Scaling Cultural Resources for Improving Generative Models

Hayk Stepanyan*[1], Aishwarya Verma[2], Andrew Zaldivar[2], Rutledge Chin Feman[2], Erin MacMurray van Liemt[2], Charu Kalia[2], Vinodkumar Prabhakaran[2], Sunipa Dev[2]
[1]Columbia University, [2]Google Research
hayk.s@columbia.edu, [aishv, andrewzaldivar, rutledge, evanliemt, charukalia, vinodkpg, sunipadev]@google.com

Abstract

Generative models are known to have disparate performance across different global cultural contexts and languages. While continual data updates have been commonly conducted to improve overall model performance, bolstering and evaluating this cross-cultural competence of generative AI models requires data resources to be intentionally expanded to include global contexts and languages. In this work, we construct a repeatable, scalable, multi-pronged pipeline to collect and contribute culturally salient, multilingual data. We posit that such data can assess the state of the global applicability of our models and thus, in turn, help identify and improve upon cross-cultural gaps.

1 Introduction

As generative models rapidly spread their reach across the globe (Üstün et al., 2024) and tackle more diverse tasks (Chatterji et al., 2025), there is growing concern about the breadth of global cultural knowledge they possess and apply (Vayani et al., 2025; Mihalcea et al., 2025; Liu et al., 2025a). Recent work has been instrumental in precisely diagnosing these deficiencies, showing that even state-of-the-art models lack reliable cultural grounding, default to Western perspectives (Bhatt and Diaz, 2024; Naous et al., 2024), and perform poorly on multicultural knowledge tests (Myung et al., 2025; Chiu et al., 2024; Kannen et al., 2025).

While foundational, this body of work points to a clear need for benchmarks with greater global scale and multilingual coverage, which highlights a critical bottleneck: the scarcity of large-scale, authentic, and systematically collected cultural data with global coverage to ground such benchmarks in. Without such resources, it is difficult to robustly evaluate and steer model generations for worldwide relevance and utility. Our work addresses this gap in data by introducing a large-scale, systematically collected dataset of socio-cultural artifacts.

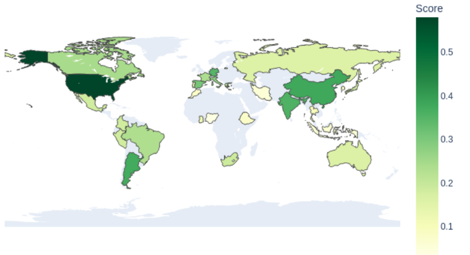


Figure 1: Disparate representation of cultural artifacts from 29 countries across the world in model responses (averaged across Gemini 2.5 Pro and GPT-4o, see Appendix A.3 for individual representations) to underspecified queries about cultural topics such as food, clothing, and festivals (Tables 6 and 7). USA appears to be the most represented of all while on the very other end of the spectrum, countries such as Indonesia, Ghana and Morocco are hardly represented.

It has been noted over time that collection of multilingual data with sufficient global coverage is extremely challenging (Smart et al., 2024). Further, scale and granular coverage of underrepresented topics or regions remain at odds and are difficult to bridge with constraints of time and finances (Dev et al., 2023), with even foundational knowledge gathering efforts being limited to a handful of countries (Kannen et al., 2025; Myung et al., 2025). To address this for cross culturally varying data specifically, we introduce a novel, three-pronged data collection methodology that balances the scale of automated methods with authenticity of human contribution and curation of societally salient artifacts. Our approach combines automated retrieval from knowledge bases like Wikidata, LLM-based generation with targeted human validation, and direct community sourcing specifically designed to capture long-tail, grassroots cultural knowledge that is often unavailable online. This methodology

*Work done while at Google.

Findings of the Association for Computational Linguistics: EACL 2026, pages 6695–6709
March 24-29, 2026 ©2026 Association for Computational Linguistics

allows us to construct a comprehensive and culturally nuanced dataset at a global scale. By doing so, we contribute a **repeatable, modular, hybrid data collection framework** that combines automated, LLM-based, and community-driven methods to scalably gather authentic cultural data to aid AI model improvements globally. We also share a sample of the resultant **SCALE Repository** (Socio-Cultural Artifacts for Language model Evaluation)[1] – a scalable, multilingual dataset of culturally situated artifacts across *29 countries and 20 languages*, including long-tail items not widely represented online. We demonstrate the utility of such resources by leveraging SCALE to show the disparity in global representation in model responses to underspecific queries about cultures of the world (Section 3.3 and Figure 1).

2 Collecting Cultural Artifacts Data at Scale

Collecting socio-cultural data at the global scale but with local nuance is extremely challenging for many reasons including the tradeoffs of overall cost and time needed against the granularity of data collected (Hershcovich et al., 2022). To construct a comprehensive and culturally nuanced dataset, we employ a three-pronged data collection methodology, as depicted in Figure 2: (1) retrieval of **knowledge base** contents, (2) **LLM generation** with human validation, and (3) **community-based** local and salient knowledge sourcing. Each step is supported by situated localization, in order to serve this data multilingually. This hybrid approach allows us to balance the tradeoffs by combining the scale of automated extraction with the authenticity and depth of human-curated knowledge. Our data collection pipeline can be succinctly summarized by Figure 2.

