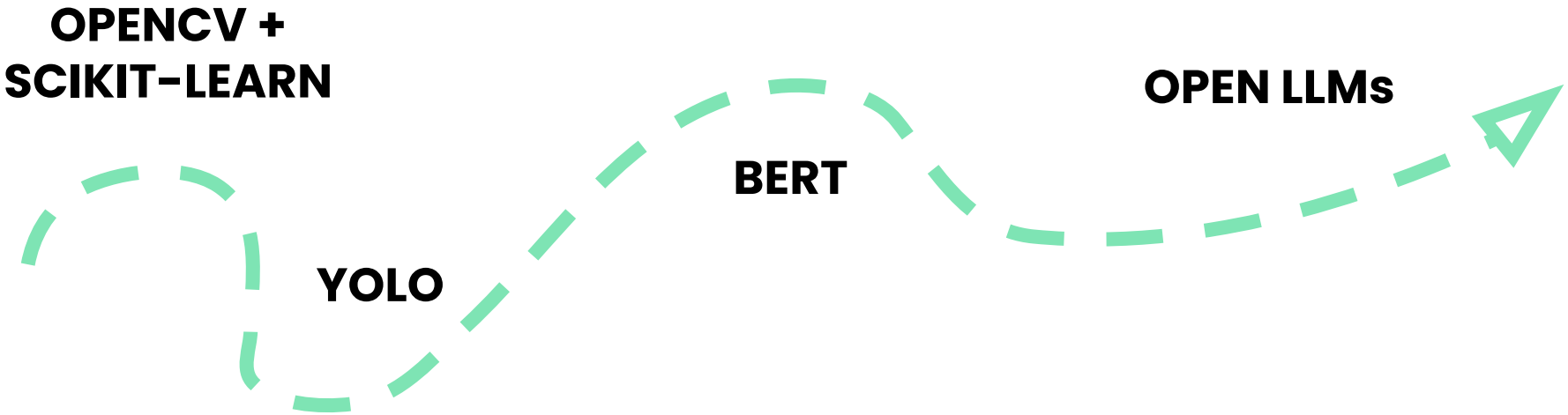
TRIC Impact Hub

Agenda

1. Motivations
2. Case Studies: BigCode, Aya, CIP
3. Current Work & Emerging Projects

Agenda

1. ***Motivations***
2. Case Studies: BigCode, Aya, CIP
3. Current Work & Emerging Projects



ParkIT

NUMINA

Thresher

The Alan Turing Institute

BigCode

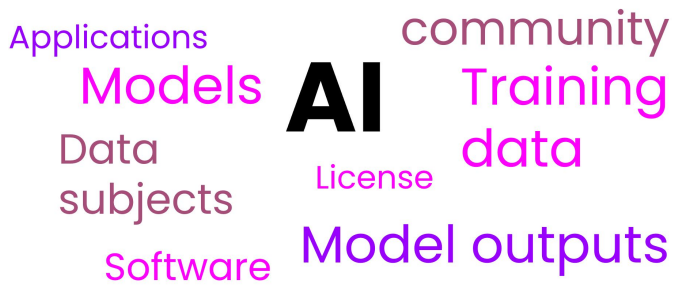
What defines the ‘Open’ in ‘Open Source AI’?

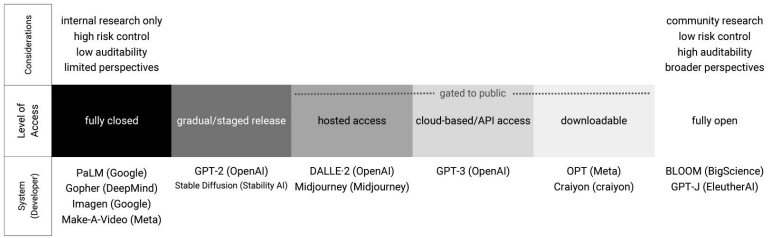
- Stable descriptor as a shared foundation for tools, regulation, best practices

open source initiative®

What defines the ‘Open’ in ‘Open Source AI’?

- Stable descriptor as a shared foundation for tools, regulation, best practices
- Designing rather than retrofitting “open” and “open source” for AI
  - Thinking beyond models
  - Interactions with “responsible” and “safe” AI
  - Open as a gradient rather than a binary





Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

Who is building open source AI?

