Policy Primer

Cohere For AI

Translating Safety

Developing AI safety for multilingual contexts

Aidan Peppin, Marzieh Fadaee, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Sara Hooker

December 2024

Summary

Global AI safety efforts have gained traction and momentum, but a critical challenge remains: how to ensure safety across diverse languages and cultures.

This challenge is often overlooked, or absent, in governance and research efforts to advance AI safety. Safety alignment efforts primarily focus on English or monolingual settings, leading to potential security flaws for other languages. This leaves many risks unaddressed or amplified for non-English speakers.

Addressing multilingual safety is complex, and involves reconciling global harms and unique local contexts. Most current approaches to improving model safety are language-specific and lack reliable datasets for evaluation beyond a few languages.

Despite these challenges, progress is being made. Many researchers around the world, including Cohere For AI, have dedicated efforts to tackle these language gaps, offering potential solutions to enhance AI safety across diverse linguistic and cultural contexts.

This Policy Primer summarises several promising avenues to addressing the language gap in AI safety. This includes: collecting robust multilingual evaluation data; distilling different safety instructions into models; adapting preference training to multilingual and multicultural contexts, merging models to increase performance, adapting evaluations across languages, and developing safety techniques for toxicity that keep pace with natural evolutions in language.

From this research, we identify five recommendations for researchers and policymakers to consider in their efforts to improve AI safety for everyone:

1. AI safety and alignment efforts should not be monolithic or monolingual.
2. Multiliningual safety should be addressed throughout the model training lifecycle.
3. Including more languages in safety mitigation can provide gains across all contexts.
4. Reporting on models' coverage of different languages is critical.
5. Curating data using human annotators with experiences and perspectives covering different languages and cultures is key.

CONTENTS

1. Introduction: challenges of AI safety in a global world

In recent months, global AI policy and industry attention has turned to focus on the question of AI safety, exploring social and technical questions about how to ensure that AI models and systems do not inadvertently cause harm to people and to society.

In 2024 alone we have seen the Seoul Frontier AI Safety Commitments signed by 16 companies who collectively operate in almost every country and territory around the world,[1] the inaugural meeting of the international network of AI Safety Institutes, representing 11 countries and regions,[2] the enshrinement of the EU's AI Act and the start of the process to draft the Act's Code of Practice for AI model providers, focused on models that pose 'systemic risk',[3] efforts led by Singapore to build capacity for AI safety testing across South East Asia,[4] and many more.

However, this global effort to improve AI safety faces a significant and often overlooked challenge: how to ensure safety across the multitude of global languages and cultural

[1] Frontier AI Safety Commitments, AI Seoul Summit (2024), https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024
[2] US Department of Commerc (2024), *International Network of AI Safety Institutes at Inaugural Convening*, https://www.commerce.gov/news/fact-sheets/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international
[3] European Commission (2024), *General-Purpose AI Code of Practice*, https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice
[4] IMDA (2024), *Singapore AI Safety Red Teaming Challenge*, https://www.imda.gov.sg/activities/activities-catalogue/singapore-ai-safety-red-teaming-challenge

contexts that exist today. This problem is an extension of the "AI Language Gap" we described in an earlier primer.[5]

Despite the international attention paid to AI safety and the fact that AI models – in particular Large Language Models (LLMs) – are currently available worldwide, their performance across languages and cultural contexts is not equal. Efforts to ensure safety alignment are primarily focused on homogeneous monolingual settings – predominantly English – or overfit to types of harm common in Western-centric datasets.[6, 7, 8] This creates a sharp cliff in performance which disproportionately amplifies risk for non-English speakers.[9, 10] It can also introduce critical security and safety flaws for all users of languages outside of English, where multilingual prompts can be used to subvert safety guardrails.[11, 12]

Not only is there a lack of adequate attention paid to multilingual safety, but addressing safety concerns across languages and cultural contexts is a non-trival challenge. Successful mitigation of multilingual harms involves reconciling differing global and local preferences, which involves grappling with some of the core tensions that characterize machine learning: data from multiple languages and geographies forms a heterogeneous distribution, which poses challenges for how to optimize models. Many approaches to improving AI model safety – particularly against harms such as generating violent, biased, false, or toxic content[13] – are largely oriented towards the English language or monolingual settings, and there is a lack of reliable datasets needed for safety evaluation outside of a small fraction of languages.[14, 15, 16] This includes the vast majority of work to-date focused on model alignment, including work on

