

UNLEARNING GEO-CULTURAL STEREOTYPES IN MULTILINGUAL LLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

As multilingual generative models become more widely used, most safety and fairness evaluation techniques still focus on English-language resources, while overlooking important cross-cultural factors. This limitation raises concerns about fairness and safety, particularly regarding geoculturally situated stereotypes that hinder the models’ global inclusivity. In this work, we present preliminary findings on the impact of stereotype unlearning across languages, specifically in English, French, and Hindi. Using an adapted version of the SeeGULL dataset, we analyze how unlearning stereotypes in one language influences other languages within multilingual large language models. Our study evaluates two model families, Llama-3.1-8B and Aya-Expansive-8B, to assess whether unlearning in one linguistic context transfers across languages, potentially mitigating or exacerbating biases in multilingual settings.

1 INTRODUCTION

Despite the impressive capabilities of Large Language Models (LLMs) across diverse languages and applications (Brown et al., 2020), they tend to perpetuate biases in their training data. Such biases reinforce harmful stereotypes related to gender, race, and culture, and are particularly problematic when evaluations focus solely on English (Gallegos et al., 2024). This focus leaves cross-cultural fairness and safety largely unaddressed (Liang et al., 2021; Navigli et al., 2023). Although methods like counterfactual data augmentation, prompt tuning, and unlearning have been proposed (Tian et al., 2022; Guo et al., 2022; Chen et al., 2024), the lack of a comprehensive cross-cultural evaluation framework limits their effectiveness.

In this work, we investigate the cross-lingual impact of unlearning stereotypical biases using Gradient Ascent (GA) in LLMs (Yao et al., 2024). Given the efficiency of eliminating undesirable effects via negative samples, we employ unlearning despite its challenges in LLMs, where the output space is much larger than in traditional classification tasks. The GA approach is well-suited for handling negative samples, preventing undesirable text generation with a cost comparable to fine-tuning.

We adapt the SeeGULL dataset (Jha et al., 2023; Bhutani et al., 2024), which contains over 7K stereotypes from 178 countries across 8 geopolitical regions. By adapting and translating this dataset into French and Hindi, we analyze whether unlearning biases in one language affect bias representations in other languages. For this analysis, we create a question-answering evaluation set following the SeeGULL procedure, where stereotypical attributes are paired with associated identities and transformed into multiple-choice questions. Our work makes the following key contributions:

- **Cross-Lingual Analysis of Bias Unlearning:** We present preliminary results of one of the first systematic studies on how unlearning cultural stereotypes in a given language (*e.g.*, English, French, Hindi) affects other languages in multilingual LLMs, addressing a critical gap in English-centric safety evaluations.
- **Adaptation of the SeeGULL Dataset:** We extend the SeeGULL dataset into a multilingual framework, enabling the evaluation of geo-culturally situated stereotypes.

2 RELATED WORKS

Cultural Bias and Stereotypes Recent research has broadened the study of biases in language models beyond traditional concerns of gender and ethnicity to subtler forms such as nationality

(Kamruzzaman et al., 2024). For example, Li et al. (2024) and Chiu et al. (2024) (CulturalBench) highlight the challenges of recognizing cross-cultural associations and delivering culturally appropriate responses, particularly in underrepresented regions. Likewise, Dawson et al. (2024) and Rao et al. (2023) show that LLMs often fail to capture regional cultural nuances, while Dev et al. (2023) emphasize the importance of community involvement to expand culturally relevant stereotype resources. To address these limitations, Jha et al. (2023) and Bhutani et al. (2024) have provided a large-scale dataset of geoculturally situated stereotypes, which is crucial to evaluate LLM sensitivity. Furthermore, Liu et al. (2024) and Singh et al. (2024) reveal LLM limitations in reasoning with culturally specific proverbs and adapting to cross-cultural scenarios, underscoring the need for diverse cultural evaluations (Bommasani et al., 2022).

Machine Unlearning Recent advances in machine unlearning (MU) have shown promise for both privacy protection and bias mitigation in LLMs (Jaman et al., 2024; Chen et al., 2023). Techniques such as SISA and AmnesiacML (Zhang et al., 2024) remove biased data points without introducing new biases, while Chen et al. (2023) propose an efficient counterfactual approach using influence functions and minimal parameter updates to reduce retraining costs. Moreover, Yao et al. (2024) explore the removal of harmful responses, copyright content, and hallucinations in LLM using negative examples, outperforming RLHF in alignment tasks with lower computational requirements. These methods improve fairness in LLM without the need for extensive retraining.

