Leveraging Wikidata for Geographically Informed Sociocultural Bias Dataset Creation: Application to Latin America

Yannis Karmim*[1,2,4], Renato Pino*[2], Hernan Contreras*[3], Hernan Lira[4], Sebastian Cifuentes[5], Simon Escoffier[6], Luis Martí[4], Djamé Seddah[1], Valentin Barriere[2,5]

[1]ALMAnaCH team, Inria Paris Center; [2]Dept. of Computer Science, Universidad de Chile;
[3]Institute of International Studies, Universidad de Chile;
[4]Inria Chile Research Center; [5]Centro Nacional de Inteligencia Artificial;
[6]School of Social Work, Pontificia Universidad Católica de Chile.

* shared first authorship; **Correspondence:** vbarriere@dcc.uchile.cl

Abstract

Large Language Models (LLMs) exhibit inequalities with respect to various cultural contexts. Most prominent open-weights models are trained on Global North data and show prejudicial behavior towards other cultures. Moreover, there is a notable lack of resources to detect biases in non-English languages, especially from Latin America (Latam), a continent containing various cultures, even though they share a common cultural ground. We propose to leverage the content of Wikipedia, the structure of the Wikidata knowledge graph, and expert knowledge from social science in order to create a dataset of question/answer (Q/As) pairs, based on the different popular and social cultures of various Latin American countries. We create the LatamQA database of over 26k questions and associated answers extracted from 26k Wikipedia articles, and transformed into multiple-choice questions (MCQ) in Spanish and Portuguese, in turn translated to English. We use this MCQ to quantify the degree of knowledge of various LLMs and find out (i) a discrepancy in performances between the Latam countries, ones being easier than others for the majority of the models, (ii) that the models perform better in their original language, and (iii) that Iberian Spanish culture is better known than Latam one.[1]

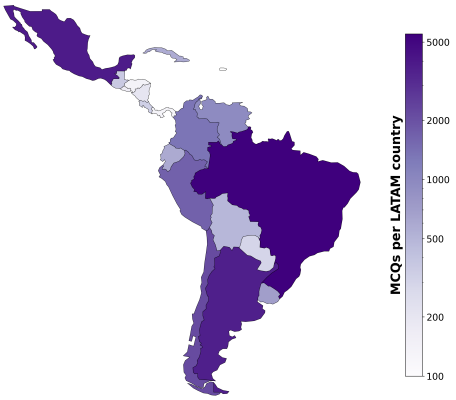

Figure 1: Geographic distribution of LatamQA cultural MCQs across Latin America composed of 23k Q/As.

1 Introduction and Related Work

Disciplinary standards for "valid" knowledge have been concentrated in Western Europe and North America (Demeter, 2020). Biases in AI systems, particularly in NLP, often originate from training data (Wiegand et al., 2019), annotation practices (Sap et al., 2022), and annotation guidelines (Parmar et al., 2023). These biases may take moral (Hämmerl et al., 2022), social (Sap et al., 2020), class-based (Curry et al., 2024), or political forms (Feng et al., 2023). Although social biases can be explicitly annotated for detection and analysis (Sahoo et al., 2023), annotation is costly and highly dependent on linguistic and cultural context (Fort et al., 2024; Barriere and Cifuentes, 2024b,a), and model's biases can be sensitive to simple context changes (Quiroga et al., 2025). Recent studies emphasize the importance of localizing dataset construction and warn against outsourcing bias-related annotation for non-English languages to actors in the Global North (Hada et al., 2024). Such practices risk overlooking culturally specific meanings and social dynamics, as many existing datasets inadequately represent the cultural contexts of the Global South. (Santy et al., 2023).

Cultural biases in LLMs are particularly underexplored for Latin America (Latam). Although many countries share Spanish or Portuguese as official languages, they differ substantially in historical, social, and cultural terms, making the region well suited for evaluating fine-grained cultural knowledge. However, current geo-cultural datasets either group countries coarsely (Czarnowska et al., 2021; Li et al., 2024), cover few Latam countries (Myung et al., 2024), merge heterogeneous regions (Adilazuarda et al., 2024), or are limited to En-

[1]Code and datasets available at https://github.com/Inria-Chile/LatamQA.

Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 177–188

Datasets	# (k)	Region	SP/PT
BLEnD (Myung et al., 2024)	2	Mixed	✓
CulFIT (Feng et al., 2025)	.08	Latam	✓
CANDLE (Nguyen et al., 2023)	2	Latam	✓
CultureBank (Shi et al., 2024)	3.3	Latam	✗
CultureAtlas (Fung et al., 2024)	1.8	Mixed	✗
GLOBAL-MMLU (Singh et al., 2024)	.01	Latam	✓
LatamQA (ours)	26	Latam	✓

Table 1: Latam-related questions in cultural Q/As datasets in terms of number of entries, region coverage, and if in Spanish and/or Portuguese.

glish (Feng et al., 2025). In Global-MMLU, only 1.6% of culturally relevant content concerns Latam (Singh et al., 2024). Table 1 summarizes existing cultural Q/As datasets with Latam content.

