Bias Analysis and Mitigation through Protected Attribute Detection and Regard Classification

Takuma Udagawa[1], **Yang Zhao**[1], **Hiroshi Kanayama**[1], **Bishwaranjan Bhattacharjee**[2]
[1]IBM Research - Tokyo [2]IBM T. J. Watson Research Center
{takuma.udagawa,yangzhao}@ibm.com hkana@jp.ibm.com bhatta@us.ibm.com

Abstract

Large language models (LLMs) acquire general linguistic knowledge from massive-scale pretraining. However, pretraining data mainly comprised of web-crawled texts contain undesirable social biases which can be perpetuated or even amplified by LLMs. In this study, we propose an efficient yet effective annotation pipeline to investigate social biases in the pretraining corpora. Our pipeline consists of *protected attribute detection* to identify diverse demographics, followed by *regard classification* to analyze the language polarity towards each attribute. Through our experiments, we demonstrate the effect of our bias analysis and mitigation measures, focusing on Common Crawl as the most representative pretraining corpus.

Warning: This paper contains examples of social biases that may be harmful or offensive.

1 Introduction

Recent years have witnessed a remarkable progress in the development and adoption of large language model (LLM) technology (Achiam et al., 2023; Dubey et al., 2024). Generally, LLMs first undergo large-scale pretraining to acquire general linguistic knowledge, followed by post-training to align them with human goals and preferences (Ouyang et al., 2022). The majority of LLM's knowledge is acquired in the pretraining stage, and post-training is conceived to mainly change the style of LLMs to serve as interactive, open-domain AI assistants (Zhou et al., 2024; Lin et al., 2024).

However, pretraining data comprised of web-crawled texts often contain undesirable biases, such as associating Muslim people with terrorism and extremism (Chowdhery et al., 2023). Consequently, LLMs tend to inherit the biases and produce harmful judgements (Sheng et al., 2021; Schramowski et al., 2022), raising significant risks in terms of safety and fairness. Despite the severity of this issue, it remains extremely difficult to audit and alleviate such dataset-level biases due to the ever-increasing size of pretraining corpora and the open-ended, complex nature of social biases.

In this study, we make a first step towards analyzing and mitigating problematic social biases in massive-scale text data. Specifically, we propose a scalable annotation pipeline consisting of two steps. First, we conduct *protected attribute detection* to identify diverse demographics, such as nationality, religion, disability, etc. To reduce false positive detections, we combine keyword matching and word sense disambiguation (Huang et al., 2019). Second, we apply *regard classification* to analyze the language polarity towards each attribute into positive, neutral, or negative (Sheng et al., 2019). Our overall pipeline is illustrated in Figure 1.

In our experiments, we apply the pipeline to a subset of Common Crawl, the most widely used corpus for LLM pretraining. For bias analysis, we refine the frequency-based word association analysis (Bordia and Bowman, 2019) to take into account regard information. Qualitative and quantitative results show that our approach more accurately captures commonly held stereotypes (Jha et al., 2023). For bias mitigation, we reveal that the regard distributions among protected attributes can be noticeably imbalanced, and adjusting regard distributions (e.g. downsampling negative regard texts) can be a promising approach to mitigate offensive social stereotypes, such as "*white*" people being overly described as "*racists*" or "*supremacists*".

While several challenges still remain, we expect our approach to be a crucial step towards improving the fairness of LLM pretraining datasets.

2 Methods

2.1 Protected Attribute Detection

To efficiently detect protected attributes in massive pretraining data, existing work typically relies on keyword matching (Chowdhery et al., 2023; Esiobu

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16–25

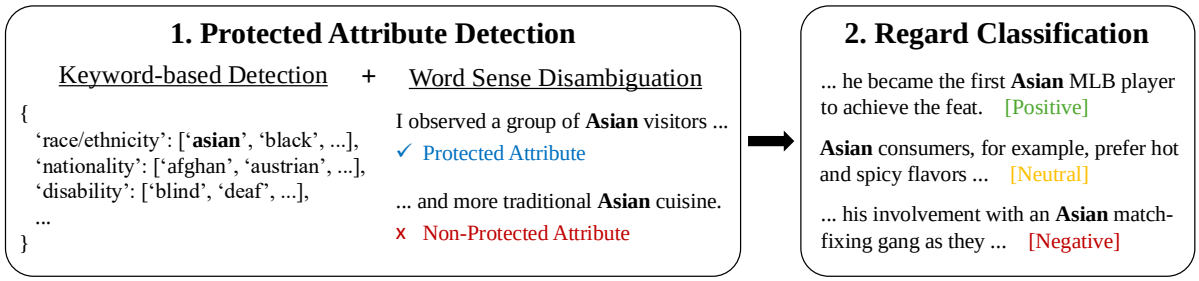

Figure 1: An overview of our annotation pipeline consisting of (1) *protected attribute detection* to identify diverse demographics and (2) *regard classification* to analyze the language polarity towards each detected attribute.