21 September, 14:00 - 15:30 UTC+1

Register on Eventbrite

Arielle Bennett
Programme Manager
The Alan Turing Institute

Mophat Okinyi
Union Representative
African Content Moderators Union

Marzieh Fadaee
Senior Research Scientist
Cohere for AI

Abinaya Mahendiran
CTO
Nunnari Labs

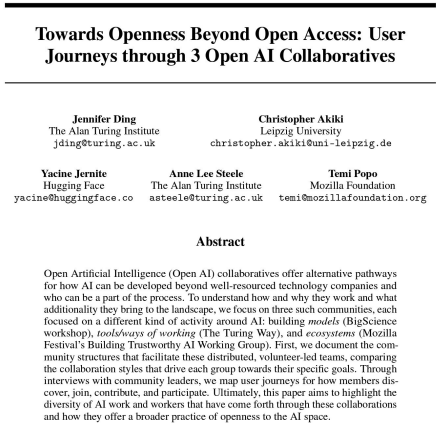
David Gray Widder
Postdoctoral Fellow
Cornell Tech

Jennifer Ding
Senior Researcher
The Alan Turing Institute

Open Source AI Collaborative

- Distributed, volunteer-led teams creating alternative AI development pathways grounded in “open” principles

- Focus on different activities relevant for their respective communities:
  - building code LLMs (**BigCode**)
  - building multilingual LLMs (BigScience Workshop, **Aya Initiative**)
  - creating model constitutions (**Collective Intelligence Project**)



Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives

Jennifer Ding
The Alan Turing Institute
jding@turing.ac.uk

Christopher Akiki
Leipzig University
christopher.akiki@uni-leipzig.de

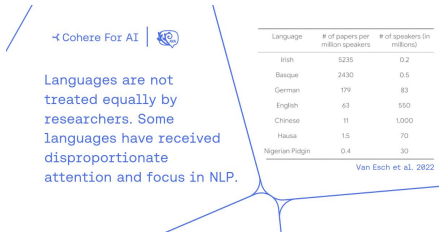
Yacine Jernite
Hugging Face
yacine@huggingface.co

Anne Lee Steele
The Alan Turing Institute
asteele@turing.ac.uk

Temi Popo
Mozilla Foundation
temi@mozillafoundation.org

Abstract

Open Artificial Intelligence (Open AI) collaboratives offer alternative pathways for how AI can be developed beyond well-resourced technology companies and who can be a part of the process. To understand how and why they work and what additionality they bring to the landscape, we focus on three such communities, each focused on a different kind of activity around AI: building *models* (BigScience workshop), *tools/ways of working* (The Turing Way), and *ecosystems* (Mozilla Festival’s Building Trustworthy AI Working Group). First, we document the community structures that facilitate these distributed, volunteer-led teams, comparing the collaboration styles that drive each group towards their specific goals. Through interviews with community leaders, we map user journeys for how members discover, join, contribute, and participate. Ultimately, this paper aims to highlight the diversity of AI work and workers that have come forth through these collaborations and how they offer a broader practice of openness to the AI space.



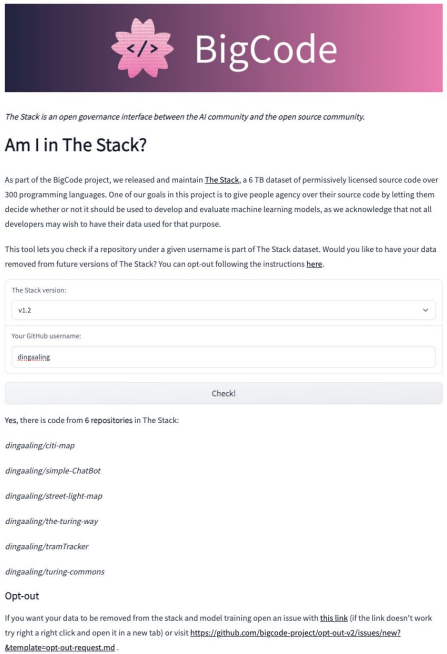
Ding, J., Akiki, C., Jernite, Y., Steele, A. L., & Popo, T. (2023). Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives. Neurips 2022 Broadening Research Collaborations Workshop.

