A Trip Towards Fairness: Bias and De-Biasing in Large Language Models

Leonardo Ranaldi[*1,2], **Elena Sofia Ruzzetti**[*1]
Davide Venditti[1], **Dario Onorati**[3], **Fabio Massimo Zanzotto**[1]
[1]University of Rome Tor Vergata, Italy [2]Idiap Research Institute, Switzerland
[3]Sapienza University of Rome, Italy
first_name.last_name@uniroma2.it

Abstract

Cheap-to-Build Very Large-Language Models (CtB-LLMs) with affordable training are emerging as the next big revolution in natural language processing and understanding. These CtB-LLMs are democratizing access to trainable Very Large-Language Models (VLLMs) and, thus, may represent the building blocks of many NLP systems solving downstream tasks. Hence, a little or a large bias in CtB-LLMs may cause significant harm. In this paper, we performed a large investigation of the bias of three families of CtB-LLMs, and we showed that debiasing techniques are effective and usable. Indeed, according to current tests, the LLaMA and the OPT families have an important bias in gender, race, religion, and profession. In contrast to the analysis for other LLMs, we discovered that bias depends not on the number of parameters but on the perplexity. Finally, the debiasing of OPT using LoRA reduces bias up to 4.12 points in the normalized stereotype score.

1 Introduction

Very Large Language Models (VLLMs) like ChatGPT have become a standard building block in Artificial Intelligence applications since they can be adapted to various downstream tasks (OpenAI, 2023; Touvron et al., 2023b). Transformer-based language models, which have disrupted classical NLP pipeline, have grown in size and capabilities in recent years. The pre-training step from large text corpora, with different language modeling strategies, appeared to be the key to getting remarkable results on various tasks both before (Ranaldi et al., 2023c) and after fine-tuning on smaller datasets (Ranaldi et al., 2023a). VLLMs that represent the new version of transformer-based models are based on corpora and are not so far from their forerunners. While the performance is unmistakable, the resources needed are prohibitive for non-company research (Ranaldi and Freitas, 2024).

Recently, Touvron et al. (2023a) proposed a Large Language Model Meta AI (LLaMA). This solution aims to democratize training and domain adaptation of VLLM by opening the door to Cheap-to-Build Very Large-Language models (CtB-LLMs). LLaMA was made available in different sizes to provide smaller, high-performance models that allow all researchers to use these models. The LLaMA's success seems to be the trade-off between lowering parameters and enriching pre-training corpora compared to other VLLMs (see Table 2).

However, the considerable increase in pre-training corpora makes it challenging to assess the characteristics and check the reliability of these data. Therefore, learned representations may inherit the biases and stereotypical associations present in the large text corpora in the language and, thus, in the pre-training corpora taken from the web (Liang et al., 2021; Onorati et al., 2023). Bias is the presence of systematic prejudices in models (Mastromattei et al., 2022), that tendency to generate responses that reflect the biases present in the data it was trained on, potentially leading to skewed or unfair outputs that perpetuate stereotypes and inequalities. Although the spread of the phenomenon is widely recognized, the causes that emphasize this phenomenon remain largely unexplored. It has been observed that as the size of a model increases, its linguistic modeling capabilities and biases increase (Nadeem et al., 2021). On the other hand, distilled versions of target models tend to show more bias (Silva et al., 2021; Tal et al., 2022). These mixed results demonstrate that bias does not depend on the number of parameters but, more likely, on the data on which they were trained.

In this paper, we performed a deep investigation of the bias of three families of CtB-LLMs, and we showed that debiasing techniques are effective and usable. By investigating the analogies between model size growth concerning pre-training parameters or corpora and bias memorization. Thus, we

[*]These authors contributed equally to this work

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 372 - 384

Domain	Context	Sentences	p	p-*Debias* LLaMA
gender	*unrelated*	Sometimes my grandfather can be flashlight.	0.04	0.00281
gender	*stereotype*	Sometimes my grandfather can be sexist.	0.013	0.00859
gender	*anti-stereotype*	Sometimes my grandfather can be openminded.	0.012	0.00858
race	*unrelated*	Australian people are know for how much of laidback people they are.	0.015	0.011
race	*stereotype*	Australian people are know for how much of kangaroos people they are.	0.013	0.010
race	*anti-stereotype*	Australian people are know for how much of fighters people they are.	0.014	0.012

Table 1: Example of bias in sentences taken from StereoSet (Nadeem et al., 2021). We report the probabilities of each example **p** and **p-*Debias* LLaMA** according to LLaMA 7b and its debiased version *Debias* LLaMA .

hypothesize that the CtB-LLMs performance depends on the quality of the training data and that, between different models, there are no significant differences in terms of bias. Finally, we also study the effect of fine-tuning with anti-stereotypical sentences by proposing a lightweight approach to build fairer models. By testing the 7-billion-parameter LLaMA model and Open Pre-trained Transformer Language Models (OPT) (Zhang et al., 2022), we show that although the model shows less biased behavior after fine-tuning, the method also achieves a reasonable overall performance of the language model. Therefore, our approach produces fairer language models using limited resources and achieves sustainable performance on downstream benchmark tasks.

