Multilingual Multicultural LLMs

Wei Xu[1], **David Ifeoluwa Adelani**[2], **Vinodkumar Prabhakaran**[3], **Sara Hooker**[4], **Alan Ritter**[1]
[1] Georgia Institute of Technology [2] Mila/McGill University [3] Google [4] Adaption Labs
{wei.xu, alan.ritter}@cc.gatech.edu david.adelani@mila.quebec vinodkpg@google.com

This tutorial proposal has been accepted and we will present the tutorial at EMNLP 2026.

Abstract

While large language models (LLMs) have substantially expanded support for non-English languages, their ability to generalize across typologically diverse languages and produce contextually appropriate, culturally grounded responses remains a significant challenge. This tutorial examines the rapidly evolving area of understanding, evaluating, and improving the multilingual capabilities and multicultural alignment of LLMs. We will primarily focus on **cutting-edge** advances and latest publications from 2025–2026, while incorporating selected **introductory** material to provide necessary background and to ensure accessibility for a broad audience. Participants will gain an up-to-date view of both methods and practical insights, including multilingual pre-, mid- and post-training, multilingual-specific optimization techniques, culturally aware evaluation, and key open problems in developing more fair, inclusive, and robust LLMs.

1 Background and Motivation

Increasingly, open-weight LLMs are expanding beyond English to support a broader range of languages, following the lead of proprietary models. Early examples include BLOOM (Scao et al., 2022) and Aya (Üstün et al., 2024). More recent models, such as Gemma (Gemma, 2025) and Qwen (Yang et al., 2025a), have made substantial progress in narrowing performance gaps between English and other languages. This marks a notable change from just a few years ago, when non-English languages consistently lagged due to the "curse of multilinguality" (Chang et al., 2024; Conneau et al., 2020).

These developments also reflect an important shift: rather than treating non-English languages as an afterthought, recent LLM research directly aims to achieve more equitable performance across a broad range of languages. More encouragingly, the past two years have seen a surge of work exploring the cultural dimensions of LLMs (see Figure 1), alongside a growing body of culturally grounded datasets spanning diverse languages and communities. Yet these advances have not fully translated into consistent, reliable performance across linguistic and cultural settings (Shani et al., 2026). We hope this tutorial will inspire more creative approaches and broader efforts toward truly multilingual, culturally grounded LLMs.

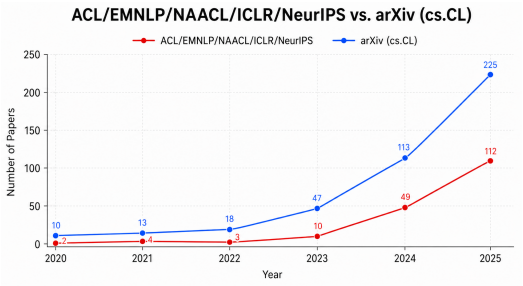

Figure 1: Rapid growth in publications on cultural aspects of LLMs from 2020 to 2025 (2026 ongoing).

2 Description of the Tutorial

In this tutorial, we will discuss five central aspects of building multilingual LLMs that aim to be not only linguistically capable but also culturally informed and inclusive:

For **LLM pre-training**, recent research (Blevins et al., 2024) has started to train cross-lingual Mixture-of-Experts (MoE) model with language-specific routings (Pfeiffer et al., 2022) that generalize to unseen languages with reduced GPU computing requirements. Other research looked into tokenization (Arnett et al., 2026; Vieira et al., 2025; Abagyan et al., 2025; Liu et al., 2025a; Limisiewicz et al., 2024; Foroutan et al., 2025), scaling laws (Chen et al., 2025b), continued pre-training (Saito et al., 2024), how LLMs handle multilingualism internally (Zhao et al., 2024; Wendler et al., 2024), implications in using multilingual pre-training data mixture (Langlais et al., 2026; Xie