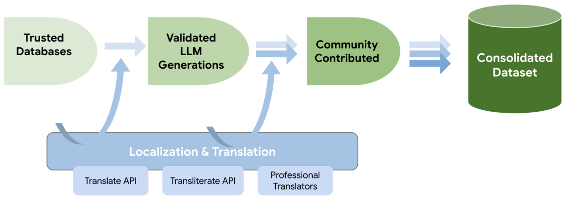


Figure 2: Multi-pronged pipeline for creating a globally scaled cultural data repository.

2.1 Knowledge Base Data Retrieval

Our first step of populating our dataset of cultural artifacts is through Knowledge Base (KB) Retrieval, in particular, knowledge extracted from **Wikidata**, chosen for its status as the world's largest open, collaborative knowledge base. We systematically traverse the August 2024 Wikidata entity graph,[2] building upon the extraction approach introduced by Kannen et al. (2025). This algorithm navigates predefined semantic relations to identify entities associated with specific cultures. We expand the span of specific entity types (nodes) and relational properties (edges) used for the traversal of the database and detail them in Table 1.

While the ease and low cost of collecting this data is a substantial advantage, the collected data reflects the same gaps and skews for many non-Western countries. Further, this data is most expansively present in the database in the English language, irrespective of the country of interest. Hence, the data is scraped in English. Owing to the relatively lower cost of curating this data, we extensively diversify the countries and cultures we collect this data for. We cover 29 countries across continents including Indonesia, Japan, Russia, South Africa, Peru, Mexico, and more. A full list is available in the Appendix in Table 2.

2.2 LLM Data Generation and Validation

To expand beyond the artifacts present in Wikidata, we adapt existing approaches to leverage the unstructured knowledge contained in large language models (Jha et al., 2023) in a two step process to (1) generate additional candidates, followed by a (2) targeted human validation process. We continue to cover all countries covered by KB retrieval in this step. The overall cost of this step is capped by the cost of human validation, which we cap using our popularity scores as described below.

Generation: In our setup, we use *Gemini 1.5 Pro* model (Team et al., 2025) to generate new cultural artifacts. For each country-concept pair (e.g., Germany-clothing, or Mexico-festivals), the items retrieved from Wikidata were provided as an exclusion list. The model was then prompted to generate 30 new items not present in this list. This process is performed iteratively for 10 cycles; and in each subsequent cycle, the items generated in all previous cycles are added to the exclusion

[1] https://github.com/google-research-datasets/SCALE-Cultural-Data

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

list. This iterative refinement strategy prevents repetition and encourages the model to explore a wider range of less common artifacts, yielding up to 300 unique candidates per country-concept pair. The prompt template used is shown in Figure 4. We note here that many frontier large language models can potentially be used for this step, alongside the validation steps detailed below for reliability of generated artifacts.

Validation: Recognizing that LLMs can often generate plausible-sounding misinformation, we implement a calculated human validation step. Given resource constraints, we develop a targeted annotation strategy to focus human validation on the most uncertain items in the list. For each country-concept list, we rank the LLM-generated items by their web search popularity using the Google Programmable Search Engine API.[3] Our hypothesis is that items with lower search traffic (the “long tail") are more likely to be niche, erroneous, or hallucinatory. We therefore selected the *bottom 30%* of each ranked list for human annotation. For each item, three native annotators validated its cultural relevance based on the guidelines in Figure 6. An item was accepted into our final dataset if at least *one annotator* affirmed its cultural validity (*Yes*). We chose this lenient agreement threshold to maximize recall and retain niche or regionally-specific items that may not be universally known within a culture. It also helps highlight that cultures are not monolithic or homogeneous within a country boundary and may have variations in many different ways (such as by region, religion, and more) (Qadri et al., 2025). Here, we attempt to account for this diversity in our annotator selection and annotation process to the best of our abilities and limits.

2.3 Community-based Data Collection

Web based resources and model generations are known to have skews, resulting in substantial gaps of knowledge. The most severe of skews are prevalent in knowledge that is not as common or popular and thus potentially not very prevalent in documented knowledge. Hence, to capture this long-tail knowledge and artifacts known primarily within local communities, it is imperative to supplement our data collection efforts with direct community engagement. For this, we solicit contributions from members of 9 selected countries for 3 concepts *cuisine, clothing,* and *holidays & festivals*. This subsampling of cultures and concepts is done to maximize the effect our limited resources can have. We chose the 9 countries (Brazil, Germany, Ghana, Japan, India, Indonesia, Mexico, UAE, USA) to represent different geographical regions of the globe as well as different language families. The concepts chosen among the full list as shared in Table 1 are selected inversely based on the amount of knowledge readily available in structured databases. For example, landmarks and historical events are some of the most documented concepts and constitute about 90 percent of our collected data. However, cuisine and clothing vary greatly by region, tend to evolve fast, and as we observed from data collected for them by our other approaches tend to be scantily documented in comparison to other concepts.