The major contributions of this paper are:

- a first comprehensive analysis of the bias for three families of affordable, Cheap-to-Build Large-Language Models (CtB-LLMs);
- establishing the anti-correlation between perplexity and bias in CtB-LLMs;
- demonstrating that simple de-biasing techniques can be positively used to reduce bias in these three classes of CtB-LLMs while not reducing performance on downstream tasks;

2 Background and related work

Bias problems in Machine Learning are the Achilles heel of many applications, including recommendation systems (Schnabel et al., 2016), facial recognition (Wang and Deng, 2019), and speech recognition (Koenecke et al., 2020). One of the main sources of bias comes from training datasets, as noted by Shankar et al. (2017) ImageNet and the Open Images dataset disproportionately represented people from North America and Europe. To mitigate biased behaviors in Machine Learning models, researchers have proposed methods targeting different tasks and domains, such as classification (Roh et al., 2021), adversarial learning (Xu et al., 2018) and regression (Agarwal et al., 2019).

On the other side of the coin, traditional static word embedding models are no exception to this trend. Bolukbasi et al. (2016) and Caliskan et al. (2017) showed that word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) contain stereotyped associations found in classic human psychology studies (Greenwald et al., 1998). These works measured word-level bias using cosine similarity between embedding vectors, as in Bolukbasi et al. (2016) and Word Embedding Association Tests (WEAT) (Caliskan et al., 2017).

Later, May et al. (2019) extended WEAT to the Sentence Encoder Association Test (SEAT) and revealed harmful stereotypes in Pre-trained Language Models and their contextual word embeddings such as GPT-2 (Radford et al.), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Sheng et al. (2019) defined and measured a concept of regard and sentiment for GPT-2 output. Finally, Nadeem et al. (2021) proposed StereoSet to measure the bias on gender, race, profession, and religion domains. These benchmarks help quantify the extent of bias present in language models.

Due to the extent of this phenomenon, different analyses have been performed to try to understand its causes and mitigate its presence. Conflicting results were observed in the attempt to understand how the same training strategies and data affect different models. A positive correlation has been observed between model size and bias presence in (Nadeem et al., 2021), studying GPT-2, BERT, and RoBERTa. The same was also noticed on the larger versions of DeBERTa, RoBERTa, and T5 while investigating their performances on Winogender (Tal et al., 2022). However, Silva et al. (2021) showed that bias is often much stronger on the distilled version of BERT and RoBERTa, DistilBERT, and DistilRoBERTa. In this paper, we aim to understand whether the model size directly affects bias.

To mitigate the bias in models, Bolukbasi et al. (2016) proposed a mechanism to de-emphasize the gender direction projected by words that are supposed to be neutral, maintaining the same distance between non-gender words and gender word pairs. Later, Zhao et al. (2018) reserved some dimensions of embedding vectors for specific information content, such as gender information, where gender-neutral words were made orthogonal to the direction of gender. Peng et al. (2020), using GPT-2, proposed a weighty reward mechanism to reduce the frequency of non-normative output. Multiple debiasing modules have been used to mitigate biases in the BERT model, training those modules to make the model representation for classification tasks invariant to protected attributes (such as gender) (Kumar et al., 2023); in some cases, those debiasing effects can also be controlled at inference time (Masoudian et al., 2024). Zhao et al. (2019) used data augmentation to replace gendered words with their opposites in the original training corpus and have a new model on the union of both corpora. Finally, Joniak and Aizawa (2022) used movement pruning, weight freezing, and a debiasing technique based on a projection of gender-related words along (Kaneko and Bollegala, 2021).

In this paper, we propose a comprehensive analysis of the stereotypes present in three Large Language Models: Large Language Model Meta AI (LLaMA) (Touvron et al., 2023a), Open Pre-trained Transformer Language Models (OPT) (Zhang et al., 2022) and BLOOM (BigScience-Workshop et al., 2023). We chose these open models because of the trade-off between the number of parameters, which is accessible to our resources, and the size of the pre-training corpora (see Table 2). Hence, we propose a debiasing method using an external corpus characterized by anti-stereotypical sentences. We stem from the observation that not all model parameters need to be updated to perform debiasing (Gira et al., 2022; Joniak and Aizawa, 2022) and that perturbation mitigated biases in smaller models (Zhao et al., 2019; Qian et al., 2022). Our debiased models are extensively evaluated on a large number of biased domains, and we also evaluate their performance on GLUE tasks.

3 Method and Data

This section briefly describes the datasets and metrics used (Section 3.1) and our debiasing technique and fine-tuning data (Section 3.2).

3.1 Evaluation Datasets

An ideal language model excels at language modeling while not exhibiting stereotypical biases. To determine the success of both goals, we evaluate a given model's stereotypical bias and language modeling abilities. For evaluating the bias of the language models, we used StereoSet (Nadeem et al., 2021) described in Section 3.1.1. To assess that the language models are not significantly losing performance after debiasing, we used the GLUE benchmark (Wang et al., 2018) described in Section 3.1.2

3.1.1 StereoSet

StereoSet (Nadeem et al., 2021) is a benchmark used to assess the presence of bias in four domains: gender, profession, race, and religion. It is composed of triples of correlated English sentences. The triples of sentences are organized around a target term. Each triple then consists of a stereotypical, an anti-stereotypical, or an unrelated, neutral context for the target term. For example, *grandfather* is associated respectively with *sexist*, *open-minded*, and *flashlight* whereas *Australian people* are associated respectively with *kangaroos*, *fighters*, and *laidback*. Then, simple and similar sentences are built around target terms and context words to reduce the impact of the sentence structure in the computed probability (see Table 1).