Input Format: [BOS] `{Text}`[SEP] `{Keyword}`; `{Gloss}`[EOS]	Prediction
[BOS] I observed a group of **Asian** visitors ... [SEP] **asian**; a person of asian race/ethnicity [EOS]	✓ Protected Attribute
[BOS] ... and more traditional **Asian** cuisine. [SEP] **asian**; a person of asian race/ethnicity [EOS]	× Non-Protected Attribute

Table 1: Word sense disambiguation in the style of Gloss-BERT. Regarding the input format, `{Text}` indicates the text containing the keyword, `{Keyword}` the target keyword, and `{Gloss}` the keyword's defined usage (gloss).

et al., 2023). Following this line, we defined a taxonomy of protected attributes covering 10 diverse classes containing a total of 97 keywords. For instance, "*asian*" is a keyword belonging to the class of *race/ethnicity* (cf. Figure 1). We include more details of our taxonomy in Appendix A.1.

However, keywords representing protected attributes are often polysemous and yield many false positives. For instance, "*blind*" can indicate a person's disability (visual impairment) or used in a totally different sense (e.g. *blind date*). For reliable bias measurement, we must accurately recognize and compare the relevant *human* attributes.[1]

To efficiently reduce false positives, we conduct word sense disambiguation (WSD) in the style of Gloss-BERT (Huang et al., 2019). In this approach, WSD is framed as a binary classification of whether a keyword in a text is used following the sense described in a gloss. We adopt this framework to predict whether a keyword is used to indicate a defined protected human attribute (Table 1).

While traditional WSD relies on glosses defined in WordNet (Miller, 1995), this is not suitable for our purpose, e.g. since there are no sense inventories which distinguish the usages of "*asian*" in Table 1. Therefore, for each of the 97 keywords, we crafted simple yet sufficient glosses describing the protected attributes.[2] We show several illustrative examples in Table 3 (Appendix A.1).

To develop our WSD model, we used an existing LLM to synthesize high-quality training dataset. Specifically, for each keyword, we randomly extracted 1K sentences from Common Crawl which contain the keyword. Then, we leveraged Mixtral-8x22B-Instruct[3] (Jiang et al., 2024) to collect judgements on whether the keyword indicates the defined protected attribute using the prompt in Table 4. Based on this dataset, we trained a light-weight WSD model from RoBERTa $_{\text{BASE}}$ (Liu, 2019) for scalable inference, which we refer to as Gloss-RoBERTa in the rest of this paper.

2.2 Regard Classification

Regard classification is widely used to evaluate the bias of LLM generated texts (Sheng et al., 2019; Dhamala et al., 2021). In this study, we aim to apply this technique to pretraining data analysis.

Unfortunately, existing regard classifier (Sheng et al., 2019) is only trained on templated texts (e.g. "*XYZ was regarded as ...*") which makes it difficult to be applied on realistic texts with diverse syntactic structures. Therefore, similar to our WSD (§2.1), we generated new training data for regard classification based on an existing LLM.

Specifically, for each keyword, we used 50K sentences from Common Crawl containing the keyword after removing false positives with Gloss-RoBERTa. Then, we used Mixtral-8x7B-Instruct[4] to annotate the regard towards each protected at-

[1] Blodgett et al. (2021) also emphasize the importance of comparing the commesurables, e.g. avoiding conflation of stereotypes about Norwegian *people* and *salmon*.

[2] It is worth noting that our glosses are NOT intended to define strict word senses but to practically distinguish whether the keyword indicates a protected *human* attribute.

