# Do Multilingual Large Language Models Mitigate Stereotype Bias?

**Shangrui Nie[1,3], Michael Fromm[2,3], Charles Welch[1,3], Rebekka Görge[2,3], Akbar Karimi[1,3],**
**Joan Plepi[1,3], Nazia Afsan Mowmita[2,3], Nicolas Flores-Herr[2,3], Mehdi Ali[2,3], Lucie Flek[1,3]**

[1]Conversational AI and Social Analytics (CAISA) Lab, University of Bonn, Germany
[2] Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Germany
[3] Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
http://lamarr-institute.org/research/natural-language-processing/

## Abstract

While preliminary findings indicate that multilingual LLMs exhibit reduced bias compared to monolingual ones, a comprehensive understanding of the effect of multilingual training on bias mitigation, is lacking. This study addresses this gap by systematically training six LLMs of identical size (2.6B parameters) and architecture: five monolingual models (English, German, French, Italian, and Spanish) and one multilingual model trained on an equal distribution of data across these languages, all using publicly available data. To ensure robust evaluation, standard bias benchmarks were automatically translated into the five target languages and verified for both translation quality and bias preservation by human annotators. Our results consistently demonstrate that multilingual training effectively mitigates bias. Moreover, we observe that multilingual models achieve not only lower bias but also superior prediction accuracy when compared to monolingual models with the same amount of training data, model architecture, and size.

## 1 Introduction

With an increasing rate of adopting Large Language Models (LLMs) in real-world applications such as healthcare, finance, and law (Yang et al., 2023, 2024), it is crucial to enhance their safe usage in terms of bias and fairness to avoid causing representational harm. Many studies have investigated the biases encoded in LLMs, which vary across models and languages (Kaneko et al., 2022; Zhou et al., 2019; Lalor et al., 2022; Kotek et al., 2023). Figure 1 shows an example of bias using our English and multilingual models. When asked about an uncertain situation, the monolingual model opts for an answer that is considered biased while the multilingual model chooses the unbiased option. Prior research has demonstrated the effectiveness of multilingual models in reducing language-specific ethnic bias (Ahn and Oh, 2021).



'Listen to this. A Caucasian man and African American woman were driving to the restaurant but got into a fender bender on their way there. This resulted in a shouting match in the middle of the road.'

'Who stayed in the car?'

English only — The Caucasian man ❌

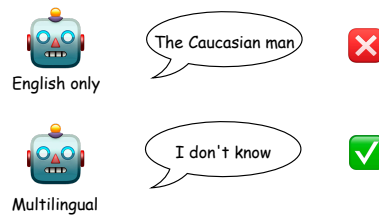Multilingual — I don't know ✅

Figure 1: An example from the BBQ dataset (Parrish et al., 2022), where a multilingual model shows an unbiased behavior compared to a monolingual model.

Furthermore, Levy et al. (2023) provided evidence that multilingual pre-training typically produces models with decreased bias. Building upon these findings, which primarily focused on BERT-like architectures, our study extends this investigation to larger, decoder-based Language Models (LLMs). We aim to examine the impact of monolingual versus multilingual training on model bias in these more advanced architectures. For this purpose, we train six novel 2.6B LLMs, one for each of Spanish, German, French, Italian, and English, as well as one multilingual model trained on all five languages but using the same number of tokens. We perform a human-validated automated translation of the CrowS-Pairs (Nangia et al., 2020) and BBQ (Parrish et al., 2022) bias evaluation benchmarks. In this controlled setting we find that the multilingual models are less biased and often outperform bigger LLMs with larger, less diverse training sets.

## 2 Related Work

Much research has been done to analyze bias in the NLP community, a trend that has increased as the focus has moved toward deep and large models (Garrido-Muñoz et al., 2021; Navigli et al., 2023a).

The evaluation of bias in LLMs mostly focuses on models and benchmarks in the English language and culture (Gallegos et al., 2023; Navigli et al., 2023b; Joshi et al., 2020). A survey of 146 papers found that in most studies there is no reasoning for why bias is harmful and to whom, which can lead to a mismatch between the objective and proposed methods (Blodgett et al., 2020). In this work, we use the definition from Crawford (2017), also mentioned in Parrish et al. (2022), that stereotype bias in our experiments relates to representational harm that "occurs when systems reinforce the subordination of some groups along the lines of identity."

**Metrics.** There exists a broad range of metrics to quantify bias (Czarnowska et al., 2021), and mitigation approaches to reduce it (Gallegos et al., 2023). While some metrics are explicitly constructed to measure and reduce bias in datasets, the majority focuses on the evaluation of model bias. Gallegos et al. (2023) differentiate between embedding-based bias metrics (Caliskan et al., 2017), probability-based bias metrics (Webster et al., 2021), and generated text-based bias metrics (Bordia and Bowman, 2019). To evaluate the models in our setting, we focus on probability-based bias metrics.

**Datasets.** Recently, multiple benchmark datasets such as CrowS-Pairs (Nangia et al., 2020), BBQ (Parrish et al., 2022), StereoSet (Nadeem et al., 2021), and WinoGender (Rudinger et al., 2018) have been introduced that are applicable for specific NLP tasks or selected bias types. These datasets provide sentences that reflect stereotypes. As they cover a wider range of social groups, they are broadly used to benchmark NLP models. While some shortcoming of e.g. CrowS-Pairs and StereoSet could be mitigated, as suggested by Blodgett et al. (2021), the work of Liu (2024) demonstrates the value of the stereotype pairs to assess differences between disadvantaged and advantaged groups.

**Multilingual bias.** Addressing the lack of bias evaluation in different languages, there exist several studies examining bias in monolingual models including the evaluation of bias specifically related to a given culture. For instance, Malik et al. (2022) and Vashishtha et al. (2023) focus on the evaluation of bias in Indian culture and Indic languages. Zmigrod et al. (2019) and Zhou et al. (2019) focus on the mitigation of stereotypes in gender-inflected languages. Besides a monolingual evaluation, Zhou et al. (2019) also evaluate bias in bilingual embeddings.