Cultural Benchmark Creation Building robust datasets that capture region-specific cultural knowledge is essential for evaluating biases in LLMs (Hershcovich et al., 2022; Liu et al., 2024; Pawar et al., 2024). Manual annotation efforts such as BLEnD (Myung et al., 2024) and CulturalBench (Chiu et al., 2025) offer high-quality data but are inherently limited in scale. In contrast, automated approaches construct cultural benchmarks by extracting data from large corpora. For example, Nguyen et al. (2023) rely on the C4 corpus, which has been shown to contain substantial noise (Fung et al., 2024). Other methods draw on social media platforms, such as CultureBank (Shi et al., 2024), which extracts from TikTok and Reddit or use curated web corpora, such as CRAFT (Wang et al., 2024), which retrieves documents from SlimPajama using keywords and LLM-generated Q/As. Wikipedia offers a middle ground with clean, curated content. CultureAtlas (Fung et al., 2024) extracts from Wikipedia using hyperlink expansion with NLI filtering; Li et al. (2024) use Wikipedia categories; Zhao et al. (2025) leverage Wikidata paths. Unlike these approaches, we exploit Wikipedia's pre-existing category ontology combined with sociologist-guided validation and LLM-based Q/A generation.

Language and Geographic Analysis The relationship between prompting language and cultural knowledge retrieval remains open. XNationQA (Tanwar et al., 2025) finds prompting language significantly impacts performance, while Ying et al. (2025) show models perform better in native languages—though findings conflict across studies (Zhao et al., 2025). Following prior work on common ground (Adilazuarda et al., 2024; Hershcovich et al., 2022), we focus on local facts at a fine-grained geographic level, particularly popular culture and sociocultural references that support shared understanding (Adams et al., 2004). Given the close relationship between language and culture (Hershcovich et al., 2022), we evaluate models using each country's native language.

Our Approach Prior datasets primarily focus on cultural norms or social practices. We target culturally grounded factual knowledge: shared references that define collective identity. We leverage Wikimedia resources together with social science expertise to construct a large-scale dataset for Latam, combining: *(i)* Portuguese and Spanish alongside English to investigate language effects; *(ii)* comparison between Latam and Spain to assess regional representation in training data; and *(iii)* fine-grained analysis across Latam countries and cultural elements. Our contributions are:

- a scalable methodology for creating geographically informed sociocultural Q/A datasets using Wikipedia categories, expert curation, and LLM-based generation,
- the LatamQA benchmark of 23,499 multiple-choice questions covering 20 Latam countries (see Figure 1) in Spanish, Portuguese, and English, and
- an empirical analysis of performance variation across countries, prompting languages, and between Latam and Iberian Spanish cultural knowledge.

2 Benchmark Creation

2.1 Raw Wikipedia Data

We apply two sociology-based filters to enhance the dataset pertinence: one at the category level and one at the article level.

Collection Our data collection method relies on the Wikipedia categories' ontology. As each category contains articles and subcategories, it is possible to scrap the content in a recursive way, and, therefore, obtain a structured list of articles with associated metadata. We start from a mother category containing cultural information about a Region of Interest (RoI) such as "*Cultura de Chile*," "*Cultura de Peru*," or other RoI, and recursively collect the links of the Wikipedia articles and subcategories (see Algorithm A.1 in Appendix A).

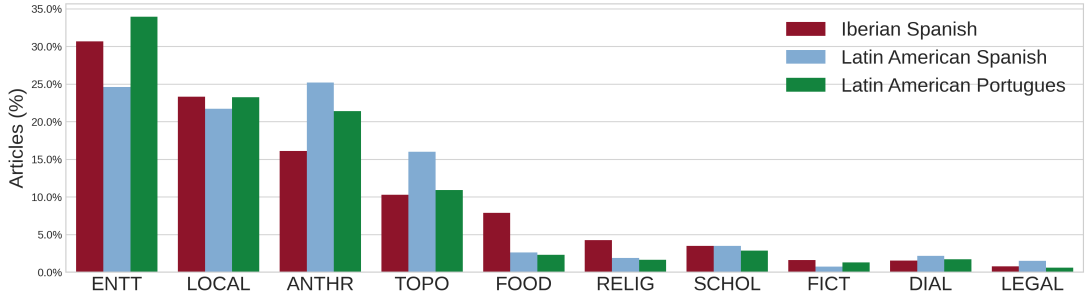

Figure 2: Distribution of the ratio of articles per cultural element per Language or Region in LatamQA. Cultural elements are: Anthroponyms (ANTHR), Forms of entertainment (ENTT), Local Institution (LOCAL), Toponyms (TOPO), Dialect (DIAL), Food and Drink (FOOD), Legal System (LEGAL), Scholastic reference (SCHOL), Religious celebration (RELIG), Fictional character (FICT).

A manual validation of the main subcategories[2] from a sociologist helps removing the categories that are not relevant for a RoI, such as "*Idioma Español*" which contains everything related to Spanish language in general, or "*Alumnados de* `[ENT]`" which contains all the people that went to the school `[ENT]`. This allowed us to obtain 154k articles. Metadata from Wikimedia was used to filter out the documents not relevant to the specific country.