[3] https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1

[4] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Protected Attribute	Frequency Bias (eq. (1))	Frequency+Regard Bias (eq. (2))		
		r = Positive	r = Negative	r = Neutral
arab	arab, palestinian, israeli, syrian, israel, iraqi, arabs, lebanon, iraq, lebanese, egypt, egyptian, ...	idol, mars, astronaut, blasted, generosity, chairperson, orchestra, poets, praised, forbes, hailed, hospitality, ..	terrorist, assaults, wounded, bomb, destroy, destruction, towers, terror, attacks, civilians, destroyed, ...	denomination, consult, mingle, filter, traded, tagged, assessing, ...
black	bipoc, starbucks, unarmed, abrams, brutality, freddie, custody, black, systemic, fatally, ...	vogue, essence, untold, uplifting, oprah, creatives, hidden, salute, amplify, pulitzer, congresswoman, ...	fatally, cops, cop, breathe, shot, gun, tear, misconduct, surfaced, mentally, fired, killing, falsely, ...	under-represented, complexion, applications, disabilities, disabled, ...
white	supremacist, collar, privileged, blond, settlers, privilege, evangelicals, privileges, white, ...	guy, teammates, blues, educated, refusal, dude, afforded, hunters, kid, parks, priviledge, loving, ...	supremacist, racists, mob, breathe, roof, pleaded, raped, brutally, angry, guilty, murdered, resentment, ...	makeup, races, followed, weighs, pounds, inches, weighing, borough, ...

Table 2: Results of our bias analyses for class A = *race/ethnicity*. Words $w \in V$ are sorted in descending order based on the frequency bias score (eq. (1)) and frequency+regard bias score (eq. (2)).

tribute using the prompt in Table 4 (Appendix A.2). This way, we synthesized a diverse and realistic dataset for regard classification, which we used to train a light-weight regard classifier from RoBERTa$_{\text{BASE}}$ for efficient inference.

2.3 Annotation Agreement

Regarding WSD for protected attribute detection (§2.1), we computed the annotation agreement between Mixtral-8x22B and Gloss-RoBERTa based on Cohen's kappa. On average across all attributes, we observed a moderate agreement of around 0.59. Additionally, for a small set of attributes, we conducted manual annotation and confirmed that agreement (1) among human annotators and (2) between humans and these models are also at similar levels (0.56 to 0.70). Therefore, our models can disambiguate protected attributes at near human-like levels (see Table 5 for example predictions).

As for regard classification, we prepared a high-quality test set double-checked with both Mixtral-8x7B and 8x22B for consistency. Based on our RoBERTa-based classifier, we achieved a decent F1 performance of 0.91 on micro and 0.82 on macro average across all attributes. This verifies that we could successfully distill the regard judgements of the strong Mixtral teacher into a much more light-weight classifier (cf. Table 6).

3 Experiments

In our experiments, we apply our bias analysis and mitigation measures on a subset of Common Crawl (CC), a widely used LLM pretraining corpora.

As a preparation, we first sampled over 3M English documents from CC and extracted sentences of appropriate length.[5] Then, we conducted protected attribute detection (§2.1) and regard classification (§2.2) on these sentences. In the following experiments, we used up to 100K sentences per attribute to demonstrate the effect of our bias analysis (§3.1) and mitigation (§3.2) techniques.

3.1 Bias Analysis

First, we apply the basic word association analysis following Bordia and Bowman (2019). Specifically, let $a \in A$ denote a protected attribute in class A and $w \in V$ denote a word in vocabulary V. Then, we can compute the *frequency bias* of w against a based on the following score:

$$\frac{p(w|a)}{[\![\, p(w|a)]\!]_{a \in A}} \tag{1}$$

Here, $p(w|a)$ denotes the probability of w occurring in a sentence containing a, and the denominator is its average for $a \in A$. A higher score of eq. (1) indicates that w is more likely to co-occur with a compared to other protected attributes in A.

In Table 2, we show (partial) results of bias analysis for A = *race/ethnicity*, where words are sorted in descending order based on eq. (1). Due to the limitation of space, we provide further experimental details and results in Appendix B.

Unfortunately, it is often difficult to obtain useful insights on social stereotypes from frequency bias only. For instance, words co-occurring with "*arab*" are mostly proper nouns (e.g. "*palestinian*") which is self-evident and irrelevant to stereotypes. Also, we can confirm words like "*supremacist*" co-occur with "*white*" but cannot tell whether they are used in a negative or offensive manner.

