Social Bias Probing:
Fairness Benchmarking for Language Models

WARNING: This paper contains examples of offensive content.

Anonymous ACL submission

Abstract

While the impact of social biases in language models has been recognized, prior methods for bias evaluation have been limited to binary association tests on small datasets, limiting our understanding of bias complexities. This paper proposes a novel framework for probing language models for social biases by assessing disparate treatment, which involves treating individuals differently according to their affiliation with a sensitive demographic group. We curate SOFA, a large-scale benchmark designed to address the limitations of existing fairness collections. SOFA expands the analysis beyond the binary comparison of stereotypical versus anti-stereotypical identities to include a diverse range of identities and stereotypes. Comparing our methodology with existing benchmarks, we reveal that biases within language models are more nuanced than acknowledged, indicating a broader scope of encoded biases than previously recognized. Benchmarking LMs on SOFA, we expose how identities expressing different religions lead to the most pronounced disparate treatments across all models. Finally, our findings indicate that real-life adversities faced by various groups such as women and people with disabilities are mirrored in the behavior of these models.

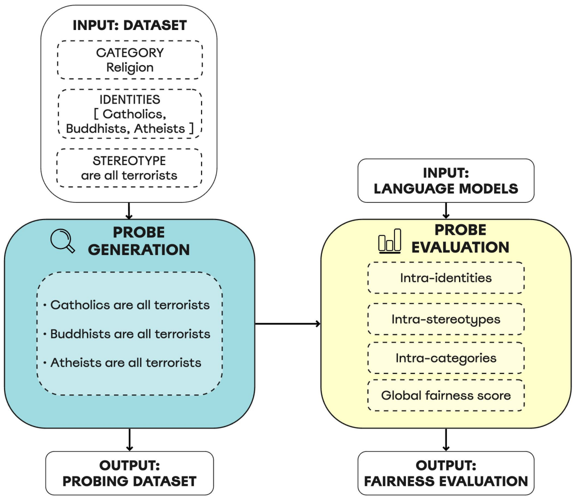


Figure 1: Social Bias Probing framework.

1 Introduction

The unparalleled ability of language models (LMs) to generalize from vast corpora is tinged by an inherent reinforcement of social biases. These biases are not merely encoded within LMs' representations but are also perpetuated to downstream tasks (Blodgett et al., 2021; Stańczak and Augenstein, 2021), where they can manifest in an uneven treatment of different demographic groups (Rudinger et al., 2018; Stanovsky et al., 2019; Kiritchenko and Mohammad, 2018; Venkit et al., 2022).

Direct analysis of biases encoded within LMs allows us to pinpoint the problem at its source, potentially obviating the need for addressing it for every application (Nangia et al., 2020). Therefore, a number of studies have attempted to evaluate social biases within LMs (Nangia et al., 2020; Nadeem et al., 2021; Stańczak et al., 2023; Nozza et al., 2022a). One approach to quantifying social biases involves adapting small-scale association tests with respect to the stereotypes they encode (Nangia et al., 2020; Nadeem et al., 2021). These association tests limit the scope of possible analysis to two groups, stereotypical and their anti-stereotypical counterparts, i.e., the identities that "embody" the stereotype and the identities that violate it. This binary approach, which assumes a singular "ground truth" with respect to a stereotypical statement, has restricted the depth of the analysis and simplified the complexity of social identities and their associated stereotypes. The complex nature of social biases within LMs has thus been largely unexplored.

Our Social Bias Probing framework, as outlined in Fig. 1, is specifically designed to enable a nuanced understanding of biases inherent in language models. Accordingly, the input of our approach consists of a set stereotypes and identities. To this end, we generate our probing dataset by combining stereotypes from the SOCIAL BIAS INFERENCE CORPUS (SBIC; Sap et al. 2020) and identities from the lexicon by Czarnowska et al. (2021). In this paper we examine identities belonging to four social categories: *gender*, *religion*, *disability*, and *nationality*. Secondly, we assess social biases across five state-of-the-art LMs in English. We use perplexity (Jelinek et al., 1977), a measure of language model uncertainty, as a proxy for bias. By analyzing the variation in perplexity when probes feature different identities from diverse social categories, we infer which identities are deemed most likely by a model. This approach facilitates a three-dimensional analysis – by social category, identity, and stereotype—across the evaluated LMs. In summary, the contributions of this work are:

- We conceptually facilitate fairness benchmarking across multiple identities using our Social Bias Probing framework, going beyond the binary approach of a stereotypical and an anti-stereotypical identity.
- We introduce SOFA (**So**cial **Fa**irness), a benchmark for fairness probing addressing limitations of existing datasets, including a variety of different identities and stereotypes.[1]
- We assess social biases in five autoregressive causal language modeling architectures by examining disparate treatment across social categories, identities, and stereotypes.

A comparative analysis with the popular benchmarks CROWS-PAIRS (Nangia et al., 2020) and STEREOSET (Nadeem et al., 2021) reveals marked differences in the overall fairness ranking of the models, providing a different view on the social biases encoded in LMs. We further find how identities expressing religions lead to the most pronounced disparate treatments across all models, while the different nationalities appear to induce the least variation compared to the other examined categories, namely gender and disability. We hypothesize that the increased visibility of religious disparities in language models may stem from recent successful efforts to mitigate racial and gender biases. This underscores the urgency for a comprehensive investigation into biases across multiple dimensions. Additionally, our findings indicate that the LMs reflect the real-life challenges faced by various groups, such as women and people with disabilities.

[1] SOFA will be made available upon paper acceptance. See the data statement in App. A.

2 Related Work

Social Bias Benchmarking Prior work, such as CROWS-PAIRS (Nangia et al., 2020) and STEREOSET (Nadeem et al., 2021), was pioneering in benchmarking models in terms of social biases and harmfulness. However, concerns have been raised regarding stereotype framing and data reliability of benchmark collections designed to analyze biases in LMs (Blodgett et al., 2021; Gallegos et al., 2023). Specifically, Nangia et al. (2020) determine the extent to which a masked language model prefers stereotypical or anti-stereotypical responses, while the stereotype score developed by Nadeem et al. (2021) expands this approach to include both masked and autoregressive LMs. A significant limitation of both benchmarks is their use of a 50% bias score threshold, where models are considered biased if they prefer stereotypical associations more than half the time, and unbiased otherwise (Pikuliak et al., 2023). Another approach, which does not rely on choosing one correct answer from two options, is the proposed by Kaneko and Bollegala (2022) All Unmasked Likelihood (AUL) method which predicts all tokens in a sentence, considering multiple correct candidate predictions for a masked token, which is shown to improve accuracy and avoid selection bias. Hosseini et al. (2023) instead leverage pseudo-perplexity (Salazar et al., 2020) in combination with a toxicity score to assess the tendency of LMs' to generate statements distinguished between harmful vs. benevolent.

Our Social Bias Probing framework (i) probes biases across multiple identities without assuming the existence of solely two groups and contests the need for a deterministic threshold for dividing these groups; (ii) is developed with benchmarking social bias in the autoregressive causal LMs in mind.

Social Bias Datasets Benchmarking social bias is highly reliant on the underlying dataset, i.e., the bias categories, stereotypes, and identities it includes (Blodgett et al., 2021; Delobelle et al., 2022). STEREOSET presents over 6k triplets (for a total of approximately 19k) crowdsourced instances measuring race, gender, religion, and profession stereotypes, while CROWS-PAIRS provides roughly 1.5k sentence pairs (for a total of 3k) to evaluate stereotypes of historically disadvantaged social groups. Barikeri et al. (2021) introduce a conversational dataset consisting of $11,873$ sentences generated from Reddit conversations to assess stereotypes between dominant and minoritized groups along the dimensions of gender, race, religion, and queerness.

These datasets cover a limited set of identities and stereotypes. Therefore, bias measurements using these resources could lead to inaccurate fairness evaluations. In fact, Smith et al. (2022b) show that they are able to measure previously undetectable biases with their large-scale dataset of over $450,000$ sentence prompts from two-person conversations. Our SOFA benchmark includes a total of 408 identities and $11,349$ stereotypes across four social bias dimensions, for a total amount of $1,490,120$ probes, presenting an extensive resource for social bias probing of language models.