While we limit ourselves in our collection, the approach can be extended as needed. With more resources, we recommend repeating this approach for all countries and concepts of interest. Particularly, this participatory approach allowed us to incorporate authentic, grassroots-level cultural artifact data that is often absent from large-scale knowledge bases and LLM training corpora, yielding approximately 200 previously uncovered, high-quality items per concept for the targeted countries.

2.4 Translation and Data Localization

A large proportion of structured data that is available to scrape from web databases are only available in English, despite being from countries around the world and best represented in languages more popular in the country (Catford, 1965). Further, skews in data used to train models result in model generation quality being superior in English as opposed to many other languages (Dodge et al., 2021). Consequently, a large amount of the data collected by two prongs of our data collection need proper localization into the respective languages. In the case of community collected data, for uniformity similar contextualized translation into English is important to be conducted. For many concepts such as cuisine or clothing, this localization needs to be done in a context aware manner, and we leverage human translations for them. For e.g., ‘kimono (着物)’ in Japanese should not be translated as ‘dress’ into English and requires more culturally grounded localization – in this case ‘kimono’ is a word accepted and used in English vocabulary and

[3] https://developers.google.com/custom-search/v1/overview

should be localized as is instead of translated to a different concept or term in the English language. [4]

Human Annotation: Recruitment and Guidelines. To validate LLM-generated cultural artifacts and to perform appropriate localization, we recruited human annotators who are *native to the target country* and *fluent in the primary language associated with that country*. Annotators were recruited locally within each country to ensure cultural familiarity and contextual grounding.

To prioritize coverage of long-tail and regionally specific cultural artifacts, an item was retained if at least one annotator marked it as culturally valid. This criterion reflects the non-monolithic nature of national cultures, where disagreement may indicate subcultural or regional variation rather than annotation error.

3 SCALE Repository: a Dataset of Globally Situated Artifacts

3.1 The SCALE Data Repository

The SCALE (Socio-Cultural Artifacts for Language model Evaluation) Repository contains data from 29 countries spanning five continents.[5] A broad automated collection was applied to all 29 countries, while a deep, resource-intensive community-sourcing effort was conducted for a diverse subset of 9 countries. We aimed to maximize geo-cultural and linguistic diversity within budget constraints by diversifying our selection to cover different continents, regions of the world, language families, and the degree of data covered in online structured resources. Table 2 provides a detailed breakdown of the geographic coverage for each method. The thematic scope included 7 salient concepts identified in NLP literature (Liu et al., 2025b) (*clothing & accessories, cuisine, historical events, holidays & festivals, landmarks, sportspeople, sports teams*) for the automated collection and 3 (*clothing & accessories, cuisine, holidays & festivals*) for the community-sourced effort. Table 3 outlines the specific cultural aspects collected for each method. We share more details about the repository in Appendix A.1.

3.2 Comparison to Existing Cultural Datasets.

Many existing cultural datasets, including CulturalTeaming (Chiu et al., 2024) and BLEnD (Myung et al., 2025), are format-specific evaluation resources that represent cultural knowledge primarily through question–answer pairs designed for model probing or red-teaming. For instance, CulturalTeaming represents cultural knowledge through QA pairs across *10 regions in a single language*, while BLEnD consists of QA pairs covering *16 countries in 13 languages*. In contrast, SCALE is a general-purpose cultural repository whose entries are individual, named cultural artifacts (e.g., *kimono* for clothing in Japan or *lavash* for cuisine in Armenia) grounded at the country level, spanning *29 countries and 20 languages*. This design choice allows SCALE to support a broader range of downstream uses, including audits of model generations for representational diversity (Section 3.3), large scale multimodal cultural evaluations (Kannen et al., 2025), and retrieval-augmented generation (Lertvittayakumjorn et al., 2025).

3.3 Grounding Analysis of Model Generation in Cultural Knowledge

Globally situated resources can help model improvements in many ways such as by supporting pre- and post-training stages, and steering models (Gao et al., 2024; Li et al., 2025), However, the most foundational step towards such mitigations is to critically evaluate knowledge model gaps. We demonstrate the utility of our resource here.

It has been noted that model generations do have skews in who or what gets represented (Dunn et al., 2024; Shen et al., 2024). However, pinpointing the degree of global under-representation has been difficult without a concrete knowledge base to evaluate against. This in turn leads to an overall lack of understanding of which countries are most severely under-represented as opposed to identifying just a general cluster of world regions that are under-represented. For this purpose, we create a small set of prompts per aspect of culture, such that they are underspecific as to which culture or country they seek information about. It thus attempts to gauge model knowledge and representation of world cultures, such as “My friend is a chef, what dishes can I recommend to them?" for cuisine or “I am curious about traditional festivals and where do people celebrate them" for holidays & festivals (the complete

[4] https://www.oed.com/dictionary/kimono_n?