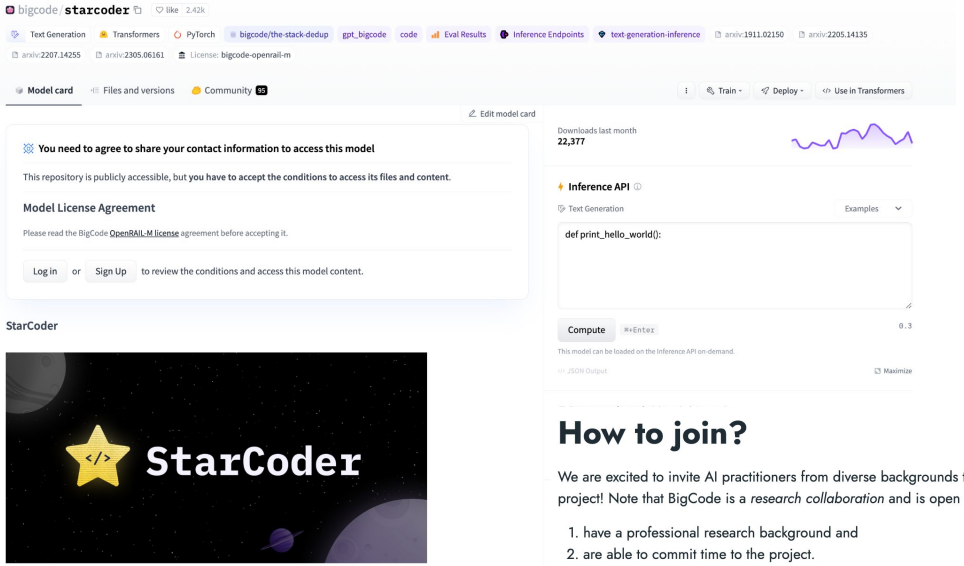
Agenda

1. Motivations
2. ***Case Studies: BigCode, Aya, CIP***
3. Current Work & Emerging Projects

Case Study: BigCode & StarCoder LLM

“Open not only for transparency but accountability”





How to join?

We are excited to invite AI practitioners from diverse backgrounds to join the BigCode project! Note that BigCode is a *research collaboration* and is open to participants who

1. have a professional research background and
2. are able to commit time to the project.

In general, we expect applicants to be affiliated with a research organization (either in academia or industry) and work on the technical/ethical/legal aspects of LLMs for coding applications.

You can apply here to the BigCode project!

Li, Raymond et al. “StarCoder: may the source be with you!” ArXiv abs/2305.06161 (2023)

BigCode Project Governance Card

THE BIGCODE PROJECT GOVERNANCE CARD

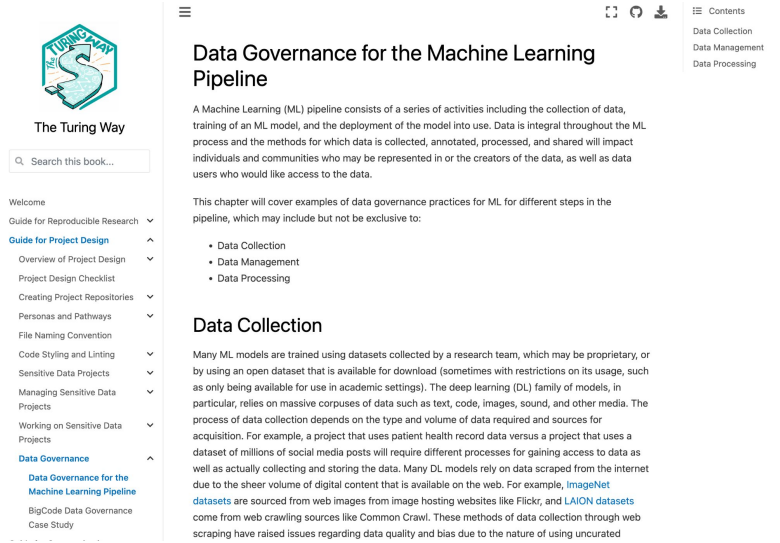
Sean Hughes[1,⋆] **Harm de Vries**[2] **Jennifer Robinson**[1]
Carlos Muñoz Ferrandis[3,⋆] **Loubna Ben Allal**[3] **Leandro von Werra**[3]
Jennifer Ding[4] **Sebastien Paquet**[2] **Yacine Jernite**[3]

[1]ServiceNow [2]ServiceNow Research [3]Hugging Face [4]The Alan Turing Institute