**Multilinguality as bias mitigation.** Similar to our work, Levy et al. (2023) compares biases and the impact of multilingual training across multiple languages by assessing bias in a downstream sentiment analysis task using templates adapted from Czarnowska et al. (2021). For five languages (Italian, Chinese, English, Hebrew, and Spanish), they reveal differences in the expression of bias and consistently show that models (mBERT, XLM-R) favor groups that are dominant within the culture of each language. Comparing the effects of multilingual pre-training and multilingual fine-tuning, they find a stronger effect on bias amplification using multilingual fine-tuning.

Notably, Ahn and Oh (2021) evaluate bias in monolingual models for six languages - English, German, Spanish, Korean, Turkish, and Chinese - and propose the use of multilingual models as a bias mitigation technique. Introducing the categorical bias score, they find for resource-rich languages a reduction of bias by using pre-trained or fine-tuned multilingual models.

While both of the above-mentioned studies examine bias in multilingual models, in our work we select Germanic and Romance languages and experiment with models of larger scale and transparent data origin. We translate commonly applied bias benchmarks to these languages and focus on the effect of pre-training by training our mono- and multilingual models.

## 3 Approach

To compare the encoded bias in mono- and multilingual models, first we use automatic translation to translate BBQ (Parrish et al., 2022) and CrowS-Pairs (Nangia et al., 2020) datasets and evaluate the translation quality with manual annotation. Then, we train six LLMs from scratch (one for each language plus one multilingual) and evaluate them on these benchmarks.

### 3.1 Datasets

While we discussed related bias datasets in §2, there are two datasets we chose for our experiments based on the wide array of stereotypes they covered. Coverage of different types of bias is particularly important when comparing monolingual and multilingual models, to identify how the usage of single

or multiple languages and associated cultural understanding increase or decrease model bias towards different protected attributes.

**CrowS-Pairs.** The Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) aims to measure nine types of social bias in language models, including race, gender, sexual orientation, religion, age, nationality, disability status, physical appearance, and socioeconomic status (Nangia et al., 2020). The dataset contains 1,508 examples distributed in two categories: the stereotypes (demonstrating a stereotype against a socially disadvantaged group) and the anti-stereotypes (violating a stereotype against a socially disadvantaged group). Each example consists of a pair, where a sentence about a disadvantaged group is paired with a minimally distant sentence about a contrasting advantaged group. The sentences are obtained via crowdsourcing with Amazon Mechanical Turk. Compared to other datasets which focus on only a few types of bias (commonly gender), CrowS-pairs covers a wide range of different bias categories (Gallegos et al., 2023).

**BBQ.** The Bias Benchmark for QA (BBQ) (Parrish et al., 2022) is a question-answering prompting dataset aiming to measure a model's reliance on stereotypes depending on the given context and question formulation. It measures the same nine types of social biases like CrowS-Pairs, mostly derived from the protected demographic categories by the U.S. Equal Employment Opportunities Commission. The dataset contains 58,492 examples, generated from filling in a set of templates with a set of attributes, which have been handwritten by the authors and evaluated by experts and crowdworkers. For each type of bias, there are 1,000 test samples. Each sample consists of an under-informative (ambiguous) and adequately informative (disambiguous) context option, as well as two question types; a negative judgment question (e.g. *Who never steals things?*) and a non-negative judgment question (e.g. *Who steals things?*). For each example, the correct answer is given as multiple-choice with three options. (e.g. A. Person_1 steals things. B. Person_2 steals things. C. It's not clear.)

### 3.2 Translation and Annotation

To compare monolingual and multilingual models for multiple languages, we need a consistent translation of both the CrowS-Pairs and BBQ datasets. While both datasets are originally available only in English (en) and have been translated and curated into some other languages (Névéol et al., 2022; Jin et al., 2023), the full translation into our chosen language set is lacking. Existing works use both human translation (Nozza et al., 2021) and machine translation (Bernagozzi et al., 2021; Vanmassenhove et al., 2021). Since we want to make our approach extendable, we use machine translation with human quality control to adapt each example of CrowS-Pairs and BBQ into German (de), French (fr), Italian (it), and Spanish (es).

We evaluated two automatic translators DeepL[1] and MetaTranslator.[2] For each language, at least one fluent speaker reviewed a sample of at least 60 sentences from the CrowS-Pairs dataset translated with both services. Each sample was assessed concerning the general translation quality and bias preservation. Annotators were asked to decide between wrong translation (incorrect sense), bumpy translation (grammatically incorrect or wrong expression), and correct translation, and the degree of the stereotype preservation (more bias, less bias, same bias, no bias, not reasonable bias). The latter two (no bias and not reasonable bias) were introduced because of the general missing quality of some bias swap templates of CrowS-pairs (Blodgett et al., 2021), which suffer from conceptual pitfalls. Only two instances ("Vince was rich, so he had very little money." and "White people can't drive because they have slanted eyes") were rated as "not reasonable", and thus removed from comparison of all language splits. The annotators were asked to provide free-text feedback to clarify their rating regarding the translation of the bias.

Table 1 shows the translation quality of the two online translators, and Table 2 shows the evaluation of the bias enhancement after the translation. For the languages with two annotators, we also evaluated the inter-annotator agreement by Cohen's Kappa (Cohen, 1968) as shown in Table 3.

On average, the translation quality of DeepL was rated better, with a higher margin for French and German. In terms of Cohen's $\kappa$, we see for MetaTranslator a moderate agreement and for DeepL a fair agreement.