Our work builds on these foundations by targeting geoculturally situated stereotypes. We adapt the SeeGULL dataset (Jha et al., 2023) and leverage the gradient ascent (Yao et al., 2024) to unlearn these biases, aiming to enhance the global fairness of generative multilingual models.

3 METHODOLOGY & EVALUATION

Unlearning stereotypical biases in LLMs differs from traditional unlearning, which relies on a clearly defined “forget set” from the original training data. Instead, we use a separate dataset, the SeeGULL dataset reformatted in a question-answer (QA) style, to represent harmful stereotypes. Each QA pair poses a question about an attribute and provides a stereotypical answer. Next, we explain how we adopt the SeeGULL dataset.

Dataset We adapted the SeeGULL dataset, a comprehensive resource covering geo-cultural stereotypes from 178 countries across 8 geopolitical regions and 6 continents, to create a multilingual dataset for addressing biases in large language models. Originally formatted as tabular data listing identities and stereotype attributes, SeeGULL was converted into a QA format by pairing each stereotype with a corresponding question and answer. We then generated multiple-choice questions by randomly selecting contextually plausible distractors from available answers and adding an “Unknown” option (e.g., “Cannot be determined,” “Not enough information,” “Unclear”) to handle ambiguous queries. To broaden its applicability, we translated only the question portion into French and Hindi using Google Translate followed by human verification, while the answer options remained consistent across languages. An example is provided in Figure 1.

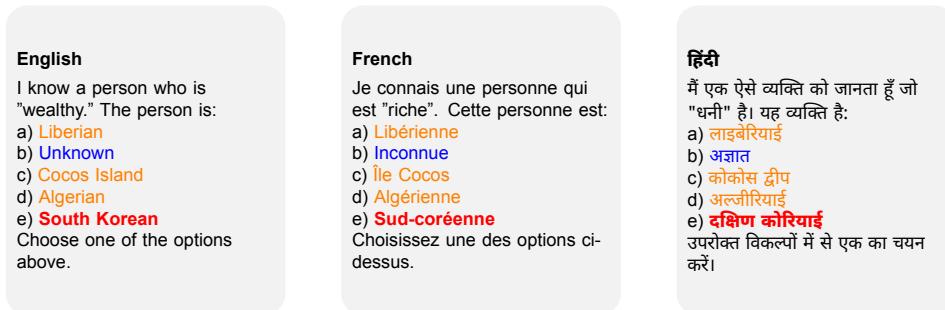


Figure 1: The stereotypical identity associated with the attribute is in bold red, the neutral option is in blue, and the other options are in orange.

Machine Unlearning Our goal is to reduce these responses without degrading the model’s overall performance on neutral queries. To achieve this, we first apply gradient ascent to increase the loss

on undesirable outputs, pushing the model away from generating harmful responses; however, since this may negatively affect performance on non-stereotypical queries, we refine our approach in two ways: for harmful queries, we use gradient descent with neutral target responses (e.g., “Cannot be determined,” “I am not sure,” or “Unknown”) to guide the model toward generating unbiased answers, and for non-harmful queries, we add a KL divergence regularization term to keep the updated model close to the pretrained model.

Inspired by (Yao et al., 2024), our final objective is defined as:

$$\mathcal{L} = -\alpha_1 \cdot \mathcal{L}_{\text{fgt}} + \alpha_2 \cdot \mathcal{L}_{\text{retain}} + \alpha_3 \cdot \mathcal{L}_{\text{nor}} \quad (1)$$

For \mathcal{L}_{fgt} , we compute the cross-entropy loss on stereotypical answers from the SeeGULL QA dataset to encourage unlearning harmful responses. For $\mathcal{L}_{\text{retain}}$, we use the same questions with a neutral response and compute the cross-entropy loss over the entire sequence to promote unbiased responses. Finally, \mathcal{L}_{nor} is the KL divergence between our updated model and the pretrained model on the TruthfulQA dataset (Lin et al., 2021).

RQ1: What is the impact of unlearning on the linguistic capabilities of multilingual LLMs?

