Policy Primer

Cohere Labs

The AI Language Gap

Considerations on the Multilingual Capabilities of AI Language Models

June 2024

Summary

More than 7000 languages are spoken around the world today, but current, state-of-art AI large language models cover only a small percentage of them and favor North American language and cultural perspectives. This is in part because many non-English languages are considered "low-resource," meaning they are less prominent within computer science research and lack the high-quality datasets necessary for training language models.

This language gap in AI has several undesirable consequences:

- Many language speakers and communities may be left behind as language models that do not cover their language become increasingly integral to economies and societies.
- The lack of linguistic diversity in models can introduce biases that reflect Anglo-centric and North American viewpoints, and undermine other cultural perspectives.
- The safety of all language models is compromised without multilingual capabilities, creating opportunities for malicious users and exposing all users to harm.

There are many global efforts to address the language gap in AI, including Cohere Lab's Aya project — a global initiative that has developed and publicly released multilingual language models and datasets covering 101 languages. However, more work is needed.

To contribute to efforts to address the AI language gap, we offer **four considerations for those working in policy and governance around the world**:

1. Direct resources towards multilingual research and development.
2. Support multilingual dataset creation.
3. Recognize that the safety of all language models is improved through multilingual approaches.
4. Foster knowledge-sharing and transparency among researchers, developers, and communities.

1. Background: The Language Gap in AI

Generative AI language models[1] are finding beneficial applications in a range of contexts across societies and economies around the world. However, the vast majority of language models are currently optimized for a small handful of languages, and the English language and North American sociocultural preferences are dominant across their design, outputs, and behavior.[2]

There are several efforts around the world to develop the multilingual capabilities of AI language models, including Cohere Lab's Aya models and dataset[3] — a family of open source, massively multilingual language models that cover 101 languages — and Cohere's Command R+, a proprietary model with open weights that covers 23 languages.[4]

However, as this policy primer describes, there still remains a gap in the availability of models that can cover non-English and minority languages, and those that do often fail to perform as highly for these languages in comparison to English.[5]

This language gap in the development, capabilities, and application of AI language models is the result of several factors, explored below.

Resources for AI language model development are biased towards English, and many non-English languages are "low-resource."
Recent breakthroughs in language models largely depend on the availability of high quality text-based datasets.[6, 7, 8] However, the most widely used datasets in natural language processing currently represent only a handful of data-rich languages, and the datasets used for instruction-fine-tuning — a key step in improving language model capability — are almost entirely focused on English.[9]

[1] Language models, or large language models (LLMs), are computational tools that use techniques from machine learning and natural language processing to process and generate text in human languages. For more detail, see: https://docs.cohere.com/docs/intro-large-language-models.
[2] Naous, T. *et al.* (2024) 'Having Beer after Prayer? Measuring Cultural Bias in Large Language Models'. arXiv. http://arxiv.org/abs/2305.14456.
[3] In February 2024, Cohere For AI launched Aya, an open language model and dataset covering 101 languages. In May 2024, we released Aya 23, a pair of more performant models fine-tuned for 23 languages. Aya is the product of a global open science initiative involving 3000 researchers from 119 countries. See more below in this report, and at: https://cohere.com/research/aya.
[4] Cohere's most advanced large language model, Command R+, covers 23 languages. It is optimized for: English, French, Spanish, Italian, German, Brazilian Portuguese, Japanese, Korean, Simplified Chinese, and Arabic; additional data has been included for: Russian, Polish, Turkish, Vietnamese, Dutch, Czech, Indonesian, Ukrainian, Romanian, Greek, Hindi, Hebrew, and Persian. For more information, see: https://docs.cohere.com/docs/command-r-plus.
[5] Li, Z. *et al.* (2024) 'Quantifying Multilingual Performance of Large Language Models Across Languages'. arXiv. http://arxiv.org/abs/2404.11553.
[6] Lee, A. *et al.* (2023) 'Beyond Scale: the Diversity Coefficient as a Data Quality Metric Demonstrates LLMs are Pre-trained on Formally Diverse Data'. arXiv. http://arxiv.org/abs/2306.13840.
[7] Marion, M. *et al.* (2023) 'When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale'. arXiv. http://arxiv.org/abs/2309.04564.
[8] Touvron, H. *et al.* (2023) 'LLaMA: Open and Efficient Foundation Language Models'. arXiv. https://doi.org/10.48550/arXiv.2302.13971.
[9] Singh, S. et al. (2024) *Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning*. arXiv. http://arxiv.org/abs/2402.06619.