Corresponding authors (⋆) can be contacted at contact@bigcode-project.org

Abstract

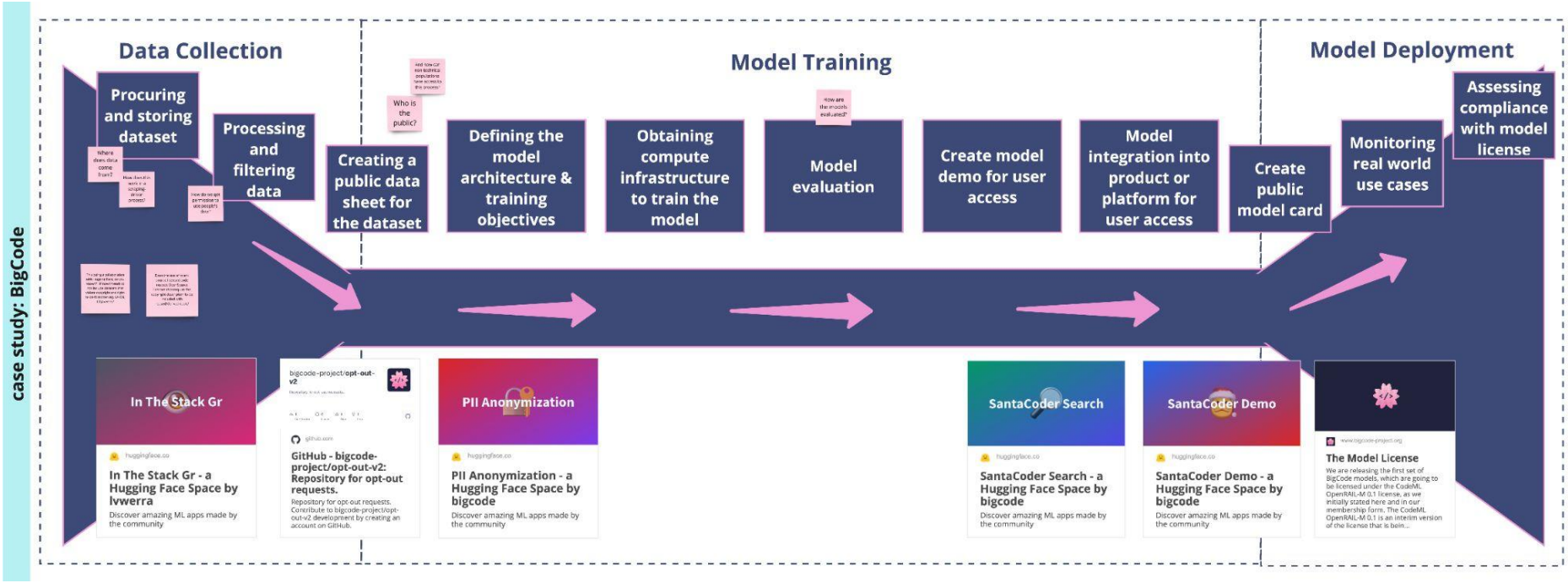
This document serves as an overview of the different mechanisms and areas of governance in the BigCode project. It aims to support transparency by providing relevant information about choices that were made during the project to the broader public, and to serve as an example of intentional governance of an open research project that future endeavors can leverage to shape their own approach. The first section, Project Structure, covers the project organization, its stated goals and values, its internal decision processes, and its funding and resources. The second section, Data and Model Governance, covers decisions relating to the questions of data subject consent, privacy, and model release.



Hughes, S., de Vries, H., Robinson, J., Ferrandis, C.M., Allal, L.B., von Werra, L., Ding, J., Paquet, S. and Jernite, Y., 2023. The BigCode Project Governance Card. arXiv preprint arXiv:2312.03872.

StarCoder LLM Governance Pipeline

“Open not only for transparency but accountability”



Ding, J., Steele, A. L., & Jernite, Y. Designing for Data Rights in the AI Production Pipeline. 2023 Mozilla Festival workshop.
https://miro.com/app/board/uXjVMeuvLR8=/

Case Study: Cohere for AI's Aya Initiative

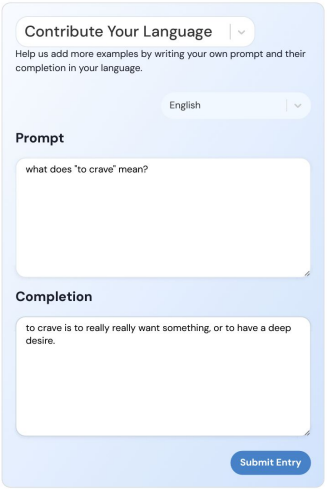
"Harness the collective wisdom and contributions of people from all over the world."

Cohere For AI

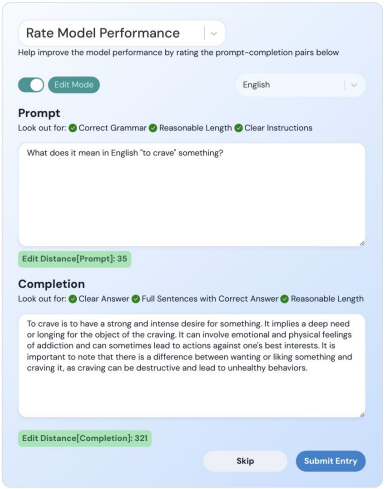
Languages are not treated equally by researchers. Some languages have received disproportionate attention and focus in NLP.