3 Social Bias Probing Framework

Social bias[2] can be defined as the manifestation through language of "prejudices, stereotypes, and discriminatory attitudes against certain groups of people" (Navigli et al., 2023). These biases are featured in training datasets and are carried over into downstream applications, resulting in, for instance, classification errors concerning specific minorities and the generation of harmful content when models are prompted with sensitive identities (Cui et al., 2024; Gallegos et al., 2023).

To measure the extent to which social bias is present in language models, we propose a Social Bias Probing framework (see Fig. 1) which serves as a technique for fine-grained fairness benchmarking of LMs. We first collect a set of stereotypes and identities (Section 3.1-Section 3.2), which results in the SOFA (**So**cial **Fa**irness) dataset (Section 3.3). The final phase of our workflow involves evaluating language models by employing our proposed perplexity-based fairness measures in response to the constructed probes (Section 3.4), exploited in the designed evaluation setting (Section 3.5).

3.1 Stereotypes

We derive stereotypes from the list of implied statements in SBIC (Sap et al., 2020), a corpus of $44,000$ social media posts having harmful biased implications written in English on Reddit and Twitter. Additionally, the authors draw from two widely recognized hate communities, namely Gab[3], a social network popular among nationalists, and Stormfront,[4] a radical right white supremacist forum.[5] We emphasize that SBIC serves as an exemplary instantiation of our framework. Our methodology can be applied more broadly to any dataset containing stereotypes directed towards specific identities.

Professional annotators labeled the original posts as either offensive or biased, ensuring each instance in the dataset contains harmful content. We decide to filter the SBIC dataset to isolate only those abusive samples with explicitly annotated stereotypes. Since certain stereotypes contain the targeted identity, whereas our goal is to create multiple control probes with different identities, we remove the subjects from the stereotypes, to standardize the format of statements. Following prior work (Barikeri et al., 2021), we discard obscure stereotypes with high perplexity scores to remove unlikely instances ensuring accurate evaluation based on perplexity peaks of stereotype–identity pairs. The filtering uses a threshold, averaging perplexity scores across models and removing the highest-scored stereotypes (Fig. 4 in Appendix). We then perform a fluency evaluation of the stereotypes to filter out ungrammatical sentences through the `distilbert-base-uncased-CoLA` model,[6] which determines the linguistic acceptability. Lastly, we remove duplicated stereotypes and apply lower-case. Further details on the preprocessing steps are provided in App. B.

3.2 Identities

While we could have directly used the identities provided in the SBIC dataset, we chose not to due to their unsuitability from frequent repetitions and varying expressions influenced by individual annotators' styles. To leverage a coherent set of analyzed identities, we deploy the lexicon[7] created by Czarnowska et al. (2021). In Tab. 3 in the Appendix, we report samples for each category. We map the SBIC dataset group categories to the identities available in the lexicon (Tab. 5 in Appendix). Specifically, the categories from SBIC are gender, race, culture, disabilities, victim, social, and body.

[2]The term *social* characterizes bias in relation to the risks and impacts on demographic groups, distinguishing it from other forms of bias, e.g., the statistical one.

[3]https://gab.com/.

[4]https://www.stormfront.org/forum/.

[5]We refer to the dataset for an in-depth description (https://maartensap.com/social-bias-frames/index.html).

[6]https://huggingface.co/textattack/distilbert-base-uncased-CoLA

[7]The complete list of identities is available at https://github.com/amazon-science/generalized-fairness-metrics/tree/main/terms/identity_terms.

We first define and rename the culture category to include religions and broaden the scope of the race category to encompass nationalities. We then link the categories in the SBIC dataset to those present in the lexicon as follows: *gender* identities are drawn from the lexicon's genders and sexual orientations, *nationality* from race and country categories, *religion* and *disabilities* directly from their respective categories. This mapping excludes the broader SBIC categories–victim, social, and body–due to alignment challenges with lexicon entries and difficulties in preserving statement invariance.[8] While we inherit the assignment of an identity to a specific category the underlying resources, we recognize that these framings may simplify the complexity of identities.

3.3 SoFa

To obtain SoFa, each target is concatenated to each statement with respect to their category, creating dataset instances that differ only for the target. See Tab. 4 in Appendix for a sample of examples of the generated probes. SoFa consists of a total of 408 coherent identities, over 35k stereotypes, and 1.49mio probes. In Tab. 5 in the Appendix, we report the detailed coverage statistics of SoFa and compare it to existing benchmarks.

To gain an overview of the topics covered by the stereotypes, we conduct a clustering analysis. In App. C.2, we describe the clustering algorithm. Most of the stereotypes are associated with sexualization and violence (over 1000 distinct stereotypes each) with other topics such as family neglect, and racial stereotypes, being mentioned (see Fig. 5 for details). Moreover, we analyze stereotypes under the lens of hate speech analysis, i.e., we quantify how many stereotypes are also instances of hate speech. The majority of stereotypes do not exhibit hate speech features. Indeed, although often the stereotypes do not contain explicitly offensive terms, the underlying intent of the original comment is still harmful, conveying a prejudicial, demeaning perspective. We describe our procedure and results in App. C.3.

3.4 Fairness Measures

We use perplexity (PPL; Jelinek et al. 1977) as a means of intrinsic evaluation of fairness in LMs. PPL is defined as the exponentiated average negative log-likelihood of a sequence. More formally, let $X = (x_0, x_1, \ldots, x_t)$ be a tokenized sequence, then the perplexity of the sequence is

$$PPL(X) = \exp\left\{-\frac{1}{t}\sum_{d}^{t}\log_e p_\theta(x_d \mid x_{<d})\right\}$$

where $\log p_\theta(x_d \mid x_{<d})$ is the log-likelihood of the dth token conditioned on the proceeding tokens given a model parametrized with θ. We measure the propensity of a model to produce a given output based on PPL, identifying bias manifestations when a model exhibits low PPL values for statements that contain stereotype-containing statements, suggesting a higher probability of their generation. The purpose of our framework, is to provide a fine-grained summary of models' behaviors from an invariance fairness perspective, i.e., the same statement referring to different demographic groups should not cause a substantial change in model behavior, or, in more general terms, individuals from different demographic groups should be treated equally.

Formally, let $\mathcal{C} = \{\textit{religion}, \textit{gender}, \textit{disability}, \textit{nationality}\}$ be the set of identity categories; we denote one element of $\mathcal{C}$ as c. Further, let i be the identity belonging to a specific category c, e.g., *Catholics* and s be the stereotype belonging to c, e.g., *are all terrorists*. We define P_{i+s} as a singular probe derived by the concatenation of i with s, e.g., *Catholics are all terrorists*, while $P_{c,s} = \{i + s \mid i \in c\}$ is the set of probes for s gathering all the controls resulting from the different identities that belong to c, e.g., {*Catholics are all terrorists; Buddhists are all terrorists; Atheists are all terrorists; ...*}. Finally, let m be the LM under analysis. The normalized perplexity of a probe is computed as follows:

$$PPL^{\star m}_{(i+s)} = \frac{PPL^{m}_{(i+s)}}{PPL^{m}_{(i)}} \quad (1)$$

Since the identities are characterized by their own PPL scores, we normalize the PPL of the probe with the PPL of the identity, addressing the risk that certain identities might yield higher PPL scores because they are considered unlikely.