This is partly due to the fact that, of the 7000 languages that exist today,[10] easily available data covers only around 1500 languages,[11] and acquiring data that has a high-enough quality for use in training language models is challenging.[12, 13] Furthermore, many languages are "low-resource," meaning they are less well-studied or privileged globally, and the availability of robust datasets required for including these languages in machine learning research and computer science is scarce.[14, 15]

Where datasets are available for low-resource languages, their quality is often not sufficient to use them in language model research and development,[16] and the vast majority of the people, organizations, and teams working to develop these datasets originate from a few countries.[17, 18] Additionally, it is not a standardized practice for AI model developers to always fully report the number of languages covered or specify which languages are included, preventing reliable performance evaluations and reducing transparency around how the language gap manifests. There are also global inequities in access to the compute resources required for language model research and development, largely due to cost and availability of hardware and infrastructure.[19]

This language gap in AI risks growing deeper and wider.
The availability of resources and attention paid to a language is not proportionate to its number of speakers, as which languages are favored is often a "symptom of historical technological use and access to resources."[20] For example Dutch, a language with 29 million speakers, has 2 million Wikipedia entries, while Somali has just 5500 Wikipedia entries despite having 18 million speakers.[21]

[10] Eberhard, David M., Gary F. Simons, and Charles D. Fennig. (2024) *Ethnologue: Languages of the World.* Twenty-seventh edition. http://www.ethnologue.com.

[11] Bapna, A. *et al.* (2022) 'Building Machine Translation Systems for the Next Thousand Languages'. arXiv. https://doi.org/10.48550/arXiv.2205.03983,

[12] Adda, G. *et al.* (2016) 'Breaking the Unwritten Language Barrier: The BULB Project', *Procedia Computer Science*, 81, pp. 8–14. https://doi.org/10.1016/j.procs.2016.04.023.

[13] NLLB Team *et al.* (2022) 'No Language Left Behind: Scaling Human-Centered Machine Translation'. arXiv. https://doi.org/10.48550/arXiv.2207.04672.

[14] Magueresse, A., Carles, V. and Heetderks, E. (2020) 'Low-resource Languages: A Review of Past Work and Future Challenges'. arXiv. http://arxiv.org/abs/2006.07264.

[15] Gabriel Nicholas and Aliya Bhatia (2023) *Lost in Translation: Large Language Models in Non-English Content Analysis*. Center for Democracy and Technology. https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf.

[16] Kreutzer, J. *et al.* (2022) 'Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets', *Transactions of the Association for Computational Linguistics*, 10, pp. 50–72. https://doi.org/10.1162/tacl_a_00447.

[17] Longpre, S. *et al.* (2023) 'The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'. arXiv. http://arxiv.org/abs/2310.16787.

[18] Maslej, N. *et al.* (2024) *AI Index Report 2024*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/report/.

[19] Ahia, O., Kreutzer, J. and Hooker, S. (2021) 'The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation', in *Findings of the Association for Computational Linguistics: EMNLP 2021*. https://doi.org/10.18653/v1/2021.findings-emnlp.282.

[20] Üstün, A. et al. (2024) 'Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model'. arXiv. http://arxiv.org/abs/2402.07827. Page 1.; See also: Bird, S. (2022) 'Local Languages, Third Spaces, and other High-Resource Scenarios', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2022.acl-long.539.

[21] Joshi, P. *et al.* (2021) 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World'. arXiv. https://doi.org/10.48550/arXiv.2004.09095.

There are also higher costs of using language model-based technologies for some non-English languages, as they may require more tokens and incur higher latency for generations,[22, 23] and speakers of low-resource languages often do not have the resources to improve NLP technology for their language due to limited access to compute, data, opportunities, and skills.[24, 25] Many existing language models fail to account for social factors, such as what the speaker's perspective is or what sociocultural norms are relevant, a problem which is amplified for low-resource languages that lack data about these factors.[26] Furthermore, as the field of natural language processing makes use of synthetic data generated by language models for training and tuning other models,[27] the fact that high-quality synthetic data will be more likely available in high-resource languages, which already dominate the field of natural language processing, risks deepening the existing gap between high- and low-resource languages.