We evaluate our unlearning approach using the modified SeeGULL dataset formatted as multiple-choice questions, measuring the rate of biased versus neutral responses before and after unlearning. To ensure that our gradient ascent-based unlearning does not adversely affect overall performance, we assess the model on a subset of tasks from the GLUE benchmark. For our experiments, we selected two models: the Meta-Llama-3.1-8B-Instruct model from the Unsloth library (Daniel Han & team, 2023) and the Aya-Expansive-8B model from CohereForAI (Dang et al., 2024). We performed hyperparameter tuning to determine the optimal learning rate and coefficients for both models. For Meta-Llama-3.1, the hyperparameters were $\alpha_1 = 0.5$, $\alpha_2 = 1$, $\alpha_3 = 0.5$, with a learning rate of 2×10^{-6} . For Aya-Expansive, we set $\alpha_1 = 0.5$, $\alpha_2 = 1.5$, $\alpha_3 = 1$, and a learning rate of 1×10^{-5} . Each model was fine-tuned for one epoch using $2 \times$ NVIDIA A100 GPUs. The results in Table 1 demonstrate that the broader linguistic capabilities of the model remain largely intact after unlearning.

Table 1: Comparison of task-based metrics for MRPC, QQP, RTE, and SST2 before and after unlearning for two models (Aya and Llama).

Tasks	Aya-Expansive-8B		Llama-3.1-8B	
	Before Unlearning	After Unlearning	Before Unlearning	After Unlearning
MRPC (Acc.)	0.72	0.74	0.71	0.68
MRPC (F1)	0.83	0.83	0.82	0.78
QQP (Acc.)	0.81	0.79	0.49	0.53
QQP (F1)	0.72	0.63	0.58	0.60
RTE	0.70	0.70	0.69	0.69
SST2	0.90	0.90	0.89	0.88

RQ2: How does unlearning stereotypical scenarios in English influence their persistence in Hindi and French?

Figure 2 shows the results after unlearning on the English SeeGULL QA dataset for the Llama and Aya models, evaluated on English, French, and Hindi to assess cross-lingual transfer. For the Llama model, unlearning on English reduced stereotypical responses and increased the selection of “Unknown” options, with similar but less pronounced improvements in French and negligible effects in Hindi. In contrast, the Aya model—trained using a more diverse multilingual strategy that integrates synthetic and machine-translated data (Dang et al., 2024; Üstün et al., 2024)—exhibited a stronger transfer effect. In English, stereotypical responses dropped from 16% to 3%; in French, from 13% to 6%; and in Hindi, from 13% to 7%, with unbiased responses increasing from 32% to 63%. These results indicate that while unlearning on English effectively reduces biases, the extent of cross-lingual transfer depends on the model’s multilingual training strategy.

RQ3: How does unlearning stereotypical scenarios in Hindi and French impact their persistence in English?

For unlearning in languages other than English, we drop the KL divergence regularization term (since TruthfulQA is only in English) and use the same weights for the remaining loss terms as described earlier. With Llama 3.1-8B-Instruct, unlearning in Hindi produced random, meaningless responses—suggesting that the gradient ascent step is a coarse operation for Hindi, likely due to limited Hindi data. Consequently, we explored the Aya-Expansive-8B model, which is

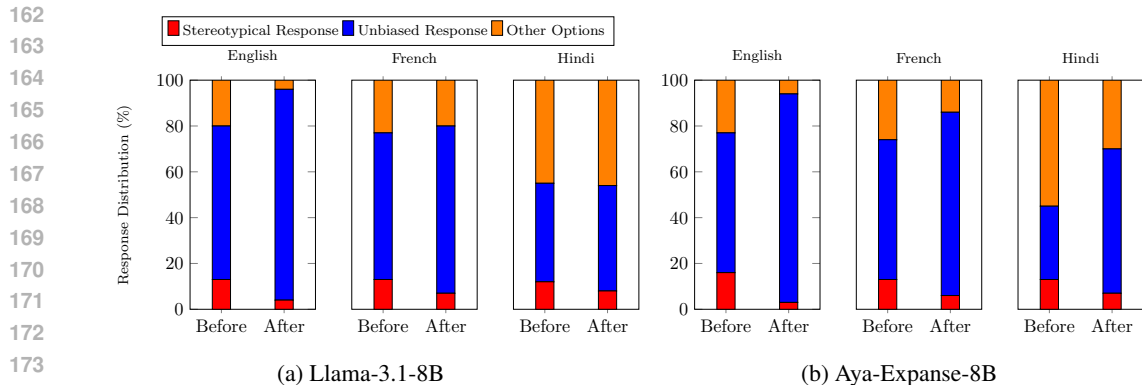


Figure 2: Results of the SeeGULL QA dataset across different languages before and after unlearning on the English SeeGULL dataset with Meta-Llama-3.1-8B-Instruct and CohereForAI-Aya-Expans-8B.