Curation Not all articles contained within the remaining subcategories are equally relevant to cultural analysis. To address this, we apply a second filter at the article level by manually annotating each article according to its socio-cultural relevance. We define three classes: positive, descriptive, and negative. The negative class includes articles that do not address any cultural elements defined in the taxonomy proposed by Espindola and Vasconcellos (2006). Articles that are culturally relevant are assigned to either the positive or descriptive classes. The descriptive class is used for articles that primarily contain technical or enumerative information details[3] with limited interpretive value, such as lists of songs from a specific music album. 500 articles were manually tagged and used to fine-tune a pre-trained multilingual Longformer (Beltagy et al.), reaching a precision of 87.5% with respect to the positive class, and 100% when merging the positive and descriptive classes. Details are available in Appendix B.

Cultural Elements Distribution Wikipedia articles are generally associated with metadata specifying their entity type within an ontology (Vrandečić, 2012). We obtained 2,169 distinct entities across the entire dataset. Using an LLM (Qwen3-Max), we tagged the entities in a zero-shot in-context-learning way with rapid manual verification, mapping each Wikidata entity type to one of our predefined cultural elements. The resulting distribution of articles across cultural elements is presented in Figure 2.

2.2 Questions and Answers Generation

We leverage the filtered Wikipedia articles database to generated article-grounded questions and associated answers. Several prompting strategy were evaluated, with respect to a topic-dependant definition of culture that would apply the most to extract interesting knowledge from the Wikipedia data. `gpt-oss-120b` was used during this phase. Examples of questions and answers are available in Table 2.

General Prompts We compared several prompts to generate questions grounded with various definitions of cultures, based on: anthropology, general cultural exploration, psychological and symbolic significance, sociology, or on an integrative cultural definition. To select the definition leading to the most pertinent questions, they were manually validated by a sociologist with respect to their simplicity, quality and sociological pertinence. Details in Appendix D.

Questions Generation and Validation Once the culture definition fixed, we designed a prompt for article-grounded extraction of questions and associated answers in a structured way. We quantitatively validated the socio-cultural pertinence of the questions using a three-dimension notation based on (Geertz, 1973; Hudson et al., 2009; Páez Rovira et al., 2007) that : (i) symbolic, (ii) social practices, and (iii) social representations, memory and iden-

[2] up to three layers of depth inside the ontology

[3] List of the football teams, statistics, transfer dates of a player *vs.* political history, details on the anthem, rivalries with opponents of a club.

Category	Questions and Answers
FICT	**Q:** What role did the *Bacab* play in Maya beekeeping? **A:** They were the primary protectors of bees and founders of apiculture. **Q:** What is *glíglico*, and in which literary work does it appear? **A:** *Glíglico* is a fictional language created by Julio Cortázar and appears in his novel *Rayuela*.
FOOD	**Q:** In which Mexican state is the *memela* considered a traditional dish? **A:** The *memela* is a traditional dish from the state of Puebla. **Q:** What cultural origins does Lima's *mazamorra morada* have in Peru? **A:** Lima's *mazamorra morada* has Afro-Peruvian roots and is part of Peru's culinary identity.
DIAL	**Q:** What old expression used in Mexico City means "it seems to me" and remains in everyday speech? **A:** The expression "*se me hace*" is used in Mexico City to mean "it seems to me." **Q:** In Chile, which social group is the term "*flaite*" disparagingly directed toward? **A:** It refers to lower-class individuals who are socially inadapted and aggressive, and also to any vulgar behavior regardless of social origin.
RELI	**Q:** What is the role of the *machi* in Mapuche culture? **A:** The *machi* is a Mapuche medicine central figure, healing physical and spiritual ailments, with religious and social functions **Q:** On which day of the month is the tradition of eating *ñoquis* observed in Argentina, Uruguay, Brazil, and Paraguay? **A:** It is celebrated on the 29th of each month.
ENTTT	**Q:** According to the article, what is the origin of *cumbia*? **A:** *cumbia* results from a mixture of Indigenous and African influences. **Q:** In which Mexican state is the novel *Falsa liebre* set? **A:** The novel is set in Veracruz, showing the most marginalized and violent side of the region.

Table 2: Example of questions and answers from different cultural elements.

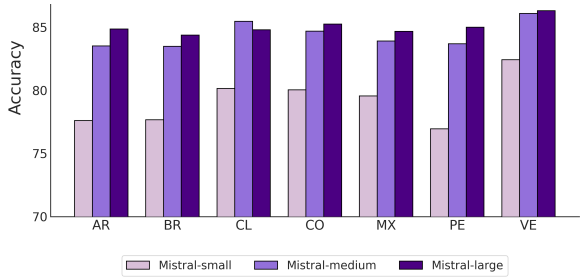

Figure 3: Cross-country performance of Mistral models on LatamQA. Scaling from Small to Large yields consistent improvements (+5 – +8% accuracy).