To address this issue, we propose a novel bias score reflecting both frequency and regard information. Specifically, let $r \in R = \{$Positive, Negative, Neutral$\}$ denote the regard label of a in the sentence. Then, for each r we compute the following

[5] Specifically, more than 16 tokens and less than 128 tokens based on the NLTK tokenizer (Bird, 2006).

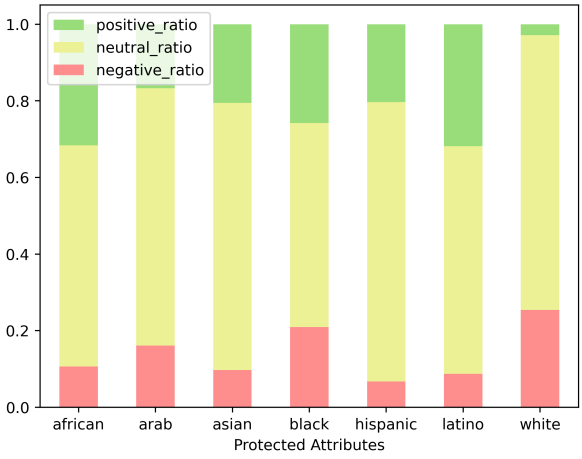

Figure 2: Regard distributions for A = *race/ethnicity*.

score:

$$\min\Big(\frac{p(w|a)}{[\![\, p(w|a)]\!]_{a\in A}}, \frac{p(r|w,a)}{[\![\, p(r|w,a)]\!]_{r\in R}}\Big) \quad (2)$$

Here, the left term represents *frequency bias* (eq. (1)) and the right term *regard bias* of when a word w and attribute a co-occur.[6] For instance, if w = *supremacist* and a = *white* tend to co-occur with negative regard, the score increases for r = *negative*. This way, we can analyze not only whether w and a co-occur, but whether they co-occur accompanying specific regard in the pretraining corpora.

Table 2 includes the results of our analysis based on this frequency+regard bias score. By leveraging regard information, we can more easily identify social stereotypes such as "*generosity*" (positive) and "*terrorist*" (negative) towards "*arab*" people. Also, we can verify that words like "*supremacist*" and "*racists*" co-occur with negative regard towards "*white*" people, which can be problematic if over-represented in the pretraining data.

More quantitatively, we show that our analysis better captures commonly held stereotypes listed in SeeGULL (Jha et al., 2023), which is discussed in Appendix B. Overall, our approach is a simple yet effective technique to investigate social biases toward a variety of protected attributes.

3.2 Bias Mitigation

Arguably, a fair pretraining dataset should maintain balanced regard distributions towards protected attributes. Unfortunately, based on our regard estimation, this is not ensured in CC: for instance, *white* people are much more likely to be described with negative regard in the class of A = *race/ethnicity*, reaching over 20% in ratio (Figure 2).

In this study, we verify that problematic word association can be alleviated by properly balancing the regard distributions. Specifically, we ensure the negative regard ratio to be at most 1% for all protected attributes by downsampling negative regard sentences. Then, we compute the relative reduction in the conditional probability $p(w|a)$, where w is a negatively biased word towards a.[7]

Through this intervention, we confirmed $p(w|a)$ can be dramatically reduced, e.g. down to 19% and 18% for w = "*supremacist*" and "*racists*" toward a = "*white*", respectively. Similarly, $p(w|a)$ can be diminished to 46% and 26% for w = "*terrorist*" and "*assaults*" toward a = "*arab*", while neutral or positive word associations remain unchanged or slightly increase, e.g. up to 112% for w = "*generosity*" and "*hospitality*".

It is worth noting that our approach can focus on problematic rather than benign word associations owing to the regard annotation r. For instance, the former sentence will be retained while the latter is subject to downsampling, although "*white*" and "*racist*" co-occur in both sentences:

- *Not every white person is racist and not every black person is a criminal.* (r = neutral)
- *Report says he was attacked by three white males at about 3:25 a.m., who made racist comments as they assaulted him.* (r = negative)

Overall, we expect our approach of regard distribution balancing to be a promising technique for mitigating undesirable social biases.

4 Discussion and Conclusion

In this study, we proposed a novel pipeline to probe social biases in large-scale pretraining corpora. In contrast to existing methods, our approach employs WSD to accurately detect protected attributes while maintaining efficiency. Furthermore, we apply regard classification to analyze the language polarity towards each attribute, which is crucial yet overlooked in the existing study of pretraining datasets (Dodge et al., 2021; Penedo et al., 2024).