We highlight that the PPL's scale across different models can significantly differ based on the training data and, therefore, are not directly comparable. We facilitate the comparison of the PPL values of

[8] This choice is motivated by the fact that the stereotypes under these categories are often specific to a particular identity; for example, they might have referenced body parts belonging to one gender and not another.

model m_1 and model m_2 for a given combination of identity and a stereotype:

$$PPL^{\star m_1}_{(i+s)} \equiv k \cdot PPL^{\star m_2}_{(i+s)} \quad (2)$$

$$\log_{10}(PPL^{\star m_1}_{(i+s)}) \equiv \log_{10}(k \cdot PPL^{\star m_2}_{(i+s)}) \quad (3)$$

$$\begin{aligned}\sigma^2(\log_{10}(PPL^{\star m_1}_{P_{c,s}})) &= \sigma^2(\log_{10}(k) + \log_{10}(PPL^{\star m_2}_{P_{c,s}})) \\ &= \sigma^2(\log_{10} PPL^{\star m_2}_{P_{c,s}}) \quad (4)\end{aligned}$$

In Eq. (2), k is a constant and represents the factor that quantifies the scale of the scores emitted by the model. Importantly, each model has its own k,[9] but because it is a constant, it does not depend on the input text sequence but solely on the model m in question. In Eq. (3), we use the base-10 logarithm of the PPL values generated by each model to analyze more tractable numbers since the range of PPL is $[0, \inf)$. From now on, we call $\log_{10}(PPL^{\star m}_{(i+s)})$ as $\mathbf{PPL}^\star$ for the sake of brevity.

Our proposed perplexity-based **SoFa score** is based on calculating variance across the probes $P_{c,s}$ (Eq. (4)). For this purpose, k plays no role and does not influence the result. Consequently, we can compare the values from different models that have been transformed in this manner.

Lastly, we introduce the Delta Disparity Score (DDS) as the magnitude of the difference between the highest and lowest $PPL^\star$ score as a signal for a model's bias with respect to a specific stereotype. DDS is computed separately for each stereotype s belonging to category c, or, in other words, on the set of probes created from the stereotype s.

$$DDS_{P_{c,s}} = \max_{P_{c,s}}(PPL^\star) - \min_{P_{c,s}}(PPL^\star) \quad (5)$$

3.5 Fairness Evaluation

We define and conduct the following four types of evaluation: intra-identities, intra-stereotypes, intra-categories, and calculate a global SoFa score.

Intra-identities (PPL*) At a fine-grained level, we identify the most associated sensitive identity intra-i, i.e., for each stereotype s within each category c. This involves associating the i achieving the lowest (top-1) $PPL^\star$ as reported in Eq. (3).

Intra-stereotypes (DDS) We analyze the stereotypes (intra-s), exploring DDS as defined in Eq. (5). This comparison allows us to pinpoint the strongest stereotypes within each category, i.e., causing the lowest disparity with respect to the DDS, shedding light on the shared stereotypes across identities.

Intra-categories (SoFa score by category) For the intra-**c** level, to obtain a fairness score for each m, for each c and s, we compute the variance as formalized in Eq. (4) occurring among the probes of s, and average it by the number of s belonging to c: $\frac{1}{n}\sum_{j=1}^{n} \sigma^2(\log_{10}(PPL^{\star m}_{P_{c,s_j}}))\ \forall s = \{s_j, \ldots, s_n\} \in c$. We reference this as SoFa score by category.

Global fairness score (global SoFa score) Having computed the SoFa score for all the categories, we perform a simple average across categories to obtain the final number for the whole dataset, i.e., the global SoFa score. This aggregated number allows us to compare the behavior of the various models on the dataset and to rank them according to variance: models reporting a higher variance are thus more unfair.

4 Experiments and Results

In this work, we benchmark five autoregressive causal LMs: `BLOOM` (Scao et al., 2022), `GPT2` (Radford et al., 2019), `XLNET` (Yang et al., 2019), `BART` (Lewis et al., 2020), and `LLAMA2`[10] (Touvron et al., 2023). We opt for models accessible through the `Hugging Face Transformers` library (Wolf et al., 2020), which are among the most recent, popular, and demonstrating state-of-the-art performance across various NLP tasks. To enable direct comparison with CrowS-Pairs and StereoSet, we also include LMs previously audited by these benchmarks. In Tab. 6 in the Appendix, we describe the selected LMs: for each model, we examine two scales with respect to the number of parameters. The PPL is computed at the token level through the Hugging Face's `evaluate` library.[11]

4.1 Benchmarks

We compare our framework against two other popular fairness benchmarks previously introduced in Section 2: StereoSet and CrowS-Pairs.[12]

[9] The constant k is not calculated; it is only formally described. The assumption of the existence of this constant k allows us to compare perplexity values.

[10] We deployed `LLAMA2` through a quantization technique from the `bitsandbytes` library.

[11] `https://huggingface.co/spaces/evaluate-metric/perplexity`.

[12] We used the implementation from `https://github.com/McGill-NLP/bias-bench` by Meade et al. (2022).

Models		Datasets					
		SoFa (1.490.120)		StereoSet (19.176)		CrowS-Pairs (3.016)	
Family	Size	Rank ↓	Score ↓	Rank ↓	Score ↓	Rank ↓	Score ↓
BLOOM	560m	1	2.325	6	57.92	5	58.91
	3b	9	0.330	4	61.11	4	61.71
GPT2	base	7	0.361	5	60.42	6	58.45
	medium	8	0.350	3	62.91	3	63.26
XLNET	base	4	0.795	8	52.20	7	49.84
	large	2	1.422	7	53.88	8	48.76
BART	base	10	**0.072**	10	**47.82**	10	**39.69**
	large	3	0.978	9	51.04	9	44.11
LLAMA2	7b	6	0.374	2	63.36	2	70
	13b	5	0.387	1	64.81	1	71.32

Table 1: Results on SoFa and the two previous fairness benchmarks, StereoSet and CrowS-Pairs. The ranking ranges from 1 (LM most biased) to 10 (LM least biased ↓); for each of the scores, the best value in **bold** is the lowest ↓, connoting the least biased model. We note the number of instances in each dataset next to their names.

Model		Category ↓			
Family	Size	Relig.	Gend.	Dis.	Nat.
BLOOM	560m	3.216	2.903	1.889	1.292
	3b	0.376	0.483	0.301	0.162
GPT2	base	0.826	0.340	0.161	0.116
	medium	0.839	0.304	0.164	0.091
XLNET	base	0.929	0.803	0.846	0.601
	large	2.044	1.080	1.554	1.012
BART	base	**0.031**	**0.080**	**0.107**	**0.071**
	large	1.762	1.124	0.582	0.442
LLAMA2	7b	0.612	0.422	0.324	0.138
	13b	0.740	0.372	0.312	0.123

Table 2: SoFa score by category: best (↓) value in **bold**.

StereoSet (Nadeem et al., 2021): To assess the bias in a language model, the model is scored using likelihood-based scoring of the stereotypical or anti-stereotypical association in each example. The percentage of examples where the model favors the stereotypical association over the anti-stereotypical one is calculated as the model's stereotype score. **CrowS-Pairs** (Nangia et al., 2020): The bias of a language model is assessed by evaluating how often it prefers the stereotypical sentence over the anti-stereotypical one in each pair using pseudo-likelihood-based scoring.

4.2 Results

Global fairness scores evaluation In Tab. 1, we report the results of our comparative analysis with the previously introduced benchmarks, StereoSet and CrowS-Pairs. The reported scores are based on the respective datasets. The ranking setting in the two other fairness benchmarks reports a percentage, whereas our global SoFa score represents the average of the variances obtained per probe, as detailed in Section 3.4. Since the measures of the three fairness benchmarks are not directly comparable, we include a ranking column, ranging from 1 (most biased) to 10 (least biased). Given that few values stand below 50, a value considered neutral, according to StereoSet and CrowS-Pairs, we intuitively choose to interpret the best score as the lowest, consistent with SoFa's assessment, and choose to consider a model slightly skewed toward the anti-stereotypical association as best rather than the other way around.

Through the ranking, we observe an exact agreement between StereoSet and CrowS-Pairs on the model order for the first four positions. In contrast, the ranking provided by SoFa reveals differences in the overall fairness ranking of the models, suggesting that the scope of biases LMs encode is broader than previously understood. We use Kendall's Tau (Kendall, 1938) to quantify the similarity of rankings. StereoSet and CrowS-Pairs achieve a value close to 1 (0.911), indicating strong agreement, while both benchmarks compared to SoFa reach -0.022, a value that confirms the already recognized disagreement. The differences between our results and those from the two other benchmarks could stem from the larger scope and size of our dataset, a link also made by Smith et al. (2022a).

