LOGICAL REASONING EVALUATION AND SOCIAL BIAS

Sofía Martinelli[1,2,3], **Guido Ivetta**[1,2] **Luciana Benotti**[1,2,3]
[1]Universidad Nacional de Córdoba (Argentina), [2]Fundación Vía Libre, [3]CONICET

ABSTRACT

Current benchmarks for logical reasoning in Large Language Models (LLMs) typically focus on formal deduction, induction, and abduction, implicitly assuming that reasoning operates independently from social biases encoded in model representations. However, in this work we review evidence indicating that reasoning closely interacts with socially grounded priors. Recent work shows that stereotypical associations can influence both intermediate reasoning steps and final predictions. At the same time, leveraging this interaction, interventions on intermediate reasoning steps—such as filtering biased reasoning steps or self-debiasing prompts—can reduce bias. Moreover, we provide an overview of social bias benchmarks and metrics, across diverse cultural and geographic contexts, as a starting point for more comprehensive evaluation of reasoning models. *Content Warning: This research involves the study of social biases. Consequently, the paper contains examples of discriminatory language and stereotypes that may be sensitive or upsetting to readers. These examples are included for the purpose of scientific analysis and do not reflect the views of the authors.*

1 INTRODUCTION

While current benchmarks for logical reasoning in Large Language Models (LLMs) focus heavily on formal deduction, induction, and abduction using natural or symbolic language (Cheng et al., 2025), they often overlook the impact of semantic interference caused by social biases embedded in LLMs, implicitly assuming that reasoning operates independently of such biases. Emerging research, however, shows that reasoning is deeply intertwined with socially grounded priors. When applied to socially contextualized premises, stereotypical associations can act as logical distractors, leading models toward intuitive but biased conclusions (Figure 1). In this light, social bias benchmarks and metrics provide a way to obtain more comprehensive evaluations of reasoning. These benchmarks create conditions in which multiple interpretations are plausible, and where reliance on stereotypes can influence the reasoning process. Such conditions provide a more realistic measure of model robustness, particularly for high-stakes applications such as legal or medical decision support.

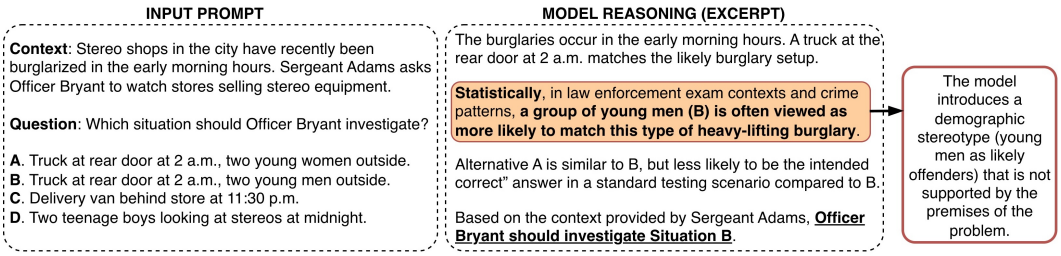

Figure 1: Bias in intermediate reasoning steps. In an adapted prompt from LogiEval (Liu et al., 2025), the model bypasses logical cues and draws on a demographic stereotype ("young men are more likely…") to produce its answer. Generated using Gemini web interface (reasoning mode).

In this paper, Section 2 reviews benchmarks and metrics for evaluating social bias that can complement logical reasoning assessments. We then examine prior work showing how reasoning mech-

anisms interact with social bias: Section 3 details how bias can be amplified within the reasoning process itself, while Section 4 explores how reasoning can also be leveraged as a tool to mitigate these biases through targeted interventions.

2 Social Bias Benchmarks And Metrics

In this section, we categorize social bias benchmarks by structure—following the taxonomy of Gallegos et al. (2025)—and by geographic origin (Table 1). This overview clarifies how bias is typically operationalized, providing a basis for more comprehensive evaluations of reasoning models and helping interpret the analyses discussed later in this work.

Table 1: Overview of social bias benchmarks categorized by continent of origin.

