Commentary

Toward Better Automated Content Moderation in Low-Resource Languages

Gabriel Nicholas and Aliya Bhatia

1 Introduction

Social media companies have learned the hard way that poor moderation of content in languages other than English can have grave consequences. Leaving harmful content up, particularly in regions where social media platforms are primary news and communication channels, has fueled conspiracy theories, violence, and even genocide around the world (Fink 2018; Iyengar 2018). American social media companies have long faced criticism for underinvesting in regions outside the US and Europe. In response, Meta, Google, and others have begun to deploy new multilingual language models that they claim can effectively detect and take action on harmful content in dozens if not hundreds of languages (Jigsaw 2021; Meta AI 2021b).

In previous work, we have argued that these models have critical shortcomings that limit their ability to perform highly language- and context-specific tasks, such as content moderation (Nicholas and Bhatia 2023b). One shortcoming is that these multilingual language models are trained predominantly on English-language text, which leads them to apply an Anglocentric lens onto their analysis of texts from non-English linguistic and cultural contexts. This is due in large part to what natural language processing (NLP) researchers call the "resourcedness gap," or the gap between the quantity, quality, and diversity of training data available in English and every other language (Joshi et al. 2020). In low-resource languages there are few, if any, high-quality examples of digitized text, which hampers developers of these models from training and evaluating models on high-quality examples of speech in those languages.

The resourcedness gap exists due to many factors, including British colonialism, which has driven mass production of English-language text, and the hegemony of American technology companies, which has further contributed to English as the language of the internet and digital exchange. This gap may widen if funders and those with the mandate for public interest do not intervene (Nicholas and Bhatia 2023a). However, Western technology companies are not financially incentivized to close this gap, and global academic institutions also tend to prioritize and privilege research and development of technologies in English at the expense of other, lower-resource languages (Bender 2019). Without key investments, the current incentive structures will continue to perpetuate the preferential treatment of the English language in computer science and, in turn, in automated trust and safety systems. This commentary argues that what we now face is a free-rider problem, where technology companies, academic institutions, and the public at large would benefit greatly from increased investments into low-resource language development, but no one actor is currently incentivized to do so. But with a few modest investments, governments, grantmaking organizations, and social media companies can

doi:10.54501/jots.v2i1.150

play an important role in igniting a virtuous cycle of research to bridge the gap between English and non-English AI capabilities.

2 Language models struggle to moderate content in non-English languages

Today, language models are used to enforce content policies on search engines, social media platforms, and even chatbot services like ChatGPT (OpenAI 2023). This represents a paradigm shift: five years ago, social media companies generally built their automated content moderation systems by training individual classifiers for every language and type of harmful content they sought to detect (Bommasani et al. 2021; Meta AI 2021c); today, many companies fine-tune more general purpose pretrained language models to the task of content moderation, taking advantage of their broader exposure to language. Since 2022, a whole cottage industry has sprung up of vendors offering content-moderation-as-a-service using GPT-4 (Bernard 2023).

Companies also use multilingual large language models for multilingual content moderation. For instance, Google's Perspective API uses a large language model to detect "rude, disrespectful, or unreasonable" content that affects participation, which works in over a dozen different languages (Lees et al. 2022); Meta claims their XLM-R language model can detect harmful content in over 100 languages (Meta AI 2021a); and even the dating app Bumble uses a large language model to detect unwanted messages in Thai, Portuguese, Russian, and a dozen other languages (Belloni 2021). As companies seek to cut costs and expand into new global markets, we will likely see more platforms adopt models for moderating content in the world's languages.

While state-of-the-art language models are able to convincingly analyze and generate text across dozens, if not hundreds of different languages, they are not necessarily up to the task of moderating content in those languages. As we have discussed in previous work, the reasons are twofold. First, many languages do not have enough high-quality digitized text available for models to adequately "learn" those languages. Second, language models learn lower-resource languages by drawing on their connections to higher-resource ones—usually English—and may thereby end up importing English-language assumptions and viewpoints when used for content moderation (Nicholas and Bhatia 2023b).