Based on our annotation, we can conduct regard-aware bias analysis to obtain valuable insights on social stereotypes associated with each attribute. Moreover, we can compute and balance the regard

[6] While we ignore the difference in the scale of these bias scores, this can be easily adjusted (e.g. by introducing scaling factors) if necessary.

[7] To be precise, we report $100\times\frac{p'(w|a)}{p(w|a)}$%, where p and p' are probabilities before and after intervention, respectively.

distributions to improve the fairness of the dataset, e.g. ensuring negative or offensive descriptions are not overly represented in the data.

While we acknowledge several difficulties and challenges still remain (as we discuss in the following section), we hope this work will be a fundamental step towards understanding and addressing inherent ethical risks of LLM pretraining.

Limitations

While our taxonomy of protected attributes covers a reasonably wide range of demographics, it does not currently support attribute classes such as *age*, *marital status*, and *political belief*, nor our list of attributes in each class is comprehensive, e.g. many attributes are still missing in the class of *nationality*, *disability*, etc. Expanding and refining the taxonomy is a subject of future study, which we expect can be approached by leveraging LLMs and human-curated resources (Smith et al., 2022).

Regarding WSD, we found that Gloss-RoBERTa can make accurate predictions in many cases but still struggle in disambiguating several attributes. In fact, there may also be genuine ambiguity which are difficult to be resolved even by humans. Establishing clearer criteria and improving the robustness of WSD also remain as open problems.

In terms of regard classification, our current classifier is distilled from (i.e. relies on the judgements of) Mixtral-8x7B-Instruct, which themselves may be biased in some undesirable ways. While we expect this issue can be alleviated by collecting multiple judgements from a diverse pool of human annotators or LLMs, developing a more reliable regard classier remains an important future work.

As for bias mitigation, we obtained promising results on reducing negative word associations through regard balancing: however, we've not conducted a full ablation study of pretraining LLMs from scratch on the regard balanced datasets. Due to the extreme cost of the above experiments, we leave it as an opportunity for future study.

Finally, due to the limitation of space, we provide an in-depth discussion on related work in Appendix C. While bias analysis and mitigation is a widely studied topic in NLP and AI safety, we believe our work proposes a unique and promising approach in addressing these problems in the challenging area of LLM pretraining.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Stefan Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl. In The 2024 ACM Conference on Fairness, Accountability, and Transparency, pages 2199–2208.

Steven Bird. 2006. NLTK: the natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 69–72.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1004–1015, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. 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, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. 2024. Safeguarding large language models: A survey. arXiv preprint arXiv:2406.02622.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3764–3814, Singapore. Association for Computational Linguistics.

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 Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. arXiv preprint arXiv:2302.07459.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv preprint arXiv:2312.06674.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. 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 (Volume 1: Long Papers), pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In The Twelfth International Conference on Learning Representations.

Yinhan Liu. 2019. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday, pages 189–202.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the Box? An analysis of undesirable content in the Common Crawl corpus. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 182–189, Online. Association for Computational Linguistics.

George A Miller. 1995. WordNet: A lexical database for English. Communications of the ACM, 38(11):39–41.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv preprint arXiv:2406.17557.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. Transactions of the Association for Computational Linguistics, 9:1408–1424.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence, 4(3):258–268.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. " I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. arXiv preprint arXiv:2205.09209.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. LIMA: Less is more for alignment. Advances in Neural Information Processing Systems, 36.

A Method Details

A.1 Taxonomy of Protected Attributes

In Table 3, we provide an overview of our protected attribute taxonomy. In this study, we collected a total of 97 keywords which are categorized into 10 attribute classes. For the curation of the taxonomy, we referenced the diversity & inclusion policies of multiple international companies to define attribute classes that must be protected from discrimination. Then, we leveraged existing resources (e.g. Mixtral models) to expand the list of candidate keywords followed by manual curation. The gloss for each keyword is generated based on hand-crafted templates followed by human verification.

A.2 Mixtral Prompts

In Table 4, we show the actual prompts used for data annotation based on Mixtral models.

Regarding WSD for protected attribute detection (§2.1), we applied chain-of-thought (CoT) prompting (Wei et al., 2022) to first explain the keyword usage in the context, which consistently improved the prediction quality. Based on Mixtral's response, we considered the answer of *yes* as protected attribute and *unsure/no* as non-protected attribute. In Table 5, we show several examples of WSD results where both Mixtral-8x22B and Gloss-RoBERTa agreed on the predictions.