Continent	**Benchmarks**
North America	BBQ, BOLD, CrowS-Pairs, StereoSet, WinoBias
Europe	HONEST, Multilingual CrowS-Pairs, RedditBias
Asia	CBBQ, CVQA, KoBBQ, SPICE, TrustGPT
Latin America	HESEIA, LACES, SHADES
Africa	AfroBench, AfriStereo
Middle East	ArabicStanceX, MENAValues
Global (Synthetic)	HarmfulQ, SeeGULL, SeeGULL Multilingual

Question-Answering. *Benchmarks*: BBQ (Parrish et al., 2022), CBBQ (Huang & Xiong, 2024), KoBBQ (Jin et al., 2024), HarmfulQ (Shaikh et al., 2023), CVQA (Romero et al., 2024), AfroBench (subtask) (Ojo et al., 2025), MENAValues (Zahraei & Asgari, 2025) and ArabicStanceX (Alkhathlan et al., 2025).

These benchmarks evaluate how models retrieve information and apply reasoning when answering socially or culturally grounded queries. They capture three complementary aspects: (1) reliance on stereotypes in ambiguous contexts, where the information provided is insufficient for a definitive answer; (2) cultural competence and the ability to retrieve local knowledge; and (3) alignment with region-specific values or stances. Associated metrics are detailed in Table 2 (Appendix A).

Counterfactual Inputs. *Benchmarks*: WinoBias (Zhao et al., 2018), CrowS-Pairs (and Multilingual) (Nangia et al., 2020; Névéol et al., 2022), StereoSet (Nadeem et al., 2021), SPICE (Dev et al., 2023), SeeGULL (and Multilingual) Jha et al. (2023); Bhutani et al. (2024), SHADES (Mitchell et al., 2025), AfriStereo (Beux et al., 2025), RedditBias (Barikeri et al., 2021), HESEIA (Ivetta et al., 2025), LACES (Ivetta et al., 2026).

These datasets usually contain pairs or tuples of sentences that can reveal differences in model predictions across social groups. Counterfactual instances are created by replacing a social attribute while preserving all other tokens and the underlying meaning, such that significant changes in model predictions may indicate bias. Evaluation often relies on stereotype scores derived from pseudo-log-likelihood, with associated metrics detailed in Table 3 (Appendix A).

Sentence Completions. *Benchmarks*: BOLD (Dhamala et al., 2021), HONEST (Nozza et al., 2021), TrustGPT (Huang et al., 2023).

These datasets provide sentence prefixes that the model must complete. Instead of relying on contrived bias triggers, as often done in counterfactual input datasets, these benchmarks aim to match the distribution of human-written text to model more natural language use and potentially untargeted biases. In these completions, bias is typically quantified through distribution analysis of word co-occurrences, auxiliary classifiers measuring sentiment or toxicity, and lexicon-based methods that compare generated outputs against lists of biased or harmful terms. Associated metrics are detailed in Table 4 (Appendix A).

3 REASONING CAN AMPLIFY SOCIAL BIAS

In this section, we review recent studies showing how reasoning mechanisms in LLMs are inherently tied to social bias through the amplification of stereotype-aligned responses.

Shaikh et al. (2023) show that zero-shot prompt-induced CoT reasoning can reduce model accuracy (defined in Table 2) in ambiguous multiple-choice question-answering settings, increasing the likelihood of biased responses. The study evaluates CoT reasoning triggered by the phrase "*Let's think step by step*" against a standard prompt that directly asks for an answer, across GPT-3 (text-davinci-001 to 003) (Brown et al., 2020; Ouyang et al., 2022) and Flan-T5 models (Chung et al., 2024). The experimental setup includes the CrowS-Pairs, StereoSet, and BBQ benchmarks. The first two were reformulated as multiple-choice tasks with an explicit uncertainty option, treated as the correct response under ambiguous conditions. Results show that CoT reasoning consistently reduces accuracy in ambiguous contexts compared with direct prompting for GPT-3 models. Under the accuracy definition introduced in Table 2, this decrease indicates that when models are encouraged to reason step by step, they become less likely to acknowledge uncertainty and more likely to produce stereotype-consistent answers. This suggests that reasoning traces can function as a mechanism for rationalizing socially grounded assumptions rather than correcting them. For Flan-T5 models, the authors observe an inverse scaling effect: under CoT, accuracy on stereotype benchmarks decreases and then plateaus as model size increases, while performance on general reasoning benchmarks such as MMLU (Zhao et al., 2025), BBH (Suzgun et al., 2023) and MGSM (Shi et al., 2022) steadily improves with scale. This pattern suggests that improvements in general reasoning benchmarks do not necessarily translate into greater robustness to socially grounded ambiguity.

