Identity-Aware Large Language Models require Cultural Reasoning

Alistair Plum[1], Anne-Marie Lutgen[1], Christoph Purschke[1], Achim Rettinger[2]

[1]University of Luxembourg, Luxembourg [2]Trier University, Germany

References

[1] Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings, 2024. arXiv:2309.08591 [cs].

[2] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of EMNLP*, 2024.

[3] 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. CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of ACL*, 2025.

[4] Felix Stalder. *The Digital Condition*. John Wiley & Sons, 2018.

[5] Lee Su Kim. Exploring the Relationship Between Language, Culture and Identity. *GEMA Online Journal of Language Studies*, 3(2), 2003. Publicly Available Content Database.

[6] Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In *Proceedings of ACL*, pages 428–446, 2023.

[7] Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of NAACL-HLT*, pages 2373–2403, 2025.

[8] Abdullah Mushtaq, Imran Taj, Rafay Naeem, Ibrahim Ghaznavi, and Junaid Qadir. WorldView-Bench: A Benchmark for Evaluating Global Cultural Perspectives in Large Language Models, 2025. arXiv:2505.09595 [cs].

[9] Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output, 2024. arXiv:2411.06032 [cs].

[10] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of ACL*, 2020.

[11] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns*, 2(11):100336, 2021.

[12] Tarek Naous and Wei Xu. On The Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of NAACL-HLT*, April 2025.

[13] Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75, 2025.

[14] Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. GIMMICK – Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking, 2025. arXiv:2502.13766 [cs].

[15] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. 2024.

[16] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus*, 3(9), September 2024.

[17] Daniel Van Niekerk, María Peréz-Ortiz, John Shawe-Taylor, Davor Orlic, Jackie Kay, Noah Siegel, Katherine Evans, Nyalleng Moorosi, Tina Eliassi-Rad, Leonie Maria Tanczer, et al. Challenging systematic prejudices: An investigation into bias against women and girls. 2024.

[18] Aleksandra Sorokovikova, Pavel Chizhov, Iuliia Eremenko, and Ivan P. Yamshchikov. Artificial Intelligence Gives Women Lower Salary Advice. Technical University of Applied Sciences Würzburg–Schweinfurt press release, 2025. https://www.thws.de/en/research/institutes/cairo/releases/thema/artificial-intelligence-gives-women-lower-salary-advice/.

[19] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. AI Generates Covertly Racist Decisions about People Based on Their Dialect. *Nature*, 633(8028):147–154, September 2024.

[20] Khyati Khandelwal, Manuel Tonneau, Andrew M Bean, Hannah Rose Kirk, and Scott A Hale. Indian-bhed: A dataset for measuring india-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 231–239, 2024.

[21] Reece Rogers. AI Is Spreading Old Stereotypes to New Languages and Cultures. *Wired*, April 2025. Accessed August 27, 2025.

alistair.plum@uni.lu