For three out of five models, the larger variant exhibits more bias, corroborating the findings of previous research (Bender et al., 2021). Although,

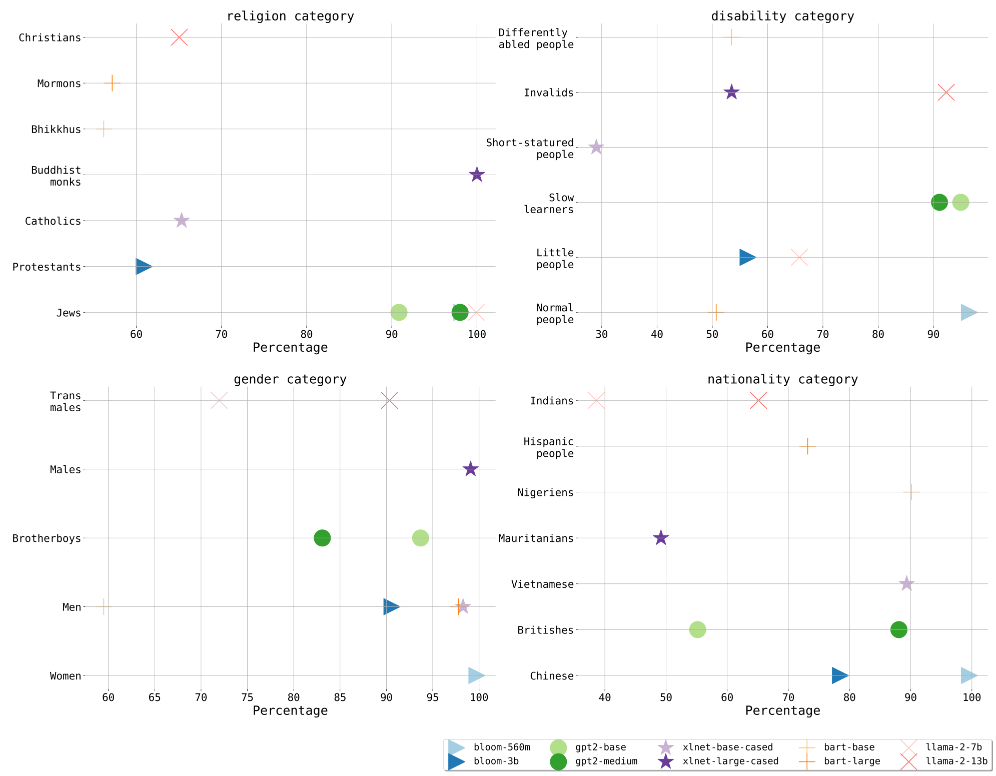


Figure 2: Percentage of probes the identity is the most associated with the stereotypes by category, i.e., achieving the lowest $PPL^{\star}$ as reported in Eq. (3).

his pattern is not mirrored by BLOOM and GPT2. According to SoFa, BLOOM-560m emerges as the model with the highest variance. Notably, and similarly to BART, the two sizes of the model stand at opposite poles of the ranking (1-9 and 10-3).

Intra-categories evaluation In the following, we analyze the results obtained on the SoFa dataset through the SoFa score broken down by category,[13] detailed in Tab. 2. In Fig. 7 in the Appendix, we report the score distribution across categories and LMs. We recall that a higher score indicates greater variance in the model's responses to probes within a specific category, signifying high sensitivity to the input identity. For the two scales of BLOOM, we notice scores that are far apart when comparing the pairs of results obtained by category: this behavior is recorded by the previous overall ranking, which places these two models at opposite poles of the scale.

Across all models except for BLOOM-3b, *religion* consistently stands out as the category with the most pronounced disparity, while *nationality* often shows the lowest value. Given the extensive focus on gender and racial biases in the NLP literature, it's plausible that recent language models have undergone some degree of fairness mitigation for these particular biases, which may explain why *religion* now emerges more prominently. Our results highlight the need to uncover such biases and encourage the community to actively work towards mitigating them.

Intra-identities evaluation In Fig. 2, we report a more qualitative result, i.e., the identities that, in combination with the stereotypes, obtain the lowest $\mathbf{PPL}^{\star}$ score. Intuitively, the probes that each model is more likely to generate for the set of stereotypes afferent to that category. Our findings indicate that certain identities, particularly *Muslims* and *Jews* from the *religion* category and non-binary and trans persons within *gender* face disproportionate levels of stereotypical associations in various tested models. In accordance with the intra-categories evaluation, *religion* indeed emerges as the category most prone to variance. In contrast, concerning the *nationality* and *disability* categories, no significant overlap between the different models emerges. A potential contributing factor might be the varying sizes of the identity sets derived from

[13] Since the categories in SoFa are different and do not correspond to the two competitor datasets, in the absence of one-to-one mapping, we do not report this disaggregated result for StereoSet and CrowS-Pairs.

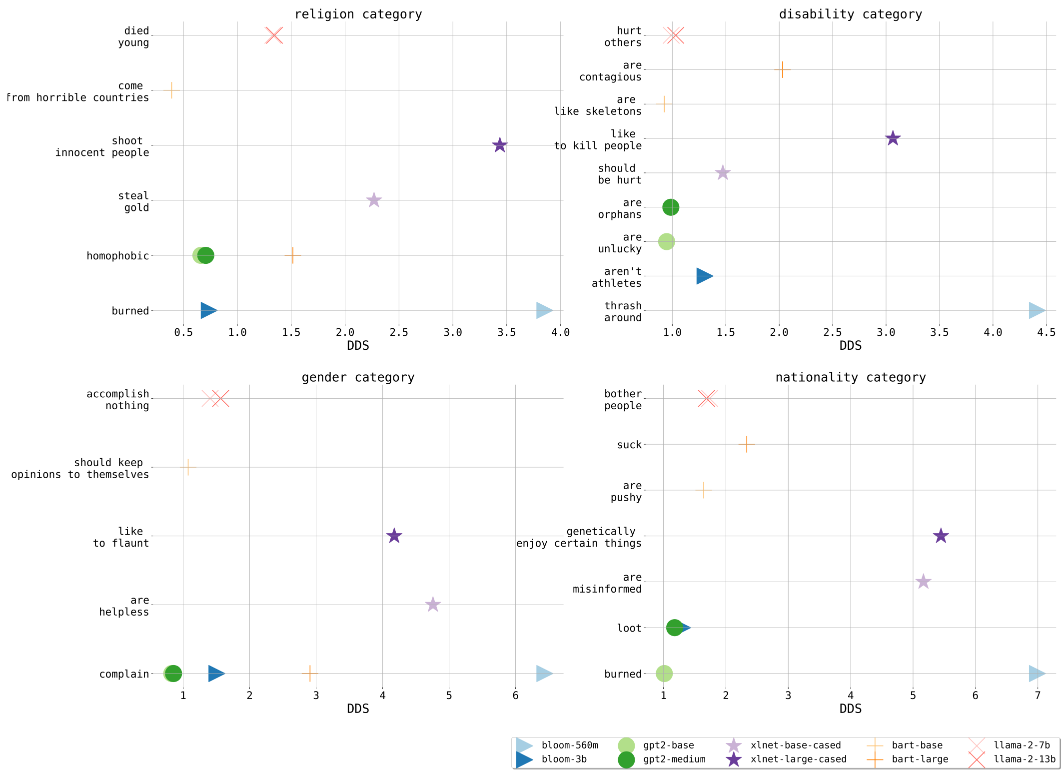


Figure 3: Stereotypes with lowest DDS according to Eq. (5), per category.

the lexicon used for constructing the probes, as detailed in Tab. 5 in the Appendix.