Model	Brazilian PT		Latam SP		Spain	
	PT	EN	SP	EN	SP	EN
Small models						
Llama 3.1-8B	65.9	66.2	69.2	64.5	76.0	80.5
Mistral-small	77.0	74.3	78.5	76.1	84.3	81.4
Medium models						
Qwen2.5-14B	65.1	62.1	68.8	67.5	79.1	78.2
GPT-4.1-mini	80.0	76.1	81.5	78.2	**88.0**	85.1
Llama-3.1-70B	72.3	73.1	76.3	77.8	83.7	82.5
Mistral-medium	82.6	81.8	83.9	80.5	87.1	85.4
Large models						
Qwen3-430B	70.8	71.4	75.8	74.0	83.7	82.4
Kimi-K2-thinking	69.6	70.5	71.6	70.9	81.0	76.1
Mistral-large	**84.3**	**83.0**	**85.4**	**81.8**	87.6	**86.4**
Latam-focused models						
PatagonIA	81.5	76.8	82.0	79.2	86.9	84.9
LatamGPT-1.0-70B	72.3	74.0	76.4	78.1	72.3	83.0

Table 3: Performance of various LLMs on the LatamQA benchmark (accuracy %). We evaluate models with both native languages (PT and SP) and MT English translations.

tity. We found that 98% of the questions were at least relevant in two of the three dimensions. We also validated quantitatively the answers' grounding to the article on 100 examples, and found no case of hallucination. Details in Appendix E.

Distractors Generations We generated distractors as Fung et al. (2024), generating alternative answers with the same LLM that generated the questions. Details in Appendix D.2.

3 Experiments and Results

Global Results and Prompting Language The performances of various size models are visible in Table 3 and in the same range of other culture-related datasets (Myung et al., 2024; Ying et al., 2025). All the models are performing better in their native language (ES or PT), which is consistent with past results as Spanish and Portuguese are high-resource languages (Myung et al., 2024; Ying et al., 2025) but contradictory with other works (Tanwar et al., 2025; Zhao et al., 2025). We believe that might be due to the nature of the source document to create the Q/A (graph triplet) or because of heterogenous generation capabilities between languages (Kabir et al., 2025). We also include PatagonIA (Instituto Sistemas Complejos de Ingeniería and WideLabs, 2026) and LatamGPT-1.0-70B (Latam-GPT, 2026). PatagonIA is an LLM specialized in Chilean Spanish, reportedly based on a Sparse-MoE architecture. LatamGPT-1.0-70B comes from a Llama 3.1 70B that has been pretrained again over 300 billion tokens spanning Spanish, English, and Portuguese, with a significant portion of the data sourced directly from various countries across the Latam region.

Despite the regional specialization, neither PatagonIA nor LatamGPT outperform general-purpose models of medium size such as Mistral-medium or GPT-4.1-mini. It is interesting to note that LatamGPT reaches higher performances in English than Spanish or Portuguese. This could be attributed to the Latam-specific data fine-tuning, which likely triggered catastrophic forgetting and degraded performance in these languages.

Model Size *vs.* Performance While performances are heterogenous with respect to the LLMs,

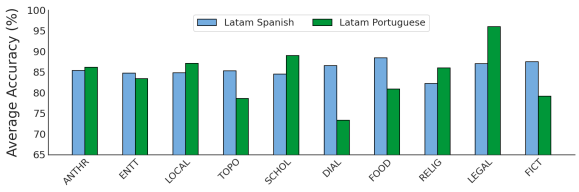

Figure 4: Performance of Mistral-large in Latam Spanish and Portuguese with respect to the different cultural elements.

it is notable that they are homogenous with respect to the size of the model. We can notice consistent improvements, with the exception of Mexico, on the biggest countries for various Mistral models in Figure 3.[4]

Iberian *vs.* Latam Spanish Using a similar process, we extracted a set of Q/As from Spain to compare the performances of the models. All the models performed better on the Iberian Spanish subset. The results are coherent with Myung et al. (2024), where the models reached higher performance on questions from Spain than questions Mexico subset. It is interesting to note that even if Mistral models still perform very well, now the best results are obtained with `GPT-4.1`.

Cultural Element-level Analysis Leveraging Wikidata's type of entity ontology, we automatically map every article and its associated question to its cultural element in the taxonomy of Espindola and Vasconcellos (2006). Performance gaps (see Figure 4) between Latam Spanish and Brazilian Portuguese are higher when the cultural elements contain few examples such as "Fictional character" and "Dialect" where we observed between 75% and 80% of accuracy for Brazilian Portuguese.

4 Conclusion

This work introduces a new sociocultural benchmark focused on Latam. The benchmark contains more than 23k distinct multiple-choice questions (MCQs), each paired with a ground-truth answer. We construct this large-scale, structured dataset by combining information from Wikipedia and Wikidata with domain expertise from the social sciences. First, LLM performance varies substantially across models, although it remains consistent across different scales within the Mistral family. Second, using the native language of the target culture leads to better performance for Spanish and Portuguese. Third, all evaluated LLMs perform better in Iberian Spanish than in Latin American Spanish. Fourth, an analysis at the level of cultural elements shows that performance varies depending on the type of knowledge being tested. Taken together, the results indicate that LLMs can operationalize epistemic injustice: they deliver systematically higher reliability for contexts already more visible in dominant information infrastructures while degrading for underrepresented ones even when language is held constant. Because LLMs are increasingly used as default knowledge interfaces, these asymmetries risk automating gatekeeping and amplifying existing global gradients in recognition and authority.