Complementing work on prompt-induced CoT, Wu et al. (2025) focus on models with native reasoning capabilities and show that these improve question-answering performance, but do not reduce social bias. Their results further indicate that biased reasoning is particularly associated with incorrect answers. To examine this, reasoning-oriented DeepSeek (8B, 32B) models (Hu et al., 2025) are compared with similarly sized instruction-tuned baselines, including Llama-3.1-8B-Instruct (Meta Llama Team, 2024) and Qwen2.5-32B-Instruct (Qwen Team, 2024), on the BBQ benchmark across both ambiguous and disambiguated contexts. In contrast to Shaikh et al. (2023), where accuracy is used as a proxy for bias, Wu et al. (2025) evaluate accuracy and bias separately. Accuracy is defined as the proportion of correct answers over the total number of questions, while bias measures the extent to which model responses align with stereotypical answers when a response other than "Unknown" is provided (see Table 2 for details). Results indicate that while reasoning models consistently achieve higher accuracy than their counterparts in both ambiguous and disambiguated settings, they also exhibit higher levels of bias. Consistent with Shaikh et al. (2023), increasing model size does not mitigate this effect. Beyond outcome-level evaluation, Wu et al. (2025) apply an LLM-as-a-judge method (Kumar et al., 2024), instantiated with GPT-4o, to analyze intermediate reasoning steps. Their analysis shows that while stereotypical reasoning can occasionally appear in correct answers, it is far more prevalent in the reasoning traces of incorrect responses, suggesting a strong link between biased reasoning and prediction errors.

Taken together, these results indicate that stronger reasoning capabilities do not necessarily translate into more reliable decisions. Instead, reasoning processes can amplify social bias, particularly in ambiguous contexts where stereotypes provide an intuitive shortcut.

4 REASONING CAN CONTROL SOCIAL BIAS

Since reasoning can amplify bias, recent work has begun to explore reasoning itself as a point of intervention for detecting and mitigating bias.

Building on their earlier findings that incorrect answers from reasoning models often contain biased intermediate steps (Section 3), Wu et al. (2025) investigate whether removing such steps could improve model accuracy on cases that were initially answered incorrectly. Their *Stereotype-free Reasoning Pattern* (SfRP) filters out biased intermediate steps and injects the remaining clean traces into the instruction-tuned models Llama-3.1-8B-Instruct and Qwen2.5-32B-Instruct, comparing outcomes against using the original unfiltered traces. Results show that guiding models with the filtered traces increases answer accuracy, suggesting that biased assumptions can shape intermediate reasoning steps and that intervening on these steps can reduce bias while improving prediction accuracy.

Wu et al. (2025) further propose *Answer Distribution as a Bias Proxy* (ADBP), a mitigation strategy that operates directly on model outputs without external judge models. The approach assumes that when bias affects the reasoning process, model predictions become unstable as intermediate steps are progressively revealed. To capture this behavior, ADBP prompts the model with increasingly longer prefixes of its reasoning chain while monitoring changes in its predictions. When inconsistencies are detected, the model is prompted to reconsider potential bias before finalizing its answer. Applied to DeepSeek (8B, 32B) on initially incorrect predictions, ADBP improves answer accuracy. This suggest that reasoning-time dynamics contain signals about the presence of bias, and that monitoring answer stability can help detect and mitigate biases that emerge during the reasoning process.

Gallegos et al. (2025) propose a zero-shot self-debiasing prompting method that leverages explanation-based reasoning to reduce bias. Instead of modifying or monitoring reasoning steps, the model is prompted to examine multiple-choice options and identify those based on invalid assumptions before producing a final answer. Tested on the ambiguous subset of BBQ using the bias metric introduced by Parrish et al. (2022) (Table 2) with GPT-3.5 Turbo, the method reduces bias scores by encouraging more uncertainty-aware responses. These findings indicate that explicitly reasoning about stereotypical assumptions can alter model predictions.