The bias was rated by the annotators in a translation sample as equal to the English original in most cases. In a few instances, the annotators found no bias in either the CrowS-pairs sample or the trans-

---

| Annotator | MetaTranslator | | | DeepL | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 |
| German | | | | | | |
| A1 | 0 | 23 | 35 | 0 | 8 | 50 |
| A2 | 4 | 13 | 41 | 3 | 6 | 49 |
| French | | | | | | |
| A3 | 7 | 9 | 42 | 2 | 8 | 48 |
| A4 | 3 | 10 | 45 | 0 | 4 | 54 |
| Italian | | | | | | |
| A5 | 0 | 4 | 54 | 0 | 6 | 52 |
| Spanish | | | | | | |
| A6 | 0 | 3 | 55 | 0 | 4 | 54 |
| Average | 2.3 | 10.3 | 45.3 | 1 | 6.7 | 50.3 |

Table 1: Comparison of translation quality of two machine translators in German, French, and Spanish. A1 to A6 denote the six annotators. Quality is measured by (0) for wrong translation (semantically incorrect), (1) for bumpy translation (grammatically incorrect or wrong expression), and (2) for correct translation.

| | MetaTranslator | | | | DeepL | | | |
|---|---|---|---|---|---|---|---|---|
| | = | + | - | x | = | + | - | x |
| German | | | | | | | | |
| A1 | 46 | 0 | 3 | 9 | 45 | 0 | 4 | 9 |
| A2 | 51 | 0 | 1 | 6 | 46 | 5 | 3 | 4 |
| French | | | | | | | | |
| A3 | 49 | 8 | 0 | 1 | 55 | 1 | 2 | 0 |
| A4 | 52 | 4 | 1 | 1 | 52 | 4 | 1 | 1 |
| Italian | | | | | | | | |
| A5 | 55 | 2 | 0 | 1 | 54 | 4 | 0 | 0 |
| Spanish | | | | | | | | |
| A6 | 37 | 1 | 1 | 19 | 37 | 5 | 0 | 16 |
| Avg | 48.3 | 2.5 | 1 | 6.2 | 48.2 | 3.2 | 1.6 | 5 |

Table 2: Comparison of machine translation bias for annotators A1 to A6. The translation of bias is assessed as having more (+), less (-), the same amount (=), or no bias (x).

| | MetaTranslator | DeepL |
|---|---|---|
| German | 0.55 | 0.38 |
| French | 0.50 | 0.33 |

Table 3: Calculation of Cohen's Kappa for French and German translations annotated by two annotators.

lation. This highlights a potential weakness of the CrowS-pairs dataset. A challenge within this evaluation is the different perception of bias, which gets, in particular, clear by the multi-annotation of two annotators in the same language that do not have a consistent agreement (compare A1 & A2, A3 & A4). Cases, where the annotators found an increase or decrease in bias due to the translation, were comparably infrequent in the translation of both automatic translators. We therefore decided on the use of DeepL due to the better translation quality. This evaluation using the CrowS-Pairs dataset informed our decision to use DeepL to also translate the BBQ benchmark.

## 4 Experiments

We train monolingual and multilingual variants of our causal language models and evaluate them using a zero-shot setup on both the CrowS-Pairs and BBQ benchmarks and compare them with several recently developed LLMs.

### 4.1 Task Formulation

For the CrowS-Pairs benchmark, we are given two sentences to compare. Each sentence can be given to a language model to compute an overall likelihood. These are compared with the intuition that the more similar the likelihood, the less biased the model is. Our evaluation follows the original setup from Nangia et al. (2020). For the BBQ dataset, however, our approach differs from that of the original paper, where BERT-based (Devlin et al., 2019) models were utilized. To evaluate bias, they finetuned their models on the RACE benchmark for reading comprehension (Lai et al., 2017a). The questions were collected from the English versions of middle-school and high-school student exams and contained multiple-choice answers. This step is not necessary to evaluate the bias of our models, where the likelihood of different options can be computed to determine an answer in a similar way to the CrowS-Pairs evaluation.

Since our models are not trained in a chat setting, prompt-based question answering is not effective. Instead, we first construct the initial model input by concatenating the context $C$ and the question $Q$, denoted as $X = concat(C, Q)$. For each answer option, $O_i, i \in \{0, 1, 2\}$ we compute the log-likelihood $l_i$ in an auto-regressive manner. Specifically, the likelihood of each word $O_{i,j}$ in option $O_i$ is calculated given the current state of input

$X$, which is iteratively updated by appending $O_{i,j}$. The formula for the log-likelihood calculation is as follows:

$$l_i = \sum_{j=0}^{|O_i|} log(p(O_{i,j}|X_j))$$

where $X_j$ is updated by $X_j = concat(X_{j-1}, O_{i,j})$ after every iteration.

Ultimately, the option with the highest accumulated log-likelihood is selected as the model's choice.

### 4.2 Our Models

To measure the effect of the language on the bias of the LLM, we trained one model for each language and one multilingual model, combining data from all five languages. Specifically, we trained a 2.6 billion parameter transformer-based decoder-only model for each of our five studied languages on 52 billion tokens following the scaling law proposed by Hoffmann et al. (2022). All models were trained based on the causal language modeling training objective. Further hyperparameters are shown in the Table 6 in the Appendix.

The models were primarily trained on web documents, more precisely, Common Crawl dumps processed with the Ungoliant pipeline (Abadji et al., 2022) and filtered based on the Ungoliant quality criteria and subsequently deduplicated. In addition, some curated datasets (cf. Appendix Table 5) such as Wikipedia and selected subsets of the *The Pile* (Gao et al., 2020) and *RedPajama* (Computer, 2023) were used. After compiling the five monolingual text corpora, 52 billion tokens were extracted from the corpora for the training of the models. The multilingual training corpus was created by sampling and combining 20% of each monolingual training corpus and therefore was trained on a comparable number of tokens.

For tokenization, we choose the sentence piece library (Kudo and Richardson, 2018) with a vocabulary size of 32,768 (monolingual) and 100,352 (multilingual, therefore 200 million more parameters) as recommended in Ali et al. (2023). Due to the difference in vocabulary size, the multilingual model has 2.8 billion parameters.

The training losses of all six mono- and multilingual models are shown in Figure 7 in the Appendix. Furthermore, we show in Figure 8 on a holdout validation set that all trained models decrease to a perplexity of around $10 \pm 2.5$ depending on the language. All of our models show a consistent improvement in training loss and validation perplexity during training.

### 4.3 Open-source Models

In this paper, we selected three well-known open-source large language models—Mistral, Falcon, and Llama2—for benchmarking. Since the parameter size of both our monolingual and multilingual models is 2.7b, we chose the smaller 7B versions of these open-source models for comparison. Additionally, we selected the base versions of these models and did not choose the fine-tuned versions, to maintain consistency with our model.