---

[5] Cohere For AI (2024) *The AI Language Gap - a Policy Primer*, https://cohere.com/research/papers/the-ai-language-gap.pdf
[6] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032.
[7] Sambasivan, N. et al. (2021) 'Re-imagining Algorithmic Fairness in India and Beyond'. arXiv. https://doi.org/10.48550/arXiv.2101.09995.
[8] 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.
[9] Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. (2023) Casteist but not racist? quantifying disparities in large language model bias between india and the west, 2023. https://ar5iv.labs.arxiv.org/html/2309.08573.
[10] Üstün, A. et al. (2024) 'Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model'. ACL 2024. https://aclanthology.org/2024.acl-long.845/ .
[11] 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.
[12] Deng, Y. *et al.* (2024) 'Multilingual Jailbreak Challenges in Large Language Models'. ICLR 2024. https://openreview.net/pdf?id=vESNKdEMGp .
[13] Weidinger, L. *et al.* (2021) 'Ethical and social risks of harm from Language Models'. arXiv. http://arxiv.org/abs/2112.04359.
[14] 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.
[15] Talat, Z. *et al.* (2022) 'You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings', *BigScience 2022*, virtual+Dublin: Association for Computational Linguistics, pp. 26–41. https://doi.org/10.18653/v1/2022.bigscience-1.3.
[16] Gehman, S. *et al.* (2020) 'RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models'. arXiv. https://doi.org/10.48550/arXiv.2009.11462.

Reinforcement Learning from Human Feedback (RLHF)[17, 18, 19] or AI Feedback (RLAIF)[20, 21] which are core components of improving model safety.

Solving these language gaps across AI safety is not a wholly-ignored field, however. Over recent years, we have spent considerable time working on these problems – alongside many other researchers and groups. In the next section we provide an overview of these approaches and how they can offer steps towards improving AI safety across languages and cultural contexts.

2. Extending safety guardrails across languages

2.1 Data: collecting evaluation data from both local and global contexts

Languages are deeply rooted in the cultural and social fabric of a community and they evolve to capture the unique nuances and perspectives of groups of people.[22] The majority of current language models reflect the world as seen through anglo-centric and predominantly North American texts and datasets, which introduces a skew away from languages and cultural perspectives that are not included during model training.[23, 24, 25, 26] This is largely because many datasets used in natural language processing represent only a handful of data-rich languages, and the datasets used for instruction-fine-tuning or preference training are almost entirely

[17] Stiennon, N. *et al.* (2022) 'Learning to summarize from human feedback'. arXiv. https://doi.org/10.48550/arXiv.2009.01325.
[18] Christiano, P. *et al.* (2023) 'Deep reinforcement learning from human preferences'. arXiv. https://doi.org/10.48550/arXiv.1706.03741.
[19] Dai, J. *et al.* (2023) 'Safe RLHF: Safe Reinforcement Learning from Human Feedback'. arXiv. https://doi.org/10.48550/arXiv.2310.12773.
[20] Bai, Y. *et al.* (2022) 'Constitutional AI: Harmlessness from AI Feedback'. arXiv. https://doi.org/10.48550/arXiv.2212.08073.
[21] Tunstall, L. *et al.* (2023) 'Zephyr: Direct Distillation of LM Alignment'. arXiv. https://doi.org/10.48550/arXiv.2310.16944.
[22] Ramezani, A. and Xu, Y. (2023) 'Knowledge of cultural moral norms in large language models', ACL 2023,. https://doi.org/10.18653/v1/2023.acl-long.26.
[23] 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.
[24] Kunchukuttan, A., Jain, S. and Kejriwal, R. (2021) 'A Large-scale Evaluation of Neural Machine Transliteration for Indic Languages'. EACL 2021, https://doi.org/10.18653/v1/2021.eacl-main.303.
[25] 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.
[26] 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.

focused on English.[27, 28, 29, 30] Acquiring data that has a high-enough quality for use in training language models is challenging.[31, 32] This is because 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.[33, 34, 35]

To bridge this gap requires building multilingual datasets explicitly intended for use in improving model safety. This motivated our work to create and openly release multilingual datasets for both training and evaluation in relation to toxicity, extending existing English datasets commonly employed for toxicity mitigation and evaluation studies by incorporating translations of these datasets into other 8 languages. These expanded datasets are used for training and evaluation of multilingual toxicity mitigation, while also establishing a foundational benchmark for future research in this field.[36]