5 Limitations

This work represents an initial step toward estimating the cultural knowledge of large language models (LLMs) in South America. However, cultural knowledge cannot be adequately captured through simple prompt-based question answering alone (Zhou et al., 2025; Kabir et al., 2025). Future work should therefore move beyond basic multiple-choice question answering (MCQA) benchmarks (Oh et al., 2025). Promising directions include directly involving human participants in benchmark construction (Ivetta et al., 2025a,b) and analyzing interactional data, such as discussions in Wikipedia Talk Pages associated with the target articles.. Similarly, we are aware of the possible preference biases (Wataoka et al.) that might be introduced by using only one LLM.

Acknowledgments

This work was partially financed with the grant U-INICIA 2024 from the Vicerrectoría de Investigación y Desarrollo (VID) number UI-011/24 "Estudios de sesgos sociales en modelos de lenguajes largos," and by the Franco-Chilean Binational Center of Artificial Intelligence, ANID Strengthening R&D capabilities Program CTI230007 Inria Chile. This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0180616119 made by GENCI. We thank the Patagonia IA and LatamGPT teams for sharing the result of their models.

References

Glenn Adams, Stephanie L Anderson, and Joseph K Adonu. 2004. The cultural grounding of closeness and intimacy. In *Handbook of closeness and intimacy*, pages 331–350. Psychology Press.

[4] Same phenomena observed for `Qwen2.5` and `Qwen3`.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. *EMNLP*.

Valentin Barriere and Sebastian Cifuentes. 2024a. Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method with Least Confounding Variables. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1511–1518, Torino, Italia. ELRA and ICCL.

Valentin Barriere and Sebastian Cifuentes. 2024b. A study of nationality bias in names and perplexity using off-the-shelf affect-related tweet classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 569–579, Miami, Florida, USA. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The Long-Document Transformer.

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 ON MEASURING THE (LACK OF) CULTURAL KNOWLEDGE OF LLMS. In *ACL*, pages 1–26.

Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, Mohamed Bin Zayed, and Dirk Hovy. 2024. Classist Tools: Social Class Correlates with Performance in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 12643–12655.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Márton Demeter. 2020. *Academic knowledge production and the global south: Questioning inequality and under-representation*. Springer.

Elaine Espindola and María-Lucia Vasconcellos. 2006. Two facets in the Subtiling Process: foreignisation and/or domestication procedures in unequal cultural encounters. *Fragmentos: revista de língua e literatura estrangeiras*, (30):43–66.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis. volume 1, pages 22413–22430.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *ACL*, volume 1, pages 11737–11762.

Karën Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S Schmeisser-nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, and 6 others. 2024. Your Stereotypical Mileage may Vary : Practical Challenges of Evaluating Biases in Multiple Languages and Cultural Contexts. In *LREC-COLING*, 2, pages 17764–17769.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. No Culture Left Behind: Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking.

Clifford Geertz. 1973. Thick Description: Toward an interpretive theory of culture. *The interpretation of cultures: Selected essays*, pages 3–30.

H. P. Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology. In *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, volume 1, pages 1926–1939. Association for Computing Machinery.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin A. Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking Multiple Languages Affects the Moral Bias of Language Models. In *Findings of ACL: ACL 2023*, pages 2137–2156.

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. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:6997–7013.

Scott Hudson, C Smith, M Loughlin, and S Hammerstedt. 2009. Symbolic and interpretive anthropologies. *Retrieved December 2025*, 6:2013.

Instituto Sistemas Complejos de Ingeniería and WideLabs. 2026. Patagonia ia.

Guido Ivetta, Marcos J Gomez, Sofía Martinelli, Pietro Palombini, M Emilia Echeveste, Nair Carolina Mazzeo, Beatriz Busaniche, and Luciana Benotti. 2025a. HESEIA : A community-based dataset for

evaluating social biases in large language models , co-designed in real school settings in Latin America. *EMNLP*, pages 25107–25129.

Guido Ivetta, Pietro Palombini, Sofía Martinelli, Marcos J Gomez, Sunipa Dev, Vinodkumar Prabhakaran, and Luciana Benotti. 2025b. Adaptive Data Collection for Latin-American Community-sourced Evaluation of Stereotypes (LACES).

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the Checkbox : Challenging Closed-Style Evaluations of Cultural Alignment in LLMs. In *EMNLP*, pages 25–52.

Latam-GPT. 2026. Latam-GPT.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024. How Well Do LLMs Identify Cultural Unity in Diversity? In *CoLM*, pages 1–23.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*.

Luis Josué Lugo Sánchez. 2025. Innovación social académica en tiempos de capitalismo cognitivo: El caso de la Biblioteca de Prompts Colaborativos. *Teknokultura: Revista de Cultura Digital y Movimientos Sociales*, 22(2):185–196.

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. *NeurIPS Datasets and Benchmarks Track*, pages 1–36.