There are significant disparities in the quantity and quality of text data available across the world's languages. English is by far the most high-resource language, with multiple orders of magnitude more high-quality data and from a wider range of sources than any other language (Joshi et al. 2020), although other languages such as Spanish, German, and Mandarin are also very high-resource. Lower-resource languages, such as Vietnamese, Bengali, Haitian Creole, and Farsi, have far fewer and often lower-quality datasets available, despite having tens or hundreds of millions of speakers. In even lower-resource languages, such as Tigrinya, Navajo, and Uyghur, longer-form examples of text that may exist on the web are either few and far between, such as a handful of Wikipedia articles or parliamentary proceedings, or are of low quality, replete with profanity, bias, gibberish, and words that native speakers don't use machine-translated from English (Kreutzer et al. 2022). As a result of these disparities, most language models, even multilingual ones, are trained predominantly on English text, some on upwards of 90% (Touvron et al. 2023).

The reason large language models are able to perform basic content generation and analysis in lower-resource languages at all is because the language models draw connections between them and high-resource languages, usually English (Artetxe, Ruder,

and Yogatama 2020; Muller et al. 2021; Wu and Dredze 2020). For instance, a model may not encounter the Turkish word for "rain" (*yağmur*) and the Turkish word for "umbrella" (*şemsiye*) in its training data, but it can learn the connection between the two words transitively, if it can learn that *rain* and *yağmur*, and *şemsiye* and *umbrella* are related. This allows language models to produce impressive-appearing text even in languages for which it does not have much data, but means that they struggle when used for more language- and locality-specific tasks, like content moderation, because they are still parsing this text through an Anglocentric frame (Nicholas and Bhatia 2023a).

The lack of data in low-resource languages and models' overreliance on English-language connections means that they struggle in highly context and language-specific tasks like content moderation. Limited, low-quality training data means these models have a limited and incomplete worldview of a particular language or dialect, and this can have outsized implications for a company's ability to moderate content.

If a models' understanding of a language is refracted through the lens of English, it may apply Anglocentric norms of language onto cultures where there are different shared norms. For example, if a model has learned that *dove* is often associated with connotations of peace, it may apply that understanding onto the Basque word for "dove," *uso*, which is often used in a derogatory and homophobic context. This poor understanding of Basque may impede a model from detecting derogatory or outright hateful speech and leave up content that violates a service's policies and contributes to potentially hostile discourse.

When models are trained, there is a tendency toward aiming for broad coverage rather than cultural and linguistic nuance, which is necessary for content moderation (Bergman and Diab 2022). This is particularly important in low-resource, polyglossic languages, those like Arabic where a word in a particular language has many meanings depending on the region and context in which it is used. A model may learn one dialect of Arabic from a more dominant culture, for which there is more training data available, and apply it to different regional and cultural contexts. In non-language model-based content moderation, this broad coverage has led to damaging mistakes. For example, training a classifier with only a few examples of an Arabic dialect misclassified words like "gay" and "lesbian," resulting in at-scale takedowns of Arabic speech by and for the LGBTQ+ community (Hassine 2016). Palestinian activists have asserted that their posts and tweet campaigns documenting human rights abuses were removed because part of their tweets used a word that was associated with a terror group in another Arabic-speaking community (Debre and Akram 2021). By applying only a limited construction of the Arabic language at scale, language models may also exacerbate existing efforts to silence marginalized communities.

The low availability of high-quality digitized text in many of the world's widely spoken languages also impedes companies from testing their language models and strengthening their safety measures. For example, OpenAI conducts adversarial testing by domain experts to test the strength of their tools and safety measures to reduce harmful outputs. "Red teamers" for OpenAI's tools, for instance, found that when a model was prompted to create recruitment propaganda for terrorist groups in English, the model refused; however, when the model was given the same prompt in Farsi, it fulfilled the request (Murgia 2023). Research has also found that language models more often produce hallucinations, errors, bias, and malicious outputs in languages other than English (Lin et al. 2021; Muller et al. 2021). This kind of gap in the functionality of safety measures can be harmful to all users if a malicious actor is seeking to generate malicious code or an at-scale influence operation to undermine access to the ballot.

The asymmetric focus on English resourcing for language models closely resembles

companies' inequitable distribution of content moderation capacity in the past. Lifting the bar of available training resources in low-resource languages can strengthen companies' ability to host important content and mitigate the spread of content that can have tragic offline consequences.

