Leveraging Wikimedia Data for Geographically-Informed Sociocultural Bias Dataset Creation

Valentin Barriere
CS Department
Universidad de Chile | CENIA

Aidan Hogan
IMFD | CS Department
Universidad de Chile

Andres Abeliuk
CS Department
Universidad de Chile | CENIA

Hernan Contreras
Institute of International Studies
Universidad de Chile

Abstract

Large Language Models (LLMs) exhibit inequalities with respect to various cultural contexts. Trained on Global North data, they can show prejudicial behavior towards other cultures. Moreover, there is a notable lack of resources to detect biases in non-English languages, especially in Latin American dialects. 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 Questions/Answers (Q/As) pairs, based on the different popular and social cultures of various Latin American countries. We propose to work on the definition of sociocultural bias such that computing methods can be used for both detecting and quantifying its associated valence. We will focus on general methods adapted to multilingual models in various contexts and propose to apply this to Latin America, a continent containing various cultures, even though they share a common cultural ground.

Introduction

Biases in AI, especially in Natural Language Processing (NLP), are pervasive and multi-faceted. They originate from multiple sources, including the data used for training (Wiegand et al., 2019), annotation processes (Santy et al., 2023; Sap et al., 2022), and even the instructions provided during annotation campaigns (Parmar et al., 2023). These biases can manifest as moral (Hämmerl et al., 2022), social (Sap et al., 2020), class-related (Curry et al., 2024), or political biases (Feng et al., 2023), influencing LLM behavior in ways that may perpetuate or exacerbate societal inequalities and cultural colonization (Amsler, 2007; Tomlinson, 2001).

Social biases can also be explicitly annotated for detection within sentences, whether these biases are overt or implicit (Sahoo et al., 2023). However, annotation efforts are costly and heavily dependent on language and cultural context (Fort et al., 2024). The subtle nuances contained in the source languages render machine translation inadequate for such tasks due to lack of cultural sensitivity, irony, etc. Moreover, Hada et al. (2024) emphasize the need for localizing the creation of quality data and the dangers of outsourcing the creation of non-English bias detection datasets in the Global North. Indeed, current datasets are poor at assessing real bias in Global South countries, failing to capture the intricacies of the local cultures, dialects, and knowledge (Santy et al., 2023). Indeed, assessments rely on databases limited to the Global North, which leads to an incomplete understanding of how biases manifest in diverse contexts, whether cultural or linguistic. This over-reliance risks

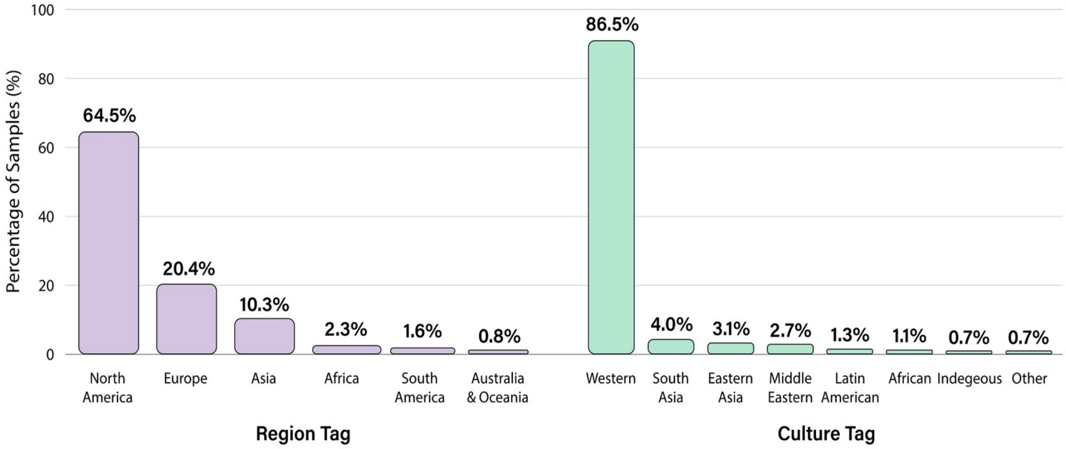


Figure 1: The distribution of cultural knowledge of MMLU (Hendrycks et al., 2021) is very unbalanced with respect to the geography. The dataset LATAM-related part is ridiculously small, representative of current datasets for cultural knowledge and bias assessment.

perpetuating systemic inequalities, as models trained on such datasets are unable to address the unique challenges and sensitivities of the Global South (Hada et al., 2024). To ensure a comprehensive and fair approach, it is critical to prioritize the creation of high-quality, localized datasets that reflect the lived realities of underrepresented regions.