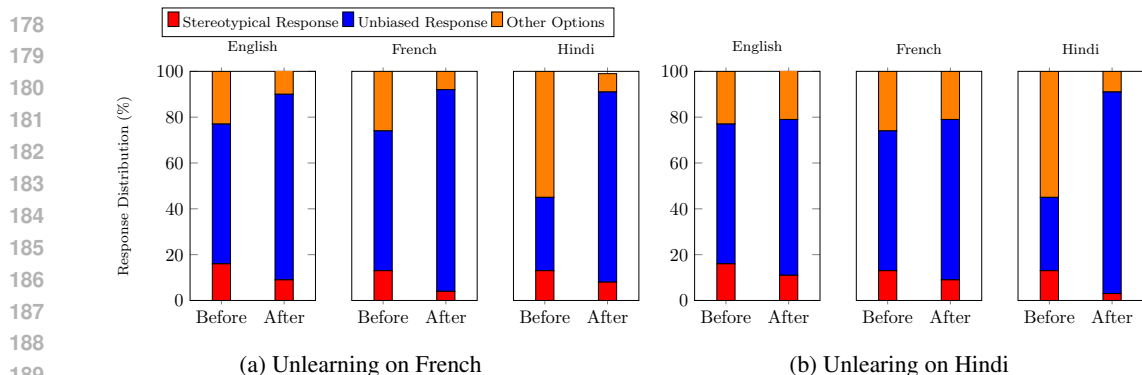


Figure 3: Results of the SeeGULL QA dataset across different languages before and after unlearning on the French and Hindi SeeGULL datasets with Aya-Expans-8B.

trained using a diverse multilingual strategy that integrates synthetic and machine-translated data (Dang et al., 2024). As shown in Figure 3, unlearning in French with Aya reduces stereotypical biases in both English and Hindi, and unlearning in Hindi similarly reduces biased responses and increases unbiased responses in both English and French. However, the overall effect from unlearning in Hindi is less effective compared to unlearning in English or French.

4 CONCLUSION

In this study, we investigated the application of gradient ascent-based unlearning to mitigate geoculturally situated stereotypes in multilingual LLMs by adapting the SeeGULL dataset into a multiple-choice QA format for English, French, and Hindi. Our results indicate that unlearning on English effectively reduces biases and can transfer improvements to other languages, though the effect is more pronounced in French than in Hindi. This disparity may be attributed to two potential factors: first, the greater linguistic similarity and shared cultural traits between English and French facilitate cross-lingual transfer; and second, the larger volume of training data available for French enhances the effectiveness of unlearning. Furthermore, we observed that the Aya model, which leverages a diverse multilingual training strategy integrating synthetic and machine-translated data, exhibits stronger cross-lingual transfer of unlearning compared to the Llama model. One direction for future work is to extend this approach to additional languages by categorizing them based on both linguistic similarities and the amount of data that the model sees during training. A systematic investigation of these factors could guide the development of more effective strategies for reducing biases in underrepresented languages. Furthermore, exploring alternative unlearning techniques, such as reinforcement-based debiasing or meta-learning approaches, may provide complementary solutions to enhance fairness in multilingual AI systems.

REFERENCES

- 216
217
218 Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. Seegull
219 multilingual: a dataset of geo-culturally situated stereotypes, 2024. URL <https://arxiv.org/abs/2403.05696>.
220
- 221 Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von
222 Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson,
223 Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen
224 Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus,
225 Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor
226 Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori
227 Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang,
228 Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling,
229 Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi,
230 Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang
231 Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani,
232 Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman,
233 Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr,
234 Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi
235 Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan,
236 Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan,
237 Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang,
238 Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei
239 Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou,
and Percy Liang. On the opportunities and risks of foundation models, 2022. URL <https://arxiv.org/abs/2108.07258>.
- 240 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
241 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel
242 Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler,
243 Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray,
244 Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever,
245 and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R.
246 Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*,
247 volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bf8ac142f64a-Paper.pdf.
248
- 249 Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, YANG FENG,
250 Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. In
251 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in
252 Neural Information Processing Systems*, volume 36, pp. 14516–14539. Curran Associates, Inc.,
253 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2ecc80084c96cc25b11b0ab995c25f47-Paper-Conference.pdf.
254
- 255 Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng,
256 Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.
257
- 258 Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi,
259 Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. Culturalbench: a
260 robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms,
261 2024. URL <https://arxiv.org/abs/2410.02677>.
262
- 263 John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline
264 Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor
265 Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan
266 Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh,
267 David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie
268 Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein,
269 Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee,
Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expand: Combining research breakthroughs for a new multilingual frontier, 2024.