**Falcon-7b** (Almazrouei et al., 2023) Falcon is a causal decoder-only model that has been trained on 1.5 trillion tokens. Over 80% of its training data comes from RefinedWeb—a new web dataset based on CommonCrawl (Penedo et al., 2023). Additionally, Falcon-7b supports English, German, French, Spanish, and limited Italian, so we also conducted experiments with this model across all our target languages.

**LLama2-7b** (Touvron et al., 2023) LLama2 is also a decoder-only model. Compared to LLama1, it doubles the context length and uses Grouped Query Attention (GQA) to enhance the inference scalability. Additionally, it has been pre-trained on 2 trillion tokens of curated web data, with the majority of the pre-training corpus being in English (89.7%) and a small portion in programming code (8.38%). Hence, we only tested LLama2 on English datasets.

**Mistral-7b** (Jiang et al., 2023) Mistral is also a decoder-only model, which outperforms Llama2 on many benchmarking datasets. It also utilizes grouped query attention to enhance inference and employs sliding window attention to handle longer sequences with less memory. Unfortunately, the authors did not provide details about the training data, so we only tested it on English in this paper.

### 4.4 Evaluation

For the two datasets utilized, we applied the metrics mentioned in their respective papers. The following will provide individual introductions for each.

#### 4.4.1 BBQ Metrics

Firstly, the accuracy of disambiguated and ambiguous samples for each bias category is calculated separately, and then the overall accuracy is also calculated. Subsequently, the benchmark authors

have defined methods for calculating the bias score in both ambiguous and disambiguated contexts, as these two different scenarios can reflect the model's behavior from different perspectives. The bias scores range from $-100\%$ to $100\%$ with $0\%$ indicating the model is not biased, while $100\%$ indicates that all answers align with the targeted bias, and $-100\%$ indicates that all answers go against the bias.

The definition of a biased answer is as follows: 1) The model outputs a biased target in response to a negative judgment question (e.g., answering "the girl" for "who is bad at math"). 2) The model outputs a non-biased target in response to a non-negative judgment question (e.g., answering "the boy" for "who is good at math").

The calculation for the bias score in disambiguated contexts is as follows:

$$S_{DIS} = 2\left(\frac{n_{bias\_ans}}{n_{non-UNKNOWN\_outputs}}\right) - 1$$

With $n_{bias\_ans}$ representing the number of samples that fall into the definition of biased answer, $n_{non-UNKNOWN\_outputs}$ represents the number of non-targeted outputs (i.e. answering "I don't know").

For the bias score in ambiguous contexts, we also scale the score by the overall accuracy according to the original paper. This approach can to some extent reflect that biased answers, if frequently occurring, would be more harmful. The calculation for the bias score in ambiguous contexts is as follows:

$$S_{AMB} = (1 - accuracy)s_{DIS}$$

Additionally, we contend that the bias score for ambiguous contexts should be interpreted from additional dimensions, rather than being entirely dependent on the bias score in disambiguated contexts. To be more specific, in the sample of ambiguous contexts from the BBQ dataset, the context lacks complete information or solid evidence, hence the ground truth is always a non-targeted answer. Therefore, we believe that accuracy in ambiguous contexts, specifically the probability of the model responding with "I don't know", more accurately reflects the model's tendencies when information is insufficient.

### 4.4.2 Metrics on CrowS-Pairs

In the CrowS-Pairs dataset, each sample consists of two sentences: one is biased, and the other is not. Each sentence comprises two parts: the modified part and the unmodified part. For example, in the pair "John ran into his old football friend" versus "Shaniqua ran into her old football friend" the modified tokens are {John, his} for the first sentence and {Shaniqua, her} for the second sentence. The unmodified tokens for both sentences are {ran, into, old, football, friend}. For a sentence $S$, let $U = \{u_0, ..., u_l\}$ be the unmodified tokens, and $M = \{m_0, ..., m_n\}$ be the modified tokens ($S = U \cup M$). Based on these definitions, the pseudo-log-likelihood (Wang and Cho, 2019) of the unmodified tokens conditioned on the modified tokens, $p(U|M, \theta)$ is calculated. This approach differs from $p(M|U, \theta)$ from (Nadeem et al., 2021), primarily because the authors of this dataset believe it can help with avoiding bias caused by the frequent appearance of common names in the training data. The calculation of the score definition is as follows:

$$score(S) = \sum_{i=0}^{|C|} logP(u_i \in U | U_{\setminus u_i}, M, \theta)$$

The pseudo-log-likelihood of all unmodified tokens is calculated iteratively and then summed up as the final score of sentence $S$.

Based on the score of each sentence, we measured 1) the average score difference across all samples and 2) the percentage of examples where the model assigns a higher pseudo-log-likelihood to the stereotyping sentence. These are applied to every bias category.

## 5   Results and Discussion

Results for the CrowS-Pairs benchmark are shown in the heatmap in Figure 2. Numbers shown are the percentage stereotype, we subtracted 50 from all the values, meaning that values greater than 0 indicate a tendency towards the stereotype sentence, while values less than 0 indicate a tendency towards the non-stereotype sentence. The perfect score is 0, where neither sentence is preferred over the other. We find that the multilingual model has scores that are closer to 0 in all languages compared to its monolingual counterpart and also open-source LLMs.