Dataset	Topics	Manually-Curated	Human Preferences	Multi modal	Lang uages	Format
FORK (Palta and Rudinger, 2023)	food-related customs	✓	✗	✗	en	multi-choice QA
CulturalBench (Chiu et al., 2025)	daily life, social etiquette, wider society	✗	✗	✗	en	multi-choice QA
CANDLE (Nguyen et al., 2023)	geography, religion, occupation, food, clothing, etc.	✗	✗	✗	en	sentence, entity
CultureBank ((Shi et al., 2024))	social norms, food, communication, festivals, etc.	✗	✗	✗	en	free-text QA, descriptions
PRISM (Kirk et al., 2024)*	work, religion, family, politics, culture, value, etc.	✓	✓	✗	en	multi-turn conversations
GeoMLAMA (Yin et al., 2022)	cultural commonsense	✓	✗	✗	5	masked prompt
BLEnD (Myung et al., 2024)	food, sports, family, education, holidays, work-life	✓	✗	✗	13	multi-choice & short-answer QA
Include (Romanou et al., 2025)*	exams from schools, occupational licenses, etc.	✓	✗	✗	44	multi-choice QA
CAMeL (Naous and Xu, 2025)	beverage, clothing, food, location, religion, sports, etc.	✓	✗	✗	ar, en	masked prompt / entity
Camellia (Naous et al., 2025)	literature, beverage, food, location, names, sports, etc.	✓	✗	✗	10	masked prompt / entity
SHADES (Mitchell et al., 2025)	culture-specific stereotypes	✓	✗	✗	16	paragraph
JMMMU (Onohara et al., 2025)*	art, heritage, history, business, science, medicine, etc.	✓	✗	✗	ja	multi-choice QA
Aya (Singh et al., 2024)*	news, stories, recipes, scientific texts, etc.	✓	✗	✗	65	free-text QA
NativQA (Hasan et al., 2025)	animals, business, clothing, education, events, food, etc.	✓	✗	✗	7	free-text QA
CaLMQA (Arora et al., 2025)	governance, literature, food, economy, customs, health, etc.	✓	✗	✗	23	free-text QA
M-RewardBench (Gureja et al., 2025)*	general-purpose instructions, queries, safety tests	✓̃	✓	✗	23	free-text QA
HelpSteer3-M (Wang et al., 2025b)*	instruction, literacy, factoid, sports, opinions, etc.	✓	✓	✗	13	free-text QA
CARE (Guo et al., 2025)	cultural entities, opinion, norms, commonsense, literacy	✓	✓	✗	ar, ja, zh	free text QA
CVQA (Mogrovejo et al., 2024)	vehicles, people, sports, plants, animals, brands, etc.	✓	✗	✓	31	multi-choice QA
ALM-bench (Vayani et al., 2025)	traditional festivals, foods, rituals, landmarks, etc.	✓̃	✗	✓	100	multi-choice & free-text QA
CulturalVQA (Nayak et al., 2024)	food, drink, clothing, rituals, traditions, etc.	✓	✗	✓	en	short-answer QA
CII-Bench (Zhang et al., 2025a)	life, art, society, politics, environment, traditional culture	✓	✗	✓	zh	multi-choice QA
Afri-MCQA (Tonja et al., 2026)	public figures, food, sports, traditions, brands, etc.	✓	✗	✓	15	multi-choice & free-text QA
C^3B (Song et al., 2026)	cultural comic concepts	✓̃	✗	✓	5	free-text QA

Table 1: A representative subset of existing datasets for studying cultural awareness and regional localization in LLMs. Most of these datasets contain only 100–1,500 examples per language, underscoring the difficulty of collecting culture-specific data. In contrast, larger datasets often mix culture-agnostic (*) and culture-specific content or are generated through automatic translation followed by human verification (✓̃).

et al., 2025) or translated pre-training data (Wang et al., 2025a). There has also been renewed interest in encoder-based architectures, highlighted by mmBERT (Marone et al., 2025), which adopts a 3-stage pre-training pipeline supporting more than 1,000 languages. These efforts reflect steady progress on improving multilingual embeddings (Wang et al., 2024; Lee et al., 2025) and scaling language coverage across increasingly diverse linguistic settings. Beyond briefly surveying these developments, we plan to discuss in depth three state-of-the-art works focused on pre-training data, methods for multilingual mid-training or continual pre-training (Fujii et al., 2024; Yoo et al., 2026; Dorkin et al., 2026), tokenization, and multilingual embeddings, with the latter two also carrying important implications for mid- and post-training.