As for regard classification (§2.2), we generally followed the criteria of Sheng et al. (2019) and explicitly included them in the prompt. In Table 6, we show several illustrative examples where both Mixtral-8x7B and our RoBERTa-based classifier agreed on the predictions.

B Bias Analysis Details

To conduct our bias analysis, we must first define the set of vocabulary V. For the purpose of identifying social stereotypes, we focused on common words which are likely to appear with any protected attribute. Specifically, we first computed 20K most frequent words V_a which appear with each attribute $a \in A$. Then, we took the intersection among them as the final set, i.e. $V = \bigcap_{a \in A} V_a$.[8]

In Table 7, we show (partial) results of our bias analysis (§3.1) for the class $A =$ *nationality*. Due to the large number of attributes in this class, we

[8]Note that we focus on comparing attributes in the same class A and avoid comparing among different classes (e.g. *female* and *blind* people) to ensure that attributes are measurable by the same standard (Blodgett et al., 2021).

Attribute Class	Attribute Keywords	Example Gloss ("a person ...")
dietary habits	vegan, vegetarian	who is a vegan
disability	autistic, blind, deaf, depression, disabled, wheelchair	who is autistic
economic status	poor, rich	who is economically poor
fertility status	fertile, infertile	who is able to conceive children
gender/sexuality	female, male, nonbinary, queer, trans	of female gender
nationality	afghan, argentine, armenian, australian, austrian, belgian, brazilian, bulgarian, canadian, chilean, chinese, colombian, croatian, cuban, danish, dominican, egyptian, ... (46 more)	of Afghan nationality
physical traits	overweight, underweight	who is overweight
race/ethnicity	african, arab, asian, black, hispanic, latino, white	of African race/ethnicity
religion	buddhist, christian, hindu, jewish, muslim	who believes in Buddhism
residence	rural, suburban, urban	who lives in rural area

Table 3: Our taxonomy of protected attributes. Each of the 97 attribute keywords corresponds to a protected attribute, which are categorized into 10 attribute classes. Each gloss used for word sense disambiguation (§2.1) is crafted as a continuation of the template phrase: "*a person ...*".

restricted our analysis to 53 nationalities which are covered in the SeeGULL dataset (Jha et al., 2023).[9] SeeGULL is a high quality and broad coverage resource containing prevalent stereotypes, which we use for the verification of our analysis.

Specifically, we utilize their *offensiveness score* annotations and consider the stereotype in their list as positive if the mean score is -1 (e.g. "*wise*" for *chinese*) and negative if it is greater than or equal to 1 (e.g. "*greedy*" for *japanese*). Considering these stereotypes as the ground truths, we measured the recall@k for the 53 nationalities based on the vocabulary sorted by the frequency bias and frequency+regard bias scores.[10]

In Table 8, we show the results of the alignment with SeeGULL stereotypes. From these results, we can verify that frequency+regard bias scores can identify positive/negative stereotypes more accurately, as (partially) illustrated in Table 7. In fact, our analysis can detect potential stereotypes that may be even missing in SeeGULL (e.g. "*technological*" or "*racist*" against *chinese* people), demonstrating the effectiveness of our approach.

C Related Work

Regarding the safety/fairness of LLM pretraining datasets, *toxicity* (Gehman et al., 2020) is a related yet distinct topic from social biases. For instance, toxic contents include expressions such as "*White people are racists*", which are directly harmful by themselves and can be removed at the instance-level based on toxicity classifiers (Longpre et al., 2024). In contrast, the problem of social biases include expressions which may be harmless by themselves but would be collectively harmful if overrepresented in the data. For instance, expressions such as "*They were protesting racist white police officers ...*" are observed in high-quality texts (e.g. news articles) and seem harmless by themselves, but they may promote undesirable stereotypes if excessively represented in the corpora. Unlike toxic contents, such expressions are difficult to be removed at the instance-level and are more desirable to be properly balanced, e.g. through downsampling.

For relatively small-scale datasets (e.g. for fine-tuning), existing works propose data augmentation methods to enhance the fairness among protected attributes (Lu et al., 2020; Qian et al., 2022). However, such approaches are often difficult to be applied on massive-scale datasets, and developing scalable and effective methods for bias analysis and mitigation remain as open challenges.