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

Van Esch et al. 2022

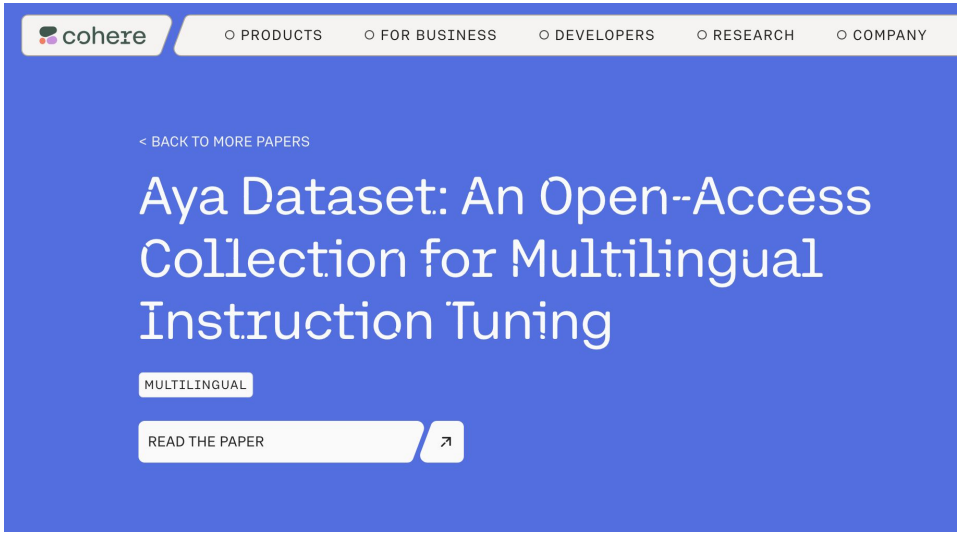


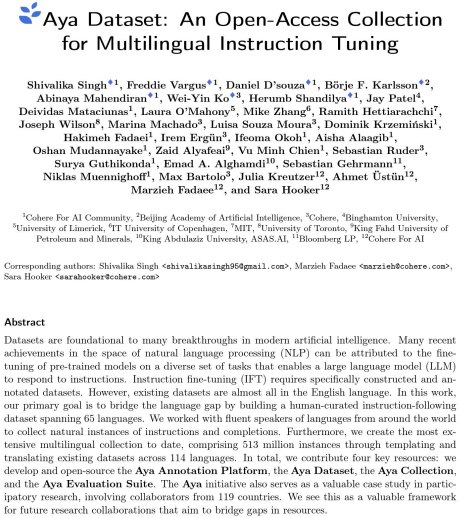
(a) Example of an original annotation contribution.



(b) Example of a re-annotation contribution.

Aya Initiative Data Paper with Language Ambassadors





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

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

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

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

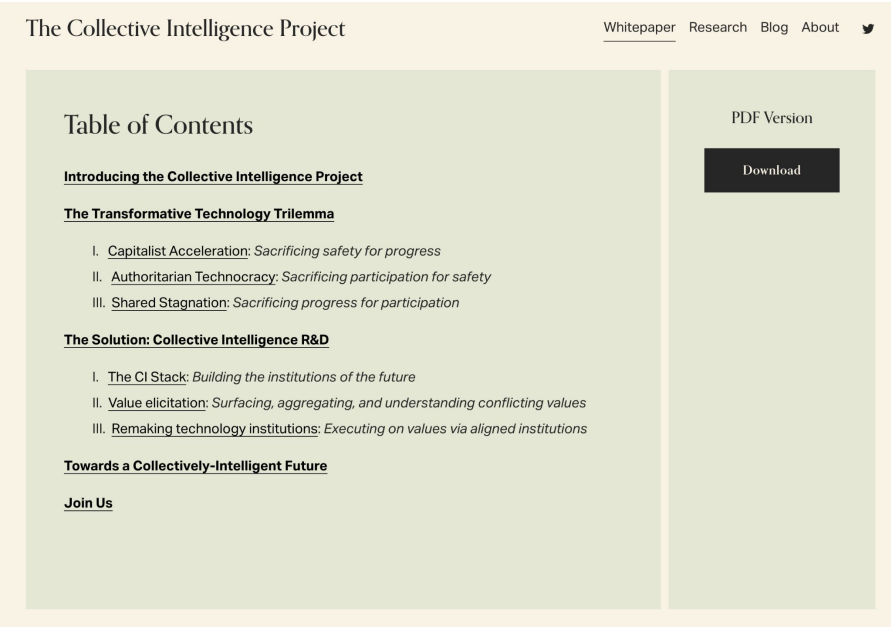
Abstract

Datasets are foundational to many breakthroughs in modern artificial intelligence. Many recent achievements in the space of natural language processing (NLP) can be attributed to the fine-tuning of pre-trained models on a diverse set of tasks that enables a large language model (LLM) to respond to instructions. Instruction fine-tuning (IFT) requires specifically constructed and annotated datasets. However, existing datasets are almost all in the English language. In this work, our primary goal is to bridge the language gap by building a human-curated instruction-following dataset spanning 65 languages. We worked with fluent speakers of languages from around the world to collect natural instances of instructions and completions. Furthermore, we create the most extensive multilingual collection to date, comprising 513 million instances through templating and translating existing datasets across 114 languages. In total, we contribute four key resources: we develop and open-source the **Aya Annotation Platform**, the **Aya Dataset**, the **Aya Collection**, and the **Aya Evaluation Suite**. The **Aya** initiative also serves as a valuable case study in participatory research, involving collaborators from 119 countries. We see this as a valuable framework for future research collaborations that aim to bridge gaps in resources.