Additionally, the Aya Evaluation Suite is a diverse evaluation suite for multilingual, open-ended generation quality.[37] It consists of 250 human-written prompts for each of 7 languages (English, Portuguese, Chinese (simplified), Arabic, Telugu, Turkish, and Yoruba), 200 automatically translated but human-selected prompts for 101 languages (including 114 dialects), and human-edited prompts of the latter. This dataset is tailored for assessing capabilities of LLMs such as brainstorming, planning, and other unstructured, long-form responses, which are key to understanding models' capabilities that are relevant for safety. Having both human annotations and translations provides more robust evaluations, but often, it is expensive to rely solely on human annotation. Our work has shown that complementing rare human annotations

---

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

[28] 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. p12.

[29] Maxwell, M. and Hughes, B. (2006) 'Frontiers in Linguistic Annotation for Lower-Density Languages', Association for Computational Linguistics. https://aclanthology.org/W06-0605.

[30] Joshi, P. *et al.* (2019) 'Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities', NLP Association of India. https://aclanthology.org/2019.icon-1.25.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] Pozzobon, L. *et al.* (2024) 'From One to Many: Expanding the Scope of Toxicity Mitigation in Language Models'. Findings of ACL 2024. https://aclanthology.org/2024.findings-acl.893/ .

[37] Singh, S. et al. (2024) 'Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning'. ACL 2024 https://aclanthology.org/2024.acl-long.620/ .

with synthetic and translated evaluations can be effective, and allows for one-to-one comparisons across languages.

However, while automatic translation approaches like those outlined above are a popular starting point to address data scarcity, it does not solve all multilingual challenges. This is because automatic translation relies on parallel data – i.e. examples of the same text in two or more languages – which is often harder to find in high quantity and quality. Additionally, translations can introduce erroneous artifacts, and nuances of the original script can be hard to capture in translation.[38, 39, 40. 41, 42, 43] This is particularly true for translated prompts used in safety evaluations, where they can lose their harmful intent or become meaningless through translation errors.[44]

To address this, we developed the Aya Red Teaming dataset with the help of compensated native speakers in 8 different languages: English, Hindi, French, Spanish, Russian, Arabic, Serbian and Filipino.[45] To build this data set, we worked with annotators with native language skills to craft prompts around a list of harmful categories. The annotators also provided the corresponding English translations, identified the categories of harm present, and assigned a label indicating whether the harm is "global" or "local" in nature. Here, "global" harm refers to model outputs that are understood and recognized as harmful across global contexts, whereas "local" harm is tied to specific cultural or historical contexts.

In addition to developing datasets, multilingual safety can be improved by extending technical approaches to cover multilingual contexts. We outline approaches to this below, drawing on our recent research efforts.

---

[38] Vanmassenhove, E., Shterionov, D. and Gwilliam, M. (2021) 'Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation'. EACL 2021, https://doi.org/10.18653/v1/2021.eacl-main.188.
[39] Hartung, K. et al. (2023) 'Measuring Sentiment Bias in Machine Translation'. arXiv. https://doi.org/10.48550/arXiv.2306.07152.
[40] Savoldi, B. et al. (2021) 'Gender Bias in Machine Translation', Transactions of the Association for Computational Linguistics. Edited by B. Roark and A. Nenkova, 9, pp. 845–874. https://doi.org/10.1162/tacl_a_00401.
[41] Meng Ji, Meng Ji, Pierrette Bouillon, and Mark Seligman. (2023) Cultural and Linguistic Bias of Neural Machine Translation Technology, pp. 100–128. Studies in Natural Language Processing. Cambridge University Press.
[42] Chen, P. *et al.* (2024) 'Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models?' arXiv. https://doi.org/10.48550/arXiv.2406.12822.
[43] Choenni, R. *et al.* (2024) 'On the Evaluation Practices in Multilingual NLP: Can Machine Translation Offer an Alternative to Human Translations?' arXiv. https://doi.org/10.48550/arXiv.2406.14267.
[44] Agrawal, A.S., Fazili, B. and Jyothi, P. (2024) 'Translation Errors Significantly Impact Low-Resource Languages in Cross-Lingual Learning'. arXiv. https://doi.org/10.48550/arXiv.2402.02080.
[45] Aakanksha *et al.* (2024) 'The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm'. arXiv. http://arxiv.org/abs/2406.18682.