[22] Ahia, O. *et al.* (2023) 'Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2023*, Singapore: Association for Computational Linguistics, pp. 9904–9923. https://doi.org/10.18653/v1/2023.emnlp-main.614.

[23] Ji, Y. *et al.* (2023) 'Towards Better Instruction Following Language Models for Chinese: Investigating the Impact of Training Data and Evaluation'. arXiv. http://arxiv.org/abs/2304.07854.

[24] Ahia, O., Kreutzer, J. and Hooker, S. (2021) 'The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation', in *Findings of the Association for Computational Linguistics: EMNLP 2021*. https://doi.org/10.18653/v1/2021.findings-emnlp.282.

[25] OECD (2023) AI language models: Technological, socio-economic and policy considerations. OECD Digital Economy Papers. https://doi.org/10.1787/13d38f92-en.

[26] Hovy, D. and Yang, D. (2021) 'The Importance of Modeling Social Factors of Language: Theory and Practice', in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. *NAACL-HLT 2021*, Online: Association for Computational Linguistics, pp. 588–602. https://doi.org/10.18653/v1/2021.naacl-main.49.

[27] Anaby-Tavor, A. *et al.* (2020) 'Do Not Have Enough Data? Deep Learning to the Rescue!', *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), pp. 7383–7390. https://doi.org/10.1609/aaai.v34i05.6233.

2. Why Multilingual Matters: Consequences of the Language Gap

The language gap in AI means that speakers of low-resource languages face a growing divide in the availability of high-quality language models and the resources required to develop them, compared to English and other dominant or high-resource languages. Left unaddressed, this risks negative feedback loops whereby language model capability and availability continues to grow for some languages, and diminish for others. This could have a range of undesirable consequences.

Many language speakers and communities risk being left behind.
The obvious consequence of the language gap in natural language processing is that as language models become increasingly integral across our economies and societies, the people and communities whose languages are not covered will be left behind and excluded. This is evident in research that demonstrates gaps between how poorly existing language models perform for low-resource languages in comparison to performance across high-resource languages.[28] This could worsen existing inequities in the provision of services and products across global communities as language models become more embedded within them.

Diversity across cultures, societies, and communities could be reduced.
Machine learning models' outputs can only reflect the world based on the data and information on which they have been trained and given access. Given this, the language gap means that the majority of current language models reflect the world as seen through anglo-centric and predominantly North American texts, which introduces biases against languages and cultural perspectives that are not included during model training.[29, 30, 31, 32] Research has also found that this lack of linguistic diversity means that the abstract "concept space" that underpins many language models' functionality is more oriented towards English than to other languages.[33]

The consequence of this bias in language models risks a regression to a Western, anglo-centric mean where products and systems rely on these models; for example, models may give answers to historical questions based only on Western values and world-views, diminishing or ignoring other global perspectives. Researchers working on the social impacts of AI systems have argued that this could "further embed the erosion of global multilingualism, undermine

[28] Adelani, D.I. *et al.* (2024) 'IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2406.03368.
[29] Schwartz, R. *et al.* (2022) *Towards a standard for identifying and managing bias in artificial intelligence*. NIST SP 1270. Gaithersburg, MD: National Institute of Standards and Technology (U.S.), p. NIST SP 1270. https://doi.org/10.6028/NIST.SP.1270.
[30] Kunchukuttan, A., Jain, S. and Kejriwal, R. (2021) 'A Large-scale Evaluation of Neural Machine Transliteration for Indic Languages', in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. *EACL 2021*, Online: Association for Computational Linguistics, pp. 3469–3475. https://doi.org/10.18653/v1/2021.eacl-main.303.
[31] Kotek, H., Dockum, R. and Sun, D. (2023) 'Gender bias and stereotypes in Large Language Models', *Proceedings of The ACM Collective Intelligence Conference*, pp. 12–24. https://doi.org/10.1145/3582269.3615599.
[32] Khandelwal, K. *et al.* (2023) 'Casteist but Not Racist? Quantifying Disparities in Large Language Model Bias between India and the West'. https://doi.org/10.48550/ARXIV.2309.08573.
[33] Wendler, C. et al. (2024) 'Do Llamas Work in English? On the Latent Language of Multilingual Transformers'. arXiv. https://doi.org/10.48550/arXiv.2402.10588.

the right to language and culture, and further marginalize the necessity for widespread multilingual education."[34]

Language model safety and security is extremely limited without multilingual capabilities and context.