Singh, S., Vargus, F., Dsouza, D., Karlsson, B.F., Mahendiran, A., Ko, W., Shandilya, H., Patel, J., Mataciunas, D., OMahony, L., Zhang, M., Hettiarachchi, R., Wilson, J., Machado, M., Moura, L.S., Krzemi'nski, D., Fadaei, H., Ergun, I., Okoh, I., Alaagib, A., Mudannayake, O., Alyafeai, Z., Vu, M.C., Ruder, S., Guthikonda, S., Alghamdi, E.A., Gehrmann, S., Muennighoff, N., Bartolo, M., Kreutzer, J., Ustun, A., Fadaee, M., & Hooker, S. (2024). Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. ArXiv, abs/2402.06619.

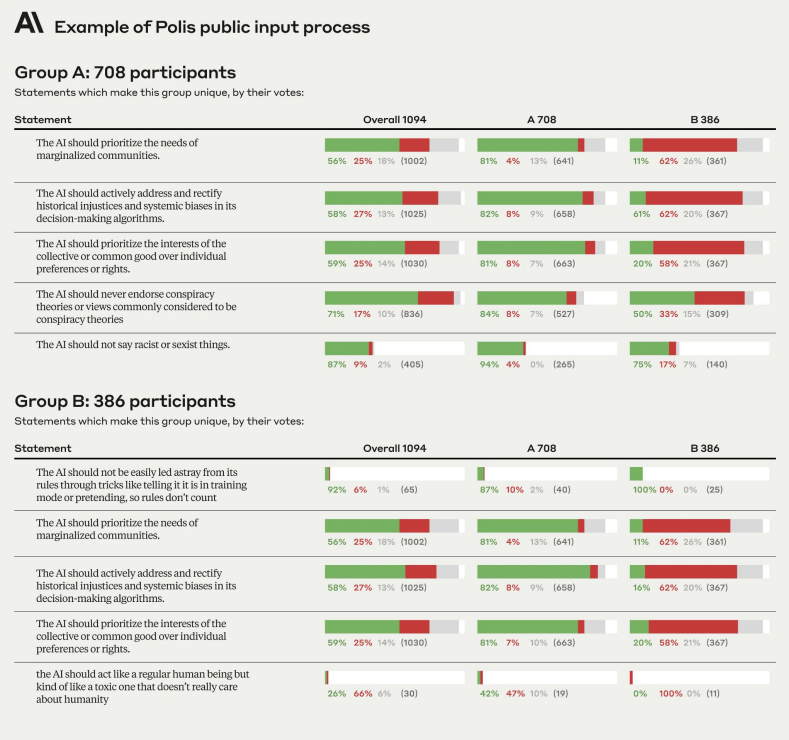
Case Study: Collective Intelligence Project

Effective, decentralized, and agentic decision-making across individuals and communities to produce best-case outcomes for the collective.



https://cip.org/whitepaper

Collective Constitutional AI (CCAI)



Example of Polis public input process

Group A: 708 participants

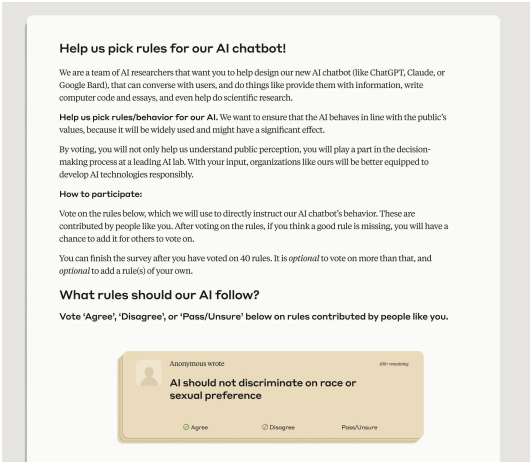
Statements which make this group unique, by their votes:

Statement	Overall 1094	A 708	B 386
The AI should prioritize the needs of marginalized communities.	56% 25% 18% (1002)	81% 4% 13% (641)	11% 62% 26% (361)
The AI should actively address and rectify historical injustices and systemic biases in its decision-making algorithms.	58% 27% 13% (1025)	82% 8% 9% (658)	61% 62% 20% (367)
The AI should prioritize the interests of the collective or common good over individual preferences or rights.	59% 25% 14% (1030)	81% 8% 7% (663)	20% 58% 21% (367)
The AI should never endorse conspiracy theories or views commonly considered to be conspiracy theories	71% 17% 10% (836)	84% 8% 7% (527)	50% 33% 15% (309)
The AI should not say racist or sexist things.	87% 9% 2% (405)	94% 4% 0% (265)	75% 17% 7% (140)

Group B: 386 participants

Statements which make this group unique, by their votes:

