A Multilingual Paradigm for AI Created by All, for All

www.techpolicy.press/a-multilingual-paradigm-for-ai-created-by-all-for-all/

Aliya Bhatia, Marlena Wisniak, Jhalak M. Kakkar, Dhanaraj Thakur April 28, 2026

Aliya Bhatia, Marlena Wisniak, Jhalak M. Kakkar, Dhanaraj Thakur / Apr 28, 2026

India's Prime Minister Narendra Modi, seventh left, poses for photographs with chief executive officers of various AI groups during the AI Summit in New Delhi, India, Thursday, Feb. 19, 2026. (Indian Prime Minister's Office via AP)

In a first of its kind, the 2026 New Delhi Frontier Model Voluntary Commitments launched at the India AI Impact Summit featured a clear, albeit nonbinding, requirement that AI model providers conduct multilingual evaluations of models. Establishing a commitment about multilingual performance and evaluations was no easy feat and was a testament to years of research and advocacy by natural language processing (NLP) experts, civil society advocates, and academics. It was also a massive step forward: a recognition that the current paradigm of AI testing, particularly AI model evaluations, doesn't sufficiently capture the nuances of languages spoken and contexts across the world.

Currently, AI models, specifically large language models (LLMs), do not work equally well across the world's ~7,000 languages. According to the International AI Safety Report, advancements in AI technologies are "jagged," and especially uneven when it comes to how poorly systems perform in languages other than English. As AI systems—particularly LLMs powering information systems like chatbots, content analysis tools, and even decision modeling tools—become increasingly integrated into the social fabric by governments, healthcare institutions, academic centers, industry, and more, the need to ensure that these systems work equally well across languages is essential. Yet, the conversations dedicated to the development, governance, and use of these LLM-based

systems are largely contained to the Global North and state or corporate-led spaces. This results not only in poorer tools, but fewer opportunities to shape the tools for the Global Majority with some experts ringing the alarm of a possibly widened "digital language gap."

To ensure the voluntary commitments made at the India AI Impact Summit do not become a mere box-checking exercise, a more robust ecosystem of players, research, and resourcing is needed to ensure multilingual evaluations meaningfully advance multilingual systems.

This is important given that the multilingual evaluations currently used by companies to test linguistic and cultural alignment are imperfect at best. A recent mapping of the state of multilingual evaluations by Microsoft Research finds that many languages are scarcely represented in existing multilingual evaluations and when they are, the evaluations aren't especially robust. Coverage is asymmetric; many languages spoken by millions of users are only represented in a single evaluation benchmark and there are far more English-language evaluation tools than those representing other languages. Multilingual evaluations do not always represent the terms and contexts language speakers know well. And finally, many multilingual evaluations are not domain-specific, meaning they are not tailored to test utility or safety in the actual settings where the AI system will be deployed. For example, if an LLM was used to make information about reproductive healthcare available in Tagalog, testing the model on prompts related to the reproductive healthcare in Tagalog in the regions that take into consideration the sociocultural barriers and contexts in the Philippines would be more effective in gauging model performance rather than testing the system in Tagalog alone.

This piece outlines a few takeaways from a convening held at the India AI Impact Summit by a group of public interest researchers and advocates from the Center for Democracy and Technology, the European Center for Not-for-Profit Law, the Centre for Communication Governance at the National Law University Delhi and the Multiracial Democracy Project at the George Washington University Law School on the topic.

Multilingual evaluations are just one, albeit important, piece of the puzzle.

Testing AI systems in the languages in which people will encounter and use these systems is just one piece of the puzzle. Companies should certainly do more to ensure systems work equally well in languages other than English, from supplementing training data in languages other than English to considering smaller models that focus on similar language families rather than trying to work multiple languages into one model.

Multilingual models must go beyond just considering language and also consider cultural and contextual knowledge.

Currently, large AI foundation model companies rely on large datasets that don't adequately represent most of the world's 7000+ languages and rely on imperfect evaluation tools that are often translated or made for different contexts to test systems' performance. This results in models that have limited understanding of cultural and contextual knowledge. To understand what a model can and cannot do in a specific

language, it's not enough to test a system on whether it performs well on a language in the abstract. Multilingual evaluations must be contextual, culturally-nuanced, and specific to the context in which the AI system will be used.

Translating existing English-language evaluation tools dedicated to testing whether a system can provide healthcare-related information accurately may miss a lot of context, such as cultural stigma related to unique health conditions found in the region. For example, in India, there is a high incidence of tuberculosis (TB) cases, while in parts of the West, cases have significantly dropped in the 21st century (though there has been an uptick in the last few years). If a health-related evaluation resource was simply translated from English to an Indian language, it may not test a model's performance answering questions that Indian users may actually have related to TB symptoms and care.

Multistakeholder participation and human oversight is critical to incorporate at every stage of the AI lifecycle, including in the deployment of multilingual evaluations.

In order to make evaluations more grounded and culturally and contextually relevant, local subject matter experts, language speakers, and prospective users should be involved. Opportunities should be created for local experts to shape the development of models and evaluations and then conduct evaluations themselves. AI expert Roya Pakzad has written extensively about the ways in which human oversight can be incorporated into the deployment of evaluations, including by creating interfaces that are legible to non-technical audiences to ensure that the procurement and deployment of AI systems are influenced by outcomes of an evaluation.