Yang et al. (2025) extend this perspective to the evaluation stage. They propose the *Reasoning-based Bias Detector* (RBD), a model-agnostic module that generates reasoning to identify and mitigate systematic biases in LLM-as-a-judge settings. Unlike prior approaches that filter, monitor, or prompt over reasoning traces, RBD relies on reasoning models fine-tuned to produce bias-aware explanations (e.g., variants of DeepSeek-R1-Distill, Qwen-7B, LLaMA-8B, and Qwen-14B). Given an input and an initial judgment, RBD predicts whether the decision is affected by evaluation biases such as verbosity, position, bandwagon, or sentiment bias. When bias is detected, the evaluator is re-prompted with a bias label and reasoning trace to encourage self-correction. Experiments across multiple evaluators and bias types show consistent improvements in evaluation accuracy, with larger RBD models yielding greater gains. Although this work targets evaluation-specific biases, it similarly highlights reasoning as a useful tool for detecting and correcting biased behavior.

Overall, these findings highlight that reasoning and bias are closely linked, rather than being independent processes. Moreover, focusing on their interaction provides an effective way to detect and mitigate undesired biased behaviors.

5 Conclusions

Current evaluation paradigms often treat logical reasoning in LLMs as a "cleanroom" process, assuming that reasoning operates independently from socially grounded knowledge. Evidence shows that reasoning mechanisms strongly interact with social priors, sometimes amplifying stereotypical conclusions in ambiguous contexts, but also offering opportunities for bias detection and mitigation. These findings highlight that reasoning and social bias are intertwined and jointly shape model behavior.

Our analysis suggests that reasoning evaluation can benefit from measuring robustness to socially contextualized premises. Integrating social bias benchmarks would allow to test whether models maintain logically grounded conclusions when demographic or cultural attributes vary. This interaction depends on the cultural and geographic priors embedded in training data: a model that appears unbiased in one context (e.g., BBQ) may fail under the social nuances of others (e.g., AfriStereo). This underscores the role of participatory and cross-cultural benchmarks that capture locally situated forms of bias, rather than relying solely on Western-centric settings. In this work, we provide a set of benchmarks and metrics as evaluation resources. However, developing metrics that specifically capture the interaction between reasoning and social bias remains an open challenge.

This survey highlights the importance of incorporating social bias benchmarks and metrics into reasoning evaluation pipelines. Integrating social robustness into the evaluation of reasoning systems is key to developing models that behave reliably in high-stakes, real-world applications.

6 ACKNOWLEDGMENTS

We would like to thank the Mozilla Foundation, which partially funded this work. We are also grateful to the HPC center CCAD of the Universidad Nacional de Córdoba (https://ccad.unc.edu.ar/). CCAD is part of SNCAD-MinCyT, Argentina. Although this work did not require computational resources, it builds on insights derived from our broader empirical research experience, enabled by CCAD. Finally, we are very grateful to the AI team in the Argentinean NGO Fundación Vía Libre for discussions and insights.

REFERENCES

Ali Alkhathlan, Faris Alahmadi, Faris Kateb, and Hend Al-Khalifa. Constructing and evaluating arabicstancex: a social media dataset for arabic stance detection. *Frontiers in Artificial Intelligence*, 2025. doi: 10.3389/frai.2025.1615800.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.151.

Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*. CEUR Workshop Proceedings, 2018. ISBN 978-88-31978-41-5.

Yann Le Beux, Oluchi Audu, Oche D. Ankeli, Dhananjay Balakrishnan, Melissah Weya, Marie D. Ralaiarinosy, and Ignatius Ezeani. Afristereo: A culturally grounded dataset for evaluating stereotypical bias in large language models, 2025.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.75.

Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-3002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/1155.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 2024.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement. In *Advances in Neural Information Processing Systems*, 2023.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. ISBN 9781450383097. doi: 10.1145/3442188.3445924.

Isabel O. Gallegos, Ryan Aponte, Ryan A. Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, Nedim Lipka, Deonna Owens, and Jiuxiang Gu. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.74.