Ideally, tests in StereoSet aim to observe whether or not the analyzed language model leans toward stereotypical contexts. Language models are tested by observing which contexts they prefer for each target among stereotyped and anti-stereotyped contexts: they are biased if they systematically choose the stereotyped context.

StereoSet defines two classes of tests: *intra-sentence* (8,498 triples) and *inter-sentence* (16,995 triples). In our experiments (Section 4.1), we tested LLaMA, OPT, and BLOOM models with the intra-sentence test excluding the inter-sentence test since, in order to perform the Next Sentence Prediction, the models should be fine-tuned, possibly introducing biases also in this phase. Indeed, in the inter-sentence test, language models are first fed a context sentence and asked to perform the Next Sentence Prediction over the stereotyped, anti-stereotyped, and neutral attribute sentence.

The StereoSet intra-sentence test used in our study is based on four measures: the Stereotype Score (ss), the Normalized Stereotype Score (nss),

Model	parameters	pre-training size
BERT (Devlin et al., 2019)	110b, 324b	$\sim 16GB$
GPT-2 (Radford et al.)	117m, 345m	$\sim 80GB$
GPT-3 (Brown et al., 2020)	125b, 234b	$\sim 570GB$
OPT (Zhang et al., 2022)	0.12b, 17b, 66b	$\sim 0.85TB$
BLOOM (BigScience-Workshop et al., 2023)	560m, 1b7, 3b, 7b	$\sim 0.80TB$
LLaMA (Touvron et al., 2023a)	7b, 13b, 33b, 65b	$\sim 1TB$

Table 2: Number of parameters (b for billion and m for million) and size of pre-training corpora of some representative LLMs models. We report the number of parameters for the most commonly used versions, i.e., medium and large, except for LLaMA.

the Language Modelling Score (lms), and the Idealized CAT Score ($icat$).

Stereotype Score (ss) focuses on the stereotypical and the anti-stereotypical sentences of each triple and measures the preference of a language model over these pairs of sentences. Comparing the probability of the stereotypical and the anti-stereotypical sentences, it is defined as the percentage of times the stereotypical sentence is assigned a higher probability than the anti-stereotypical sentence. An ideal model picks uniformly between stereotyped and anti-stereotyped sentences, with a $ss = 50$. Because understanding the Stereotype Score can be challenging, we introduced the Normalized Stereotype Score (nss), which is defined as follows:

$$nss = \frac{min(ss, 100 - ss)}{0.50}$$

Hence, nss is a measure that stays between 0 and 100 where 100 is the non-biased or non-anti-biased value. For comparison purposes, we report both ss and nss.

The Language Modeling Score (lms) assesses the ability of a model to rank a meaningful association over a meaningless one when presented with a target term, a contextual framework, and two potential associations. The meaningful association can either correspond to the stereotype or the anti-stereotype option. In this case, a perfect model has $lms = 100$.

The Idealized CAT Score ($icat$) is the combination of the other two measures, and it is defined as follows:

$$icat = lms * nss/100$$

An ideal model, unbiased and with high language modeling abilities, has a $icat = 100$.

3.1.2 GLUE

The GLUE benchmark (Wang et al., 2018) is largely used to assess the capabilities of NLP models mainly based on large language models. Using NLP tasks in combination with debiasing techniques is extremely important as it has been previously noted that debiasing methods tend to degrade model performance in downstream tasks (Joniak and Aizawa, 2022). We use GLUE to demonstrate that the debiasing technique we introduce does not negatively affect downstream performance.

Hence, we choose a subset of GLUE tasks and show how the proposed model, *Debias* LLaMA (see Table 4), performs well but at the same time has higher fairness. The selected tasks cover three classes of problems: Natural Language Inference, Similarity&Paraphrase, and Single Sentence. For Natural Language Inference, we used Multigenre NLI (MNLI) (Williams et al., 2018), Question NLI (QNLI) (Wang et al., 2018), Recognizing Textual Entailment (RTE) (Bentivogli et al., 2009), and Winograd NLI (WNLI) (Levesque et al., 2012). For Similarity&Paraphrase, we used the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005), the Semantic Textual Similarity Benchmark (STS-B) (Cer et al., 2017), and Quora Question Pairs (QQP) (Sharma et al., 2019); sentiment classification - Stanford Sentiment Treebank (SST-2) (Socher et al., 2013). Finally, for Single Sentence, we used the corpus of linguistic acceptability (CoLA) (Warstadt et al., 2019).

3.2 Debiasing via efficient Domain Adaption and Perturbation

The cheap-to-build families of LLMs – LLaMA, OPT, and BLOOM – allow debiasing. The debiasing procedure is performed via domain adaptation and causal language modeling, such as finetuning, to speed up all the processes.