For **LLM post-training**, one of the most common and scalable approaches is to translate English datasets into other languages (Dang et al., 2024; Peter et al., 2025; Gureja et al., 2025; Le et al., 2024; Parekh et al., 2024; Liu et al., 2025b). However, this often introduces "translationese" (Li et al., 2025a)[1] and provides limited coverage of regional knowledge and cultural diversity. Other work constructs multilingual preference data by native speakers (Guo et al., 2025; Wang et al., 2025c; Zhang et al., 2026). In this tutorial, we will focus on methods that introduce cross-lingual chains of thought at inference time (Son et al., 2026; Weihua et al., 2025; Li et al., 2025b) and multilingual training algorithms that better leverage cross-lingual supervision and reasoning (She et al., 2024; Yang et al., 2025b; Zhao et al., 2025; Hwang et al., 2025; Zhang et al., 2025b; Guo et al., 2026). One caveat is that many of these approaches still assume that high-resource languages, particularly English, produce higher-quality generations. We will discuss why this assumption does not always hold, as different tasks and domains may favor different languages (Huang et al., 2024; Guo et al., 2026), along with potential solutions and future directions.

At the same time, newer benchmarks are being developed for **multicultural and multimodal evaluation**, that ranges from helpfulness (Wang et al., 2025c), values (Kirk et al., 2024), morality (Jin et al., 2025), to Western cultural favoritism (Naous et al., 2024, 2025), and exams (Romanou et al., 2025) in LLMs. Many of the multicultural evaluation are often focused on questions on cultural categories such as food, clothing, celebrations across various countries and languages (Myung et al., 2024; Mogrovejo et al., 2024). However, they often tend to equate culture with specific countries, some make use of translated benchmarks (Vayani et al., 2025), and many of the questions often do not require multi-step reasoning (Zheng et al., 2025) which cause these benchmarks to lose their usefulness or "disappear" shortly after release.

[1]Translationese refers to linguistic patterns and artifacts that distinguish translated text from naturally written text in the target language, and can arise in both human and machine translations (Wein and Schneider, 2024; Graham et al., 2020).

Finally, **the social and ethical considerations** opens up new challenges in the context of multicultural LLMs. Technologies with relatively monocultural development contexts interacting with increasingly multi-cultural usage contexts with divergent values and interpretive practices gives rise to cultural incongruencies (Prabhakaran et al., 2022) that can lead to systemic barriers, value imposition, safety gaps, violation of cultural norms, and even cultural erasure. Addressing these issues requires moving beyond multilingual to multicultural approaches, that account for value pluralism, i.e., the existence of diverse and sometimes conflicting ethical and social values across the globe. This is especially important in AI safety — what is considered harmful, offensive, or unsafe is deeply embedded in specific cultural contexts (Aroyo et al., 2023; Rastogi et al., 2025; Davani et al., 2024). Another core consideration is more community-engaged and non-extractive participatory research (Birhane et al., 2022) as well as a more nuanced thick evaluation approaches (Qadri et al., 2025) that provides more deep, qualitative, and contextual understanding over purely quantitative metrics. Finally, we must also account for differences in preferences and perceptions of AI across the globe; for instance, tendencies towards anthropomorphism and the expectations for AI interaction differ widely (Schimmelpfennig et al., 2025), impacting user trust, adoption, and the technology's overall societal benefit.

3 Tutorial Outline

We will build in 10%-15% of time for Q&A in each part of the tutorial.

A. Introduction (10 min.)

B. Pre-training of Multilingual LLMs (40 min.)

- The "curse of multilinguality" and recent strategies to mitigate it
- Multilingual pre-training corpora
- Tokenization
- Mulitlingual embeddings

C. Post-training of Multilingual LLMs (50 min.)

- Multilingual human preference learning
- Multilingual reward models for reinforcement learning
- Multilingual reasoning
- Techniques for improving cultural awareness and performance on non-English languages
- Cross-lingual training strategies
- The role of translation

D. Building Culturally-aware LLM Benchmarks (30 min.)

- Multicultural evaluation of LLMs and VLMs
- Scaling multicultural evaluation to diverse regions and languages
- From translated to multicultural benchmarks
- Towards more challenging multilingual benchmarks

F. Social and Ethical Considerations (30 min.)

- Multilingual $\neq$ Multicultural
- Value Pluralism and AI
- Evaluating AI safety cultural contexts
- Community-engaged evaluations and data
- The idea of *Thick Evaluations*
- Preferences, perceptions of AI across the globe (e.g., Anthropomorphism)

G. Discussions and Future Directions (20 min.)

4 Tutorial Information

Type of the Tutorial. The tutorial is primarily ***cutting-edge***, focusing on recent research published in 2025 and 2026, while also incorporating selected ***introductory*** material to provide necessary background and ensure accessibility.