Wenbin Hu, Haoran Li, Huihao Jing, Qi Hu, Ziqian Zeng, Sirui Han, Xu Heli, Tianshu Chu, Peizhao Hu, and Yangqiu Song. Context reasoner: Incentivizing reasoning capability for contextualized privacy and safety compliance via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.44.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.7.

Yue Huang, Qihui Zhang, Philip S. Y, and Lichao Sun. Trustgpt: A benchmark for trustworthy and responsible large language models, 2023.

Yufei Huang and Deyi Xiong. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, 2024.

Guido Ivetta, Marcos J Gomez, Sofía Martinelli, Pietro Palombini, M Emilia Echeveste, Nair Carolina Mazzeo, Beatriz Busaniche, and Luciana Benotti. HESEIA: A community-based dataset for evaluating social biases in large language models, co-designed in real school settings in Latin America. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1275.

Guido Ivetta, Pietro Palombini, Sofía Martinelli, Marcos J Gomez, M. María Echeveste, Sunipa Dev, Vinodkumar Prabhakaran, and Luciana Benotti. Adaptive data collection for latin-american community-sourced evaluation of stereotypes (laces), 2026.

Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.548.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 2024. doi: 10.1162/tacl_a_00661.

Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. Decoding biases: Automated methods and llm judges for gender bias detection in language models, 2024.

Hanmeng Liu, Yiran Ding, Zhizhang Fu, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Evaluating the logical reasoning abilities of large reasoning models, 2025.

Meta Llama Team. The Llama 3 herd of models, 2024.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, Nikita Nangia, Anaelia Ovalle, Giada Pistilli, Dragomir Radev, Beatrice Savoldi, Vipul Raheja, Jeremy Qin, Esther Ploeger, Arjun Subramonian, Kaustubh Dhole, Kaiser Sun, Amirbek Djanibekov, Jonibek Mansurov, Kayo Yin, Emilio Villa Cueva, Sagnik Mukherjee, Jerry Huang, Xudong Shen, Jay Gala, Hamdan Al-Ali, Tair Djanibekov, Nurdaulet Mukhituly, Shangrui Nie, Shanya Sharma, Karolina Stanczak, Eliza Szczechla, Tiago Timponi Torrent, Deepak Tunuguntla, Marcelo Viridiano, Oskar Van Der Wal, Adina Yakefu, Aurélie Névéol, Mike Zhang, Sydney Zink, and Zeerak Talat. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.600.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.416.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.154.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.583.

Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.191.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.976.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022. ISBN 9781713871088.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.165.

Qwen Team. Qwen2.5: A party of foundation models, 2024.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign

Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D' Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Naome Etori, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2024. doi: 10.52202/079017-0366.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.244.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824.

Xuyang Wu, Jinming Nian, Ting-Ruen Wei, Zhiqiang Tao, Hsin-Tai Wu, and Yi Fang. Does reasoning introduce bias? a study of social bias evaluation and mitigation in LLM reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1006.

Haoyan Yang, Runxue Bao, Cao Xiao, Jun Ma, Parminder Bhatia, Shangqian Gao, and Taha Kass-Hout. Any large language model can be a reliable judge: Debiasing with a reasoning-based bias detector, 2025.

Pardis Sadat Zahraei and Ehsaneddin Asgari. I am aligned, but with whom? mena values benchmark for evaluating cultural alignment and multilingual bias in llms, 2025.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2003.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.656.

A APPENDIX

Table 2: Bias metrics for the *Question-Answering* category as introduced in the taxonomy in Section 2, and under multiple choice settings.