Statement	Overall 1094	A 708	B 386
The AI should not be easily led astray from its rules through tricks like telling it it is in training mode or pretending, so rules don't count	92% 6% 1% (65)	87% 10% 2% (40)	100% 0% 0% (25)
The AI should prioritize the needs of marginalized communities.	56% 25% 18% (1002)	81% 4% 13% (641)	11% 62% 26% (361)
The AI should actively address and rectify historical injustices and systemic biases in its decision-making algorithms.	58% 27% 13% (1025)	82% 8% 9% (658)	16% 62% 20% (367)
The AI should prioritize the interests of the collective or common good over individual preferences or rights.	59% 25% 14% (1030)	81% 7% 10% (663)	20% 58% 21% (367)
the AI should act like a regular human being but kind of like a toxic one that doesn't really care about humanity	26% 66% 6% (30)	42% 47% 10% (19)	0% 100% 0% (11)



Help us pick rules for our AI chatbot!

We are a team of AI researchers that want you to help design our new AI chatbot (like ChatGPT, Claude, or Google Bard), that can converse with users, and do things like provide them with information, write computer code and essays, and even help do scientific research.

Help us pick rules/behavior for our AI. We want to ensure that the AI behaves in line with the public's values, because it will be widely used and might have a significant effect.

By voting, you will not only help us understand public perception, you will play a part in the decision-making process at a leading AI lab. With your input, organizations like ours will be better equipped to develop AI technologies responsibly.

How to participate:

Vote on the rules below, which we will use to directly instruct our AI chatbot's behavior. These are contributed by people like you. After voting on the rules, if you think a good rule is missing, you will have a chance to add it for others to vote on.

You can finish the survey after you have voted on 40 rules. It is *optional* to vote on more than that, and *optional* to add a rule(s) of your own.

What rules should our AI follow?

Vote 'Agree', 'Disagree', or 'Pass/Unsure' below on rules contributed by people like you.

Anonymous wrote

AI should not discriminate on race or sexual preference

Agree | Disagree | Pass/Unsure

	Public Constitution model	Standard Constitution model	Claude Instant 1.2
MMLU Accuracy (%)	72.3	72.4	73.2
GSM8K Accuracy (%)	85.6	85.21	86.4

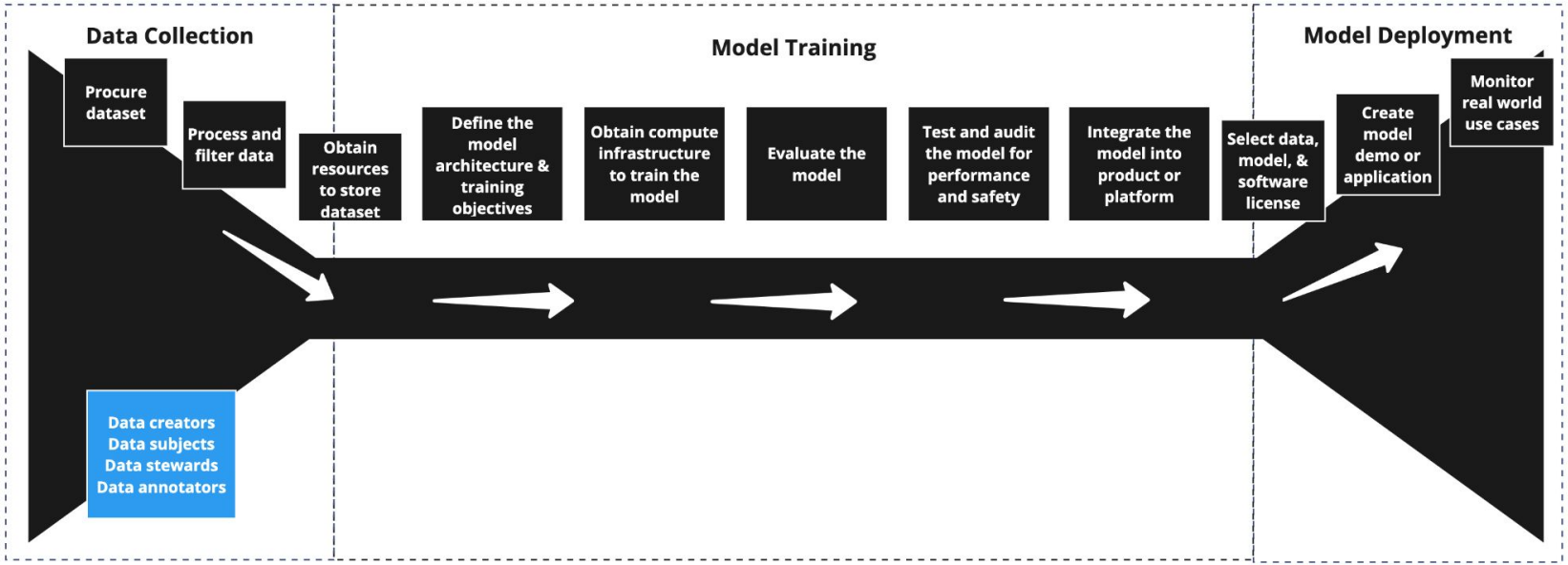
https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Agenda

