A report from

cdt | Research

EXECUTIVE SUMMARY - ENGLISH

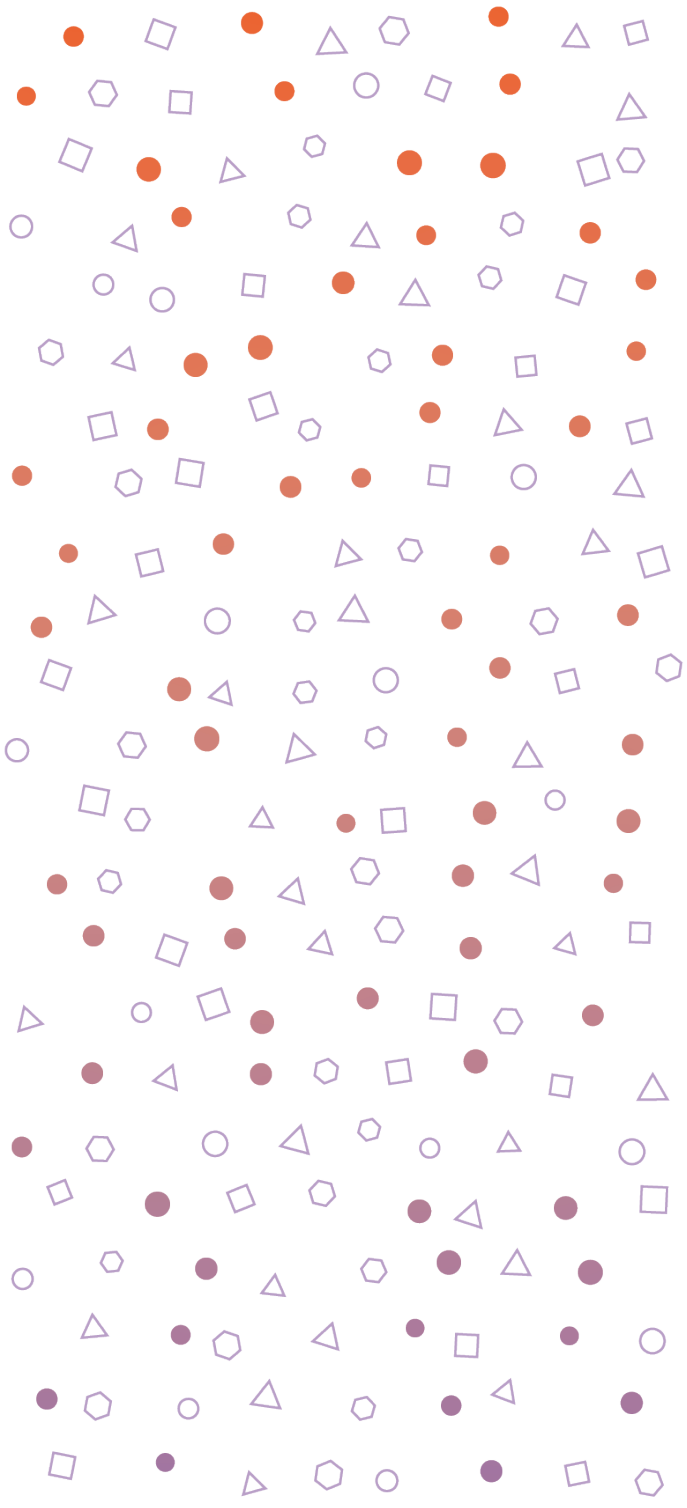
Lost in Translation_

Large Language Models in Non-English Content Analysis

Gabriel Nicholas
Aliya Bhatia

May 2023

Executive Summary



The internet is the primary source of information, economic opportunity, and community for many around the world. However, the automated systems that increasingly mediate our interactions online — such as chatbots, content moderation systems, and search engines — are primarily designed for and work far more effectively in English than in the world's other 7,000 languages.

In recent years, large language models have become the dominant approach for building AI systems to analyze and generate language online, but again, they have been built primarily for the English language. A large language model (e.g., Open AI's GPT-4, Meta's LLaMa, Google's PaLM) is a machine learning algorithm that scans enormous volumes of text to learn which words and sentences frequently appear near one another and in what context. Large language models can be adapted to perform a wide range of tasks across different domains. They are most known for being used to build chatbots, such as ChatGPT, but researchers and technology companies also use them for content analysis tasks, such as sentiment analysis, text summarization, and hate speech detection. Google, Meta, Microsoft, and other companies have already incorporated large language models into their core product functions, such as content moderation and search. Other vendors soon may incorporate them into automated decision-making systems, such as resume scanners.

Recently though, researchers and technology companies have attempted to extend the capabilities of large language models into languages other than English by building what are called *multilingual language models*. Instead of being trained on text from only one language, multilingual language models are trained on text from dozens or hundreds of languages at once. Researchers posit that multilingual language models infer connections between languages, allowing them to apply word associations and underlying grammatical rules learned from languages with more text data available to train on (in particular English) to those with less. In some applications, multilingual language models outperform models trained on only one language — for instance, a model trained on lots of text from lots of languages, including Hindi, might perform better in Hindi contexts than a model just trained on Hindi text.