Length. This is a 3-hour tutorial.

Venues. EMNLP 2026. All instructors will be able to attend EMNLP in-person, but not AACL-IJNLP.

Breadth of Research Covered. This tutorial is intended to cover recent research advances from both academia and industry. We estimate that 20% of the work covered in this tutorial will be from the presenters and the remaining 80% by others.

Target Audience. The tutorial will be accessible to anyone who has a basic knowledge of natural language processing. We think the topic will be of interest to researchers, students, and practitioners in both academia and industry. We will primarily target at the CL/NLP community while bridging to a broader audience from HCI, Linguistics, Social Science, and other interdisciplinary fields.

Diversity Considerations. The tutorial presenters bring diverse perspectives, representing four different institutions (two universities and two leading industry AI labs) and at different career stages. They also reflect broad geographic and cultural diversity, having been born in Asia, America, Europe,

and Africa. **David Adelani is a long-time member and a Director at the Masakhane**,[2] a grassroots organisation whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans. Wei Xu has coordinated cross-national research collaborations involving researchers across multiple Asian countries. We will leverage our networks to further broaden participation. We will also reach out to academic and community organizations such as WiNLP,[3] LatinX in AI,[4] EOSC-Nordic,[5] and Southeast Asian NLP to encourage engagement from diverse audiences.

Estimated Size of Audience. We anticipate that the tutorial will attract an audience of at least 100-150 people, and potentially more. This estimation is based on many relevant workshops and events at *ACL/EMNLP conferences, such as *AfricaNLP* and *Slavic NLP* Workshops at ACL 2025, *Workshop on Cross-Cultural Considerations in NLP (C3NLP)* at NAACL 2025, Bird-of-Feathers events *Southeast Asian NLP Community, Projects, and Beyond* and *Multilingualism* at ACL 2025.

Open Access. We will build a website for the tutorial and share all teaching material freely online. We also agree to allow the publication of slides and video recordings by ACL.

Technical Equipment. Standard AC setup and access to the Internet.

5 Ethics Statement

This tutorial aims to advance the fairness, inclusivity, and accessibility of LLMs across diverse linguistic and cultural contexts (Adilazuarda et al., 2024; Blevins et al., 2025). We recognize that multilingual and multicultural AI systems can reflect and amplify existing societal biases (Li et al., 2024), cultural stereotypes (Nadeem et al., 2021; An et al., 2023), and global power imbalances (Ahia et al., 2023). There are also other works (Chiu et al., 2024; Kirk et al., 2024; Huang et al., 2025; Choenni et al., 2024) that explored diverse values and subjective viewpoints (Zhong et al., 2021). We will highlight relevant research papers that either uncovered (Tan and Lee, 2025; Naous et al., 2025) or addressed (Limisiewicz et al., 2024; Guo et al., 2025) these relevant issues related to these topics as part of the tutorial.

[2] https://www.masakhane.io/
[3] https://www.winlp.org/
[4] https://www.latinxinai.org/
[5] https://www.eosc-nordic.eu/