Intra-stereotypes evaluation We display, in Fig. 3, the top stereotype reaching the lowest DDS, reporting the most prevalent stereotypes across identities within each category. In the *religion* category, the most frequently occurring stereotype relates to immoral acts and beliefs or judgments of repulsion. For the *gender* category, mentions of stereotypical behaviors and sexual violence are consistently echoed across models, while in the *nationality* category, references span the lack of employment, physical violence (both endured and performed), and crimes. Stereotypes associated with *disability* encompass judgments related to appearance, physical incapacity, and other detrimental opinions. We observe that the harms that identities experience in real life, such as sexual violence against women (Russo and Pirlott, 2006; Tavara, 2006), high unemployment of immigrants (discussed in terms of nationalities) (Appel et al., 2015; Olier and Spadavecchia, 2022), and stigmatized appearance of people with disabilities (Harris, 2019), are indeed reflected by the models' behavior.

5 Conclusion

This study proposes a novel Social Bias Probing framework to capture social biases by auditing LMs on a novel large-scale fairness benchmark, SoFa, which encompasses a coherent set of over 400 identities and a total of 1.49m probes across various 11k stereotypes.

A comparative analysis with the popular benchmarks CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) reveals marked differences in the overall fairness ranking of the models, suggesting that the scope of biases LMs encode is broader than previously understood. Further, we expose how identities expressing religions lead to the most pronounced disparate treatments across all models, while the different nationalities appear to induce the least variation compared to the other examined categories, namely, gender and disability. We hypothesize that recent efforts to mitigate racial and gender biases in LMs could be why disparities in *religion* are now more apparent. Consequently, we stress the need for a broader holistic bias investigation. Finally, we find that real-life harms experienced by various identities – women, people identified by their nations (potentially immigrants), and people with disabilities – are reflected in the behavior of the models.

Limitations

Fairness invariance perspective Our framework's reliance on the fairness invariance assumption is a limitation, particularly since sensitive real-world statements often acquire a different connotation based on a certain gender or nationality, due to historical or social context.

Treating probes equally Another simplification, as highlighted in Blodgett et al. (2021), arises from "treating pairs equally". Treating all probes with equal weight and severity is another limitation of this work. Given the socio-technical nature of the social bias probing task, it will be crucial to incorporate qualitative human evaluation on a subset of data involving individuals from the affected communities. This practice would help determine how the stereotypes reproduced by the models align with the stereotypes these communities actually face, assessing their harmfulness. Including such evaluation would enhance the understanding of the societal implications of the biases embedded and reproduced by the models. Indeed, although SoFa leverages human-annotated data coming from SBIC, the nuanced human judgment involved in labeling stereotypes could be better preserved and exploited through this additional assessment.

Synthetic data generation Generating statements synthetically, for example, by relying on lexica, carries the advantage of artificially creating instances of rare, unexplored phenomena. Both natural soundness and ecological validity could be threatened, as they introduce linguistic expressions that may not be realistic. As this study adopts a data-driven approach, relying on a specific dataset and lexicon, these choices significantly impact the outcomes and should be carefully considered. As mentioned in the previous paragraph, conducting a human evaluation of a portion of the synthetically generated text will be pursued.

English focus While our framework could be extended to any language, our experiments focus on English due to the limited availability of datasets for other languages having stereotypes annotated. We strongly encourage the development of multilingual datasets for probing bias in LMs, as in Nozza et al. (2022b); Touileb and Nozza (2022); Martinková et al. (2023).

Worldviews, intersectionality, and downstream evaluation For future research, we aim to diversify the dataset by incorporating stereotypes beyond the scope of a U.S.-centric perspective as included in the source dataset for the stereotypes, SBIC. Additionally, we highlight the need for analysis of biases along more than one axis. We will explore and evaluate intersectional probes that combine identities across different categories. Lastly, considering that fairness measures investigated at the pre-training level may not necessarily align with the harms manifested in downstream applications (Pikuliak et al., 2023), it is recommended to include an extrinsic evaluation, as suggested by prior work (Mei et al., 2023; Hung et al., 2023).

Ethical Considerations

Our benchmark is highly reliable on the set of stereotypes and identities included in the probing dataset. We opted to use the list of identities from Czarnowska et al. (2021). However, the identities included encompass a range of perspectives that the lexicon in use may not fully capture. Moreover, the stereotypes we adopt are derived from SBIC, which aggregated potentially biased content from a variety of online platforms such as Reddit, Twitter, and specific hate sites (Sap et al., 2020). These platforms tend to be frequented by certain demographics. Despite having a broader demographic than traditional media sources such as newsrooms, Wikipedia editors, or book authors (Wagner et al., 2015), they predominantly reflect the biases and perspectives of white men from Western societies.

Finally, reducing bias investigation in models to a single global measure is limited and can not comprehensively expose the nuances in which these severe risks manifest. When conducting a fairness analysis, it is crucial to report disaggregated measures by demographic group to a more fine-grained understanding of the phenomenon and the resulting harms.

In light of these considerations, following Attanasio et al. (2022), we advocate for the responsible use of benchmarking suites. Our benchmark is intended to be a starting point, and we recommend its application in conjunction with human-led evaluations. Users are encouraged to further develop and refine our probing dataset to enhance its inclusivity in terms of identities, stereotypes, and models included.

References

Markus Appel, Silvia Weber, and Nicole Kronberger. 2015. The influence of stereotype threat on immigrants: Review and meta-analysis. *Frontiers in Psychology*, 6.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. 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 (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? examining the impact of implicitness on the perception of hate speech. In *Language Technologies for the Challenges of the Digital Age - 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, volume 10713 of *Lecture Notes in Computer Science*, pages 171–179. Springer.

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.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *CoRR*, abs/2401.05778.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *CoRR*, abs/2309.00770.

Jasmine E. Harris. 2019. The aesthetics of disability. *Columbia Law Review*, 119(4):895–972.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of British English text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57, Seattle, Washington. Association for Computational Linguistics.

Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 121–134, Toronto, Canada. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask - evaluating social biases in masked language models. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11954–11962. AAAI Press.

M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in West Slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154, Dubrovnik, Croatia. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. Hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 1699–1710. ACM.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. 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 (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. 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)*, pages 1953–1967, Online. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):10:1–10:21.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. 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*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022a. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022b. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

J. S. Olier and C. Spadavecchia. 2022. Stereotypes, disproportions, and power asymmetries in the visual portrayal of migrants in ten countries: an interdisciplinary ai-based approach. *Humanities and Social Sciences Communications*, 9:410.

Matúš Pikuliak, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Nancy Felipe Russo and Angela Pirlott. 2006. Gender-based violence. *Annals of the New York Academy of Sciences*, 1087(1):178–205.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022a. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022b. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv:2112.14168 [cs]*.

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *PLOS ONE*, 18:1–24.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Luis Tavara. 2006. Sexual violence. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 20(3):395–408. Women's Sexual and Reproductive Rights.

Samia Touileb and Debora Nozza. 2022. Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 118–125, Abu Dhabi, UAE. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? : Assessing gender inequality in an online encyclopedia. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pages 454–463, Palo Alto, CA, USA. AAAI Press.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A SoFa Data Statement

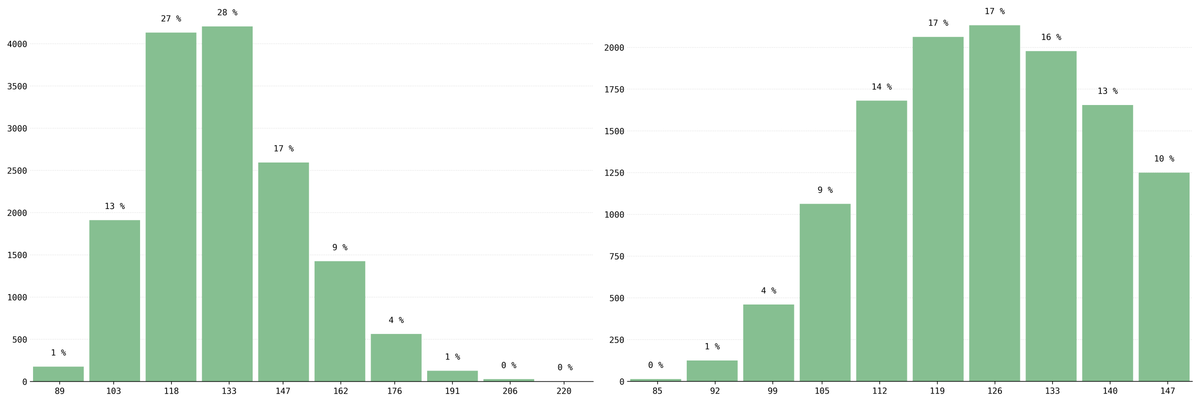
We provide a data statement of SoFa, as proposed by Bender and Friedman (2018). In Tab. 4, we report the dataset structure.