3 There is a lack of incentives to invest in low-resource language research

Despite the benefit to the public at large, neither private companies nor NLP researchers are incentivized to develop training and evaluation sets for low-resource languages. For private companies, this can be classified as a free-rider problem. In economic terms, this means that the data used to train language models (i.e., data scraped from the web) is a non-rivalrous, non-excludable public good, so while all companies would benefit from it, no individual company stands to exclusively reap the rewards of investing to improve it.

Most state-of-the-art language models today are built using huge volumes of web-scraped data.[1] Scraping the web is relatively cheap, scalable, and technically straightforward for companies to do. In English and other high-resource languages, it can help developers amass colossal data sets on many different topics with high relevance to the real world. However, web scraping is not as effective for obtaining lots of high-quality text in low-resource languages. Web data in low-resource languages is more often machine translated, misidentified by language detection software, about less diverse topics, and of lower quality along many other dimensions (Kreutzer et al. 2022). Building high-quality datasets in these languages therefore requires the more costly measures of finding, creating, and scanning texts to build new language corpora (Perrigo 2023). Building evaluation datasets requires the further step of hiring and training native language speakers to label data (Nguyen et al. 2022; Salganik 2019).

For private companies, it is not clear that they will recoup their costs from the investment. As popular wisdom goes, languages besides English offer smaller market opportunities relative to the cost, and many lucrative opportunities for language models, such as in scientific research and global commerce, already leverage English as the lingua franca. Although it may be financially worthwhile for companies to invest in training and testing their models in some languages spoken in larger, wealthier countries, it may not be commercially worthwhile for them to invest in lower-resource languages spoken predominantly in economically weaker countries, even if they have hundreds of millions of speakers. In other words, the same market forces that have led social media companies to underinvest in content moderation in the global majority world also undermine the effectiveness of language model-based moderation in those contexts as well (Amrute, Singh, and Guzmán 2022).

Academics are also not incentivized to address the resourcedness gap between languages. Although many datasets and benchmarks in the field of natural language processing do come from academia, academics focus far more on English than any other language. Between May 2022 and January 2023, there were likely 100 times more NLP publications about English than the next highest language (German).[2] Many of the

1. Social media companies could in theory train their language models with user data that is not available on the public web, but this raises a whole host of potential legal, privacy, and public relations concerns outside the scope of this commentary.

2. Papers that do not explicitly mention any language in its abstract are almost always about English (Bender 2019). Of the 5,290 papers the Association for Computational Linguistics published between May 2022 and January 2023, 4,720 mentioned no language in its abstract and 311 mentioned English. The next highest language mentioned was German, with 27 (ACL Rolling Review Dashboard, n.d.).

lowest-resource languages are overlooked by NLP researchers altogether.

This is at least in part because many languages do not have their own academic publications or conferences, and certainly not ones with the same reputational status as those focused on English. English and a handful of other high-resource languages experience a virtuous cycle of investment: researchers collect data, create benchmarks, and build models in these languages to publish their results in conferences and journals, burnishing the reputations of both themselves and their outlets and making it easier for other NLP researchers to do work in the future. Lower-resource languages, however, experience this as a vicious cycle: research is not only difficult for all the reasons mentioned above, but it is difficult to get attention—and funding—for their work from the larger, more English-centric NLP community (Nicholas and Bhatia 2023a).

4 The right kind of investment in NLP research can address the resourcedness gap

In order to improve automated content moderation in more of the world's languages, social media companies need access to more training and testing data in those languages. Perhaps social media companies should make this investment on their own, but they are not financially incentivized to, and public pressure and the law are limited in how much they can convince them otherwise. A lower-hanging fruit is for funders, including the federal government, grantmaking organizations, and larger tech companies' external research funding arms, to invest in low-resource language NLP research initiatives. Financial investments should be aimed at growing grassroots efforts, so as to shift low-resource language research communities from a vicious to a virtuous cycle of research. We argue that successful efforts will be managed by local language experts, help build self-sustaining research ecosystems, and include parallel social science work.