6 Reading list

- The Roots of Performance Disparity in Multilingual Language Models: Intrinsic Modeling Difficulty or Design Choices? (Shani et al., 2026)
- Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models (Blevins et al., 2024)
- Explaining and Mitigating Crosslingual Tokenizer Inequities (Arnett et al., 2026)
- mmBERT: A Modern Multilingual Encoder with Annealed Language Learning (Marone et al., 2025)
- All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages (Vayani et al., 2025)
- Learning to Route Languages for Multilingual Policy Optimization (Guo et al., 2026)
- CARE: Multilingual Human Preference Learning for Cultural Awareness (Guo et al., 2025)
- Chameleon: A Flexible Data-mixing Framework for Language Model Pretraining and Finetuning (Xie et al., 2025)
- DialUp! Modeling the Language Continuum by Adapting Models to Dialects and Dialects to Models (Bafna et al., 2025)
- Having Beer after Prayer? Measuring Cultural Bias in Large Language Models (Naous et al., 2024)
- INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge (Romanou et al., 2025)
- M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (Chen et al., 2025a)
- Parity-Aware Byte-Pair Encoding: Improving Cross-lingual Fairness in Tokenization (Foroutan et al., 2025)
- Hire your Anthropologist! Tethinking Culture Benchmarks through an Anthropological Lens (AlKhamissi et al., 2026)
- Scaling Cultural Resources for Improving Generative Models Lens (Stepanyan et al., 2026)
- Challenges and strategies in cross-cultural NLP (Hershcovich et al., 2022)
- Extrinsic evaluation of cultural competence in large language models (Bhatt and Diaz, 2024)
- A Unified Framework to Quantify Cultural Intelligence of AI (Dev et al., 2026)

7 Tutorial Instructors

Wei Xu (https://cocoxu.github.io/) is an Associate Professor in the College of Computing and Machine Learning Center at the Georgia Institute of Technology. Her research interests include evaluation, robustness, multilingualism, cultural alignment, user simulation, and reasoning of LLMs. She is a recipient of the NSF CAREER Award, faculty research awards from Google, Sony, Crowdflower, and Criteo, paper awards and honorable mentions at COLING'18, ACL'23, and ACL'24, including the Best Social Impact Paper Award at ACL 2024, titled "*Having Beer after Prayer? Measuring Cultural Bias in Large Language Models*." She is confirmed to serve as a **Program Chair for EMNLP 2027** and has regularly served as a senior area chair for *ACL conferences. She has given a tutorial on "*Automatic and Human-AI Interactive Text Generation*" at ACL 2024 (https://acl2024-text-generation-tutorial.github.io/).

David Ifeoluwa Adelani (https://dadelani.github.io/) is an Assistant Professor at the McGill University School of Computer Science, a Core Academic Member at Mila - Quebec AI Institute, a Canada CIFAR AI Chair, and an IVADO Professor. His research interests include multilingual natural language processing and speech processing with a focus on low-resource languages. Some of his research on AfricaNLP have received Paper Awards including Best Paper Award (Global Challenges) at COLING 2022 for AfroXLMR, a multilingual pre-trained language model for African languages, Area Chair Award at IJCNLP-AACL 2023 for MasakhaNEWS topic classification dataset, Outstanding Paper Award at NAACL 2025 for "*IrokoBench—an LLM evaluation benchmark on African languages*", and the Best Theme Paper Award for "*WorldCusines–a multicultural multilingual Visual QA dataset*" at NAACL 2025. Previously, he gave a tutorial on "*A Journey Through the Opportunity of Low Resourced Natural Language Processing — An African Lens*" at NeurIPS 2021 with Dr. Vukosi Marivate (https://neurips.cc/virtual/2021/tutorial/21898).

Vinodkumar Prabhakaran (https://research.google/people/vinodkumarprabhakaran/) is a Senior Staff Research Scientist at Google Research, where he co-leads the interdisciplinary Technology, AI, Society and Culture (TASC) team. His research focuses on questions at the intersection of AI and society, with a focus on global and cross-cultural considerations. Before Google, he was a postdoc at Stanford University, and obtained his PhD from Columbia University. His prior research focused on building scalable ways using language technologies to identify and address large-scale societal issues such as racial disparities in policing, workplace incivility, and online abuse. He has published over 50 articles in top-tier venues such as the PNAS, ACL, TACL, NAACL, EMNLP, NeurIPS, and FAccT.