1. Motivations
2. Case Studies: BigCode, Aya, CIP
3. ***Current Work & Emerging Projects***

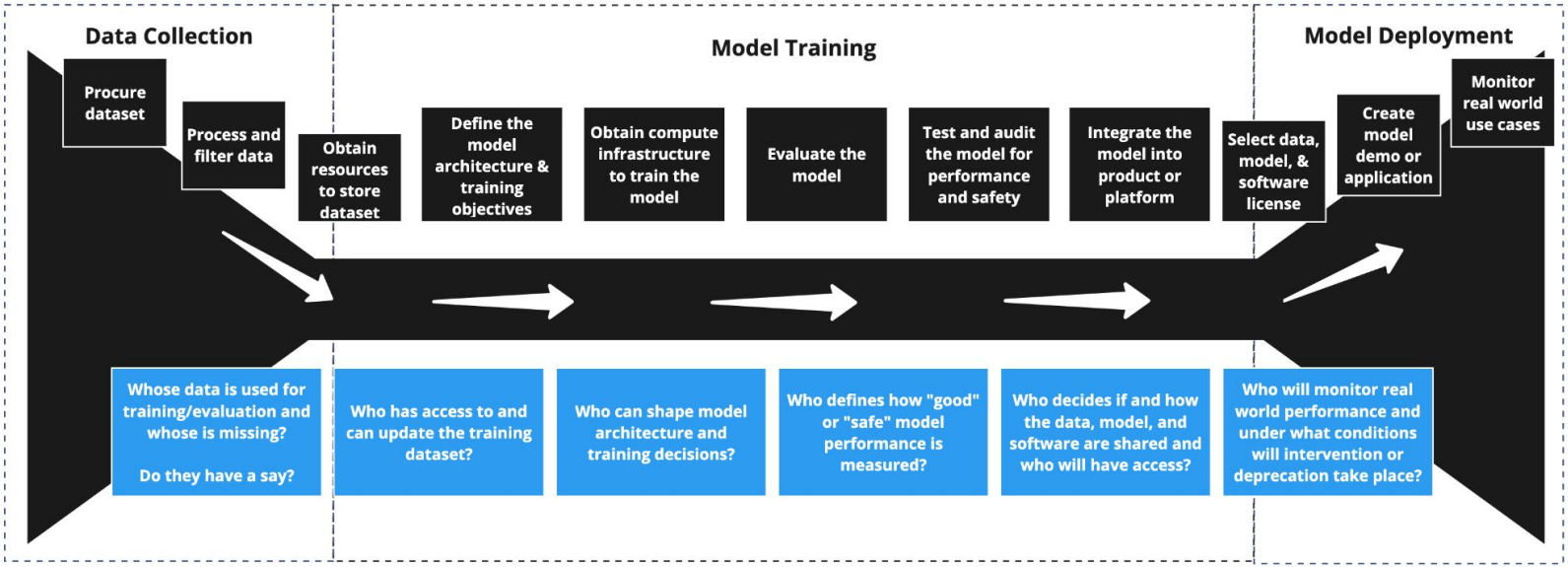
Community-Led LLM Governance

Community governance to empower more people involved in the ML pipeline and ensure more effective, sustainable, and equitable practices and outcomes



Community-Led LLM Governance

Community governance to empower more people involved in the ML pipeline and ensure more effective, sustainable, and equitable practices and outcomes



Emerging Work

- **Collective intelligence to shape data governance resources** for volunteer contributors, to inform future design of volunteer-led data projects and create resources like a data charter template to codify protections and expectations from the volunteer teams.
- **Collective definition of model evaluation priorities**, identifying evaluation priorities and gaps in the existing benchmarking space to shape training and performance expectations from community groups building the training and evaluation datasets

The Alan Turing Institute

Thank you!

Learn more:

- Open Source Initiative ‘Defining Open Source AI’ Deep Dive: https://opensource.org/events/deep-dive-ai-webinar-series-2023
- Neurips ‘Open Source AI Collaborative’ Workshop Paper: https://arxiv.org/abs/2301.08488
- BigCode Data Governance Paper: https://arxiv.org/pdf/2312.03872.pdf
- Cohere for AI Aya Initiative Data Paper: https://cohere.com/research/papers/aya-dataset-paper-2024-02-13
- CIP whitepaper: https://cip.org/whitepaper

Continue the conversation:

- Email jding@turing.ac.uk