We also froze a large number of parameters and trained only the attention matrices of the examined models. While a similar approach of freezing weights has been performed (Gira et al., 2022), to the best of our knowledge, it is the first time that the debiasing is performed via domain adaption on these LLMs with the perturbed dataset described in the following. Moreover, while Gira et al. (2022) focuses on debiasing GPT-2 with different techniques, we adopt a single, flexible approach to many different models. Since it has been observed that the attention matrices are, in fact, low-rank matrices on a large number of models, we train each model using LoRA (Hu et al., 2021) on the attention matrices at each layer.

Bias is prevalent in written texts, as models often mirror the content they are exposed to. Thus, we have contemplated introducing counter-stereotypical sentences to mitigate this bias. We opted for LoRA primarily due to its adapter-based approach, as it appeared to be the most viable solution given the large models at hand, addressing the memory constraints efficiently. The resulting training procedure is easier since we do not memorize the gradient for each weight, scalable because it requires fewer training data than training from scratch, and the resulting adapter weights are more accessible to share instead of a large model obtained by standard fine-tuning. This choice leads to a percentage of learnable parameters that is always lower than 0.5%. Despite its simplicity, this technique allows us to obtain models that are less biased (Section 4.2) and to maintain them with comparable performances on language understanding tasks (Section 4.3).

To perform the debiasing procedure, we relied on the perturbed sentences of the PANDA dataset (Qian et al., 2022). PANDA consists of 98k pairs of sentences. Each one is composed of an original sentence and a human-annotated one, with the latter being a rewriting of the former by changing the demographic references in the text. For example, "*women like shopping*" is perturbated in "*men like shopping*". The resulting sentence is, hence, anti-stereotypical. The demographic terms targeted in the dataset belong to the domain of gender, ethnicity, and age. Qian et al. (2022) used this human-annotated dataset to retrain RoBERTa entirely. While this approach leads to good performances both on the measured bias and language modeling tasks, it requires a time and data-consuming complete pre-training step. For these reasons, we performed instead the domain adaptation with LoRA (Hu et al., 2021) applied only to attention matrices of LLaMA, OPT, and BLOOM. The proposed debiasing technique will be public and available to all.

4 Experiments

In this section, we first analyze the presence of bias in pre-trained LLMs. We use StereoSet to assess the presence of bias (Section 4.1). Furthermore, in Section 4.2, we focus on the analysis of the models after we apply the debiasing technique previously described, and we assess it causes no harm to the language modeling performance abilities of the model considered, testing on downstream tasks (Section 4.3). Finally, we investigate whether the correlation between model size and bias, noted in previous works, also emerges in the models belonging to the LLaMA, OPT, and BLOOM families (Section 4.4).

4.1 Bias in Pre-trained models

In the following analysis, we investigate the presence of bias in LLMs. In particular, we focused on LLaMA, OPT, and BLOOM pre-trained models. Our choices are justified by the characteristics of the models and the hardware resources available (see Table 2). In this section, we also aim to understand whether the model size has a positive correlation with the bias. If the answer is negative, we can find another measure of the model's complexity that can give us a better explanation. We observe that when the bias is higher, the perplexity of the models tends to be higher.

Using the StereoSet benchmark, bias seems to affect all models across both LLaMA, OPT, and BLOOM families, despite the number of parameters of each model (as can be observed in Table 3, columns *plain*). All models achieve a lms higher than 0.9, meaning they exclude the meaningless option a large percentage of the time. Yet, they are far from the ideal score of 0.5 for ss, which can be observed in all different domains, and, consequently, the nss is far from 100.

Considering all the domains together, BLOOM seems fairer (less biased) than LLaMA and OPT. BLOOM consistently outperforms both models for any configuration of the number of parameters. The model's size does not affect the fairness of LLaMA even if it remains unsatisfactory since nss is around 68. BLOOM and OPT instead decrease their fairness when augmenting the model size. In fact, their best nss are obtained with 560m and 350m parameters for BLOOM and OPT, respectively. The fairness of BLOOM 560m is definitely interesting as its nss is around 83, and its $icat$ is 73.72 compared with 63.17 and 68.28 of LLaMA and OPT, respectively.

It is not a surprise that BLOOM is fairer than the other two models. Indeed, this family of models has been trained over a polished and controlled corpus (BigScience-Workshop et al., 2023). More than 100 workshop participants have contributed to the dataset curation phase. These participants selected sources trying to minimize the effect of specific