Like many technologies, language models pose some well-studied risks and potential harms, including generating violent, biased, false, or toxic content.[35] A range of best practices in testing and improving the safety of language models is emerging and becoming adopted across the field of natural language processing, however a growing body of research shows how these approaches are largely oriented towards the English language, and there is a lack of reliable datasets needed for safety evaluation outside of a small fraction of languages.[36, 37, 38, 39, 40, 41]

This lack of multilingual safety testing and mitigation means that language models can produce harmful outputs when prompted in languages for which they are not optimized or safety-tested.[42] For example, a well-known consumer-facing chatbot has shown stereotypical gender biases when translating into languages such as Bengali and Turkish,[43] and it more frequently exhibits unsafe behavior when prompted in or asked to generate outputs in low-resource languages.[44]

Harms such as this can occur intentionally — for example when a malicious user intends to generate harmful output — meaning that the lack of multilingual safety approaches leaves "backdoors" that bad actors can exploit. Harms can also occur unintentionally, meaning users from language communities that are underserved by existing language models may be unknowingly more exposed to harms due to the lack of effective safeguards for their

[34] Solaiman, I. et al. (2023) 'Evaluating the Social Impact of Generative AI Systems in Systems and Society'. https://doi.org/10.48550/arXiv.2306.05949, p.15.

[35] Weidinger, L. *et al.* (2021) 'Ethical and social risks of harm from Language Models'. arXiv. http://arxiv.org/abs/2112.04359.

[36] Pozzobon, L. *et al.* (2024) 'From One to Many: Expanding the Scope of Toxicity Mitigation in Language Models'. arXiv. http://arxiv.org/abs/2403.03893.

[37] Talat, Z. *et al.* (2022) 'You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings', in A. Fan et al. (eds) *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. *BigScience 2022*, virtual+Dublin: Association for Computational Linguistics, pp. 26–41. https://doi.org/10.18653/v1/2022.bigscience-1.3.

[38] Gehman, S. *et al.* (2020) 'RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models'. arXiv. https://doi.org/10.48550/arXiv.2009.11462.

[39] Christiano, P. *et al.* (2023) 'Deep reinforcement learning from human preferences'. arXiv. https://doi.org/10.48550/arXiv.1706.03741.

[40] Chung, H.W. *et al.* (2022) 'Scaling Instruction-Finetuned Language Models'. arXiv. https://doi.org/10.48550/arXiv.2210.11416.

[41] Üstün, A. et al. (2024) 'Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model'. arXiv. http://arxiv.org/abs/2402.07827.

[42] Anwar, U. et al. (2024) 'Foundational Challenges in Assuring Alignment and Safety of Large Language Models'. arXiv. http://arxiv.org/abs/2404.09932.

[43] Ghosh, S. and Caliskan, A. (2023) 'ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages'. arXiv. https://doi.org/10.48550/arXiv.2305.10510.

[44] Yong, Z.X., Menghini, C. and Bach, S. (2023) 'Low-Resource Languages Jailbreak GPT-4', in. *Socially Responsible Language Modelling Research*. https://openreview.net/forum?id=pn83r8V2sv.

language.[45, 46, 47] Recent work to develop a first-of-its-kind multilingual red-teaming dataset shows a promising step towards addressing these safety gaps, but future work and greater adoption of such approaches are needed.[48]

[45] Shen, L. *et al.* (2024) 'The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts'. arXiv. https://doi.org/10.48550/arXiv.2401.13136.

[46] Deng, Y. *et al.* (2024) 'Multilingual Jailbreak Challenges in Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2310.06474.

[47] Yong, Z.X., Menghini, C. and Bach, S. (2023) 'Low-Resource Languages Jailbreak GPT-4', in. *Socially Responsible Language Modelling Research*. https://openreview.net/forum?id=pn83r8V2sv.