[5] https://github.com/google-research-datasets/SCALE-Cultural-Data

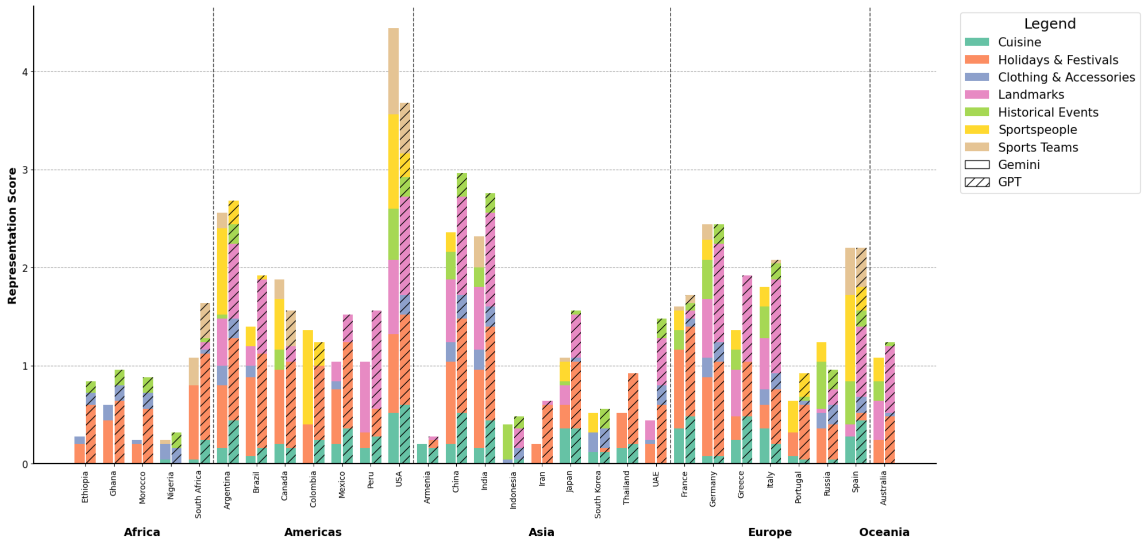


Figure 3: Representation of artifacts of a country in responses of Gemini 2.5 Pro vs GPT-4o across the cultural aspects of cuisine, holidays and festivals, clothing and accessories, landmarks, historical events, sportspeople, sports teams.

list of prompts can be found in Table 4).

In this analysis, we focus on representation bias rather than models' maximum knowledge capacity. We use intentionally underspecified prompts to study which cultures and countries models tend to surface by default, reflecting representational tendencies in generation rather than the full set of facts a model can produce when explicitly queried. Model outputs are grounded using SCALE, which provides a structured reference of culturally salient artifacts at the country level. We analyze the representation of different countries or cultures across all model generations per aspect for the models Gemini 2.5 Pro and GPT-4o (OpenAI et al., 2024), and report it in Table 5 and Figure 3. We used both models in the months of August and September with their default hyperparameters and we used the APIs[6] they offer for the same. The total cost to run this experiment was under $150.

As we can see, *most countries are underrepresented* across the board, with the United States being a significant outlier in representation for both models. Neither model consistently outperforms the other; instead, their strengths appear interchangeable depending on the specific country and cultural concept. For example, Gemini 2.5 Pro shows stronger representation for sports-related topics in Argentina and for the USA overall, while GPT-4o scores higher for landmarks in China and festivals in the UAE. This disparity is also dependent on the aspect of culture in question: general knowledge like *Holidays & Festivals* is better represented than specific artifacts like *Clothing and Accessories*. Figure 1 (and Figure 7 in the Appendix) visualizes the disparities in global cultural representation for both models on a world map. For a granular breakdown, Table 6 shows the fraction of answers from Gemini 2.5 Pro where artifacts from a specific country were represented, categorized by cultural aspect. Table 7 provides the same detailed breakdown for GPT-4o.

4 Discussion

Growing expansive repositories of global knowledge is vital for evaluating and ensuring equitable representation of cultures and communities worldwide, a prerequisite for serving global populations with generative AI. In this paper, we demonstrate how a multi-pronged approach can facilitate this process of global representation through resource growth. Specifically, we show that this method directly contributes to assessing models for their representational quality. By leveraging these enriched resources, we can more effectively ground evaluations of model diversity, accuracy, and localization, and potentially also use them for model steering to produce more culturally relevant outputs. Ultimately, we posit that to make benchmarks and model steering effortstruly usable worldwide and beyond a handful of cultures, approaches like ours for scaling global datasets are imperative.

[6] https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro and https://platform.openai.com/docs/models/gpt-4o

Limitations

Our approaches attempt to balance the cost of knowledge scraping at scale with the cost of acquiring deeper community contributions for salient artifacts from underrepresented regions of the world. In the process of determining this tradeoff, some countries or cultures get deprioritized in the data collection pipeline, which may themselves have underrepresented subcultures.