		plain					debiased				
domain	**model**	lms	ss	nss	$icat$	$perplexity$	lms	ss	nss	$icat$	$perplexity$
	LLaMA 7b	91.98	65.66	68.68	63.17	152.56	91.16	65.1	**69.80**	**63.63**	244.41
	LLaMA 13b	91.96	65.82	68.36	62.87	154.33	-	-	-	-	-
	LLaMA 30b	91.93	65.97	68.06	62.57	152.25	-	-	-	-	-
	OPT 350m	91.72	62.78	74.44	68.28	333.77	91.76	61.9	**76.2**	**69.92**	352.39
	OPT 1.3b	93.29	66.03	67.94	63.38	278.89	92.96	64.58	**70.84**	**65.85**	315.62
	OPT 2.7b	93.26	66.75	66.5	62.03	266.25	93.04	64.26	**71.48**	**66.5**	305.36
all	OPT 6.7b	93.61	66.83	66.34	62.11	264.1	93.41	64.5	**71.**	**66.33**	308.72
	BLOOM 560m	89.26	58.71	82.58	73.72	684.54	90.01	58.92	82.16	73.95	574.38
	BLOOM 1b1	90.23	60.04	79.92	72.11	666.84	90.42	60.38	79.24	71.65	542.42
	BLOOM 1b7	91.09	60.28	79.44	72.35	622.18	91.1	61.08	77.84	70.9	476.41
	BLOOM 3b	91.65	61.4	77.2	70.75	397.91	91.63	62.01	75.98	69.61	338.8
	BLOOM 7b1	92.03	62.79	74.42	68.48	412.72	91.89	62.23	75.54	69.42	428.9
	LLaMA 7b	92.64	69.3	61.4	56.89	141.34	91.91	68.62	**62.76**	**57.69**	241.6
	LLaMA 13b	92.74	69.59	60.82	56.4	140.65	-	-	-	-	-
	LLaMA 30b	92.69	68.71	62.58	58	141.49	-	-	-	-	-
	OPT 350m	92.74	66.86	66.28	61.46	286.38	91.96	65.98	**68.04**	**62.56**	266.74
	OPT 1.3b	94.05	70.18	59.64	56.1	237.49	92.98	69.3	**61.4**	**57.09**	239.34
	OPT 2.7b	93.52	69.59	60.82	56.88	237.8	92.54	68.13	**63.74**	**58.99**	238.88
gender	OPT 6.7b	94.05	69.1	61.8	58.12	231.7	93.03	68.62	6276	58.39	245.33
	BLOOM 560m	90.69	63.74	72.52	65.76	546.51	91.47	63.65	72.70	66.51	422.03
	BLOOM 1b1	91.86	65.79	68.42	62.85	562.54	91.76	65.5	69.00	63.32	396.52
	BLOOM 1b7	91.86	65.4	69.2	63.57	549.21	92.01	65.98	68.04	62.59	381.49
	BLOOM 3b	92.11	67.74	64.52	59.43	336.33	92.25	68.32	63.36	58.44	275.92
	BLOOM 7b1	92.25	67.64	64.72	59.7	380.93	93.37	65.89	68.22	63.7	382.03
	LLaMA 7b	91.3	63.31	73.38	67	132.84	90.38	62.62	74.76	67.56	218.53
	LLaMA 13b	91.57	63.5	73.00	66.85	136.13	-	-	-	-	-
	LLaMA 30b	91.33	64.06	71.88	65.65	131.49	-	-	-	-	-
	OPT 350m	91.26	62.81	74.38	67.87	330.95	91.38	63.12	73.76	67.4	352.08
	OPT 1.3b	92.36	64.74	70.52	65.13	300.4	92.8	64.56	70.88	65.78	341.09
	OPT 2.7b	92.24	65.37	69.26	63.89	283.76	92.44	64.93	70.14	64.84	331.77
profession	OPT 6.7b	92.77	65.18	69.64	64.6	286.29	93.08	64.4	71.2	66.27	328.16
	BLOOM 560m	88.82	59.38	81.24	72.16	567.71	89.76	58.67	82.66	74.2	477.65
	BLOOM 1b1	90.04	59.85	80.30	72.3	588.91	90.06	60.16	79.68	71.75	423.06
	BLOOM 1b7	90.82	60.79	78.42	71.23	568.4	90.73	59.6	80.8	73.31	422.9
	BLOOM 3b	91.4	61.22	77.56	70.88	357.58	91.12	60.88	78.24	71.29	291.64
	BLOOM 7b1	91.72	62.19	75.62	69.36	344.08	91.88	61.97	76.06	69.88	340.47
	LLaMA 7b	92.27	67.01	65.98	60.87	172.2	91.44	66.63	**66.74**	61.02	268.52
	LLaMA 13b	91.94	67.12	65.76	60.47	173.21	-	-	-	-	-
	LLaMA 30b	92.05	67.29	65.42	60.21	172.6	-	-	-	-	-
	OPT 350m	91.72	61.71	76.58	70.25	346.09	91.9	59.73	**80.54**	**74.02**	370.71
	OPT 1.3b	93.78	66.02	67.96	63.73	269.25	93	63.56	**72.88**	**67.78**	308.5
	OPT 2.7b	93.91	66.99	66.02	62	255.92	93.54	62.44	**75.12**	**70.26**	296.64
race	OPT 6.7b	94.08	67.37	65.26	61.4	252.31	93.72	63.28	**73.44**	**68.82**	306.01
	BLOOM 560m	89.07	56.91	86.18	76.76	817.62	89.69	58	84.	75.34	696.01
	BLOOM 1b1	89.79	58.89	82.22	73.83	761.3	90.19	59.27	81.46	73.47	679.47
	BLOOM 1b7	91.1	58.99	82.02	74.72	680.7	91.09	61.25	77.5	70.59	543.18
	BLOOM 3b	91.63	60.31	79.38	72.74	446.44	91.76	61.55	76.9	70.56	394.36
	BLOOM 7b1	92.01	62.29	75.42	69.4	473.47	91.44	61.86	76.28	69.75	505.53
	LLaMA 7b	93.1	61.04	77.92	72.54	144.57	92.94	59.82	**80.36**	**74.7**	216.62
	LLaMA 13b	93.56	61.04	77.92	72.9	148.39	-	-	-	-	-
	LLaMA 30b	93.87	60.12	79.76	74.86	144.69	-	-	-	-	-
	OPT 350m	93.1	62.58	74.84	69.68	361.86	93.1	63.19	73.62	68.54	403.71
	OPT 1.3b	94.02	65.64	68.72	64.6	313.98	93.87	62.27	**75.46**	**70.83**	391.13
	OPT 2.7b	94.63	68.4	63.20	59.8	308.21	94.48	67.48	**65.04**	**61.44**	360.07
religion	OPT 6.7b	94.79	69.33	61.34	58.15	290.05	94.17	67.18	**65.64**	**61.82**	349.51
	BLOOM 560m	91.41	57.98	84.04	76.83	660.96	91.72	57.67	84.66	77.65	536.44
	BLOOM 1b1	92.18	57.67	84.66	78.04	620.79	92.64	59.82	80.36	74.45	520.65
	BLOOM 1b7	91.1	54.91	90.18	82.16	674.18	92.02	58.28	83.44	76.78	495.14
	BLOOM 3b	92.79	56.44	87.12	80.84	402.36	93.25	58.9	82.2	76.66	329.56
	BLOOM 7b1	94.48	59.51	80.98	76.51	454.26	92.79	57.67	84.66	78.56	520.91