2.2 Safety Context Distillation

A core safety guardrail for language models is the ability to refuse to respond to potentially harmful prompts. For example, when a model is prompted to produce hate speech, it will refuse to do so. To develop the Aya 101 model and ensure its ability to refuse harmful prompts across different languages, we used 'safety context distillation' to teach the model in which contexts refusals are appropriate.[46] This method distilled refusals across different languages from a "teacher" model into the Aya 101 model. Instead of manually defining refusal templates for specific safety contexts – which is resource and time intensive – through this approach we generated a dataset of diverse refusals based on previously published harmful prompts.

We expanded the language coverage of this dataset with automatic translation to 101 languages, and the generated (safe) responses were then paired with the original prompts to finetune our Aya model. We found this step reduced harmful generations from adversarial prompts by 78–89% as judged by human experts. However, this came with the cost of negatively impacting general quality of outputs. This trade-off between model safety and general performance is not uncommon.

2.3 Preference Training

Preference optimization techniques have become a standard final stage for training state-of-art LLMs. These techniques provide models with feedback on their outputs, so they can learn what a high quality output looks like. However, again the vast majority of work to-date on preference optimization has focused on globally dominant languages like English and Chinese. This year we introduced a novel, scalable method for generating high-quality multilingual feedback data.[47] We created a dataset of prompts translated from English into 22 languages and generated completions for each language using multiple LLMs of varying multilingual capabilities. We used this dataset to compare preference optimization techniques. Interestingly, we found that increasing the number of languages in the training data improved overall performance, again highlighting the need for multilingual data to improve LLMs.

Additionally, using the Aya Red-teaming dataset (see above), we investigated the effectiveness of two common optimization techniques, Direct Preference Optimization (DPO) and Supervised Fine-tuning (SFT), for multilingual safety alignment.[48] We demonstrated that these techniques

---

[46] Üstün, A. et al. (2024) 'Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model'. arXiv. http://arxiv.org/abs/2402.07827.
[47] Dang, J. *et al.* (2024) 'RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs'. arXiv. https://doi.org/10.48550/arXiv.2407.02552.
[48] Aakanksha *et al.* (2024) 'The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm'. arXiv. http://arxiv.org/abs/2406.18682.

can significantly reduce harmful model generations across various languages while maintaining general performance.

Through this research we explored the interplay between mitigating global and local harms, finding that training on local harms improves the mitigation of global harms, suggesting shared learning between these categories.

2.4 Model Merging

In another research effort, we explored merging specialized models in a diverse multi-task setting, combining safety and general-purpose tasks within a multilingual context.[49] Model merging is a technique where we combine the strengths of different specialized models to create a more capable and balanced system, particularly for handling multiple languages.

We compared different approaches to merging models with a method called "data mixing", where the model is trained on a "mix" of multilingual data. Our findings revealed that merging models generally worked better than mixing data alone, leading to noticeable improvements in both safety and overall performance.

These results highlight how model merging can help build stronger and safer multilingual systems, offering clear advantages for handling complex tasks in diverse languages.

2.5 Multitask benchmarks

Cultural biases in multitask datasets limit their utility as global benchmarks. These biases arise not only from differences in language but also from the cultural knowledge required to interpret and understand questions effectively. Translations often introduce artifacts that distort meaning or clarity, further complicating evaluations.

We analyzed the Massive Multitask Language Understanding (MMLU) benchmark – a commonly used benchmark for assessing LLM capability – and found significant Western-centric biases.[50] Notably, 28% of questions require specific knowledge of Western culture, while 84.9% of geography-related questions focus exclusively on North American or European regions. These findings underscore how existing benchmarks prioritize Western concepts, distorting evaluations of multilingual models. To address these challenges, we developed Global-MMLU (G-MMLU), an enhanced multilingual test set covering 42 languages.

---

[49] Aakanksha *et al.* (2024) 'Mix Data or Merge Models? Optimizing for Diverse Multi-Task Learning'. arXiv. https://doi.org/10.48550/arXiv.2410.10801.

[50] Singh, S. et al. (2024) 'Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation'. arXiv. https://arxiv.org/abs/2412.03304

Concurrently with G-MMLU, we curated another multilingual evaluation set, INCLUDE, to focus on capturing regional and cultural knowledge across 44 languages.[51] The benchmark is constructed from a diverse set of 197,243 multiple-choice questions from local exams, professional certifications, and academic tests. This approach addresses the limitations of existing multilingual benchmarks that often rely on translations from English, which fail to adequately capture regional nuances and knowledge. Using this dataset, we found that the performance of LLMs can vary significantly across languages and regions, particularly on questions that require specific cultural or regional understanding. For example, GPT-4o performs better on global history exams than regional history exams, suggesting a gap in local knowledge. This highlights the need for more comprehensive evaluation resources to ensure the equitable and effective deployment of LLMs across diverse linguistic communities.