As AI systems, especially LLMs, become an integral part of critical domains, addressing bias and ensuring fairness remain key challenges. While significant progress has been made in understanding and mitigating bias, several open questions remain: How can we comprehensively address the trade-off between fairness and interpretability, especially in multilingual and multicultural contexts? What methods can ensure equitable representation of underrepresented groups in resource-limited settings? Finally, how can these efforts be scaled to keep pace with the rapid development and use of LLMs in diverse applications? Addressing these questions will require interdisciplinary collaboration, robust evaluation frameworks, and an ongoing commitment to aligning AI systems with ethical and societal values.

Finally, in a South American context, current geo-cultural bias detection datasets for LLMs are under development, but the grid of bias detection is either coarse, regrouping countries in huge groups based on their GDP (Czarnowska et al., 2021), or very sparse, containing only a very few countries (Myung et al., 2024). A thorough analysis on Latin America is omitted from the study for lack of fine-grained data (Singh et al., 2024; see Figure 1).

According to Adilazuarda et al. (2024) and Hershcovich et al. (2022), in the axes to consider when assessing the cultural knowledge of a model, common ground is the shared knowledge that can be assumed as known by others. Factual local knowledge is a subpart of this, on which we are focusing at a fine-grained level with respect to the geographic region. We are targeting popular culture and sociocultural knowledge as this is a strong common basis for individuals of the same social group and important to understand the local humor, political discussions, and popular references. Even though studies state that culture is not restrained to trivia (Zhou et al., 2025), references and knowledge of popular culture are important basics for common ground in the way people are communicating with each other, helpful to reach a higher degree of intimacy when two entities are discussing (Adams et al., 2004). Even though some events can be specific to social groups and can be told in-between the groups, some others are the basis of a cultural foundation that enables fluid communication. The way we refer to certain events can define social groups, hence knowing what these events are about is essential.

To summarize

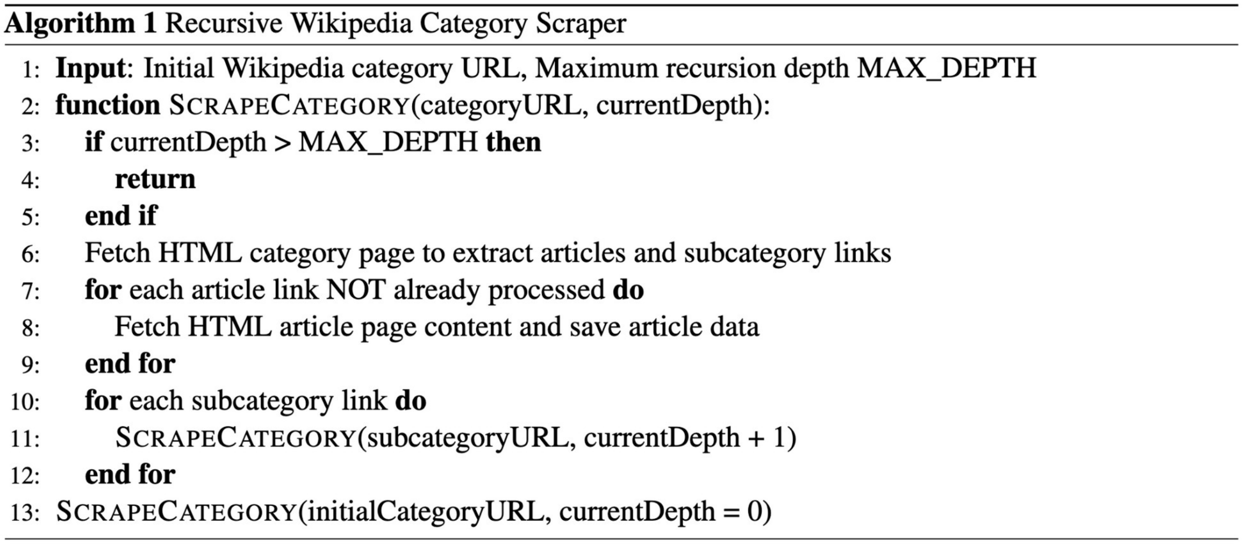
- We would like to detect biases in LLMs, in terms of their factual and socio-cultural knowledge on specific regions.
- This is an important issue, and especially for Inclusion, as LLMs are increasingly adopted for critical applications.
- Our solution would take the form of a database of Q/As, created from high-quality Wikipedia and Wikidata content that is semi-automatically curated. This dataset can serve many more purposes than just bias assessment (see Section 3.5).
- Our project aligns with three of the Wikimedia 2030 Movement Strategy Recommendations. By using Wikipedia's structured knowledge and engaging Wikimedia Chile, the project promotes are recommendations **[3] Provide for Safety and Inclusion**, **[7] Manage Internal Knowledge** and **[8] Identity Topics for Impact**.
- We hypothesize that Wikimedia data can be leveraged to extract high-quality data related to socio-cultural knowledge.