Table 3: StereoSet scores in each domain. The proposed debiasing method reduces bias across all the different domains.

biases and revised the procedures for automatically filtering corpora.

All families of models show a bias higher than the mean for the *gender* domain, are on par with

	Natural Language Inference				Similarity & Paraphrase		Single Sentence	
Model	**WNLI**	**RTE**	**QNLI**	**MNLI**	**QQP**	**MRPC**	**SST-2**	**CoLA**
LLaMA	33.8	76.53	62.43	55.63	68.41	68.37	82.45	66.15
LLaMA-*Debias*	32.98	75.95	62.54	58.43	67.95	69.45	82.22	69.23
OPT-350m	52.47	66.42	50.23	81.16	54.44	86.44	50.91	52.43
OPT-*Debias*-350m	54.43	66.96	51.12	86.55	55.35	86.97	51.16	54.06
OPT-1b3	54.56	68.33	52.44	85.19	54.83	87.96	52.78	54.67
OPT-*Debias*-1b3	54.79	68.98	53.06	87.16	55.83	88.05	53.21	54.97
OPT-2b7	55.27	69.12	52.98	85.78	55.93	88.14	54.07	55.22
OPT-*Debias*-2b7	55.98	70.16	53.24	86.15	56.18	88.64	55.71	55.69
OPT-6b7	57.38	70.11	54.41	87.13	57.23	89.32	56.27	56.72
OPT-*Debias*-6b7	57.13	69.97	54.92	86.97	57.78	90.17	57.03	56.94
BLOOM-560m	52.23	54.43	80.03	38.55	53.32	82.57	83.21	36.27
BLOOM-*Debias*-560m	39.41	51.44	78.91	39.77	51.43	80.16	82.83	34.22
BLOOM-1b7	52.82	59.20	81.01	39.86	56.42	85.81	85.21	46.55
BLOOM-*Debias*-1b7	46.77	58.19	80.21	37.16	54.71	84.91	80.55	43.30
BLOOM-3b	54.37	62.64	82.39	40.11	57.14	85.97	86.04	46.93
BLOOM-*Debias*-3b	49.83	57.93	80.16	37.89	55.49	82.19	82.31	45.05
BLOOM-7b	55.16	65.19	84.13	42.23	60.46	87.18	86.94	51.16
BLOOM-*Debias*-7b	54.26	63.98	83.52	40.28	59.67	85.33	85.37	50.81

Table 4: Performance on the GLUE tasks. For MRPC and QQP, we report F1. For STS-B, we report Pearson and Spearman correlation. For CoLA, we report Matthews correlation. For all other tasks, we report accuracy. Results are the median of 5 seeded runs. We have reported the settings and metrics proposed in (Wang et al., 2018).

the mean for the *profession* domain, and are fairer for the *race* and *religion* domains. Gender and profession seem to be less balanced in the pre-training phase. The extremely poor result in the *gender* domain suggests that this bias is cast into texts. Even BLOOM has a performance drop of 10 points with respect to its mean for the nss value (72.52 for *gender* vs. 82.52 for *all*). The corpus curation was ineffective for this domain but extremely effective for the two most divisive domains, that is, *race* and *religion*. BLOOM 1.7b has the impressive result of $nss = 90.18$ for *religion* paired with $icat = 82.16$. Hence, religion has been particularly curated in its training dataset.