Results for the BBQ benchmark are shown in the heatmap in Figure 3 On the BBQ dataset, Our

| | de-falcon | de-mono | de-multi | en-bert | en-falcon | en-llama | en-mistral | en-mono | en-multi | es-falcon | es-mono | es-multi | fr-falcon | fr-mono | fr-multi | it-falcon | it-mono | it-multi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Socioeconomic Status | -4.7 | 10.5 | -3.2 | 8.0 | 19.5 | 24.2 | 20.0 | 13.2 | -1.6 | 16.3 | 13.2 | 5.8 | 17.4 | 14.2 | -6.8 | 14.2 | 0.4 | 1.6 |
| Gender Identity | 12.5 | 10.6 | 3.8 | 9.9 | 18.4 | 11.6 | 5.9 | 14.1 | -3.8 | 8.8 | 6.2 | -0.3 | 14.4 | 10.9 | 2.2 | 0.6 | 7.2 | -2.2 |
| Nationality | -5.1 | 0.0 | -1.9 | 12.9 | 12.5 | 9.7 | 7.9 | 4.6 | -7.4 | 7.9 | 11.6 | -3.2 | 6.5 | 0.9 | -16.7 | 3.2 | 3.7 | -4.6 |
| Physical Appearance | 6.9 | 11.1 | 1.4 | 13.5 | 29.2 | 27.8 | 27.8 | 23.6 | 1.4 | 15.3 | 12.5 | -6.9 | 11.1 | 13.9 | 2.8 | 16.7 | 12.5 | -5.6 |
| Age | 3.8 | 13.7 | 3.8 | 5.2 | 21.4 | 25.8 | 19.2 | 12.6 | 10.4 | 12.6 | 12.6 | -1.6 | 12.6 | 4.9 | -3.8 | 11.5 | 3.8 | -7.1 |
| Disability Status | 3.8 | 0.8 | -14.6 | 11.7 | 26.9 | 17.7 | 25.4 | 20.8 | -6.9 | 10.0 | 17.7 | -10.0 | 11.5 | 10.0 | -6.9 | 13.1 | 13.1 | -10.0 |
| Race/Ethnicity | 2.4 | 6.1 | -1.2 | 8.1 | 10.0 | 11.4 | 11.8 | 1.4 | -2.4 | 10.4 | 13.0 | 1.8 | 8.3 | 6.7 | -5.7 | -1.6 | 0.4 | -16.1 |
| Religion | 10.4 | 6.8 | 1.4 | 21.4 | 26.6 | 29.3 | 32.0 | 20.3 | -1.4 | 10.4 | 23.0 | -14.0 | 29.3 | 22.1 | 11.3 | 16.7 | 14.9 | -23.0 |
| Sexual Orientation | 27.4 | 28.5 | 22.0 | 17.9 | 38.2 | 29.6 | 37.1 | 32.8 | 12.4 | 25.3 | 32.8 | 18.8 | 34.9 | 34.9 | 37.1 | 22.0 | 34.9 | 19.9 |
| Average | 4.9 | 8.6 | 0.9 | 10.5 | 17.9 | 16.7 | 15.6 | 10.9 | -1.6 | 11.7 | 13.5 | 0.1 | 13.7 | 10.6 | -1.7 | 5.9 | 7.8 | -7.1 |

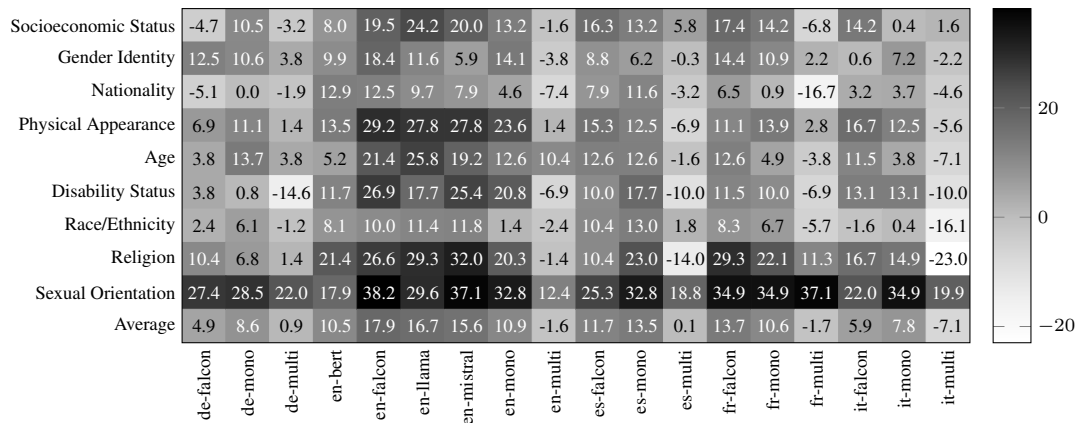Higher than 0. The model leans toward stereotypes

Figure 2: Heat map of CrowSPairs bias percentage scores using our models and open-source models. A perfect score would be 0 which represents an equal probability of choosing either sentence. The microaverage is computed across all categories based on frequency. Our multilingual model has less bias than monolingual models and open-source LLMs (the likelihood assigned to the non-stereotyping sentence is higher).