Sara Hooker (https://www.sarahooker.me/) served as Vice President of Research at Cohere from 2022 to 2025, where she led Cohere Labs, a fundamental AI research lab that seeks to solve complex machine learning problems. Since 2026, she has been the Co-Founder and CEO of Adaptation Labs. Her research interests include model efficiency and AI accessibility, with a focus on multilingual, multimodal, model compression, ethical AI practices and fair LLM evaluation. She is a cofounder of the Trustworthy ML Initiative, serves on Kaggle's ML Advisory Board and the World Economic Forum Council on the Future of Artificial Intelligence, and is a member of the MLC research group. Sara was also named one of TIME's 100 Most Influential People in AI in 2024, and named one of AI's "Top 13 Innovators" by Fortune in 2023. She received the Best Paper Award at ACL 2024, titled "*Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model*".

Alan Ritter (https://aritter.github.io/) is an Associate Professor in the College of Computing at the Georgia Institute of Technology. His research focuses on natural language processing and machine learning. He received his Ph.D. in Computer Science and Engineering from the University of Washington and subsequently completed a postdoctoral fellowship in the Machine Learning Department at Carnegie Mellon University. Alan served as **Program Chair for NAACL 2025** and is the recipient of an NSF CAREER Award, an Ama-

zon Research Award, a Sony Faculty Innovation Award, and a Google Academic Research Award, as well as Paper Awards at ACL 2023, 2024, and 2025. His research has been featured in media outlets including *WIRED*, *TNW*, *Bloomberg*, and *VentureBeat*.

Acknowledgement

We thank Geyang Guo for contributing background material that informed portions of Table 1 and the discussion of LLM post-training. These sections build upon and extend material she previously prepared for papers published at EMNLP 2025 (Guo et al., 2025) and ICML 2026 (Guo et al., 2026). Wei Xu is supported by NSF CAREER Award IIS-2144493 and a Google Academic Research Award; Alan Ritter is supported by NSF CAREER Award IIS-2052498. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the U.S. Government.

References

Diana Abagyan, Alejandro R. Salamanca, Andres Felipe Cruz-Salinas, Kris Cao, Hangyu Lin, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2025. One tokenizer to rule them all: Emergent language plasticity via multilingual tokenizers. arXiv.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 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, pages 9904–9923, Singapore. Association for Computational Linguistics.

Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2026. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens. In Findings of the Association for Computational Linguistics: EACL 2026, pages 1218–1235.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1573–1596, Dubrovnik, Croatia. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, Stella Biderman, and Ben Bergen. 2026. Explaining and mitigating crosslingual tokenizer inequities. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. Advances in Neural Information Processing Systems, 36:53330–53342.

Niyati Bafna, Emily Chang, Nathaniel Romney Robinson, David R. Mortensen, Kenton Murray, David Yarowsky, and Hale Sirin. 2025. DialUp! modeling the language continuum by adapting models to dialects and dialects to models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20188–20233, Vienna, Austria. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16055–16074.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pages 1–8.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In EMNLP.

Terra Blevins, Susanne Schmalwieser, and Benjamin Roth. 2025. Do language models accommodate their users? a study of linguistic convergence. Preprint, arXiv:2508.03276.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In EMNLP.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025a. M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv 2402.03216.

William Chen, Jinchuan Tian, Yifan Peng, Brian Yan, Chao-Han Huck Yang, and Shinji Watanabe. 2025b. OWLS: Scaling laws for multilingual speech recognition and translation models. In ICML.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalTeaming: AI-assisted interactive red-teaming for challenging LLMs'(lack of) multicultural knowledge. arXiv preprint arXiv:2404.06664.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25663–25701.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In ACL.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18511–18526.

Sunipa Dev, Vinodkumar Prabhakaran, Rutledge Chin Feman, Aida Davani, Remi Denton, Charu Kalia, Piyawat Lertvittayakumjorn, Madhurima Maji, Rida Qadri, Negar Rostamzadeh, Renee Shelby, Romina Stella, Hayk Stepanyan, Erin van Liemt, Aishwarya Verma, Oscar Wahltinez, Edem Wornyo, Andrew Zaldivar, and Saška Mojsilović. 2026. A unified framework to quantify cultural intelligence of ai. Preprint, arXiv:2603.01211.

Aleksei Dorkin, Taido Purason, Emil Kalbaliyev, Hele-Andra Kuulmets, Marii Ojastu, Mark Fišel, Tanel Alumäe, Eleri Aedmaa, Krister Kruusmaa, and Kairit Sirts. 2026. Estllm: Enhancing estonian capabilities in multilingual llms via continued pretraining and post-training. arXiv preprint arXiv:2603.02041.