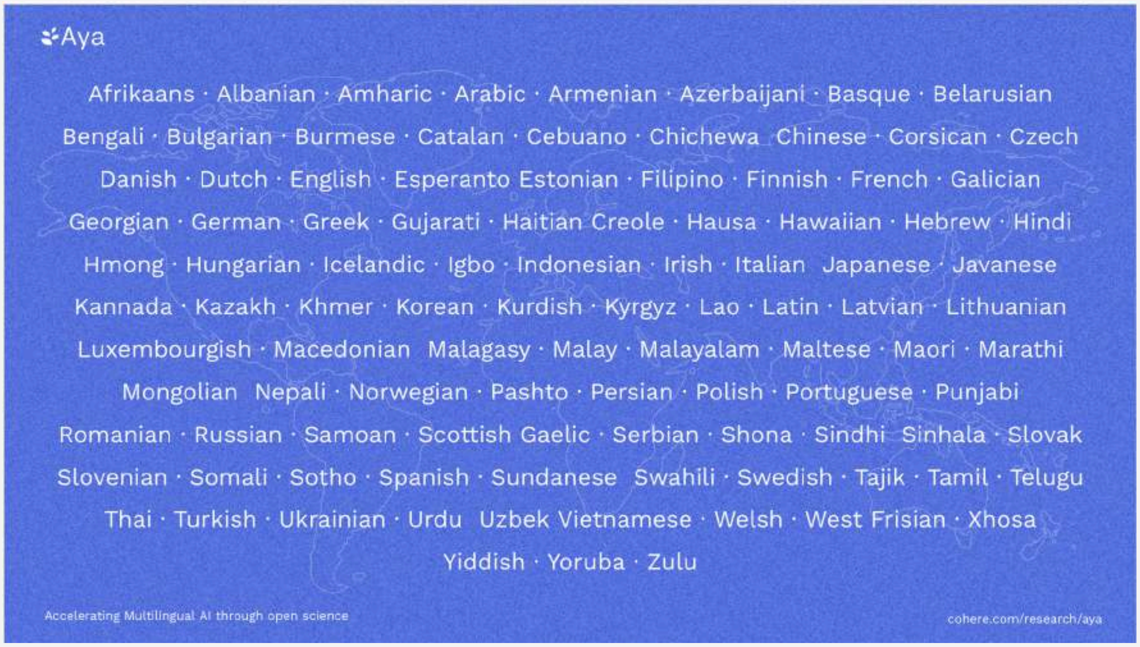
[48] Aakanksha *et al.* (2024) 'The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm'. arXiv. http://arxiv.org/abs/2406.18682.

3. Closing the AI Language Gap

The research outlined above makes clear that increasing the multilingual capabilities and safety of language models is crucial to ensuring and improving language model performance, usage, and safety for everyone.

At Cohere Labs, working to solve the AI language gap is a core focus.

In February 2024, Cohere Labs launched Aya, a new state-of-the-art, open-source, multilingual, large language research model and dataset that cover 101 different languages — more than double the number of languages covered by existing open-weights models.[49, 50, 51] In May 2024, we released Aya 23, a pair of more performant models for 23 languages within the Aya dataset, expanding state-of-art language modeling capabilities to approximately half of the world's population.[52]



Aya

Afrikaans · Albanian · Amharic · Arabic · Armenian · Azerbaijani · Basque · Belarusian
Bengali · Bulgarian · Burmese · Catalan · Cebuano · Chichewa Chinese · Corsican · Czech
Danish · Dutch · English · Esperanto Estonian · Filipino · Finnish · French · Galician
Georgian · German · Greek · Gujarati · Haitian Creole · Hausa · Hawaiian · Hebrew · Hindi
Hmong · Hungarian · Icelandic · Igbo · Indonesian · Irish · Italian Japanese · Javanese
Kannada · Kazakh · Khmer · Korean · Kurdish · Kyrgyz · Lao · Latin · Latvian · Lithuanian
Luxembourgish · Macedonian Malagasy · Malay · Malayalam · Maltese · Maori · Marathi
Mongolian Nepali · Norwegian · Pashto · Persian · Polish · Portuguese · Punjabi
Romanian · Russian · Samoan · Scottish Gaelic · Serbian · Shona · Sindhi Sinhala · Slovak
Slovenian · Somali · Sotho · Spanish · Sundanese Swahili · Swedish · Tajik · Tamil · Telugu
Thai · Turkish · Ukrainian · Urdu Uzbek Vietnamese · Welsh · West Frisian · Xhosa
Yiddish · Yoruba · Zulu

Accelerating Multilingual AI through open science

cohere.com/research/aya

The Aya 101 model and dataset cover 101 languages

[49] Üstün, A. et al. (2024) 'Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model'. arXiv. http://arxiv.org/abs/2402.07827.

[50] Singh, S. et al. (2024) 'Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning'. arXiv. http://arxiv.org/abs/2402.06619.

[51] Cohere For AI (2024) *Introducing Aya*, *Cohere*. https://cohere.com/research/aya.