The starting date of our proposal is July 1, 2025 and it will end a year later.

Related work

Culturally aware Natural Language Processing (NLP) is an emerging topic that has drawn a lot of attention in recent years (Hershcovich et al., 2022; Liu et al., 2024; Pawar et al., 2024). Fung et al. (2024) propose CultureAtlas, which uses Wikipedia content in order to create a dataset of assertions that are true or false. They focus on cultural norms and practices, but not on local factual knowledge from popular culture, useful to understand inside jokes or references. BLEnD (Myung et al., 2024) is a dataset of questions and answers (Q/As), from 16 countries/regions and 13 languages containing 15,000 short answers questions on topics such as sports, food, family, education, work-life, and holidays. As the dataset has been created manually, its size and topics are limited to 1,000 Q/As per region. CulturalBench (Chiu et al., 2024) is another example of a manually created dataset, containing 1,227 Q/As on 17 diverse cultural topics from 45 global regions.

Nguyen et al. (2023) extract knowledge from large corpora such as C4, which is noisy and leads to a loss of specific information with respect to individual subregions (Fung et al., 2024). Wang et al. (2024) propose CRAFT, a method that retrieves culture-related data from the 600B tokens SlimPajama dataset. They first used keywords to extract text segments and off-the-shelf LLMs to generate Q/As using or not the context, hence relying on non-verified content from the dataset or (non-verified) parametric knowledge of another LLM model. CultureBank (Shi et al., 2024) is a dataset created from TikTok and Reddit, constructed by automatically extracting culture-related comments and associated cultural descriptors using an LLM. The descriptors expressed in natural language are clustered, and then described using another LLM. They evaluate the LLM in a grounded evaluation with a persona and an action in a contextual situation. Unlike prior datasets, our proposal focuses on culturally grounded factual