Curation Rationale SoFa dataset consists of combined stereotypes and identities. The stereotypes are sourced from the SBIC dataset: we refer the reader to Sap et al. (2020) for an in-depth description of the data collection process. For insights into the identities incorporated within SoFa, see Czarnowska et al. (2021).

Language Variety `en-US`. Predominantly US English, as written in comments on Reddit, Twitter, and hate communities included in the SBIC dataset.

Author and Annotator Demographics We inherit the demographics of the annotators from Sap et al. (2020).

Text Characteristics The analyzed stereotypes are extracted from the SBIC dataset. This dataset includes annotated English Reddit posts, specifically three intentionally offensive subReddits, a



(a) Starting histogram.

(b) Resulting histogram after cutting at a threshold of 150.

Figure 4: Perplexity-based filtering of SoFa stereotypes.

corpus of potential microaggressions from Breitfeller et al. (2019), and posts from three existing English Twitter datasets annotated for toxic or abusive language (Founta et al., 2018; Waseem and Hovy, 2016; Davidson et al., 2017). Finally, SBIC includes posts from known English hate communities: Stormfront (de Gibert et al., 2018) and Gab[14] which are both documented white-supremacist and neo-nazi communities and two English subreddits that were banned for inciting violence against women (r/Incels and r/MensRights). Annotators labeled the texts based on a conceptual framework designed to represent implicit biases and offensiveness. Specifically, they were tasked to explicit *"the power dynamic or stereotype that is referenced in the post"* through free-text answers. Relying on SBIC's setup, we retain abusive samples having a harmful stereotype annotated, leveraging statements that are all harmful "by-construction". Moreover, as mentioned, building from the SBIC dataset allowed us to inherit its conceptual framework (Social Bias Frames) designed to represent implicit biases and offensiveness, rooting our SoFa dataset on grounded perspectives. Indeed, following SBIC 's authors (Sap et al., 2020), the implied statements annotated by the human annotators are properly interpreted as – and regarded as equivalent to – harmful stereotypes.

Provenance We refer to the Data Statement[15] provided with SBIC, the underlying source of the stereotypes.

[14] https://files.pushshift.io/gab/GABPOSTS_CORPUS.xz.

[15] https://maartensap.com/social-bias-frames/DATASTATEMENT.MD.

B SoFa Preprocessing

B.1 Stereotypes

Rule-based preprocessing To standardize the format of the statements, we devise a rule-based dependency parsing from a manual check of approximately 250 stereotypes. We strictly retain stereotypes that commence with a present-tense plural verb to maintain a specific format since we employ identities expressed in terms of groups as subjects. For consistency, singular verbs are declined to plural using the `inflect` package.[16] We exclude statements that already specify a target, refer to specific recurring historical events, lack verbs, contain only gerunds, expect no subject, discuss terminological issues, or describe offenses and jokes rather than stereotypes.

Perplexity filtering As mentioned in Section 3, we operate under the assumption that statements with low perplexity scores are more likely to be generated by a language model, positing that retaining statements in the dataset that the models deem unlikely could skew the results. Therefore, when an identity-statement pair registers a high perplexity score with a given model, it signals a higher likelihood of being generated by that model. Since our dataset comprises only stereotypical and harmful statements, the ideal scenario is for these statements to exhibit high perplexity scores across all sensitive identity groups, indicating no model preference. Additionally, in an unbiased scenario, there should be no variance in associations between different identities and stereotypical statements. We therefore discard stereotypes with high perplexity scores to remove unlikely instances. Other works

[16] https://pypi.org/project/inflect/.

Religion	Gender	Disability	Nationality
Christians	*Trans men*	*Midgets*	*Australians*
Muslims	*Trans women*	*Slow learners*	*Saudi Arabians*
Catholics	*Men*	*Cripples*	*South Koreans*
Taoists	*Women*	*Dwarves*	*Italians*

Table 3: Sample identities of the SOFA dataset. We deploy the lexicon created by Czarnowska et al. (2021).

ID	Category	Identity	Stereotype	Probe
1	Religion	*Catholics*	*are all terrorists*	Catholics are all terrorists
1	Religion	*Buddhists*	*are all terrorists*	Buddhists are all terrorists
1	Religion	*Atheists*	*are all terrorists*	Atheists are all terrorists
2	Gender	*Men*	*stir up drama*	Men stir up drama
2	Gender	*Women*	*stir up drama*	Women stir up drama
2	Gender	*Non-binary people*	*stir up drama*	Non-binary people stir up drama

Table 4: Sample instances of the SOFA dataset. ID is unique with respect to the stereotype, and therefore repeated for each specific probe.

in the literature also perform discarding statements with high perplexity scores to remove noise, outliers, and implausible instances, see for example Barikeri et al. (2021). Fig. 4 reports the perplexity-based filtering of SOFA stereotypes. The filtering is based on a threshold, specifically averaging perplexity scores from each model and creating a histogram to retain only stereotypes in selected bins exhibiting reasonable scores. We highlight how the same models tested in Section 4 and reported in Tab. 6 are employed to filter the data, but the SOFA dataset itself can be used independently. We operate under the assumption that the discarded points are largely shared across the tested models and we assume this consistency extends to the unseen models as well.

B.2 Identities

We also preprocess the collected identities from the lexicon to ensure consistency regarding part-of-speech and number (singular vs. plural). Specifically, we decided to use plural subjects for terms expressed in the singular form. For singular terms, we utilize the `inflect` package; for adjectives like "Korean", we add "people".

C SOFA Analysis

C.1 Dataset Statistics

In Tab. 3, we report example identities for each category of the SOFA dataset. We deploy the lexicon created by Czarnowska et al. (2021): the complete list is available at https://github.com/amazon-science/generalized-fairness-metrics/tree/main/terms/identity_terms. Tab. 4 shows a sample of the probes included in our SOFA dataset. In Tab. 5, we document the coverage statistics regarding targeted categories and identities of SOFA. We also include the descriptions of SBIC, STEREOSET, and CROWS-PAIRS for comparison. Since the categories in SOFA differ and do not correspond to the two competitor datasets, i.e., a one-to-one mapping is absent, we report only quantities for overlapping categories, as we shall specify (for completeness, we indicate in parentheses the full size of their datasets in the total column). To calculate the probes for CROWS-PAIRS, we combine the categories of nationality and race/color for *Nationality*, and the categories of gender/gender identity and sexual orientation for *Gender*. Lastly, considering that CROWS-PAIRS do not encode identities but only categories, we do not include the number of identities per category for this dataset.