Tuan Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting Cultural Commonsense Knowledge at Scale. *ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023*, pages 1907–1917.

Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. 2025. Culture is Everywhere : A Call for Intentionally Cultural Evaluation. In *EMNLP*, pages 19156–19168.

Darío Páez Rovira, Elza María Techio, and José Marques. 2007. Memoria colectiva y social. In *Psicologia social*, pages 693–716. McGraw-Hill USA.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1771–1781.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of Cultural Awareness in Language Models: Text and Beyond. pages 1–87.

Tamara Quiroga, Felipe Bravo-Marquez, and Valentin Barriere. 2025. Adapting bias evaluation to domain contexts using generative models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28055–28066, Suzhou, China. Association for Computational Linguistics.

Taffy E Raphael. 1986. Teaching Question Answer Relationships, Revisited. *Reading Teacher*, 39(6):516–522.

Nihar Ranjan Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With Prejudice to None : A Few-Shot , Multilingual Transfer Learning Approach to Detect Social Bias in Low Resource Languages. In *Findings of ACL: ACL 2023*, pages 13316–13330.

Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, volume 1, pages 9080–9102.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics*, pages 5477–5490.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5884–5906.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua 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 ACL: EMNLP 2024*, pages 1–32.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang, and Wang Tanmoy. 2025. Do You Know About My Nation ? Investigating Multilingual Language Models ' Cultural Literacy Through Factual Knowledge. In *EMNLP*, pages 14968–14991.

Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*, pages 1063–1064.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F Chen. 2024. CRAFT: Extracting and Tuning Cultural Instructions from the Wild. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 42–47.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: The problem of biased datasets. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:602–608.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling Language and Culture for Evaluating Multilingual Large Language Models. In *ACL*, volume 1, pages 22230–22251.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A Hedderich. 2025. MAKI EVAL: A Multilingual Automatic WiKIdata-based Framework for Cultural Awareness Evaluation for LLMs. In *Findings of ACL: EMNLP 2025*, pages 23104–23136.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is Not Trivia: Sociocultural Theory for Cultural NLP.

A Scraping Algorithm

The scraping algorithm is described in Algorithm A.1. We collected articles from the category pages titled "*Cultura de* `[Country]`" for the 20 countries listed in Table C.1 using each country's main language. We set MAX_DEPTH to 5 as empirical testing showed that greater depths reduced the relevance of the retrieved articles. For Spain, we used `MAX_DEPTH = 3` due to the substantially larger initial number of articles.

B Automatic Filtering

B.1 Elements of Culture

We base our filtering on the elements of culture from Espindola and Vasconcellos (2006) used to filter out articles in the negative class.

B.1.1 Definitions

- **TOPO** (Toponyms): a place name, a geographical name, a proper name of locality, region, or some other part of Earth's surface or its natural or artificial feature.
- **ANTHR** (Anthroponyms): ordinary and famous people's names and nick-names and names referring to regional background which acquire identification status; it also includes animals that have been given human qualities, symbols, or political meanings in social representations.
- **ENTT** (Forms of entertainment): amusement or diversion including public performances or shows, it also encompasses hospitality provided, such as dinners, parties, business lunches, etc.; it also includes artistic expressions.
- **FICT** (Fictional character): a person in a novel, play, or a film who is related to fiction, works of imagination.
- **LEGAL** (Legal System): rules of conduct inherent in human nature and essential to or binding upon human society.
- **INST** (Local Institution): an organization that helps or serves people in a certain area - health, education, work, political, administrative, religious, artistic; it also includes national symbols.
- **FOOD** (Food and Drink): any solid or liquid substance that is used by hu-man beings as a source of nourishment.
- **SCHOL** (Scholastic reference): related to school or studying.
- **RELIG** (Religious celebration): to do something special to mark a religious occasion.
- **DIAL** (Dialect): user-related variation, which determines speaker's status as regards social class, age, sex, education, etc.; it also includes slang.

B.1.2 Distribution

The distributions of the articles with respect to its cultural element relevance is shown in Figure 2. The biggest difference lies within the ratio of articles about Food and Drink: they are more dominant in Spain than in Latam.

Algorithm A.1 Recursive Wikipedia Category Scraper

```
Input: Initial Wikipedia category URL, Maximum recursion depth MAX_DEPTH
function SCRAPECATEGORY(categoryURL, currentDepth):
    if currentDepth > MAX_DEPTH then
        return
    end if
    Fetch HTML category page to extract articles and subcategory links
    for each article link NOT already processed do
        Fetch HTML article page content and save article data
    end for
    for each subcategory link do
        SCRAPECATEGORY(subcategoryURL, currentDepth + 1)
    end for
SCRAPECATEGORY(initialCategoryURL, currentDepth = 0)
```

B.2 Classifier

We fine-tuned and validated a pre-trained XLM-RoBERTa Longformer[5] on 500 3-class examples. When merging the descriptive and positive classes, the classifier reaches an accuracy of 97.8%. The confusion matrix obtained from cross-validation is the following ($c_{i,j} = c_{y,\hat{y}}$):

$$\begin{bmatrix} 198 & 24 & 0 \\ 43 & 98 & 19 \\ 11 & 40 & 131 \end{bmatrix}$$

C Per-Country Distribution

The distribution of the dataset questions per country is shown in Table C.1.