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut, and Rico Sennrich. 2025. Parity-aware byte-pair encoding: Improving cross-lingual fairness in tokenization. Preprint, arXiv:2508.04796.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In First Conference on Language Modeling.

Gemma. 2025. Gemma 3 technical report. arXiv preprint 2503.19786.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 72–81, Online. Association for Computational Linguistics.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Assessing the impact of multilingual human preference learning on cultural awareness. In EMNLP.

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2026. Learning to route languages for multilingual policy optimization. In Proceedings of the Forty-third International Conference on Machine Learning (ICML).

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-rewardbench: Evaluating reward models in multilingual settings. arXiv preprint 2410.15522.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In Findings of the Association for Computational Linguistics: ACL 2025, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6997–7013.

Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. 2025. Values in the Wild: Discovering and analyzing values in real-world language model interactions. Preprint, arXiv:2504.15236.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024. 1+1>2: Can large language models serve as cross-lingual knowledge aggregators? In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13394–13412.

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. arXiv preprint arXiv:2507.05418.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. Language model alignment in multilingual trolley problems. In ICLR.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In NeurIPS (Datasets and Benchmarks Track).

Pierre-Carl Langlais, Pavel Chizhov, Catherine Arnett, Carlos Rosas Hinostroza, Mattia Nee, Eliot Krzysztof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. 2026. Common corpus: The largest collection of ethical data for LLM pre-training. In The Fourteenth International Conference on Learning Representations.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. In The Twelfth International Conference on Learning Representations.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In International Conference on Learning Representations, volume 2025, pages 79310–79333.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025a. Lost in literalism: How supervised training shapes translationese in LLMs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12875–12894, Vienna, Austria. Association for Computational Linguistics.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. 2025b. The impact of language mixing on bilingual LLM reasoning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32531–32548, Suzhou, China. Association for Computational Linguistics.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

Alisa Liu, Jonathan Hayase, Valentin Hofmann, Sewoong Oh, Noah A Smith, and Yejin Choi. 2025a. SuperBPE: Space travel for language models. In Second Conference on Language Modeling.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025b. Is translation all you need? a study on solving multilingual tasks with large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning. Preprint, arXiv:2509.06888.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in

large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In Advances in Neural Information Processing Systems, volume 37, pages 78104–78146.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? Measuring cultural bias in large language models. In ACL.

Tarek Naous, Anagha Savit, Geyang Guo, Jaehyeok Lee, Kyungdon Lee, Neel Kothari, Sarah Masud, Sahajpreet Singh, Tanish Patwa, Mengyu Ye, Alan Ritter, Yuki Arase, Tanmoy Chakraborty, JinYeoung Bak, Keisuke Sakaguchi, and Wei Xu. 2025. Camellia: Benchmarking cultural biases in LLMs for asian languages. In In preperation.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In Proceedings of the ACM Web Conference 2023, pages 1907–1917.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. JMMMU: A Japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, New Mexico. Association for Computational Linguistics.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9952–9962.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual label projection for cross-lingual structured prediction. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5738–5757.

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. 2025. Mind the gap... or not? how translation errors and evaluation details skew multilingual results. arXiv preprint 2511.05162.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In NAACL.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. arXiv preprint arXiv:2211.13069.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The case for" thick evaluations" of cultural representation in ai. ACM AIES.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Díaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, and 1 others. 2025. Whose view of safety? a deep dive dataset for pluralistic alignment of text-to-image models. Advances in Neural Information Processing Systems.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In The Thirteenth International Conference on Learning Representations.

Koshiro Saito, Sakae Mizuki, Masanari Ohi, Taishi Nakamura, Taihei Shiotani, Koki Maeda, Youmi Ma, Kakeru Hattori, Kazuki Fujii, Takumi Okamoto, Shigeki Ishida, Hiroya Takamura, Rio Yokota, and Naoaki Okazaki. 2024. Why we build local large language models: An observational analysis from 35 japanese and multilingual llms. Preprint, arXiv:2412.14471.

Teven Le Scao, Angela Fan, Christopher Akiki, ..., David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. ArXiv.