The closest example to this kind of initiative was DARPA's Low Resource Languages for Emergent Incidents (LORELEI) research program, which ran from 2015 to 2020 (Christianson, Duncan, and Onyshkevych 2018). The program was aimed at building translation and information extraction software to help US humanitarian efforts operate and communicate in low-resource language areas. The project led to the creation of many high-quality datasets in medium- and low- resource languages that today bolster models' performance in those languages. However, LORELEI was particularly interested in rapid deployment, meaning it wanted its participants to build models that could be quickly repurposed to learn any new language from a small amount of text (Strassel and Tracey 2016). It also directed its $26 million of funding only to large American universities and defense contractors, rather than directly to researchers from the language communities they sought to serve (Diño 2015). Today, LORELEI's approach of building models optimized to scale across as many languages as possible is industry standard. This may work for basic translation, but as we discussed earlier in this commentary, this trade-off of breadth versus depth hurts models' ability to perform the deeply language- and region-specific task of content moderation.

In order to build deeper NLP expertise in low-resource languages, funders can help support existing local collectives of language- and language family-specific NLP research networks who are working to digitize and build tools for some of the lowest-resource languages. Collectives such as Masakhane (African languages) (Orife et al. 2020), IndoNLP (Indonesian languages) (Aji et al. 2022), AmericasNLP (Indigenous languages of the Americas) (Mager et al. 2021), and ARBML (Arabic dialects) (Alyafeai and Al-Shaibani 2020) often have deep knowledge of where the largest gaps are in their language's specific research but sorely lack the funds necessary to address them. Cohere AI's Aya

project too may act as a central node of language development experts to work with to facilitate connections with local language NLP groups around the world (Cohere AI, n.d.).

Funders can target these efforts more toward trust and safety by sponsoring specific research agendas both in the US and abroad. The most direct way for funders to do this is to sponsor the creation of new publications, conferences, academic and industry research collaborations, and competitions in specific low-resource languages. One model for how this can work is exemplified by EVALITA, an event hosted by the Italian Association for Computational Linguistics. In it, researchers submit datasets for new language tasks and benchmarks, such as dating documents or identifying misogyny. Then, researchers compete to train models to maximize those benchmarks and publish the best results in conference proceedings, thereby driving interest and attention toward Italian NLP and creating resources companies and external stakeholders can use to evaluate Italian-language hate speech detection systems (Basile et al. 2020). EVALITA has led to the creation of many trust and safety-related datasets, some of which at least Google (and perhaps others) have used to evaluate and improve their Italian-language moderation systems (Lees et al. 2022).

Finally, funders should also consider supporting social science research to better understand where the limitations and shortcomings are in low-resource language content moderation. Better understanding is needed at both a micro- and macro-level. At the micro-level, research into the negative effects and disparate impacts of poor content moderation in specific language contexts is necessary to help social media companies better allocate resources and aid policymakers, academics, and civil society in developing solutions. At the macro-level, research is needed to equip the NLP field with an understanding of how to balance the positive effects of making language models operate better in more lower-resource languages (e.g., economic inclusion, protection from linguistic erasure, stopping scams and propaganda), from the negatives (e.g., labor displacement, dual-use problems with language models generating dangerous content).

Today, the divide between high- and low-resource languages is a structural one. Power is concentrated in a few languages, particularly English, and companies continue to invest disproportionately in those languages. This colonial divide trickles down into our language models and, in turn, our content moderation systems, where a lack of high-quality data in many languages spoken by the majority of the world's people leaves those speakers disenfranchised and, worse, exposed to all of the internet's greatest dangers. Addressing this gap at a structural level requires direct involvement and input from local language experts, and funding local NLP research in low-resource languages is one step toward ensuring social media companies are serving everyone.

References

ACL Rolling Review Dashboard. n.d. "Papers mentioning >0 languages." Accessed July 23, 2023. https://stats.aclrollingreview.org/submissions/linguistic-diversity/.

Aji, Alham Fikri, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, et al. 2022. "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 7226–49. Dublin, Ireland: Association for Computational Linguistics, May. https://doi.org/10.18653/v1/2022.acl-long.500.