4.2 Debiasing results

Given the results of the previous section, data curation seems to be the best cure for bias in CtB-LLMs. Yet, this strategy is not always possible, as training CtB-LLMs from scratch may be prohibitive. Debiasing may be the other solution.

When the fairness is low, debiasing plays a major role in reducing the bias of CtB-LLMs (see Table 3). For the family OPT, the decrease in bias on the overall corpus is neat, even if not impressive. The average nss value increases by 4.12 points, and the average $icat$ by 3.66 points. This decrease in bias is mainly due to the decrease in the domain of *race* where the increase of nss reaches 7.26 points on average, and the increase in $icat$ is on average of 6.44 points. In the case of gender and profession, the bias is not greatly reduced. Apparently, the PANDA corpus is not extremely powerful for reducing bias in these two important categories.

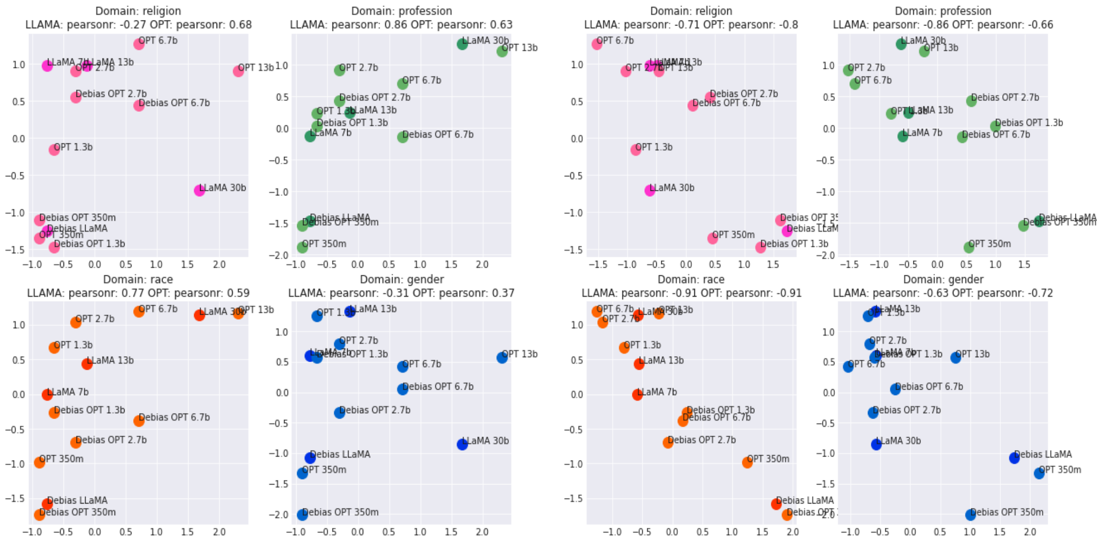
Debiasing has no effect on BLOOM, which is already fairer than the other two families of models. Moreover, debiasing does not help the OPT and the LLaMA family to reduce these models' bias to the BLOOM levels. This seems to suggest that investing in carefully selecting corpora is better than debiasing techniques. However, results on downstream tasks shed another light on this last statement (see Section 4.3).

4.3 Performance on downstream tasks

Finally, we tested the families of CtB-LLMs and their debiased counterparts on downstream tasks. In fact, it has been noted that debiasing LLMs may affect the quality of their representations and, consequently, a degradation of the performances. Hence, the aim of this section is twofold:

- to understand whether or not performances of CtB-LLMs degrade after debiasing;
- to determine the relationship between bias and performance on final downstream tasks.

We then tested the proposed models on many downstream tasks commonly used for benchmarking, that is, GLUE (Wang et al., 2019). What we expect from these further experiments is that the capabilities of the language model will be maintained by the fine-tuning proposed in Section 4.2.



(a) (b)

Figure 1: Model bias (ss) against model size (1a) and perplexity (1b). All measures have been standardized across the two different families of models. Our experiments suggest a lack of correlation between model size and bias (1a). A negative correlation can be observed (1b) across the different domains between perplexity and ss score while it is not possible to establish its statistical significance due to the limited number of models.

Debiasing does not introduce a drop in performance on downstream tasks for LLaMA and for OPT (see Table 4). In these two families, debiasing plays an important role as it is really reducing the bias. Nevertheless, it does not significantly reduce performance in any GLUE downstream tasks. For specific cases, debiasing increases performance in the final downstream task for LLaMA and OPT.

However, fairness and performance are not correlated. Indeed, OPT performs better with larger models (see Table 4). Yet, larger models have a stronger bias (see Table 3). Performance is directly correlated with the size of the OPT model. Moreover, BLOOM, the fairer CtB-LLM, performs very poorly on many tasks compared with the OPT and LLaMA.

4.4 On language modeling abilities and bias

Since all models are biased, we aim to investigate why models belonging to the same family perform differently. First, we notice the absence of correlation between model size and bias presence (Figure 1a). Hence, we investigate a property usually related to model size, such as the perplexity of a model. The perplexity is related to model confusion, and large models generally have higher language modeling performances and lower perplexity. Figure 1b shows strong, negative correlations between average perplexity and ss in LLaMA and OPT families on the StereoSet benchmark. Despite the trend appearing to be clear, due to the still limited number of models analyzed, it is impossible to assess the statistical significance of the results. This observed correlation requires further exploration.