Finally, biased generations from LLMs can be mitigated at the post-training step (Schick et al., 2021; Ganguli et al., 2023) or through external guardrails (Inan et al., 2023; Dong et al., 2024). While these approaches can substantially suppress the exhibition of biased outputs, it remains unclear whether undesirable biases are effectively removed or merely hidden by the models (Gonen and Goldberg, 2019). Therefore, we believe understanding and alleviating social biases in the pretraining data remain as crucial problems, both from academic and industrial perspectives (Dodge et al., 2021; Luccioni and Viviano, 2021; Feng et al., 2023; Esiobu et al., 2023; Baack, 2024).

[9] https://github.com/google-research-datasets/seegull.

[10] As for the frequency+regard bias scores, we set the regard r = positive/negative in eq. (2) to compute the alignment with SeeGULL's positive/negative stereotypes, respectively.

WSD for Protected Attribute Detection (§2.1)
Below is a text including the keyword "{Keyword}". Determine whether the keyword indicates a person (or people) {Gloss}. - Focus on the keyword "{Keyword}". Briefly explain its usage in the text and then answer with yes, no, or unsure. - Answer yes if there is a clear reference to the person (or people) {Gloss}. - Answer no if there is no such reference, e.g. the keyword describes *non-human* entities (e.g. animals, objects, events) or does not indicate the individual(s) {Gloss}. - Answer unsure otherwise, e.g. if the reference is unclear, indirect, or implicit. Text: {Text} After examining the text: - Briefly justify your answer up to 100 words, and then answer with yes, no, or unsure. - Conclude your response in the format: "Therefore, the answer is ...".
Regard Classification (§2.2)
What best describes the regard towards the person (or people) {Gloss} in the following text? Focus on the individual(s) indicated by the keyword "{Keyword}" and answer from the following options: - Positive: The regard for the individual(s) in the text is predominantly positive, e.g. containing descriptors related to positive personality, high capability, significant achievements, etc. - Negative: The regard for the individual(s) in the text is predominantly negative, e.g. containing descriptors related to negative personality, lack of ability, commitment of illegal activities, etc. - Neutral: There is no clear impact of regard for the individual(s) associated with the text. Text: {Text} After examining the text: - Start your response with the answer (positive, negative, or neutral), followed by justification up to 100 words.

Table 4: Mixtral prompts used in word sense disambiguation (WSD) for protected attribute detection (§2.1) and regard classification (§2.2). Similar to the input format in Table 1, {Text} indicates the text containing the keyword, {Keyword} the target keyword, and {Gloss} the keyword's defined gloss.

Example Sentence (Keyword in **Bold**)	Prediction
My wife is **vegan** so we went to this place and she really loved the food. I know that the advocates for the **blind** and visually impaired will continue his work, ... Our aim is to help the **poor** and marginalized communities in Ghana to have access to education ... She is thought to be the youngest certified **female** yoga instructor in the United States. To be alone in secret with the Father should become one of the highest joys for a **Christian**. But **rural** children in particular are more likely to be driven to school or other activities, ...	✓ Protected Attribute
This water-based, non-toxic, **vegan** nail color has been formulated especially for kids. We're all **blind** to our own mistakes, and a fresh pair of eyes can do wonders for our manuscripts. It has received **poor** reviews from critics and viewers, who have given it an IMDb score of 6.1. Did you know that **female** fireflies can't fly? **Christian** Byrt is an award winning photographer based in Bunbury. Road connectivity of **rural** areas to urban areas is yet to be accomplished.	× Non-Protected Attribute

Table 5: Illustrative examples of word sense disambiguation (WSD) for protected attribute detection (§2.1) where both Mixtral-8x22B and Gloss-RoBERTa agreed on the predictions.

Example Sentence (Keyword in **bold**)	Prediction
Our **female** guide was calm, competent and interesting whilst our male guide was more hyperactive. We become quickly amazed by the **Japanese** people manners, politeness, and way of life from the beginning ... Although she was **overweight**, he still loved every part of her as she learned how to manage her health.	Positive
The girl is described as a white **female**, standing 5-foot-2 and weighing 120 pounds with blue eye ... Some cultures traditionally swam in their undergarments, and for the **Japanese**, that meant the fundoshi, ... Many **overweight** people lose weight by this one rule – especially if they have previously consumed a lot of ...	Neutral
Debora Green is an American physician and **female** Villain who pleaded no contest to setting a 1995 fire ... This showed horrific scenes of cruelty and deception by the **Japanese** authorities who want to hide this from ... ... and the general description that people have of her is that she is **overweight** and unattractive.	Negative

Table 6: Illustrative examples of regard classification towards protected attributes (§2.2) where both Mixtral-8x7B and our RoBERTa-based classifier agreed on the predictions.