Alyafeai, Zaid, and Maged Al-Shaibani. 2020. "ARBML: Democratizing Arabic Natural Language Processing Tools." In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS),* 8–13. Association for Computational Linguistics, November. https://doi.org/10.18653/v1/2020.nlposs-1.2.

Amrute, Sareeta, Ranjit Singh, and Rigoberto Lara Guzmán. 2022. *A Primer on AI in/from the Majority World.* Technical report. Data & Society, September. https://datasociety.net/library/a-primer-on-ai-in-from-the-majority-world/.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. 2020. "On the Cross-lingual Transferability of Monolingual Representations." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 4623–37. ArXiv:1910.11856 [cs]. https://doi.org/10.18653/v1/2020.acl-main.421.

Basile, Valerio, Maria Di Maro, Danilo Croce, and Lucia Passaro. 2020. "EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian." In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.* December 17, 2020.

Belloni, Massimo. 2021. *Multilingual message content moderation at scale.* Medium, December 8, 2021. https://medium.com/bumble-tech/multilingual-message-content-moderation-at-scale-ddd0da1e23ed.

Bender, Emily. 2019. "The #BenderRule: On Naming the Languages We Study and Why It Matters." *The Gradient* (September 15, 2019). https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

Bergman, A. Stevie, and Mona T. Diab. 2022. "Towards Responsible Natural Language Annotation for the Varieties of Arabic." In *Findings of the Association for Computational Linguistics: ACL 2022.* March 17, 2022. http://arxiv.org/abs/2203.09597.

Bernard, Tim. 2023. "The Evolving Trust and Safety Vendor Ecosystem." *Tech Policy Press* (July 24, 2023). https://techpolicy.press/the-evolving-trust-and-safety-vendor-ecosystem/.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. *On the Opportunities and Risks of Foundation Models.* Technical report. ArXiv:2108.07258 [cs] type: article. Stanford Center for Research on Foundation Models, August 18, 2021. https://crfm.stanford.edu/assets/report.pdf.

Christianson, Caitlin, Jason Duncan, and Boyan Onyshkevych. 2018. "Overview of the DARPA LORELEI Program." *Machine Translation* 32, no. 1 (June 1, 2018): 3–9. ISSN: 1573-0573. https://doi.org/10.1007/s10590-017-9212-4.

Cohere AI. n.d. "Aya." Accessed August 1, 2023. https://sites.google.com/cohere.com/aya-en/home.

Debre, Isabel, and Fares Akram. 2021. "Facebook's language gaps weaken screening of hate, terrorism." *AP News* (October 25, 2021). https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f.

Diño, Gino. 2015. "DARPA Doles out Millions to Academia & Vendors to Translate Any Language by 2019." *Slator* (December 4, 2015). https://slator.com/darpa-doles-out-millions-to-academia-and-vendors-to-translate-any-language-by-2019/.

Fink, Christina. 2018. "Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar." *Journal of International Affairs* 71 (1): 43–52. ISSN: 0022-197X.

Hassine, Wafa Ben. 2016. *How Arab Governments Use the Law to Silence Expression Online.* Technical report. Electronic Frontier Foundation. https://www.eff.org/pages/crime-speech-how-arab-governments-use-law-silence-expression-online.

Iyengar, Rishi. 2018. "WhatsApp has been linked to lynchings in India. Facebook is trying to contain the crisis." *CNN* (September 30, 2018). https://www.cnn.com/2018/09/30/tech/facebook-whatsapp-india-misinformation/index.html.

Jigsaw. 2021. *10 New Languages for Perspective API.* Technical report. Jigsaw, December 10, 2021. https://medium.com/jigsaw/10-new-languages-for-perspective-api-8cb0ad599d7c.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 6282–93. Association for Computational Linguistics, July. https://doi.org/10.18653/v1/2020.acl-main.560.

Kreutzer, Julia, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, et al. 2022. "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets." *Transactions of the Association for Computational Linguistics* 10 (January 31, 2022): 50–72. ISSN: 2307-387X. https://doi.org/10.1162/tacl_a_00447.

Lees, Alyssa, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. "A New Generation of Perspective API: Efficient Multilingual Character-level Transformers." In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 3197–207. KDD '22. New York, NY, USA: Association for Computing Machinery, August 14, 2022. ISBN: 978-1-4503-9385-0. https://doi.org/10.1145/3534678.3539147.

Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, et al. 2021. "Few-shot Learning with Multilingual Language Models." *arXiv:2112.10668 [cs]* (December 20, 2021). http://arxiv.org/abs/2112.10668.

Mager, Manuel, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, et al. 2021. "Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas." In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas,* 202–17. Association for Computational Linguistics, June. https://doi.org/10.18653/v1/2021.americasnlp-1.23.

Meta AI. 2021a. *Harmful content can evolve quickly. Our new AI system adapts to tackle it.,* December 8, 2021. https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/.

Meta AI. 2021b. *Meta's New AI System to Help Tackle Harmful Content,* December 8, 2021. https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/.

———. 2021c. *The shift to generalized AI to better identify violating content,* November 9, 2021. https://ai.facebook.com/blog/the-shift-to-generalized-ai-to-better-identify-violating-content/.

Muller, Benjamin, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. "When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 448–62. Association for Computational Linguistics, June. https://doi.org/10.18653/v1/2021.naacl-main.38.

Murgia, Madhumita. 2023. "OpenAI's red team: the experts hired to 'break' ChatGPT." *Financial Times* (April 14, 2023). https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8.

Nguyen, Cuong, Daniel Nkemelu, Ankit Mehta, and Michael Best. 2022. "Why So Inflammatory? Explainability in Automatic Detection of Inflammatory Social Media Users." In *Proceedings of Practical ML for Developing Countries at ICLR 2022.* August 21, 2022. http://arxiv.org/abs/2208.09941.

Nicholas, Gabriel, and Aliya Bhatia. 2023a. *Lost in Translation: Large Language Models in Non-English Content Analysis.* Technical report. Center for Democracy & Technology, May 23, 2023. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.

———. 2023b. "The Dire Defect of 'Multilingual' AI Content Moderation." *WIRED* (May 23, 2023). ISSN: 1059-1028. https://www.wired.com/story/content-moderation-language-artificial-intelligence/.

OpenAI. 2023. *Our approach to AI safety.* Technical report. OpenAI Blog, April 5, 2023. https://openai.com/blog/our-approach-to-ai-safety.

Orife, Iroro, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, et al. 2020. *Masakhane – Machine Translation For Africa.* Technical report. Masakhane, March 13, 2020. https://doi.org/10.48550/arXiv.2003.11529.

Perrigo, Billy. 2023. "AI By the People, For the People." *Time* (July 27, 2023). https://time.com/6297403/the-workers-behind-ai-rarely-see-its-rewards-this-indian-startup-wants-to-fix-that/.

Salganik, Matthew J. 2019. *Bit by bit: Social research in the digital age.* Princeton University Press. ISBN: 0-691-19610-9. https://press.princeton.edu/books/paperback/9780691196107/bit-by-bit.

Strassel, Stephanie, and Jennifer Tracey. 2016. "LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 3273–80. Portorož, Slovenia: European Language Resources Association (ELRA), May. https://aclanthology.org/L16-1521.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models,* July 18, 2023. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Wu, Shijie, and Mark Dredze. 2020. “Are All Languages Created Equal in Multilingual BERT?” In *Proceedings of the 5th Workshop on Representation Learning for NLP,* 120–30. Association for Computational Linguistics, July. https://doi.org/10.18653/v1/2020.repl4nlp-1.16.

Authors

Gabriel Nicholas is Research Fellow at the Center for Democracy & Technology and a Non-Resident Fellow at the NYU Information Law Institute. He can be reached at gnicholas@cdt.org.

Aliya Bhatia is a Policy Analyst on the Free Expression Team at the Center for Democracy & Technology. She can be reached at abhatia@cdt.org.

Acknowledgements

We would like to thank Dan Bateyko, Samir Jain, Emma Llansó, and Dhanaraj Thakur for their input on this commentary.

Funding

This work is made possible through a grant from the John S. and James L. Knight Foundation. The Center for Democracy & Technology receives roughly two-fifths of its funding from a mix of corporate donors, which you can find listed here. Corporate donors have no influence or control over our work products or priorities, and none were contacted as part of this commentary.

Keywords

Language models; content moderation; low-resource languages; global south