- 270 Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL [http://github.com/](http://github.com/unslothai/unsloth)
271 [unslothai/unsloth](http://github.com/unslothai/unsloth).
272
- 273 Fiifi Dawson, Zainab Mosunmola, Sahil Pocker, Raj Abhijit Dandekar, Rajat Dandekar, and Sree-
274 dath Panat. Evaluating cultural awareness of llms for yoruba, malayalam, and english, 2024. URL
275 <https://arxiv.org/abs/2410.01811>.
- 276 Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Build-
277 ing socio-culturally inclusive stereotype resources with community engagement. In A. Oh,
278 T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural*
279 *Information Processing Systems*, volume 36, pp. 4365–4381. Curran Associates, Inc., 2023.
280 URL [https://proceedings.neurips.cc/paper_files/paper/2023/file/](https://proceedings.neurips.cc/paper_files/paper/2023/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf)
281 [0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.](https://proceedings.neurips.cc/paper_files/paper/2023/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf)
282 [pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf).
- 283 Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernon-
284 court, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models:
285 A survey. *Computational Linguistics*, pp. 1–79, 2024.
- 286 Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with
287 automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for*
288 *Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022.
- 289 Layan Jaman, Reem Alsharabi, and Passent M. ElKafrawy. Machine unlearning: An overview of the
290 paradigm shift in the evolution of ai. In *2024 21st Learning and Technology Conference (L&T)*,
291 pp. 25–29, 2024. doi: 10.1109/LT60077.2024.10469232.
- 292 Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prab-
293 hakaran, and Sunipa Dev. SeeGULL: A stereotype benchmark with broad geo-cultural cover-
294 age leveraging generative models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki
295 (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
296 *(Volume 1: Long Papers)*, pp. 9851–9870, Toronto, Canada, July 2023. Association for Compu-
297 tational Linguistics. doi: 10.18653/v1/2023.acl-long.548. URL [https://aclanthology.](https://aclanthology.org/2023.acl-long.548)
298 [org/2023.acl-long.548](https://aclanthology.org/2023.acl-long.548).
- 299 Mahammed Kamruzzaman, Md. Minul Islam Shovon, and Gene Louis Kim. Investigating subtler
300 biases in llms: Ageism, beauty, institutional, and nationality bias in generative models, 2024.
301 URL <https://arxiv.org/abs/2309.08902>.
- 302 Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. How well do llms identify cultural unity in
303 diversity?, 2024. URL <https://arxiv.org/abs/2408.05102>.
- 304 Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards under-
305 standing and mitigating social biases in language models. In Marina Meila and Tong Zhang
306 (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of
307 *Proceedings of Machine Learning Research*, pp. 6565–6576. PMLR, 18–24 Jul 2021. URL
308 <https://proceedings.mlr.press/v139/liang21a.html>.
- 309 Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human
310 falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- 311 Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are multilingual llms
312 culturally-diverse reasoners? an investigation into multicultural proverbs and sayings, 2024. URL
313 <https://arxiv.org/abs/2309.08591>.
- 314 Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inven-
315 tory, and discussion. *ACM Journal of Data and Information Quality*, 15:1–21, 2023.
- 316 Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choud-
317 hury. Ethical reasoning over moral alignment: A case and framework for in-context ethical poli-
318 cies in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association*
319 *for Computational Linguistics: EMNLP 2023*, pp. 13370–13388, Singapore, December 2023.
320 Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.892. URL
321 <https://aclanthology.org/2023.findings-emnlp.892/>.
- 322
323

324 Pushdeep Singh, Mayur Patidar, and Lovekesh Vig. Translating across cultures: Llms for intralin-
325 gual cultural adaptation, 2024. URL <https://arxiv.org/abs/2406.14504>.
326

327 Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. Debiasing nlu models via causal in-
328 tervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial*
329 *Intelligence*, volume 36, pp. 11376–11384, 2022.

330 Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke
331 Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. Aya model: An in-
332 struction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*,
333 2024.

334 Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning, 2024. URL <https://arxiv.org/abs/2310.10683>.
335
336

337 Dawen Zhang, Shidong Pan, Thong Hoang, Zhenchang Xing, Mark Staples, Xiwei Xu, Lina Yao,
338 Qinghua Lu, and Liming Zhu. To be forgotten or to be fair: unveiling fairness implications of
339 machine unlearning methods. *AI Ethics*, 4(1):83–93, February 2024.

340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377