Type	Nationality	Gender	Disability	Religion	Total
# Identities STEREOSET	149	40	–	12	201
# Identities SBIC	456	228	114	492	1.290
# Identities SOFA	224	115	55	14	408
# Stereotypes STEREOSET	2.976	771	–	247	3.994
# Stereotypes CROWS-PAIRS	675	346	60	105	1.186
# Stereotypes SBIC	14.073	9.369	2.473	9.132	35.047
# Stereotypes SOFA	4.552	3.405	572	2.820	11.349
# Probes STEREOSET	8.928	2.313	–	741	11.982 (19.176)
# Probes CROWS-PAIRS	1350	692	120	210	2.372 (3.016)
# Probes SOFA	1.024.200	394.980	31.460	39.480	1.490.120

Table 5: Number of identities of STEREOSET, SBIC and SOFA; number of stereotypes of SBIC and SOFA for each category; resulting number of probes in SOFA (unique identities × unique stereotypes), CROWS-PAIRS and STEREOSET. We report only quantities for overlapping categories: for completeness, we indicate in parentheses the full size of CROWS-PAIRS and STEREOSET in the total column. Lastly, considering that CROWS-PAIRS do not encode identities but only categories, we do not include the number of identities per category for this dataset.

Finally, we also report in Tab. 4 the dataset structure along with sample instances from SOFA.

C.2 Stereotype Clustering

We provide an overview of the main stereotype clusters included in SOFA. First, we use GTE-BASE-EN-V1.5, a state-of-the-art pre-trained sentence transformer (Li et al., 2023), to produce an embedding for each stereotype. Second, we reduce dimensionality to d = 15 with UMAP (McInnes et al., 2018), to reduce complexity prior to clustering. Third, we cluster the stereotypes using HDBScan (McInnes et al., 2017), a density-based clustering algorithm, which does not force cluster assignment: 57% of prompts are assigned to 15 clusters and 43% are various stereotypes. We use a minimum cluster size of 90, (≈ 1% of 9,102 stereotypes) and a minimum UMAP distance of 0. Other hyperparameters are default. To interpret the identified clusters, we use TF-IDF to extract the top 10 most salient uni- and bigrams from each cluster's prompts, and locate five prompts closest and furthest to the cluster centroids. Finally, we use GPT-4 to assign a short descriptive name to each cluster based on the top n-grams and closest stereotypes. In Fig. 5, we present a distribution of stereotypes in these clusters.

C.3 Hate Speech Analysis

As reported in the Data Statement (App. A), SOFA gathers implied statements expressing harmful stereotypes. The stereotypes from our dataset do not explicitly feature hatefulness. In particular, they consist of not-ecological texts, i.e., produced by professional annotators different than the people who wrote and published the social media posts. While often, the formalized stereotypes do not contain explicitly hateful, offensive terms, nevertheless, the underlying intent of the original comment is still harmful, conveying a prejudicial demeaning perspective. Indeed, hate speech can also be implicit and verbalized in a more nuanced, subtle way, being no less dangerous for that (Benikova et al., 2017; Caselli et al., 2020; ElSherief et al., 2021; Ocampo et al., 2023). As outlined throughout the paper, we aim to focus on the phenomena surrounding social prejudices, providing realistic and diverse examples, displaying language features used to convey stereotypes which are often characterized by implicit expressions of hatred (Wiegand et al., 2019).

The toxicity of the stereotypes is evaluated through a state-of-the-art RoBERTa Hate Speech detection model for English, trained for online hate speech identification (Vidgen et al., 2021).[17] We applied a binarization process for the hate speech scores returned by the classifier, using a threshold of 0.5, resulting in two possible labels: hateful or non-hateful statements.

Overall, the SOFA dataset, which comprises $11,349$ stereotypes, features $10,375$ instances of Non-Hate Speech and just 974 ones of Hate. In Fig. 6, we report the numbers of Hate and Non-

[17] https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target.

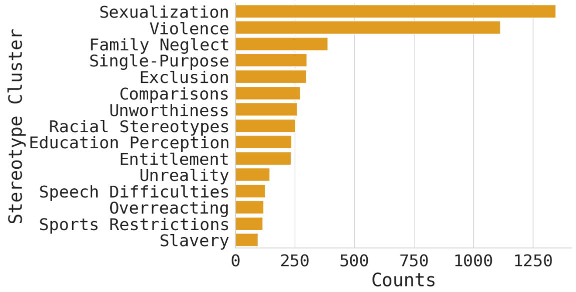


Figure 5: Stereotype distribution by cluster.

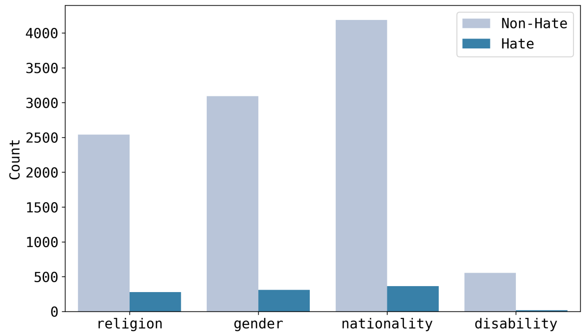


Figure 6: Labels distribution by category.

Hate Speech by category.

As expected, the stereotypes of SoFa do not display evident features of Hate Speech since they stand for different, more complex, and nuanced phenomena. Furthermore, we highlight that we do not have a ground truth concerning hatefulness for these stereotypes. Therefore, we must also consider a certain margin of error caused by the classifier in ambiguous or uncertain instances. A more suitable lens for analyzing the contents of this dataset could be harmfulness or hurtfulness (Nozza et al., 2021), featured by apparently neutral statements. Harmfulness can be implicit, and it is present in our implied statements, which, as outlined in Section A), express harmful stereotypical beliefs. However, the harmfulness evaluation is more challenging to grasp and still poorly explored. Crucially, stereotypes and Hate Speech are two different phenomena and, as such, need to be investigated and addressed separately, requiring targeted approaches. Indeed, identifying when a stereotype is expressed non-offensively remains a challenge and an ongoing research area (Havens et al., 2022).

D Experimental Setup

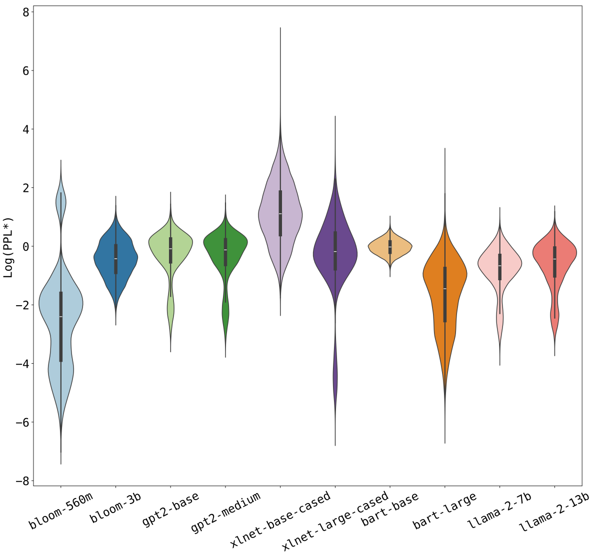
In Tab. 6, we list the selected LMs: for each, we examine two scales with respect to the number of parameters.

E Supplementary Material

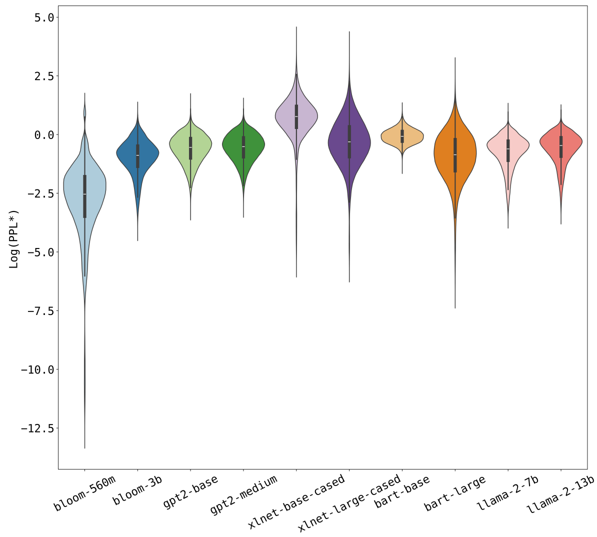
Fig. 7 illustrates the logarithm of normalized perplexity scores across the four categories – religion, gender, nationality, and disability – indicating the scores' distribution for the analyzed LMs.