[52] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032; see also: Cohere For AI (2024) *Cohere For AI Launches Aya 23, 8 and 35 Billion Parameter Open Weights Release*. https://cohere.com/blog/aya23.

The Aya dataset represents the largest collection of multilingual, instruction fine-tuned data to date, with 513 million prompts and completions covering 114 languages, including rare, human-curated annotations from fluent speakers worldwide. This data is a core contribution of Aya, providing researchers around the world with high-quality, human annotated data for instruction fine-tuning.

The Aya models and dataset have been released publicly,[53] and are intended to contribute to closing the language gap by providing resources for researchers and developers to further advance multilingual capabilities and safety.

To our knowledge, Aya is also the largest participatory machine learning research initiative to date, involving 3000 independent collaborators across 119 countries. Aya was a year-long, participatory research project that leveraged best practices from open-source and crowd-sourced science projects.[54, 55, 56] We designed and built a novel, intuitive user interface that our community of fluent language speakers used to collect human-curated instances of language model instructions and completions — a key component of the datasets required to develop language models.[57]

In addition to Aya, Cohere Labs runs several programs that aim to expand who can participate in machine learning research and how.
We do this by broadening access to the resources and collaborations needed to advance language model capability and safety beyond well-resourced environments, largely clustered in North America.

- Our **Scholars Program**[58] offers alternative routes into machine learning research by pairing aspiring machine learning researchers from around the world with Cohere staff to collaborate on an eight-month research project. The program provides an alternative point of entry into machine learning research, especially for researchers in nations that are underrepresented in machine learning research. Our inaugural cohort ran in 2023 and our second cohort started in Jan 2024, with scholars based in countries including Brazil, Nigeria, Germany, the UK, Canada, and the U.S.
- We additionally run a **global open science community** with over 3200 members across 126 countries, many of them in Europe, Asia, and Africa. We support this community to engage in machine learning research, share insights and resources, and explore collaborations that can advance their projects.
- Cohere Labl's **Research Grants Program** provides subsidized access to the Cohere API, allowing researchers to access state-of-the-art models for projects that aim to release a peer-reviewed scientific artifact or apply machine learning to public good projects. The program launched in July 2023 and, at the time of publication, we have awarded over 75+ grants to researchers based across 19 countries.[59]

[53] The Aya models and dataset are available via the Cohere For AI HuggingFace page: https://huggingface.co/CohereForAI.
[54] Lenart-Gansiniec, R. *et al.* (2023) 'Understanding crowdsourcing in science', *Review of Managerial Science*, 17(8), pp. 2797–2830. https://doi.org/10.1007/s11846-022-00602-z.
[55] Beck, S. *et al.* (2022) 'The Open Innovation in Science research field: a collaborative conceptualisation approach', *Industry and Innovation*, 29(2), pp. 136–185. https://doi.org/10.1080/13662716.2020.1792274.
[56] Franzoni, C. and Sauermann, H. (2014) 'Crowd science: The organization of scientific research in open collaborative projects', *Research Policy*, 43(1), pp. 1–20. https://doi.org/10.1016/j.respol.2013.07.005.
[57] Singh, S. et al. (2024) 'Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning'. arXiv. http://arxiv.org/abs/2402.06619.
[58] See: Cohere For AI (2023) Research Scholars Program. https://cohere.com/blog/c4ai-scholars-program.
[59] Cohere For AI (2024) 'Granting Access: Supporting Researchers to Use Large Language Models', *Cohere*. https://cohere.com/blog/granting-access.

Addressing language gaps in training datasets and model capabilities is also a focus for many other research groups around the world.
For example, Masakhane is a grassroots organization who, since 2020, has been working to strengthen natural language processing research in African languages.[60] In 2022, a global open-science initiative introduced BLOOM, an open-access language model trained on a dataset comprised of 46 languages.[61] In 2023, a team of researchers introduced NusaCrowd, a "collaborative initiative to collect and unify existing resources for Indonesian languages,"[62] with connections to a collaboration of researchers working to develop AI resources for Southeast Asian languages and cultures more broadly.[63] Researchers in India are developing benchmarks to contribute to measuring and improving LLM capabilities across Indic languages.[64] In late 2023, a team of researchers in Thailand released a series of language models focused on the Thai language.[65] These and many other ongoing efforts around the world — many of them grassroots community initiatives — are working to broaden the capabilities of language models across a wider range of languages.