Algorithm 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)
```

knowledge rather than beliefs or social norms, and it exploits Wikipedia's structured taxonomy to automate and scale the Q/A generation process across underrepresented regions

Methods

Raw Wikipedia Data Collection

Our data collection method is relying on the knowledge graph of Wikidata in several ways: to get an initial pool of candidate articles, and to further filter them. Every category contains articles and subcategories, which makes it possible to scrape the content in a recursive way. The main idea is to start from a category containing cultural information about a region such as "Cultura de Chile", "Cultura de Peru", or other Region of Interest (RoI), and recursively collect the links of the Wikipedia articles (see Algorithm 1). A manual validation of the subcategories from a sociologist helps to reduce the categories that are not relevant for our RoI, such as "Idioma Española" which contains everything related to Spanish in general, or "Alumnados de `[ENT]`" which contains all the people that went to the school `[ENT]`. Overall, this method is domain-agnostic and allows the construction of a high-quality database, enhanced with the specific metadata contained in the Wikidata knowledge graph, which helps to structure and analyze its own content.

Curation

All the articles falling in the subcategories of our mother categories are not all relevant. For this reason, we find it necessary to filter out what is judged as general interesting knowledge from what is more collateral. We collect the following metadata to enhance the characterization of each article:

- Median number of monthly views
- Number of languages of the page
- PageRank score
- Type of entity
- Neighborhoods in the graph
- Path in the category taxonomy
- Length of the article

These features will be evaluated to learn a supervised machine learning model to accept or reject any article based on labels from sociologist annotators. This process will restrict our database of articles to culturally relevant content from the RoI.

Q/As Generation

In this step, the database of socio-culturally relevant Wikipedia articles from a RoI is leveraged in order to create a set of questions and associated answers. For this objective, we are using an LLM taking the article as context in a more general prompt containing a definition of culture.

General Prompts In order to find the prompt that outputs the most relevant questions, we will try to ground them with different definitions of cultures, based on: anthropology, general cultural exploration, psychological and symbolic significance, sociology, or on an integrative cultural definition. The quality of the questions will be validated manually by human experts (i.e., sociologists) with a quantifiable methodology grounded in social sciences. Each of these prompts will be evaluated regarding the quality and sociological pertinence of the generated Q/As (see Section 3.3).

Hallucination Reduction We define criteria to assess the quality of a pair of question/answer: the information asked should be based on the content of the article, without adding any external information such as other facts or complex reasoning, even though true, that will undesirably influence the true answer with the type of LLM used to generate the dataset. 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) . A good answer responds totally to the question, uses solely the content of the article without adding outside facts, and does not add specific reasoning.

These definitions will be added in the prompt of the LLM. Additionally, specific LLM-based techniques to mitigate hallucinations will be applied to ensure that the model focuses on external content and not parametric knowledge (Sun et al., 2025; Jin et al., 2024; Li et al., 2024).

Q/As Validation

At this point, the dataset of Q/As, even though coming specifically from Wikipedia and supposedly on topics of interest, will be passed through a large-scale human validation phase. We will first ask sociologist experts to give a score to the questions, and then non-expert people from different socio-cultural backgrounds to give an interest score to the article, a score to a question, and then give a score to the answer of the model. This ensures a validation of our collected data by both a group of social scientists and a group from civil society. We will also gather human answers to the questions of the dataset in order to compare them with the ones given by Wikipedia-based LLMs.

Benchmark

Ultimately, the benchmark will be used to structurally assess the socio-cultural knowledge of LLMs on specific regions, crucial to improve inclusion within such technologies. Metadata will be used to structure the scores regarding different topics from the RoI, creating inter-country categories such as "Gastronomy", "Arts", "Sports". We plan to evaluate several state-of-the-art LLMs [3] and publicly release our dataset and code, an online platform to explore the data, and the benchmarking results in a scientific paper. As a side note, the benchmark can be extended to multimodal model assessment using the pictures of the articles, which are scraped as metadata.

Other Applications

We will investigate other potential applications of the benchmark such as:

- [7] Linking similar concepts from different regions of interest
- Creating quizzes to test the knowledge of a user on Wikipedia article pages
- [7] Automatically detecting subjective facts in Wikipedia articles
- [3] Based on similarity, recommend a Wikipedia page based on a socio-culturally related user question
- [8] Detect topics where knowledge is missing and where parametric knowledge from state-of-the-art LLMs can help. In particular, methods like the one of Liu et al. (2025) that can trace back to the original document from a trillion tokens corpus in real time, would help to find new content that is missing in the Wikipedia pages.

Expected output

The project will deliver a new methodology to assess sociocultural knowledge in language models, along with an open-source dataset of culturally-specific question-answer pairs focused on the Latin American region. Through a scientific publication in conferences such as EMNLP or ACL, the methodology will be openly documented and shared with researchers in the AI and Social Science community. These researchers will benefit from an open-source, replicable and transparent method that can be adapted for diverse geographical contexts.