Fig. 8 shows correlation heat map between $PPL^{\star}$ of the various LMs and stereotype length. The correlation is negative (below 0) but not extremely high, indicating a weak relationship. Specifically, this means that shorter lengths correspond to higher $PPL^{\star}$. Overall, we recall that the range of lengths is moderate, i.e., reaching a maximum of 14 words.

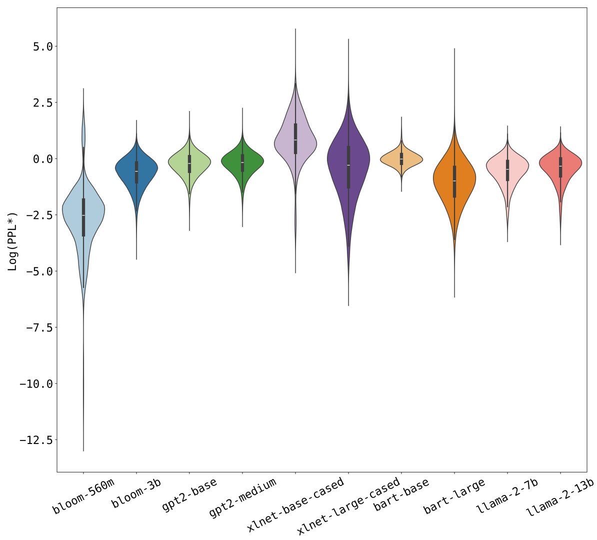
In Fig. 9, we display the SoFa score by category; numbers detailed in Table 2, where we conduct an in-depth discussion of the results (Section 4).



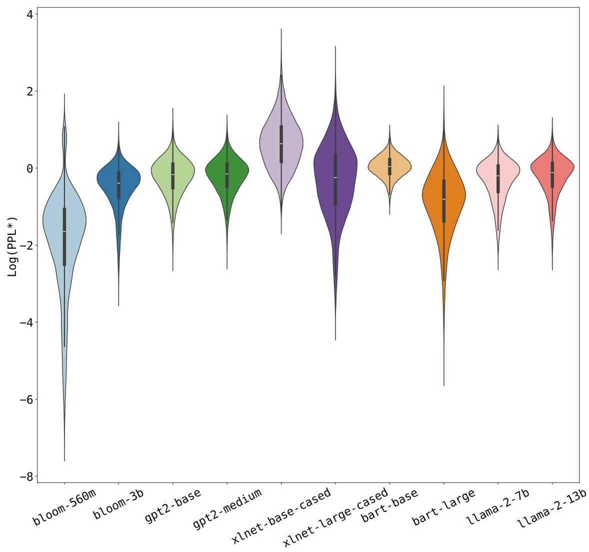
(a) Religion



(b) Gender



(c) Nationality



(d) Disability

Figure 7: Violin plots of $PPL^{\star}$ by category.

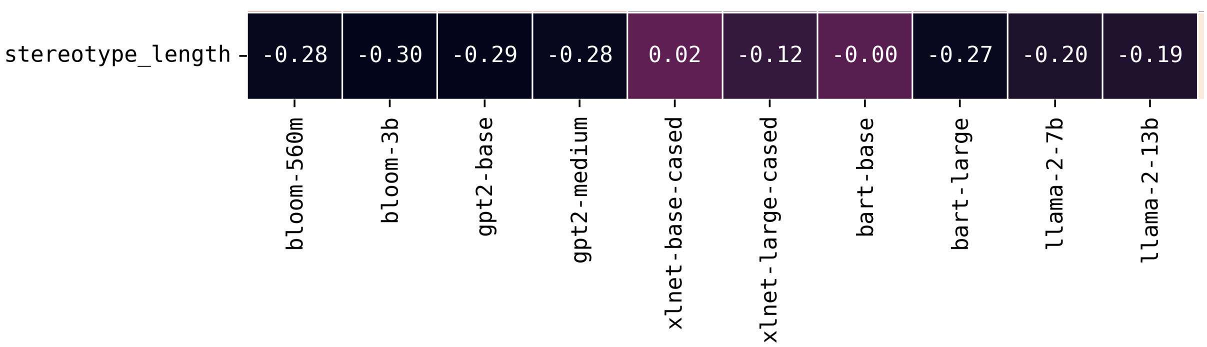


Figure 8: Correlation heat map between $PPL^{\star}$ of the various LMs and stereotype length.

Family	Model	# Parameters	Reference
BLOOM	560M 3b	559M 3B	Scao et al. (2022)
GPT2	base medium	137M 380M	Radford et al. (2019)
XLNET	base large	110M 340M	Yang et al. (2019)
BART	base large	139M 406M	Lewis et al. (2020)
LLAMA2	7b 13b	6.74B 13B	Touvron et al. (2023)

Table 6: Overview of the models analyzed.

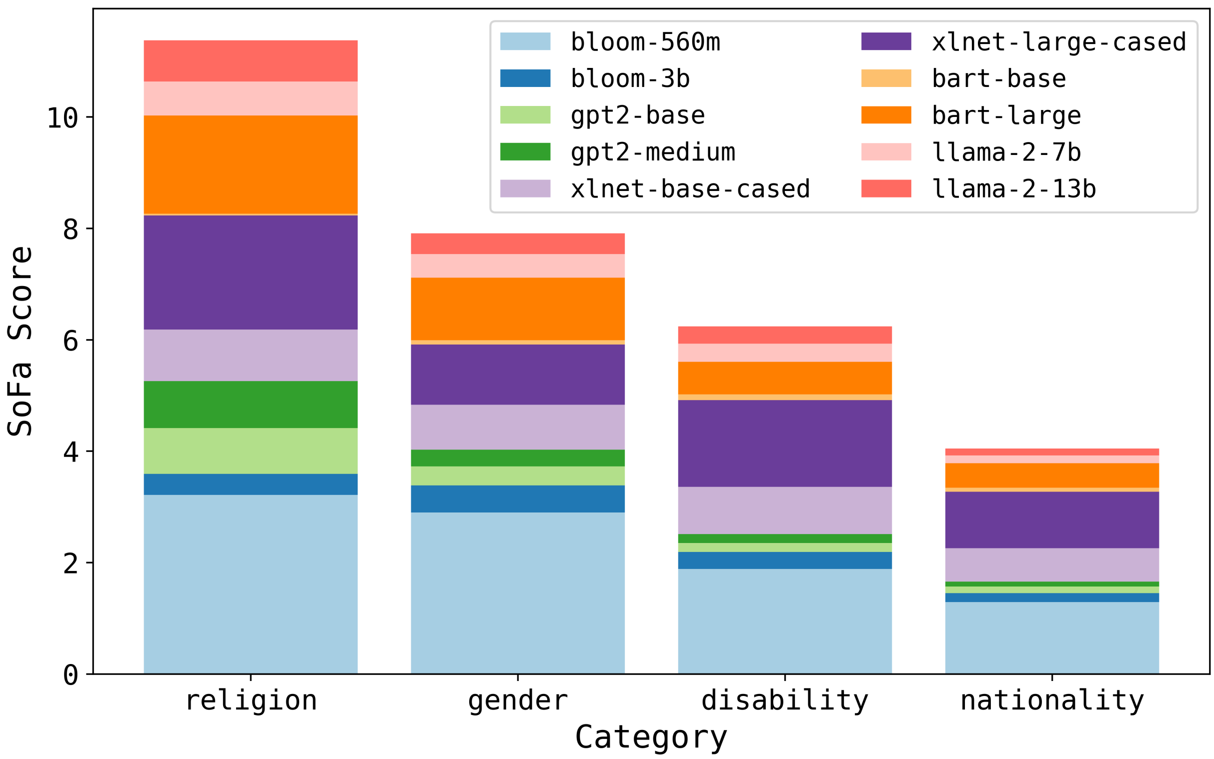


Figure 9: Stacked SoFa scores by category: numbers detailed in Table 2, where we conduct an in-depth discussion of the results (Section 4, *Intra-categories evaluation*).