Many governments around the world are supporting efforts to increase the languages covered by language models.
It's not just researchers in the machine learning field working to address the language gap in AI, national and international governments and public bodies are creating initiatives too. For example, the European Commission's "Common European Language Data Space" creates a platform to support and enable stakeholders to exchange language data and resources for a range of innovative purposes,[66] the Danish Government have invested EUR 4 million in a Danish language resource,[67] the South African Government is developing a platform to support natural language processing research across the nation's eleven official languages,[68] and the Indian Ministry of Electronics and Information Technology launched a program to facilitate human-machine interaction across Indian languages.[69] A 2023 report from the OECD outlines these efforts and those of many more national governments around the world, but also notes that the costs and resource requirements to develop and deploy language models remain a barrier for minority languages.[70]

[60] See Masakhane: A grassroots NLP community for Africa, by Africans. https://www.masakhane.io/.
[61] Workshop, B. *et al.* (2023) 'BLOOM: A 176B-Parameter Open-Access Multilingual Language Model'. arXiv. http://arxiv.org/abs/2211.05100.
[62] Cahyawijaya, S. et al. (2023) 'NusaCrowd: Open Source Initiative for Indonesian NLP Resources'. arXiv. https://doi.org/10.48550/arXiv.2212.09648.
[63] *SEACrowd · GitHub* (no date). https://github.com/SEACrowd.
[64] Guneet Singh Kohli (2024) 'SanskritiBench: Bridging NLP and Indian Cultural Richness'. https://sanskritibench.streamlit.app/.
[65] Pipatanakul, K. *et al.* (2023) 'Typhoon: Thai Large Language Models'. arXiv. http://arxiv.org/abs/2312.13951.
[66] European Commission (no date) *The Common European Language Data Space*. https://language-data-space.ec.europa.eu/about_en.
[67] Ministry of Foreign Affairs of Denmark (2021), 'Denmark to strengthen opportunities for NLP businesses', https://investindk.com/insights/denmark-to-strenghten-opportunities-for-nlp-businesses.
[68] Government of South Africa (2019), 'Government Establishes A New Digital Centre To Promote Indigenous Languages', https://www.dst.gov.za/index.php/media-room/latest-news/2885government-establishes-a-new-digital-centre-to-promote-indigenous-lang.
[69] Government of India (2021), Technology Development for Indian Languages (TDIL), https://www.meity.gov.in/content/technology-development-indian-languages-tdil.
[70] OECD (2023) 'AI language models: Technological, socio-economic and policy considerations.' OECD Digital Economy Papers. https://doi.org/10.1787/13d38f92-en.

Despite efforts across the machine learning research community and global government initiatives, several barriers must be addressed to close the AI language gap.

- **Gathering high quality, human-curated data from fluent speakers is resource-intensive.**
  Developing and evaluating the multilingual capabilities of language models relies on access to high-quality, human-curated data from fluent speakers of a diverse range of languages. However, this is a non-trivial challenge, especially for low-resource languages, as collecting and curating high-quality data for language model development requires significant investment of time and resources to engage with fluent language speakers.[71, 72] This is likely a major factor contributing to the fact that most efforts to date have focused primarily on easily-available and high-quality English datasets.

- **Access to digital tools needed to develop and use large language models is disparate.**
  Expanding the languages covered by AI language models will rely on the input of language speakers around the world. Fortunately, the global availability of internet-connected devices means that it is possible to connect with, engage, and collaborate with people across continents and timezones in real time. This was a key enabler for our Aya project, as it meant we could use online chat platforms to coordinate input across our global community. Unfortunately, the availability of devices and internet access is not equitable across the world.[73] Desktop and laptop computers with wired, high-speed internet are commonplace across households in more economically developed nations, but in many other parts of the world, particularly the Global South, mobile devices and cellular or satellite internet are more common. In our Aya project, approximately 54% of users accessed our data collection platform via desktop browsers while 46% utilized mobile browsers.[74] To enable participation of a wide range of language speakers, language model and dataset development requires the creation of tools that are accessible across different devices, operating systems, and internet connectivities.