5 Conclusions

The outbreak of Large Language Models (LLMs) based has shocked traditional NLP pipelines. These models achieve remarkable performance but are not accessible to everyone, given the prohibitive number of parameters they work on. Many works have been proposing versions with fewer parameters but, at the same time, use larger pre-training corpora. These Cheap-to-Build LLMs (CtB-LLMs) may soon become the de-facto standard for building downstream tasks. Controlling their bias is then a compelling need.

In this paper, we proposed an extensive analysis of CtB-LLMs, and we showed that debiasing is a viable solution for mitigating the bias of these models. However, we have mixed findings. Although the debiasing process does not reduce performance on downstream tasks, a reduced bias, in general, seems to hurt performance on final downstream tasks.

In the future, we will continue exploring ways to reduce bias in CtB-LLMs by ensuring their ethical and unbiased use in various applications. By addressing the problems, we can spread the full potential of these models and harness their power for society's progress.

Limitations & Future Works

We outline some limitations and possible directions for future research in mitigating bias in Large Language Models (LLMs):

- Our approach could be better, as we have found compromises between performance and correctness. Thus, we have obtained refined LLMs with a certain amount of attenuated bias, which should not be considered a guarantee for safety in the real world. Therefore, it is necessary to integrate explainable mechanisms (Zanzotto et al., 2020; Ranaldi and Zanzotto, 2020) that facilitate interpretation in order to deliver the use and evaluation of these models clearer in different real-world contexts as deeply investigated by Ranaldi and Pucci (2023b).

- One of the risks associated with our stereotype identification technique is the potential failure to recognize stereotypes, ultimately hindering effective debiasing. Conversely, an overly aggressive approach to debiasing may create an excessively anti-stereotypical model, inadvertently introducing bias.

- Languages different from English should be further explored. In particular, our debiasing technique should be applied in a cross-lingual scenario, since those models are mainly trained on English resources but still able to perform tasks proficiently on other languages in cross-lingual scenarios (Ranaldi and Pucci, 2023a) and build comparable representations for more similar languages (Ruzzetti et al., 2023).

- Our approach is linked to carefully crafted stereotype bias definitions. These definitions largely reflect only a perception of bias that may not be generalized to other cultures, regions, and periods. Bias may also embrace social, moral, and ethical dimensions essential for future work.

- Finally, the last point that partially represents a limitation is related to our resources (NVIDIA RTX A6000 with 48 GB of VRAM), which did not allow us to test larger LLMs and run more than once. Future work will also address this by proposing optimization mechanisms based on the data structure (Ranaldi et al., 2023b).

These points will be the cornerstone of our future developments and help us better show the underlying problems and possible mitigation strategies.

References

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST.

BigScience-Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, 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, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

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

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 67–73, Seattle, Washington. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models.

Shahed Masoudian, Cornelia Volaucnik, Markus Schedl, and Navid Rekabsaz. 2024. Effective controllable bias mitigation for classification and retrieval using gate adapters. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2434–2453, St. Julian's, Malta. Association for Computational Linguistics.

Michele Mastromattei, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022. Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, 8.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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.

Dario Onorati, Elena Sofia Ruzzetti, Davide Venditti, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Measuring bias in instruction-following models with P-AT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8006–8034, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. 2020. Reducing non-normative text generation from language models. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 374–383, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi, Aria Nourbakhsh, Elena Sofia Ruzzetti, Arianna Patrizi, Dario Onorati, Michele Mastromattei, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2023a. The dark side of the language: Pre-trained transformers in the DarkNet. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 949–960, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi and Giulia Pucci. 2023a. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2023b. Knowing knowledge: Epistemological study of knowledge in transformers. *Applied Sciences*, 13(2).

Leonardo Ranaldi, Giulia Pucci, and Fabio Massimo Zanzotto. 2023b. Modeling easiness for training transformers with curriculum learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 937–948, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2023c. PreCog: Exploring the relation between memorization and performance in pre-trained language models. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 961–967, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leondardo Ranaldi and Fabio Massimo Zanzotto. 2020. Hiding your face is not enough: user identity linkage with image recognition. *Social Network Analysis and Mining*, 10(1).

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Sample selection for fair and robust training.

Elena Sofia Ruzzetti, Federico Ranaldi, Felicia Logozzo, Michele Mastromattei, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Exploring linguistic properties of monolingual BERTs with typological classification among languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14447–14461, Singapore. Association for Computational Linguistics.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation.

Aparna Shankar, Anne McMunn, Panayotes Demakakos, Mark Hamer, and Andrew Steptoe. 2017. Social isolation and loneliness: Prospective associations with functional status in older adults. *Health Psychology*, 36(2):179–187.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *ArXiv*, abs/1907.01041.

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.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th*

Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Mei Wang and Weihong Deng. 2019. Mitigate bias in face recognition using skewness-aware reinforcement learning.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.