Multilingual language models give technology companies a way to scale their AI systems to many languages at once, and some have already begun to integrate them into their products. Online service providers in particular have deployed multilingual language models to moderate content: Meta uses a multilingual language model to detect harmful content on its platforms in over 100 languages; Alphabet's Perspective API uses one to detect toxic content in eighteen different languages; Bumble uses one to detect and take action on unwanted sexual messages around the world.

Multilingual language models allow technologists to attempt to build models in languages for which they otherwise might not have enough digitized text. Languages vary widely in *resourcedness*, or the volume, quality, and diversity of text data they have available to train language models on. English is the highest resourced language by multiple orders of magnitude, but Spanish, Chinese, German, and a handful of other languages are sufficiently high resource enough to build language models in. Medium resource languages, with fewer but still high-quality data sets, such as Russian, Hebrew, and Vietnamese, and low resource languages, with almost no training data sets, such as Amharic, Cherokee, and Haitian Creole, have too little text for training their own large language models. Language data in low resource languages is also often of particularly poor quality: either it is mistranslated or even nonsensical language scraped from the internet, or is limited to sources with narrow domains, such as religious texts and Wikipedia. This gap in data availability between languages is known as the *resourcedness gap*.

Multilingual language models are designed to address these gaps in data availability by inferring semantic and grammatical connections between higher- and lower-resource languages, allowing the former to bootstrap the latter. However, this architecture raises its own concerns. Multilingual language models are still usually trained disproportionately on English language text and thus end up transferring values and assumptions encoded in English into other language contexts where they may not belong. For example, a multilingual model might associate the word "dove" in all languages with "peace" even though the Basque word for dove ("uso") can be an insult. The disparity in available data also means multilingual language models work far better in higher resource languages and languages similar to them than lower resource ones. Model developers will sometimes try to fill in these gaps with machine-translated text, but translation errors may further compound language misrepresentation. And when multilingual language models do fail, their unintuitive connections between languages can make those problems harder to identify, diagnose, and fix.

Large language models' general use in content analysis raises further concerns. Computational linguists argue that large language models are limited in their capacity to analyze forms of expression not included in their training data, meaning they may struggle to perform in new contexts. They may also reproduce any biases present in their training data. Often, this text is scraped from the internet, meaning that large language models may encode and reinforce dominant views expressed online.

Companies, researchers, and governments each have a role to play in protecting the public from the potential dangers of multilingual language model content analysis systems. To ensure better public accountability, companies that deploy large language models should always be transparent about how

they use them and in which languages. Companies should deploy language models with narrow remits and adequate channels for human review.

Researchers and research funders meanwhile should invest in efforts to improve the use and performance of language models in languages other than English, in particular, to reduce failures that disparately impact speakers of lower-resourced languages. The best way to do this is by supporting language-specific research communities, who can promote the virtuous cycle of collecting data, curating datasets, training language models, publishing, and building applications. Local language speakers and context experts need to be part of each step of this process and also be curating the data and assessing the language models deployed by large, global online services.

Finally, governments need to be careful about how they use or encourage the use of large language models. Large language models should never power systems used to make high-stakes decisions without oversight, such as decisions about immigration status or healthcare, nor should governments mandate or inadvertently require by law the use of large language model-powered systems to moderate content from online services. Instead, governments should convene different stakeholders to align on what norms and guardrails should be around developing and deploying large language models.

Large language models in general and multilingual language models in particular have the potential to create new economic opportunities and improve the web for all. However, mis- or over-application of these technologies poses real threats to individuals' rights, such as undermining their right to free expression by inaccurately taking down a person's post on social media or their right to be free of discrimination by misinterpreting an individual's job or visa application. Multilingual language models specifically can inadvertently further entrench the Anglocentrism they are intended to address. In light of these limitations, technology companies, researchers, and governments must consider potential human and civil rights risks when studying, procuring, developing, or using multilingual language models to power systems, in particular when they are used to make critical information available or play a role in decisions affecting people's access to economic opportunities, liberty, or other important interests or rights.

A report from

cdt | Research

Read the full report at *cdt.org*.

CONTRIBUTIONS BY

Samir Jain, Mallory Knodel, Emma Llansó, Michal Luria, Nathalie Maréchal, Dhanaraj Thakur, and Caitlin Vogus.

ACKNOWLEDGEMENTS

We thank Pratik Joshi, Sebastin Santy, and Aniket Kesari for their invaluable feedback on the technical aspects of this report. We also thank Jacqueline Rowe, Damini Satija, and Ángel Díaz for their insightful comments and suggestions. All views in this report are those of CDT.

The translation of our executive summary is made possible by Global Voices Translations and with the help of Iverna McGowan, Maria Villamar, Ophélie Stockhem, and Tomás Pomar.

This work is made possible through a grant from the John S. and James L. Knight Foundation.

Suggested Citation: Nicholas, G. and Bhatia, A. (2023) Lost in Translation: Large Language Models in Non-English Content Analysis. Center for Democracy & Technology. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1996, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C., and has a Europe Office in Brussels, Belgium.

GABRIEL NICHOLAS

Research Fellow at the Center for Democracy & Technology.

ALIYA BHATIA

Policy Analyst, Free Expression Project at the Center for Democracy & Technology.

CENTER FOR
DEMOCRACY
& TECHNOLOGY