- **Languages are not monolithic: they contain dialect, regional, and cultural nuances.**
  Many languages, such as are spoken across multiple regions of the world and have cultural or regional dialects. Languages in existing multilingual datasets, including our Aya dataset, have limited representation of these regional nuances as often only a few human annotators are responsible for annotating the majority of any one language dataset.[75] This might mean that data for a particular language is annotated in a way that represents the perspective of a particular contributor or cultural viewpoint. For

[71] Joshi, P. *et al.* (2021) 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World'. arXiv. https://doi.org/10.48550/arXiv.2004.09095.
[72] Ouyang, L. *et al.* (2022) 'Training language models to follow instructions with human feedback'. arXiv. https://doi.org/10.48550/arXiv.2203.02155.
[73] Avle, S. *et al.* (2020) 'Research on Mobile Phone Data in the Global South: Opportunities and Challenges'. Edited by B. Foucault Welles and S. González-Bailón, pp. 487–509. https://doi.org/10.1093/oxfordhb/9780190460518.013.33.
[74] Singh, S. et al. (2024) 'Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning'. arXiv. http://arxiv.org/abs/2402.06619.
[75] Singh, S. et al. (2024) 'Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning'. arXiv. http://arxiv.org/abs/2402.06619.

example, annotations in French might "contain many examples about the history of France, its food, songs, and other cultural practices, but not contain much information about the cultural heritage of French-speaking communities in Québec, Togo, or Senegal."[76] For English language models, it is a common practice to have multiple data annotators that contribute a broader range of perspectives to the data curation. However, these are still largely oriented towards North American and British perspectives, compared to, for example, English speakers in Singapore or South Africa. This lack of regional, cultural, and dialect diversity represented in data annotation is then reflected in similarly narrow representations of languages and their associated cultures through language model outputs. Our recent research shows that approaches to address this may involve the inclusion of multilingual data throughout the model training process, including fine-tuning and preference training.[77, 78]

[76] Singh, S. et al. (2024), p.32; see also Vigouroux, C.B. (2013) 'Francophonie', *Annual Review of Anthropology*, 42(1), pp. 379–397. https://doi.org/10.1146/annurev-anthro-092611-145804.

[77] Dang, J. et al. (forthcoming) 'RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs'.

[78] Aakanksha *et al.* (2024) 'The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm'. arXiv. http://arxiv.org/abs/2406.18682.

4. Conclusion and Considerations for Policy and Governance

This Policy Briefing Note has detailed how the language gap in AI is a significant issue that risks excluding many language speakers and communities from the benefits of language models, undermining the safety of existing models, and exacerbating existing social, linguistic, and cultural inequalities, particularly for speakers of low-resource languages.

The evidence outlined in this primer suggests that closing the AI language gap will require concerted efforts across the ecosystem, from those developing and deploying AI models, to those working in policy and governance settings. To guide those working in policy and governance settings to support these efforts, we draw on the evidence above to outline the following considerations:

1. **Direct resources towards multilingual research and development.** Address the bias in resource allocation for AI language model development by investing funding in and making other resources (e.g. expertise, compute) available to initiatives that support the development of multilingual models, especially for low-resource languages where market incentives alone will not be sufficient. Partner with existing efforts and programs — such as Aya — to bring multilingual capabilities to a wider range of language communities.

2. **Support multilingual dataset creation.** Build on and expand existing efforts to increase the availability of high-quality datasets for languages around the world that are currently under-resourced in language model development. Promote the creation of high-quality datasets for low-resource languages in collaboration with fluent language speakers and communities. Ensure accurate representation of regional, cultural, and dialect variations, including by ensuring that data collection tools are accessible across a range of devices and platforms.

3. **Recognize that the safety of all language models is improved through multilingual approaches.** Address multilingual safety vulnerabilities for all language models by encouraging the development of multilingual safety evaluation tools, datasets, benchmarks, and frameworks, and support their adoption into best practices, standards, and regulation.

4. **Foster knowledge-sharing and transparency among researchers, developers, and communities.** Expand the diversity of perspectives and expertise in language model and machine learning development by creating international fora for multilingual language model development and enable the participation of underrepresented groups and regions. Encourage greater transparency around multilingual coverage by encouraging or requiring model developers to evaluate multilingual performance, and clearly report on what languages and dialects are covered by a model.

Addressing the AI language gap is a complex and ongoing endeavor that requires sustained commitment from policymakers, researchers, developers, and communities alike. By working together to implement these considerations, we can strive towards creating a more inclusive, equitable, and safe future for language models and those they serve.

Contact:

Aidan Peppin
Policy and Responsible AI Lead
Cohere Labs