The methodology will be applied to the Latin American region (it is language-agnostic), resulting in an open-source dataset of sociocultural knowledge question-answer pairs, and the scores of various LLMs on the benchmark, in order to be aware of cultural gaps and improve the models. An interactive platform will be created to explore the dataset, offer quizzes for Wikimedia users, and collect annotations from the community in order to enhance the dataset quality. Finally, the results will be disseminated through a collaborative seminar with Wikimedia Chile, engaging local communities in contributing to methods to mitigate cultural bias in growingly-mainstream LLMs.

Risks

One key risk is the potential non-relevance of the scraped Wikipedia articles, and another is the hallucinations that may occur during content generation by LLMs. To mitigate the first risk, we will leverage information from the knowledge graph of Wikidata along with sociological expertise to select the most relevant topics, then use automated metrics such as the number of languages in which an article is available, median monthly page views, and PageRank scores (Page et al., 1999) to filter and ensure the pertinence of selected content. For the second risk, we will employ specific techniques to anchor LLM-generated outputs strictly to factual information contained within the Wikipedia articles (Niu et al., 2024; Sun et al., 2025; Gao et al., 2023; Jin et al., 2024), thus reducing hallucinations and focusing on factual information to avoid unintended biases related to the reasoning patterns of specific LLMs.

Community impact plan

We plan to organize a workshop with Wikimedia Chile to present our work to the community. This half-day workshop will take place at the University of Chile and will allow: *(i)* the research team to present in detail the different technical facets of the work such as the models' architectures, the issues faced, and results obtained from the annotation platform; *(ii)* the Wikimedia Chile team to present possible applications of the dataset such as quiz generation or factuality assessment, *(iii)* a demo involving the Wikimedia volunteers from Latin America presenting the data exploration platform and the annotation recollection platform to contribute to the continuous improvement of the project.

Evaluation

A concrete measure of the project success would be the quality of the Latin American dataset. In order to assess its quality, we will ask human experts from diverse sociological backgrounds to give feedback on the dataset, using a set of scoring functions. People from different countries will give a score of pertinence to every question, answer themselves, and finally grade the gold-standard answer obtained from the Wikipedia page. Our expected results on the objective metrics are several: *(i)* people from the country should be significantly better at answering their cultural background, *(ii)* questions relating to a country should be seen as relevant for a significant part of the population of the country, *(iii)* answers should be seen as pertinent and factual. This will also help us to estimate the level of knowledge of humans when compared with LLMs.

Another concrete measure would be the engagement of the Wikimedia community, particularly as the culture evolves in time, to keep collecting data continuously after the project ends and ensure the dataset remains up-to-date. This can be quantified with the number of annotations collected and the number of visits on our dataset platform.

Budget

Total budget of the project has been evaluated of 36,925 USD. Our project does not provide incentives to the researchers, only support for master and PhD students, as students are from a socio-economic class well known to be prone to precariousness in Chile. https://docs.google.com/spreadsheets/d/1wKQAqosN34CJZZOTnNpiaVKtl2N1LT-94Agd0uQD7rY/edit?usp=sharing

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.

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*.

Sarah Amsler. 2007. Cultural colonialism. *The Blackwell encyclopedia of sociology*.

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. 2024. CULTURALBENCH: A Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs. 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)*, pages 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.

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*, volume 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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey.

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.

Dan Hendrycks, Mantas Mazeika, Collin Burns, Dawn Song, Andy Zou, Jacob Steinhardt, and UC Berkeley. 2021. Measuring Massive Multitask Language Understanding. *ICLR*.

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 Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 6997–7013.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2024. Disentangling Memory and Reasoning Ability in Large Language Models. Pages 1–22.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases.

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*.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, YenSung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Cheng, Karen Farley, Sruthi Sreeram, Taira Anderson, and 12 others. 2025. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens.

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. Submitted to *NeurIPS 2024 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.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. *ACL*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Infolab.

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*, 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.

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. NLPersonality: Characterizing Design Biases of Datasets and Models. 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 of the 58th Annual Meeting of the 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 - Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 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.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang

Song, and Han Li. 2025. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability. *ICLR,* pages 1–23.

John Tomlinson. 2001. Cultural Imperialism: A Critical Introduction. A&C Black.

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.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: The Problem of Biased Datasets. *NAACL HLT 2019 - Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* volume 1, pages 602–608.

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