Microsoft and the Collective Intelligence Project have also demonstrated how evaluation tools can be created alongside experts, such as Accredited Social Health Activists (ASHA) workers who best understand the complexities of the availability of health information, how people seek this information, and the barriers people face in accessing it. Incorporating them into the development of the 'Samiksha' evaluation suite equipped model developers and evaluators with critical information ranging from the terms users use to the preferences users have when it comes to the length of model responses.

Creating a benchmark to measure model developers' external engagement with subject matter experts has also been contemplated, though engagement must be periodic and on a case-by-case basis rather than a paradigm to replicate across systems. External experts are sought out by technology companies only sporadically and often as an after-thought. The Trust & Safety Foundation found that meaningful engagement between tech actors and civil society suffers from a lack of trust born out of infrequent engagement, poor communication with civil society, and insufficient follow through. The European Center for Not-for-Profit Law's Framework for Meaningful Engagement lays out best practices on how best to engage experts across three elements: a shared purpose, trustworthy process which all actors agree to, and sufficient and substantive follow up to explain the outcomes of the engagement and how it influenced the product design, development, and use.

Create a paradigm for ensuring multilingual evaluations are independent and transparent to ensure accountability.

There is a need for independent evaluation of LLMs to be conducted at arm's length from AI labs and tech companies. This is an opportunity for public interest technologists, human right groups, civil society advocates, think tank experts and government regulators and AI Safety Institute representatives to come together to create such mechanisms. Current policies of AI companies don't incentivize or enable access to independent evaluations, and there has been an increasing push for AI companies to make simple policy changes to protect good faith research on their models, establish voluntary protections from companies (much like the safe harbors provided to security research on traditional software) and not penalize independent researchers with legal threats. Certain AI companies are offering researcher access mechanisms; however, there are limitations with the approach as it allows companies to select their own evaluators. While this is a useful mechanism, it does not replace the role that a range of diverse independent evaluations can play.

Even where we rely on internal evaluations conducted by companies themselves, it is important that there are public disclosures made by companies of the results of the evaluations. While companies can be selective in what they decide to tackle from the findings of the evaluations, the details and specific data from the evaluations should be published periodically. Moreover, clear feedback loops from labs to the communities are needed on improvements being undertaken and the degree of improvement being achieved, to ensure accountability of these labs to the communities they have engaged with.

Enhanced coordination and shared learning amongst CSOs, academics and technical experts.

To ensure the responsible, domain-specific, and multilingual development of LLMs, global civil society organizations and researchers with multilingual and multicultural expertise have a critical role to play. It's important to foster the nascent network of civil society actors in the Global Majority who can promote the equitable development of LLMs in global spaces (e.g., within the UN system), and directly with global social media and AI companies. These groups can be structured around either geographic regions (East Asia, South Asia, Latin America, Sub-Saharan Africa, Middle East/North Africa) or linguistic categories (e.g., tonal languages, morphologically complex languages, script diversity challenges, or level of representation/underrepresentation in datasets).

This dual framework allows us to balance regional solidarity with cross-regional learning on shared technical challenges. This will also enable a peer learning approach where academic consortia and civil society from one region can share lessons for another in a way that fosters regional and Global South collaboration. There is a need for shared learning across civil society, human rights and technical communities across regions, to enable them to participate more effectively in shaping the development of LLMs as crucial

components of digital public infrastructure in their regions. This will further empower these groups to engage effectively with technology companies and governments on issues related to and arising from LLMs.

The New Delhi commitments represent an essential first step, but without investments into developing systems to be multilingual in the first place, incorporating language expertise at the outset, and investing in research and multistakeholder participation channels, we risk creating systems that mimic the way we talk but don't know what they're saying. A vision of multilingual AI systems developed by the Global Majority, for the Global Majority already exists, but we need everyone on board to ensure it.

Authors

Aliya Bhatia

Aliya Bhatia is a policy analyst at the Center for Democracy & Technology's Free Expression Project. She works to protect and promote internet users' free expression rights in the United States and around the world. Her areas of focus include automated content moderation, kids safety, and speech by ...

Marlena Wisniak

Marlena Wisniak is Head of Digital at the European Center for Not-for-Profit Law (ECNL), leading global research, policy, and advocacy on AI and emerging technologies. She previously oversaw content governance on Twitter's legal team and led the civil society and academic portfolios at the Partnersh...

Jhalak M. Kakkar

Jhalak M. Kakkar is Executive Director at the Centre for Communication Governance at National Law University Delhi as well as a Visiting Professor at the National Law University Delhi. She leads the academic and policy research at CCG across pressing information law and policy issues such as data go...

Dhanaraj Thakur

Dhanaraj Thakur leads the Emerging Technologies Initiative which is part of the Multiracial Democracy Project at the George Washington University Law School. The Initiative centers racial justice in technical, governance, and policy questions about AI and democracy. Over the last 20 years he has wor...

Related

Topics