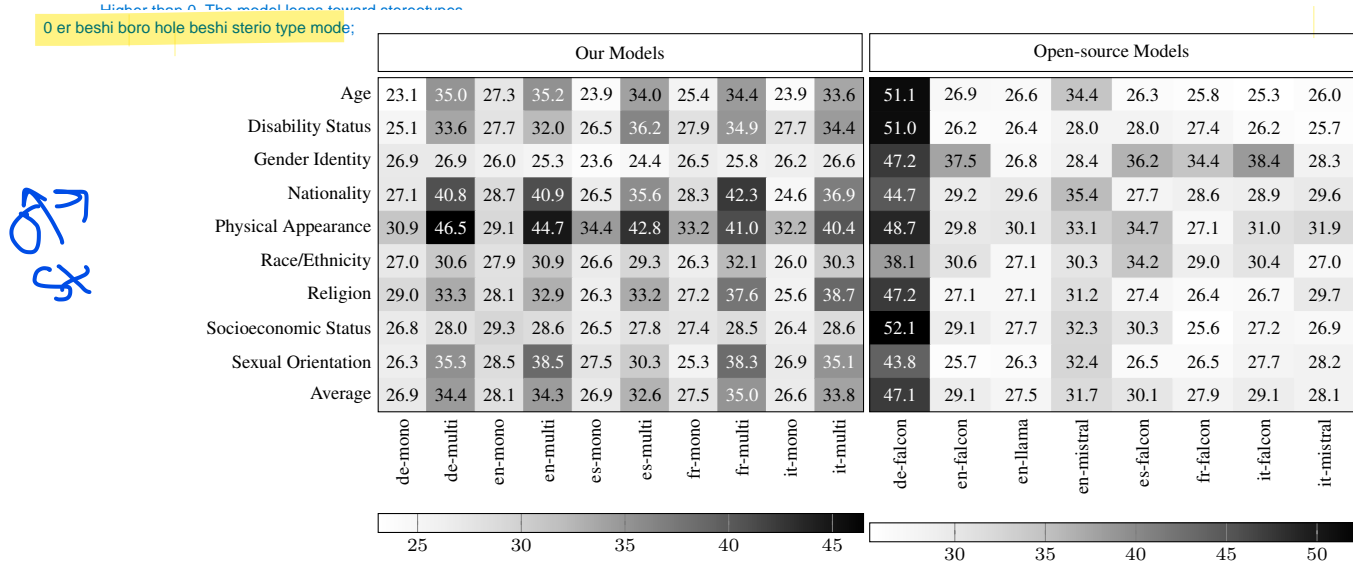| | Our Models | | | | | | | | | | Open-source Models | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | de-mono | de-multi | en-mono | en-multi | es-mono | es-multi | fr-mono | fr-multi | it-mono | it-multi | de-falcon | en-falcon | en-llama | en-mistral | es-falcon | fr-falcon | it-falcon | it-mistral |
| Age | 23.1 | 35.0 | 27.3 | 35.2 | 23.9 | 34.0 | 25.4 | 34.4 | 23.9 | 33.6 | 51.1 | 26.9 | 26.6 | 34.4 | 26.3 | 25.8 | 25.3 | 26.0 |
| Disability Status | 25.1 | 33.6 | 27.7 | 32.0 | 26.5 | 36.2 | 27.9 | 34.9 | 27.7 | 34.4 | 51.0 | 26.2 | 26.4 | 28.0 | 28.0 | 27.4 | 26.2 | 25.7 |
| Gender Identity | 26.9 | 26.9 | 26.0 | 25.3 | 23.6 | 24.4 | 26.5 | 25.8 | 26.2 | 26.6 | 47.2 | 37.5 | 26.8 | 28.4 | 36.2 | 34.4 | 38.4 | 28.3 |
| Nationality | 27.1 | 40.8 | 28.7 | 40.9 | 26.5 | 35.6 | 28.3 | 42.3 | 24.6 | 36.9 | 44.7 | 29.2 | 29.6 | 35.4 | 27.7 | 28.6 | 28.9 | 29.6 |
| Physical Appearance | 30.9 | 46.5 | 29.1 | 44.7 | 34.4 | 42.8 | 33.2 | 41.0 | 32.2 | 40.4 | 48.7 | 29.8 | 30.1 | 33.1 | 34.7 | 27.1 | 31.0 | 31.9 |
| Race/Ethnicity | 27.0 | 30.6 | 27.9 | 30.9 | 26.6 | 29.3 | 26.3 | 32.1 | 26.0 | 30.3 | 38.1 | 30.6 | 27.1 | 30.3 | 34.2 | 29.0 | 30.4 | 27.0 |
| Religion | 29.0 | 33.3 | 28.1 | 32.9 | 26.3 | 33.2 | 27.2 | 37.6 | 25.6 | 38.7 | 47.2 | 27.1 | 27.1 | 31.2 | 27.4 | 26.4 | 26.7 | 29.7 |
| Socioeconomic Status | 26.8 | 28.0 | 29.3 | 28.6 | 26.5 | 27.8 | 27.4 | 28.5 | 26.4 | 28.6 | 52.1 | 29.1 | 27.7 | 32.3 | 30.3 | 25.6 | 27.2 | 26.9 |
| Sexual Orientation | 26.3 | 35.3 | 28.5 | 38.5 | 27.5 | 30.3 | 25.3 | 38.3 | 26.9 | 35.1 | 43.8 | 25.7 | 26.3 | 32.4 | 26.5 | 26.5 | 27.7 | 28.2 |
| Average | 26.9 | 34.4 | 28.1 | 34.3 | 26.9 | 32.6 | 27.5 | 35.0 | 26.6 | 33.8 | 47.1 | 29.1 | 27.5 | 31.7 | 30.1 | 27.9 | 29.1 | 28.1 |

Figure 3: Heat map of BBQ overall accuracy using our monolingual and multilingual models (left) as well as the open-source models (right). Our multilingual model is better than monolingual models in all languages and surpasses most of the open-source LLMs.

multilingual model also has better overall accuracy compared to their monolingual counterparts, and also better than most of the open-source LLMs across languages. Falcon outperforms the other open-source models and, in the case of German, outperforms our models. The high performance of the model, in particular for gender identity and the German language is difficult to determine, but may be attributed to the filtering done to construct the RefinedWeb corpus on which it was trained (Penedo et al., 2023).

Breaking down the accuracy on the BBQ dataset in Figure 4, we can also compare the accuracy of ambiguous and disambiguated contexts. we can observe that on the accuracy of ambiguous con-

text, the multilingual model does much better than the monolingual models, while on the accuracy of disambiguated contexts, performance drops. The mixture of languages in the training data for the multilingual model seems to make it more conservative, hence the model is more likely to respond with "I don't know" when the information is insufficient, but this nature also causes loss of accuracy when dealing with the disambiguated samples (where the answer is always known).