2.6 Tackling toxic language as it evolves

Another challenge for multilingual safety relates to the crucial fact that languages evolve naturally over time.[52, 53, 54] Considerable effort has been dedicated to mitigating toxicity – the generation of offensive or harmful text-content – but existing methods often require drastic modifications to model parameters or the use of computationally intensive methods. This means keeping toxicity safety guards up-to-date as language evolves is onerous.

In research we published in 2023, we introduced Goodtriever, a flexible methodology that matches the current state-of-the-art toxicity mitigation while being more computationally efficient.[55] Building on this, we explored multilingual toxicity mitigation for text generation, moving beyond the traditional English-centric approaches.[56] Covering nine languages, our experiments yielded insights into the complexities of multilingual toxicity mitigation, offering ways for future research in this increasingly important field.

---

[51] Romanou, A. et al, (2024) 'INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge'. arXiv. https://arxiv.org/abs/2411.19799

[52] Frermann, L. and Lapata, M. (2016) 'A Bayesian Model of Diachronic Meaning Change', Transactions of the Association for Computational Linguistics. Edited by L. Lee, M. Johnson, and K. Toutanova, 4, pp. 31–45. https://doi.org/10.1162/tacl_a_00081.

[53] Horn, F. (2021) 'Exploring Word Usage Change with Continuously Evolving Embeddings', Association for Computational Linguistics, pp. 290–297. https://doi.org/10.18653/v1/2021.acl-demo.35.

[54] Jaidka, K., Chhaya, N. and Ungar, L. (2018) 'Diachronic degradation of language models: Insights from social media', Association for Computational Linguistics, pp. 195–200. https://doi.org/10.18653/v1/P18-2032.

[55] Pozzobon, L. *et al.* (2023) 'Goodtriever: Adaptive Toxicity Mitigation with Retrieval-augmented Models'. arXiv. http://arxiv.org/abs/2310.07589.

[56] 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.

3. Recommendations for safety across multiple languages

If global efforts to address AI model safety are to be successful in assessing and mitigating risks across global contexts, they must account for and address multilingual gaps. It is therefore necessary to develop a comprehensive understanding of and mitigation for the safety concerns across diverse linguistic and cultural contexts. However, data scarcity and limited efforts to address safety across languages and cultures contribute to AI models being predominantly safety-optimized for English and North American linguistic and sociocultural norms only.

This means that AI model developers and policymakers must make provisions to ensure AI models are safe for all the language and cultural contexts in which they are deployed. From our work on mitigating harm across languages we offer the following considerations:

6. **AI safety and alignment efforts should not be monolithic or monolingual:** alignment techniques, model evaluations, and the data involved in them should cover multiple cultures and languages - especially for the contexts across which a model is deployed.
7. **Multiliningual safety should be addressed throughout the model training lifecycle**, as demonstrated by the approaches we outlined above, as well as other examples of multilingual IFT,[57] self-supervised learning,[58] and training across multilingual datasets.[59]
8. **Including more languages in safety mitigation can provide gains across all contexts**. There is a strong potential for cross-lingual transfer, so even including additional limited multilingual data can raise robustness and safety for all users.
9. **Reporting on the coverage of different languages is critical** to provide clarity around models' multilingual performance and any implications for potential safety gaps or vulnerabilities as a result of gaps in coverage.
10. **Curating data using human annotators with experiences and perspectives covering different languages and cultures is key** to improving performance and safety across language contexts.

---

[57] Shaham, U. et al. (2024) 'Multilingual Instruction Tuning With Just a Pinch of Multilinguality'. arXiv. https://doi.org/10.48550/arXiv.2401.01854.
[58] Chau Tran *et al.* (2020) 'Cross-lingual Retrieval for Iterative Self-Supervised Training'. https://ai.meta.com/research/publications/cross-lingual-retrieval-for-Iterative-self-supervised-training/.
[59] Llama Team, AI @ Meta (2024) 'Introducing Meta Llama 3: The most capable openly available LLM to date'. https://ai.meta.com/blog/meta-llama-3/.