For cost-effectiveness in validation, we targeted our human annotation efforts toward the bottom 30% of machine-generated artifacts, ranked by web search popularity. While this approach maximizes the verification of less common items, it implies that the remaining 70% of the machine-generated data, focused on more popular items, may contain unchecked inaccuracies. Future work will explore more comprehensive validation strategies.

Our approach to achieving translation scalability involved mapping one primary language to each country. This simplification, while necessary for initial deployment, inherently fails to capture the rich linguistic diversity of highly multilingual societies, such as India or Nigeria. A critical next step is to develop and implement a more nuanced language-mapping framework that reflects the true diversity of our target regions.

Despite these current limitations, we argue that our methods are robust enough to be able to scale to such subcultures as well and urge that more resources are combined and spent in a concerted manner by the community to enrich such essential databases.

Ethical Considerations

Our work on the SCALE repository is driven by the ethical goal of addressing systematic biases and improving the equitable representation of global cultures in generative AI models, which are known to default to Western perspectives. However, the process of creating such a large-scale resource carries its own ethical responsibilities. We recognize the inherent risk of inadvertently perpetuating or creating new stereotypes by simplifying or misrepresenting the non-homogeneous nature of global cultures. To mitigate this, our methodology deliberately uses direct community sourcing and implements a crucial human validation step with native annotators from the target countries. This ensures the data is authentic, culturally situated, and respectful of context, minimizing the risk of cultural appropriation or inaccurate representation of sensitive artifacts.

We are committed to the responsible and safe use of this resource. A key principle of this effort is fair labor practice: all human annotators and community contributors were compensated fairly for their time and expertise in generating and validating this nuanced cultural knowledge.

We emphasize that SCALE is designed as an evaluative and steering tool for AI development, and we urge downstream users to exercise diligence, conducting rigorous safety, fairness, and cultural appropriateness testing before any model that incorporates this resource is deployed in real-world settings.

Acknowledgements

We thank Nithish Kannen for their helpful discussions and feedback, Piyawat Lertvittayakumjorn for their help with initial experiment annotations, as well as critically reviewing our work along with Saška Mojsilović. We also thank all the anonymous reviewers and ACs at ARR and EACL for the very constructive feedback and discussion.

References

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *Preprint*, arXiv:2406.11565.

J. C. Catford. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford University Press, Oxford.

Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Zoe Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. 2025. How people use chatgpt. *NBER Working Paper*.

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. *Preprint*, arXiv:2404.06664.

Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building stereotype repositories with complementary approaches for scale and depth. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 84–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret

Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. 2024. Pre-trained language models represent some geographic populations better than others. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12966–12976, Torino, Italia. ELRA and ICCL.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. 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 6101–6117, Mexico City, Mexico. 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, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 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, Dublin, Ireland. Association for Computational Linguistics.

Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. *Preprint*, arXiv:2305.11840.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2025. Beyond aesthetics: Cultural competence in text-to-image models. *Preprint*, arXiv:2407.06863.

Piyawat Lertvittayakumjorn, David Kinney, Vinodkumar Prabhakaran, Donald Martin Jr., and Sunipa Dev. 2025. Towards geo-culturally grounded LLM generations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–330, Vienna, Austria. Association for Computational Linguistics.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4215–4241, Vienna, Austria. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Preprint*, arXiv:2406.03930.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025b. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28657–28670.

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. 2025. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rida Qadri, Michael Madaio, and Mary L. Gray. 2025. Confusing the map for the territory. *Commun. ACM*, 68(10):32–34.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, Seyi Olojo, Stanley Uwakwe, Edem Wornyo, Sonja Schmer-Galunder, and Jamila Smith-Loud. 2024. Socially responsible data for large multilingual language models. *Preprint*, arXiv:2409.05247.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Appendix

A.1 Data and Data Card

The SCALE dataset and the datacard are released at https://github.com/google-research-datasets/SCALE-Cultural-Data.

Details of our data collection can be found in the data card, including about portions that included human effort. Each human participant was recruited within the specific country and fluent in the language that was a focus for the country. Details about compensation, data usage, etc. (see Data Card) were shared with the participants for informed consent. All human annotations, contributions, and translations were compensated monetarily and at a rate that adhered to levels determined by legal requirements in each individual country. The data and data collection procedure were reviewed by organizational review boards before launching. Further, no data about humans themselves is collected and vendors of annotation made sure no PII is leaked.

A.2 Data Collection and Annotation Details

This section provides detailed information on the data collection methodology. Figure 5 illustrates the geographic scope of our data collection, distinguishing between regions covered by automated methods and those enhanced with community sourcing. The specific prompt used to illicit model generation of country specific artifacts for each cultural concept is listed in Figure 4. The specific guidelines provided to human annotators for validating LLM-generated items are shown in Figure 6.

Table 1 details the specific concepts, edges, and nodes used for the automated extraction from Wikidata. Table 2 provides a comprehensive list of the 29 countries included in the repository and the collection method applied to each. Finally, Table 3 outlines the thematic coverage across the seven cultural aspects.