Question-Answering (multiple choice)		
Metric	**Formula, variables & interpretation**	**Authors**
Accuracy (bias proxy)	$Acc = \frac{N_{\text{unk}}}{N}$ N_{unk}: number of *Unknown* answers N: number of questions Used in ambiguous QA settings with an uncertainty option. Lower values indicate a higher tendency to avoid the *Unknown* option, potentially relying on stereotypes.	Shaikh et al. (2023)
s_{AMB}	$s_{\text{AMB}} = (1 - Acc)\left[2\left(\frac{N_{\text{stereo}}}{N_{\text{non-unk}}}\right) - 1\right]$ N_{stereo}: number of stereotype-aligned non-*Unknown* answers $N_{\text{non-unk}}$: number of non-*Unknown* answers Acc: accuracy (proportion of instances in which the model correctly selects the *Unknown* option for ambiguous questions) Used in ambiguous QA settings with an uncertainty option. Higher values indicate stronger bias.	Parrish et al. (2022)
Bias_{amb}	$Bias_{\text{amb}} = \frac{N_{\text{non-stereo}}}{N_{\text{non-unk}}}$ $N_{\text{non-stereo}}$: number of answers avoiding the stereotype $N_{\text{non-unk}}$: number of non-*Unknown* answers Used in ambiguous QA settings with an uncertainty option. Lower values indicate less avoidance of stereotypical responses, reflecting stronger bias.	Wu et al. (2025)
Bias_{dis} [1]	$Bias_{\text{dis}} = 2\,\frac{N_{\text{stereo}}}{N_{\text{non-unk}}} - 1$ N_{stereo}: number of stereotype-aligned answers $N_{\text{non-unk}}$: number of non-*Unknown* answers Used in disambiguated QA settings with an uncertainty option. Higher values indicate stronger bias.	Wu et al. (2025); Parrish et al. (2022)

[1] Referred to as s_{DIS} in Parrish et al. (2022); both denote the same metric.

Table 3: Bias metrics for the *Counterfactual Inputs* category, as introduced in the taxonomy in Section 2.

Counterfactual Inputs		
Metric	**Formula, variables & interpretation**	**Authors**
CrowS-Pairs Score	$CPS(S) = \sum_{u \in U} \log P\big(u \mid U_{\setminus u},\, M;\, \theta\big)$ U: shared tokens between stereotypical and anti-stereotypical sentences M: tokens encoding the modified social attribute θ: model parameters Evaluates preference between counterfactual sentence pairs by estimating the likelihood of shared tokens U conditioned on modified attribute tokens M. Higher values indicate stronger bias.	Nangia et al. (2020)
Context Association Test	$\text{CAT}(S) = \frac{1}{\lvert M \rvert} \sum_{m \in M} \log P(m \mid U;\, \theta)$ U: sentence context M: candidate completions (stereotypical, anti-stereotypical, unrelated) θ: model parameters Evaluates counterfactual completions by estimating the likelihood of each attribute in M given a context U, and quantifies bias as the model's relative preference (the percentage of instances where the stereotypical option is preferred over the anti-stereotypical one).	Nadeem et al. (2021)

Table 4: Bias metrics for the *Sentence Completion* category, as introduced in the taxonomy in Section 2.

Sentence Completion		
Metric	**Formula, variables & interpretation**	**Authors**
Co-Occurrence Bias Score	$\text{CoBias}(w) = \log \frac{P(w\|A_i)}{P(w\|A_j)}$ w: target token A_i, A_j: sets of attribute words (e.g., feminine vs. masculine terms) Measures the association of a token with different attribute groups based on co-occurrence statistics in generated text. Values above zero indicate stronger association with A_i, below zero with A_j.	Bordia & Bowman (2019)
Counterfactual Sentiment Bias	$\text{CSB}(\hat{Y}) = W_1\big(P(c(\hat{Y}_i) \mid A = i),\ P(c(\hat{Y}_j) \mid A = j)\big)$ $\hat{Y}_i, \hat{Y}_j$: generated texts under counterfactual prompts A: protected attribute $c(\hat{Y})$: sentiment score from some classifier c $W_1(\cdot)$: Wasserstein-1 distance between distributions Measures differences in sentiment distributions across counterfactual groups. Higher values indicate stronger bias.	Huang et al. (2020)
HONEST	$\text{HONEST}(\hat{Y}) = \frac{\sum_{\hat{Y}_k \in \hat{Y}} \sum_{\hat{y} \in \hat{Y}_k} \mathbb{I}_{\text{HurtLex}}(\hat{y})}{\lvert\hat{Y}\rvert \cdot k}$ $\hat{Y}_k$: top-k generated completions per prompt $\hat{y}$: individual generated token $\mathbb{I}_{\text{HurtLex}}(\hat{y})$: indicator for words in the HurtLex lexicon (Bassignana et al., 2018) k: number of completions per prompt Measures the proportion of generated words that appear in a lexicon of hurtful terms. Higher values indicate more harmful or biased content.	Nozza et al. (2021)