Robin Schimmelpfennig, Mark Díaz, Vinodkumar Prabhakaran, and Aida Davani. 2025. Humanlike ai design increases anthropomorphism but yields divergent outcomes on engagement and trust globally. arXiv preprint arXiv:2512.17898.

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. 2026. The roots of performance disparity in multilingual language models: Intrinsic modeling difficulty or design choices? arXiv preprint 2601.07220.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In ACL.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Chanuk lim, Srikant Panda, Minhyuk Kim, Nikunj drolia, Dasol Choi, Kyong-Ha Lee, and Youngjae Yu. 2026. Pushing on multilingual reasoning models with language-mixed chain-of-thought. In The Fourteenth International Conference on Learning Representations.

Yuchen Song, Andong Chen, Wenxin Zhu, Kehai Chen, Xuefeng Bai, Muyun Yang, and Tiejun Zhao. 2026. Culture in a frame: C^{3}b as a comic-based benchmark for multimodal culturally awareness. In The Fourteenth International Conference on Learning Representations.

Hayk Stepanyan, Aishwarya Verma, Andrew Zaldivar, Rutledge Chin Feman, Erin MacMurray van Liemt, Charu Kalia, Vinodkumar Prabhakaran, and Sunipa Dev. 2026. Scaling cultural resources for improving generative models. In Findings of the Association for Computational Linguistics: EACL 2026, pages 6695–6709.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Srija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A Mohamed, Debela Desalegn Yadeta, Negasi Haile Abadi, Abigail Oppong, Nnaemeka Casmir Obiefuna, and 1 others. 2026. Afri-MCQA: Multimodal Cultural Question Answering for African Languages. In ACL.

Ahmet Üstün, Viraat Aryabumi, ..., and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In ACL.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 19565–19575.

Tim Vieira, Ben LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2025. From language models over tokens to language models over characters. In ICML.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025a. Multilingual language model pretraining using machine-translated data. arXiv.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11897–11916.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Daniel Egert, Ellie Evans, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. 2025b. HelpSteer3: Human-annotated feedback and edit data to empower inference-time scaling in open-ended general-domain tasks. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25640–25662, Vienna, Austria. Association for Computational Linguistics.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025c. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. arXiv.

Zheng Weihua, Roy Ka-Wei Lee, Zhengyuan Liu, Wu Kui, AiTi Aw, and Bowei Zou. 2025. CCL-XCoT: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1768–1788, Suzhou, China. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using Abstract Meaning Representation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 753–765, St. Julian's, Malta. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In ACL.

Wanyun Xie, Francesco Tonin, and Volkan Cevher. 2025. Chameleon: A flexible data-mixing framework for language model pretraining and finetuning. In ICML.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025b. Language imbalance driven rewarding for multilingual self-improving. In ICLR.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hangyeol Yoo, ChangSu Choi, Minjun Kim, Seohyun Song, SeungWoo Song, Inho Won, Jongyoul Park, Cheoneum Park, and KyungTae Lim. 2026. ELO: Efficient layer-specific optimization for continual pretraining of multilingual LLMs. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 752–763, Rabat, Morocco. Association for Computational Linguistics.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025a. Can MLLMs understand the deep implication behind Chinese images? In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. 2026. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In The Fourteenth International Conference on Learning Representations.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025b. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. arXiv preprint arXiv:2510.07300.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. MPO: Multilingual safety alignment via reward gap optimization. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23564–23587, Vienna, Austria. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In NeurIPS.

Weihua Zheng, Zhengyuan Liu, Tanmoy Chakraborty, Weiwen Xu, Xiaoxue Gao, Bryan Chen Zhengyu Tan, Bowei Zou, Chang Liu, Yujia Hu, Xing Xie, Xiaoyuan Yi, Jing Yao, Chaojun Wang, Long Li, Rui Liu, Huiyao Liu, Koji Inoue, Ryuichi Sumida, Tatsuya Kawahara, and 16 others. 2025. MMA-ASIA:

A multilingual and multimodal alignment framework for culturally-grounded evaluation. Preprint, arXiv:2510.08608.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. WIKIBIAS: Detecting multi-span subjective biases in language. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1799–1814, Punta Cana, Dominican Republic. Association for Computational Linguistics.