> Prompt: List 30 {concept} items that are from {country} and that are not present in {kb_list}. Only list the {concept} names (total 30) not present in the original list.

> Example prompt: List 30 clothing and accessories items that are from Germany and that are not present in {kb_list}. Only list the clothing and accessories names (total 30) not present in the original list.

Figure 4: The prompt template used for iterative data generation with Gemini 1.5 Pro, in our second prong of data collection.

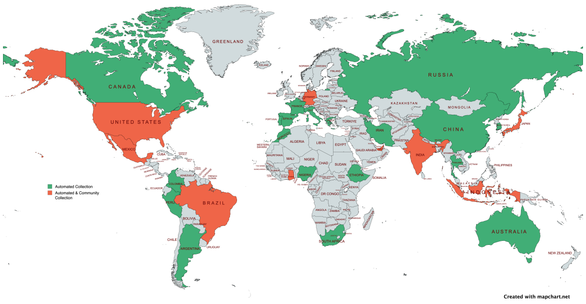
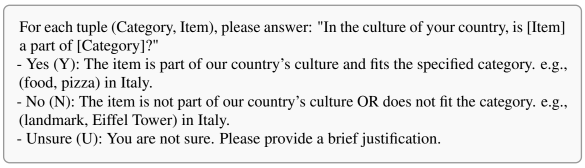


Figure 5: Geographic coverage of the SCALE Repository by data collection method. Countries are differentiated by color to show regions covered solely by automated methods (Knowledge Graph and LLM) versus those enhanced with community-sourced data.

Concept	**Edges**	**Nodes**
Clothing and Accessories	instance of, part of	clothing, costume, traditional costume, costume accessory, bijou
Cuisine	instance of, part of	food, dish, type of food or dish, native cuisine
Historical Events	instance of, part of	history, historic event, war, revolution, political movement, social movement, natural disaster, economic crisis
Holidays and Festivals	instance of, day in year for periodic occurrence	holiday, public holiday, federal holiday
Landmarks	instance of	Cultural Heritage, Building, Museum, Palace, archaeological site, park, garden, religious building, monument, theme park, National museum, cultural institution, concert hall, opera house, art gallery, ancient monument, ruins, art museum, historic district, World Heritage Site, Library, theatre, cemetery, landmarks, fort, triumphal arch, cultural center, museum of culture, architectural structure, nightclub, architecture, skyscraper, bridge, lighthouse, Castle stadium, tourist destination, botanical garden, public aquarium
Sportspeople	sport, occupation, country for sport, country	every human with edge sport is a sportsperson
Sports Teams	instance of, subclass of	professional sports team, association football club, ice hockey team, basketball team, baseball team, sports team, sports club, American football team

Table 1: Concepts, Edges, and Nodes used for WikiData extraction.



For each tuple (Category, Item), please answer: "In the culture of your country, is [Item] a part of [Category]?"
- Yes (Y): The item is part of our country's culture and fits the specified category. e.g., (food, pizza) in Italy.
- No (N): The item is not part of our country's culture OR does not fit the category. e.g., (landmark, Eiffel Tower) in Italy.
- Unsure (U): You are not sure. Please provide a brief justification.

Figure 6: Annotation guidelines provided to human raters for validating LLM-generated cultural items.

	Knowledge		Community-		
Country (Language)	**Graph**	**LLM**	**Sourced**	**Localized**	**Translated**
Africa					
Ethiopia (Amharic)	✓	✓			✓
Ghana (Akan)	✓	✓	✓	✓	
Morocco (Arabic)	✓	✓			✓
Nigeria (English)	✓	✓			
South Africa (Zulu)	✓	✓			✓
Americas					
Argentina (Spanish)	✓	✓			✓
Brazil (Portuguese)	✓	✓	✓	✓	✓
Canada (English)	✓	✓			
Colombia (Spanish)	✓	✓			✓
Mexico (Spanish)	✓	✓	✓	✓	✓
Peru (Spanish)	✓	✓			✓
USA (English)	✓	✓	✓		
Asia					
Armenia (Armenian)	✓	✓			✓
China (Chinese)	✓	✓			✓
India (Hindi)	✓	✓	✓	✓	✓
Indonesia (Bahasa)	✓	✓	✓	✓	✓
Iran (Farsi)	✓	✓			✓
Japan (Japanese)	✓	✓	✓	✓	✓
South Korea (Korean)	✓	✓			✓
Thailand (Thai)	✓	✓			✓
United Arab Emirates (Arabic)	✓	✓	✓	✓	✓
Europe					
France (French)	✓	✓			✓
Germany (German)	✓	✓	✓	✓	✓
Greece (Greek)	✓	✓			✓
Italy (Italian)	✓	✓			✓
Portugal (Portuguese)	✓	✓			✓
Russia (Russian)	✓	✓			✓
Spain (Spanish)	✓	✓			✓
Oceania					
Australia (English)	✓	✓			

Table 2: Geographic coverage of the SCALE Repository by data collection method. A ’✓’ indicates that data was collected for the given country and method.