Country/Region	Language	Count
Brazil (BR)	Portuguese	6,075
México (MX)	Spanish	4,893
Argentina (AR)	Spanish	4,243
Chile (CL)	Spanish	2,469
Perú (PE)	Spanish	1,921
Colombia (CL)	Spanish	1,752
Brazil (BR)	Spanish	1,164
Venezuela (VE)	Spanish	1,030
Cuba (CU)	Spanish	674
Ecuador (EC)	Spanish	720
Uruguay (UY)	Spanish	991
Bolivia (BO)	Spanish	750
Guatemala (GT)	Spanish	743
Costa Rica (CR)	Spanish	467
El Salvador (SV)	Spanish	306
Nicaragua (NI)	Spanish	436
Paraguay (PY)	Spanish	542
Dominican Republic (RD)	Spanish	234
Honduras (HN)	Spanish	180
Panamá (PA)	Spanish	218
Puerto Rico (PR)	Spanish	193
Total		**26,213**

Table C.1: Distribution of articles from the LatamQA dataset that are from a categories under the mother category `"Cultura de [Country]"`. In this case, an article can be associated with several countries, and possibly languages.

D Questions-Generation Prompts

D.1 Domain-specific Culture Definition

We first selected a prompt from five different prompts using different definitions of culture: an anthropological approach, general cultural exploration approach, psychological and symbolic significance approach, sociological approach and integrative cultural definition approach. The quality of the questions was assessed with respect to the clarity of language under "theoretical principles of phenomenology, which studies things as they are shown in consciousness to make them comprehensible" (Lugo Sánchez 2025; page 188), which means expressing the questions in simple terms. A good question is correctly formulated, asking for something precise and not ambiguous present in the article, not using an overly complex or specific vocabulary or concepts (such as "collective identity" or "communal expression").

The general cultural exploration approach was judged the most relevant for the benchmark creation (see Figure D.1).

D.2 MCQA Generation

Using general cultural exploration approach, a new prompt was designed to extract the pairs of Q/As,

[5] `markussagen/xlm-roberta-longformer-base-4096`

General Cultural Exploration

System: From the given Wikipedia article, identify and formulate relevant sociocultural questions and answers.

Questions should address topics such as cultural identity, popular symbols, collective memory, traditions, and practices specific to the local community or region described.

Answers must briefly summarize the sociocultural relevance of these elements

Figure D.1: General Cultural Exploration approach prompt.

Cultural QA Generation Prompt (Spanish)

System: You are an expert in cultural anthropology and educational content creation.
Your task is to generate three culturally meaningful question–answer pairs in Spanish based only on information explicitly present in the following Wikipedia article.

Goal: Identify the most culturally significant element explicitly mentioned in the article and produce three question–answer pairs about it, focusing on:

- Cultural identity
- Popular symbols
- Collective memory
- Traditions
- Practices specific to the local community

Rules:

- Use only information explicitly contained in the article.
- Select the most widely recognized cultural element mentioned.
- The question must be clear and culturally meaningful.
- The answer must be brief and concise, one or two sentences at most.
- Output must follow exactly the JSON format below.
- Do not invent information.

Article: `{content}`

Output Format:

```
{"question1": "...", "answer1": "...",
 "question2": "...", "answer2": "...",
 "question3": "...", "answer3": "..."}
```

Figure D.2: Prompt template for generating culturally grounded question–answer pairs from Spanish Wikipedia articles.

adding specific rules to force the questions and answer to be precise, explicit, pertinent and generated in a specific format (see Figure D.2). A good answer responds totally to the question, uses solely the content of the article without adding external facts, and does not add specific reasoning (Raphael, 1986; Grice, 1975).

Second, following the methodology of (Fung et al., 2024), another prompt was used to generate challenging counterfactual answers for the MCQ, which we call distractors (see Figure D.3).

E Validation of Q/As

We asked two experts to score 100 questions with a 5-point likert scale with respect to the symbolic, the social practices, and the social representations, memory and identity. Only two questions over the 100 obtained a score less than 5, which means that only 2% of the questions obtained a score of 1/1/1 or 1/1/2 and were rejected. Inter-annotator-agreement was high.

F Mistral Models Performances

Cross-country Performances The full performances of the models from the Mistral family are visible in Figure F.1. It is visible that the scale consistently helps in reaching higher performances. Except for a few countries where the medium excels slightly the large model: Costa Rica, Honduras and Ecuador.

Cultural Element-level Performances Figure F.2 shows the performances of the Mistral models for separated with respect to the cultural elements of the questions.

MCQ Distractor Generation Prompt (Spanish)

System: You are an expert in creating difficult multiple-choice questions to evaluate reading comprehension and cultural knowledge.