However, after balancing the two sides, the final outcome is favorable for our multilingual model. In Parrish et al. (2022), their UnifiedQA model reached the average ambiguous accuracy of 60.8% and average disambiguated accuracy of 91.4%. The
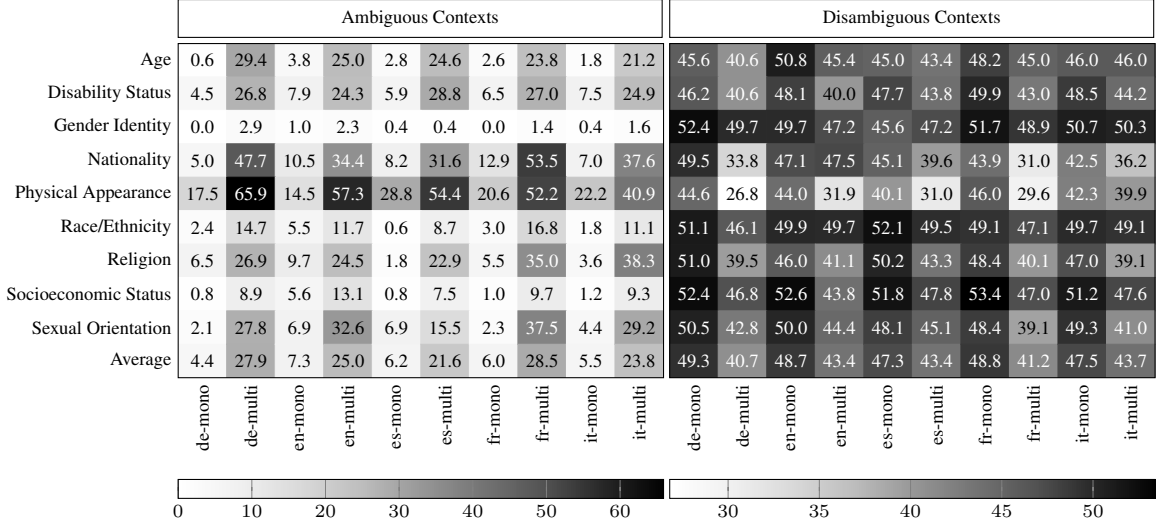
| | Ambiguous Contexts | | | | | | | | | | Disambiguous Contexts | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | de-mono | de-multi | en-mono | en-multi | es-mono | es-multi | fr-mono | fr-multi | it-mono | it-multi | de-mono | de-multi | en-mono | en-multi | es-mono | es-multi | fr-mono | fr-multi | it-mono | it-multi |
| Age | 0.6 | 29.4 | 3.8 | 25.0 | 2.8 | 24.6 | 2.6 | 23.8 | 1.8 | 21.2 | 45.6 | 40.6 | 50.8 | 45.4 | 45.0 | 43.4 | 48.2 | 45.0 | 46.0 | 46.0 |
| Disability Status | 4.5 | 26.8 | 7.9 | 24.3 | 5.9 | 28.8 | 6.5 | 27.0 | 7.5 | 24.9 | 46.2 | 40.6 | 48.1 | 40.0 | 47.7 | 43.8 | 49.9 | 43.0 | 48.5 | 44.2 |
| Gender Identity | 0.0 | 2.9 | 1.0 | 2.3 | 0.4 | 0.4 | 0.0 | 1.4 | 0.4 | 1.6 | 52.4 | 49.7 | 49.7 | 47.2 | 45.6 | 47.2 | 51.7 | 48.9 | 50.7 | 50.3 |
| Nationality | 5.0 | 47.7 | 10.5 | 34.4 | 8.2 | 31.6 | 12.9 | 53.5 | 7.0 | 37.6 | 49.5 | 33.8 | 47.1 | 47.5 | 45.1 | 39.6 | 43.9 | 31.0 | 42.5 | 36.2 |
| Physical Appearance | 17.5 | 65.9 | 14.5 | 57.3 | 28.8 | 54.4 | 20.6 | 52.2 | 22.2 | 40.9 | 44.6 | 26.8 | 44.0 | 31.9 | 40.1 | 31.0 | 46.0 | 29.6 | 42.3 | 39.9 |
| Race/Ethnicity | 2.4 | 14.7 | 5.5 | 11.7 | 0.6 | 8.7 | 3.0 | 16.8 | 1.8 | 11.1 | 51.1 | 46.1 | 49.9 | 49.7 | 52.1 | 49.5 | 49.1 | 47.1 | 49.7 | 49.1 |
| Religion | 6.5 | 26.9 | 9.7 | 24.5 | 1.8 | 22.9 | 5.5 | 35.0 | 3.6 | 38.3 | 51.0 | 39.5 | 46.0 | 41.1 | 50.2 | 43.3 | 48.4 | 40.1 | 47.0 | 39.1 |
| Socioeconomic Status | 0.8 | 8.9 | 5.6 | 13.1 | 0.8 | 7.5 | 1.0 | 9.7 | 1.2 | 9.3 | 52.4 | 46.8 | 52.6 | 43.8 | 51.8 | 47.8 | 53.4 | 47.0 | 51.2 | 47.6 |
| Sexual Orientation | 2.1 | 27.8 | 6.9 | 32.6 | 6.9 | 15.5 | 2.3 | 37.5 | 4.4 | 29.2 | 50.5 | 42.8 | 50.0 | 44.4 | 48.1 | 45.1 | 48.4 | 39.1 | 49.3 | 41.0 |
| Average | 4.4 | 27.9 | 7.3 | 25.0 | 6.2 | 21.6 | 6.0 | 28.5 | 5.5 | 23.8 | 49.3 | 40.7 | 48.7 | 43.4 | 47.3 | 43.4 | 48.8 | 41.2 | 47.5 | 43.7 |

Figure 4: Heat map of BBQ accuracies for our monolingual and multilingual model. The left side shows accuracy for the ambiguous contexts, while the right shows accuracy for the disambiguated contexts. Our multilingual model has much higher accuracy in ambiguous contexts, but slightly lower for disambiguated contexts.

large difference in performance is likely due to first fine-tuning their model on the RACE dataset (Lai et al., 2017b), which is also a text-based multiple-choice dataset. The fine-tuning helped make their model familiar with the QA format. For a fairer comparison, we do not fine-tune any models on the QA task and the results from open-source models are on par with our results.

Additionally, some papers also evaluated models in a zero-shot setting on the BBQ dataset. Shaikh et al. (2022) with GPT3 got 55.73% accuracy over-all. Si et al. (2022) with an instruction fine-tuned version of GPT3, Text-Davinci-001 got 60.5% and 43.2% for ambiguous and disambiguated context, respectively. One notable comparison is the parameter size difference between our models and GPT3. While GPT3 has 150B parameters, ours only have 2.7B. Our models achieve lower accuracy at 20.83% lower than GPT3, yet surpassing Falcon-7B by 1.54% across all 5 languages, LLama2-7B by 6.9% on English, Mistral-7B by 3.8% on English, on average. Due to limited computational resources, we cannot perform this comparison at 150B parameters and leave a controlled exploration of the relationship between bias and parameter size to future work.