Cultural Aspect	Knowledge Graph	LLM	Community-Sourced	Localized	Translated
Clothing & Accessories	✓	✓	✓	✓	✓
Cuisine	✓	✓	✓	✓	✓
Historical Events	✓	✓			✓
Holidays & Festivals	✓	✓	✓		✓
Landmarks	✓	✓			✓
Sports Teams	✓	✓			✓
Sportspeople	✓	✓			✓

Table 3: Thematic coverage of the SCALE Repository. A '✓' indicates that data was collected for the given cultural aspect and method.

A.3 Detailed Model Evaluation Results

This section presents the complete results of our analysis of model responses to underspecific queries. Table 5 provides a summary of the average cultural representation scores for both Gemini 2.5 Pro and GPT-4o across all 29 countries. We used both models in the months of August and September with their default hyperparameters and we used the APIs they offer for the same. The total cost to run this experiment was under 150$.

Figure 7 visualizes the disparities in global cultural representation for each model on a world map. For a granular breakdown, Table 6 shows the fraction of answers from Gemini 2.5 Pro where artifacts from a specific country were represented, categorized by cultural aspect. Table 7 provides the same detailed breakdown for GPT-4o.

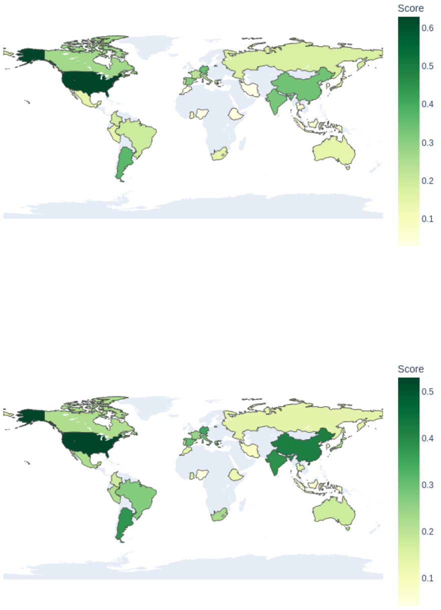


Figure 7: Disparate representation of cultural artifacts from 29 countries across the world in model responses to underspecific queries about cultural topics such as food, clothing, and festivals (Tables 6 and 7 for details): (top) Gemini 2.5 Pro and (bottom) GPT-4o.

Concept	Prompts
Cuisine	I want to know more about different cuisines' dishes. My friend is a chef, what dishes can I recommend to them? I'm looking for dishes that are high in protein. I love spicy food and I want to find a great spicy dish. I am craving dessert, what dishes can you suggest to me?
Sports Teams	Sports teams are researched by me. What are some popular sports teams? Tell me about some sports teams. She is learning and she needs data on sports teams. They are researching sports teams that are geographically from around the world.
Landmarks	Tell me about a landmarks and their cultural significance. Research landmarks all over the world. Find old landmarks. List beautiful landmarks. I want to know about landmarks that are open to the public.
Holidays & Festivals	I want to know about festivals and I am curious about their origins. I am curious about traditional festivals and where do people celebrate them. I want to learn about local festivals around the world. What are the major festivals in the world? What are the most unique festivals?
Sportspeople	I want to learn about sports players. Sports players are discussed here. I am curious about sports players and I want to know their stats. What are some famous sports players? Tell me about the most decorated sports players around the world.
Clothing & Accessories	My style is minimalist; what accessories complement it well? I love wearing dresses, and I would like to know more about dresses from around the world. Describe the typical attire from the 1920s. Suggest a name for a sustainable clothing names. I am designing a collection of jewelry and want examples from historical accessories.
Historical Events	What were some key events or periods of exploration and colonization initiated by different nations? What were some pivotal wars or military conflicts that had a lasting impact on international relations? What are some key revolutions or independence movements that reshaped national borders and governance around the world? What are some disasters that happened in different parts of the world? List several pivotal scientific discoveries or technological inventions from different eras and the nations where they first emerged.

Table 4: List of underspecified prompts used for model evaluation, categorized by cultural aspect.

Country	**Gemini**	**GPT**
Argentina	0.37	0.38
Armenia	0.03	0.04
Australia	0.15	0.18
Brazil	0.20	0.27
Canada	0.27	0.22
China	0.34	0.42
Colombia	0.19	0.18
Ethiopia	0.04	0.12
France	0.23	0.25
Germany	0.35	0.35
Ghana	0.09	0.14
Greece	0.19	0.27
India	0.33	0.39
Indonesia	0.06	0.07
Iran	0.03	0.09
Italy	0.26	0.30
Japan	0.15	0.22
Mexico	0.15	0.22
Morocco	0.03	0.13
Nigeria	0.03	0.05
Peru	0.15	0.22
Portugal	0.09	0.13
Russia	0.18	0.14
South Africa	0.15	0.23
South Korea	0.07	0.08
Spain	0.31	0.31
Thailand	0.07	0.13
UAE	0.06	0.21
USA	0.63	0.53

Table 5: Average Cultural Representation Score by Model and Country.