Context:
Article: `{content}` Question: `{question}` Correct answer: `{answer}`

Task: Generate exactly 3 incorrect answers (distractors) in Spanish that are:

- Plausible and similar in length/style to the correct answer
- Culturally believable (could reasonably be confused with the correct answer)
- Difficult to eliminate without having carefully read the article
- Grammatically correct in Spanish

Strategies to use (vary across the 3 distractors):

1. *Entity swap*: Use another cultural element, place, or name mentioned in the article
2. *Temporal/geographical confusion*: Use a date, period, or location from elsewhere in the article
3. *Partial truth*: Capture part of the cultural context but make false on the key element
4. *Cultural plausibility*: Reference a related but incorrect tradition from the same region
5. *Subtle relation inversion*: Invert origin/influence, cause/effect, or historical sequence

Constraints: Each distractor must fool a quick reader; avoid absurd answers and obvious negations; match the tone of the correct answer.

Figure D.3: Prompt template for generating challenging distractors for multiple-choice questions.

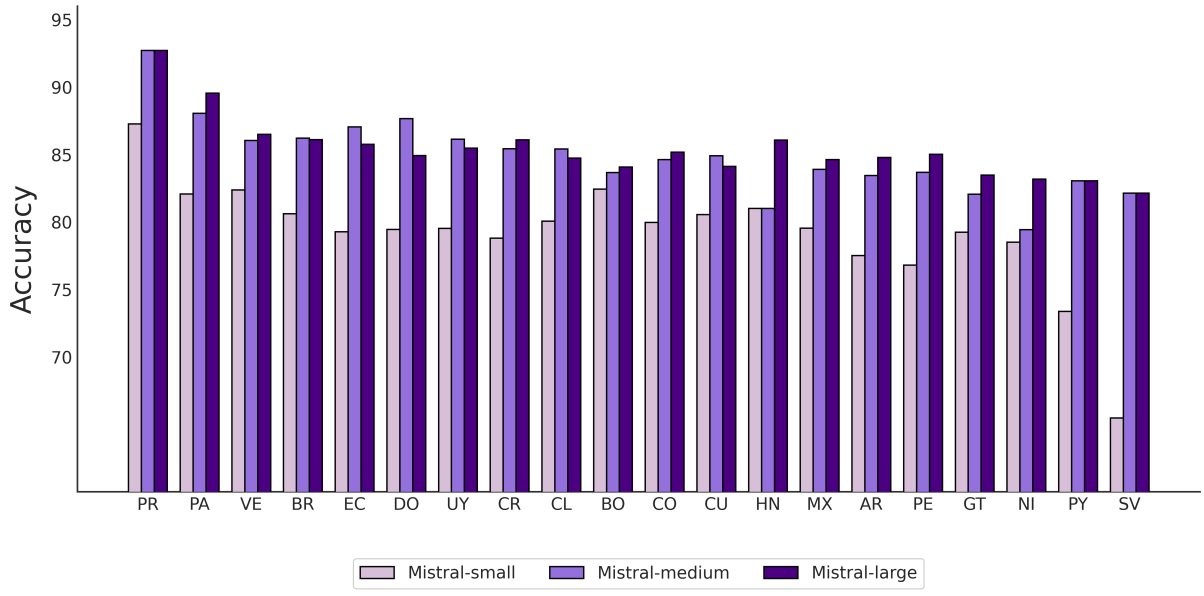

Figure F.1: Cross-country performance of Mistral models on cultural knowledge evaluation. Scaling from Small to Large yields consistent improvements (+5-8% accuracy).

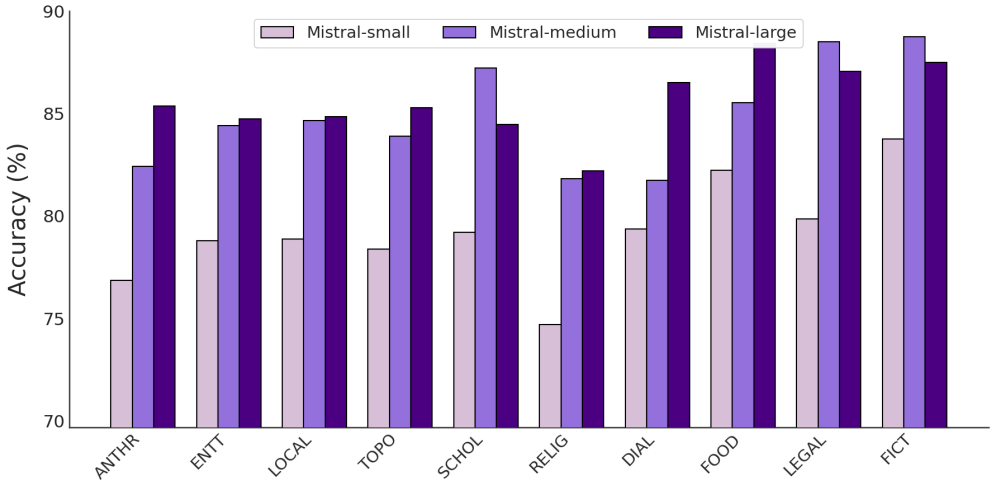

Figure F.2: Performance of Mistral models on Latam Spanish data, with respect to the cultural element.