Protected Attribute	Frequency Bias (eq. (1))	Frequency+Regard Bias (eq. (2))		
		$r =$ Positive	$r =$ Negative	$r =$ Neutral
chinese	chinese, xi, beijing, china, kong, hong, ma, apple, investors, korea, korean, consumers, epidemic, outbreak, belt, russians, accepting, ..	scientists, jack, ma, artificial, sun, richest, optimism, innovation, wise, medicine, ambitious, technological, boost, enhance, engineering, ...	sensitive, edited, selling, theft, dollars, sweeping, spreading, profit, amounts, restrictions, apple, debt, canceled, sell, impose, racist, ...	evaluate, examined, populations, examine, sidelines, backgrounds, sample, comparative, ...
iraqi	iraqi, iraq, bush, coalition, civilians, forces, civilian, fighters, destruction, combat, abu, troops, weapons, syrian, refugees, ...	sacrifice, veteran, brave, determination, minds, capable, volunteer, courage, helping, liberation, stability, democracy, optimistic, relief, ...	executed, shooting, threw, abuse, prison, murder, throwing, prisoners, fired, crimes, suicide, blamed, dead, firing, least, killed, ...	assess, mechanism, discuss, accordance, declined, request, discussed, postponed, walks, ..
mexican	mexico, slim, texas, drug, fox, carlos, california, el, cuban, diego, los, y, angeles, las, lord, immigration, vegas, san, del, ..	guitar, painter, colors, iconic, vibrant, superstar, richest, y, shape, homage, painted, actress, honor, proudly, hollywood, starring, loves, ...	lord, drug, racist, illegal, gun, agents, drugs, guns, pounds, judge, federal, customs, healine, deportation, agent, trump, wall, ...	populations, examined, cuban, ancestry, regardless, examine, origin, agreement, identify, ...
pakistani	pakistan, khan, afghanistan, firing, delhi, terror, bin, india, bangladesh, frontier, cia, muslim, dawn, posts, northwest, raid, ...	pioneer, designer, credited, culinary, nobel, education, entertainment, oxford, survived, drama, teenager, brave, actress, advocate, commentator, ..	firing, fake, violation, alleged, attacked, terrorist, blamed, banned, district, killed, lone, arrested, posts, targeted, allegedly, ...	criteria, apply, sidelines, applications, submit, eligible, counterpart, delegation, scholarship, ..
ukrainian	ukrainian, ukraine, inquiry, joe, investigate, complaint, trump, phone, russia, bride, investigations, call, russians, ...	intelligent, courage, boxing, faithful, beautiful, tender, dignity, fights, participant, software, honored, lady, reliable, sciences, resilience, ...	conspiracy, investigative, corruption, lawmaker, imprisonment, allegations, violations, positions, politically, prosecutor, scrutiny, ...	criteria, submit, discussed, accordance, telephone, presidents, depends, procedures, ...

Table 7: Results of our bias analyses for class $A =$ *nationality*. Words $w \in V$ are sorted in descending order based on the frequency bias score (eq. (1)) and frequency+regard bias score (eq. (2)). Positive stereotypes listed in SeeGULL is highlighted in green and negative stereotypes in red.

	Recall@	Frequency Bias	Frequency+Regard Bias
Positive Stereotypes (*mean offensiveness score* $= -1$)	50	2.67	**5.01**
	100	4.00	**7.79**
	200	7.23	**13.24**
	300	9.79	**16.80**
	500	15.80	**24.69**
Negative Stereotypes (*mean offensiveness score* ≥ 1)	50	3.68	**13.97**
	100	9.56	**17.65**
	200	13.97	**33.09**
	300	16.18	**38.97**
	500	25.00	**47.06**

Table 8: Results of the alignment with SeeGULL stereotypes. For all thresholds of recall@k, our frequency+regard bias analysis consistently outperforms the baseline based on the frequency bias analysis.