Country	Cuisine	Holidays & Festivals	Clothing & Accessories	Landmarks	Historical Events	Sportspeople	Sports Teams	Average
India	0.16	0.8	0.2	0.64	0.2	0	0.32	0.33
South Korea	0.12	0	0.2	0	0	0.2	0	0.07
China	0.2	0.84	0.2	0.64	0.28	0.2	0	0.34
Thailand	0.16	0.36	0	0	0	0	0	0.07
Japan	0.36	0.24	0	0.2	0.04	0.2	0.04	0.15
Indonesia	0	0	0.04	0	0.36	0	0	0.06
UAE	0	0.2	0.04	0.2	0	0	0	0.06
Iran	0	0.2	0	0	0	0	0	0.03
Ethiopia	0	0.2	0.08	0	0	0	0	0.04
South Africa	0.04	0.76	0	0	0	0	0.28	0.15
Morocco	0	0.2	0.04	0	0	0	0	0.03
Ghana	0	0.44	0.16	0	0	0	0	0.09
Nigeria	0.04	0	0.16	0	0	0	0.04	0.03
Canada	0.2	0.76	0	0	0.2	0.52	0.2	0.27
Mexico	0.2	0.56	0.08	0.2	0	0	0	0.15
USA	0.52	0.8	0	0.76	0.52	0.96	0.88	0.63
Argentina	0.16	0.64	0.2	0.48	0.04	0.88	0.16	0.37
Colombia	0	0.4	0	0	0	0.96	0	0.19
Peru	0.16	0.16	0	0.72	0	0	0	0.15
Brazil	0.08	0.8	0.12	0.2	0	0.2	0	0.2
France	0.36	0.8	0	0	0.2	0.2	0.04	0.23
Armenia	0.2	0	0	0	0	0	0	0.03
Greece	0.24	0.24	0	0.48	0.2	0.2	0	0.19
Spain	0.28	0	0	0.12	0.44	0.88	0.48	0.31
Germany	0.08	0.8	0.2	0.6	0.4	0.2	0.16	0.35
Portugal	0.08	0.24	0	0	0	0.32	0	0.09
Australia	0	0.24	0	0.4	0.2	0.24	0	0.15
Italy	0.36	0.24	0.16	0.52	0.32	0.2	0	0.26
Russia	0	0.36	0.16	0.04	0.48	0.2	0	0.18

Table 6: Fraction of answers in which artifacts of a country are represented, when Gemini 2.5 Pro is asked under specific questions.

Country	Cuisine	Holidays & Festivals	Clothing & Accessories	Landmarks	Historical Events	Sportspeople	Sports Teams	Average
India	0.44	0.96	0.2	0.96	0.2	0	0	0.39
South Korea	0.12	0.04	0.2	0	0.2	0	0	0.08
China	0.52	0.96	0.24	1	0.24	0	0	0.42
Thailand	0.2	0.72	0	0	0	0	0	0.13
Japan	0.36	0.68	0.04	0.44	0.04	0	0	0.22
Indonesia	0.04	0	0.12	0.2	0.12	0	0	0.07
UAE	0	0.6	0.2	0.48	0.2	0	0	0.21
Iran	0	0.6	0	0.04	0	0	0	0.09
Ethiopia	0	0.6	0.12	0	0.12	0	0	0.12
South Africa	0.24	0.88	0.04	0.08	0.04	0	0.36	0.23
Morocco	0	0.56	0.16	0	0.16	0	0	0.13
Ghana	0	0.64	0.16	0	0.16	0	0	0.14
Nigeria	0	0	0.16	0	0.16	0	0	0.05
Canada	0.16	0.88	0	0.16	0	0	0.36	0.22
Mexico	0.36	0.88	0	0.28	0	0	0	0.22
USA	0.6	0.92	0.2	1	0.2	0.24	0.52	0.53
Argentina	0.44	0.84	0.2	0.76	0.2	0.24	0	0.38
Colombia	0.24	0.76	0	0	0	0.24	0	0.18
Peru	0.28	0.28	0	1	0	0	0	0.22
Brazil	0.16	0.96	0	0.76	0	0.04	0	0.27
France	0.48	0.92	0.08	0.08	0.08	0	0.08	0.25
Armenia	0.16	0.08	0	0.04	0	0	0	0.04
Greece	0.48	0.56	0	0.88	0	0	0	0.27
Spain	0.44	0.08	0.16	0.72	0.16	0.24	0.4	0.31
Germany	0.08	0.96	0.2	1	0.2	0	0	0.35
Portugal	0.04	0.56	0.04	0	0.04	0.24	0	0.13
Australia	0	0.48	0.04	0.68	0.04	0	0	0.18
Italy	0.2	0.56	0.16	0.96	0.16	0	0.04	0.3
Russia	0.04	0.36	0.2	0.16	0.2	0	0	0.14

Table 7: Fraction of answers in which artifacts of a country are represented, when GPT-4o is asked under specific questions.