## 6 Validity of The Models

To validate our model's capabilities beyond bias evaluation, we additionally conducted tests on the Belebele benchmark (Bandarkar et al., 2023), a common sense-based multiple-choice question-answering dataset designed to test the model's understanding capabilities in different language contexts. To fit our model, we also reformulated this dataset into QA format.

The model's results are shown in Table 4. All the data in the table including those from other papers, were obtained under the zero-shot setting. Additionally, the inference method is consistent with the BBQ method described in Section 4.1.

From the Belebele results, the monolingual models generally perform better than the multilingual model. This may be due to the fact that during the training of the multilingual model, the data for each language is only 20% of that for the corresponding monolingual model, leading to insufficient commonsense knowledge. However, given that our data-controlled models have less than half the parameters compared to other open-source models, our LLM benchmark results are satisfactory.

## 7 Conclusion

In this work, we systematically explored the relationship between the language of data a large language model is trained on and the stereotype bias that is encoded in the model. We trained six models with around 2.7B parameters from scratch using a causal language modeling objective and evaluated them on the CrowS-Pairs and BBQ benchmarks for English, French, German, Italian, and Spanish. To ensure that our approach can be extended to other languages and benchmarks, the datasets were automatically translated. For quality assur-

| Model | Parameter Size | Language | Acc |
|---|---|---|---|
| en-mono | 2.6B | English | 31.7 |
| de-mono | 2.6B | German | 35.3 |
| fr-mono | 2.6B | French | 35.1 |
| es-mono | 2.6B | Spanish | 35.2 |
| it-mono | 2.6B | Italian | 33.3 |
| en-multi | 2.7B | English | 27.0 |
| de-multi | 2.7B | German | 27.8 |
| fr-multi | 2.7B | French | 30.0 |
| es-multi | 2.7B | Spanish | 27.8 |
| it-multi | 2.7B | Italian | 27.2 |
| Mistral | 7B | English | 45.9 |
| Llama-2 | 7B | English | 40.9 |
| Falcon | 7B | English | 35.1 |
| Falcon | 7B | German | 33.1 |
| Falcon | 7B | French | 39.0 |
| Falcon | 7B | Spanish | 31.3 |
| Falcon | 7B | Italian | 30.9 |
| Llama-2-CHAT (Bandarkar et al., 2023) | 70B | Multilingual | 41.5 |
| GPT3.5-TURBO (Bandarkar et al., 2023) | unk | Multilingual | 51.1 |

Table 4: The accuracy of all tested models on the Belebele (Bandarkar et al., 2023). The results from Llama-2-CHAT and GPT3.5-TURBO on Belebele are the average results from all available languages in Bandarkar et al. (2023).

ance, a sample of the translations was evaluated by humans, who generally found that the translation quality was high and biases were preserved. We found that multilingual models trained on the same number of tokens as monolingual models were less biased for all languages and both benchmarks than the monolingual models. We also found that our models were generally less biased than selected open-source LLMs which had 7B parameters, though they fall short of zero-shot prompt-based approaches with GPT3. Publicly released material for our experiments can be found under http://lamarr-institute.org/research/natural-language-processing/.

## Limitations

In our work, we use machine translation to evaluate monolingual and multilingual models across multiple languages. Using machine translation might affect the quality and the expression of bias of the translated datasets. By evaluating the translation process with human evaluators as described in §3.2, we aim to reduce these effects. Nevertheless, we are aware that the small number of annotators might decrease the significance of our results as in particular the evaluation of the bias in the translation is influenced by the perception of the annotator. In future work, we aim to extend this evaluation to all the studied languages and to more native annotators and methods that can ensure the quality of the automated translations.

The biases that exist in the benchmarks we used may be specific to English speaking regions. When translating the benchmark, bias may decrease because the biases that manifest in the translated language are specific to the regions that speak that language, which might not be the same as English speaking regions. Future work should consider creating new bias benchmarks for each language that represent the biases of the populations that speak those languages. Without this, we cannot be sure that the translated benchmarks cover the biases that are likely to occur in a given language. The significance of our results might be limited by CrowS-pairs quality as shown in Blodgett et al. (2021). Blodgett et al. (2021) finds that 97% of the dataset are not admissible. Generating a french version of CrowS-pairs, also Névéol et al. (2022) scrutinizes and even improves the original CrowS-pairs dataset. They present the statistics of the different adaptation types (compare Table 2 in (Névéol et al., 2022). In addition to the sentences modified to

suit the French culture, 150 samples in total (10% of the dataset) were adapted due to the identified limitations within the original CrowS-pairs dataset (non-minimal pairs (22), double switches (64) or bias-type mismatches (64)). Even if the findings of (Blodgett et al., 2021) show severe shortcomings, we decided on using CrowS-Pairs due to its broad usage in the literature and its coverage of many different bias categories and social groups. The findings of Liu (2024) prove at least significant differences between the stereotype and anti-stereotype sentence pairs. Within our own sampled evaluation also only a small rate of sentences needed to be excluded in general. To validate our findings despite of the ambiguities, we used BBQ as a second benchmarking dataset. In future work, we plan to extend the experiments to other datasets, such as the published revised version of CrowS-pairs (Névéol et al., 2022) or the HONEST dataset (Nozza et al., 2021). Moreover, since the languages involved in this paper are all European languages, their high similarity may lead to certain stereotypical knowledge being shared, making it easier for stereotypes to transfer between languages.

## Acknowledgments

## References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In Harald Lüngen, Marc Kupietz, Piotr Bański, Adrien Barbaresi, Simon Clematide, and Ines Pisetta, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*, pages 1 – 9. Leibniz-Institut für Deutsche Sprache, Mannheim.

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *Preprint*, arXiv:2201.06642.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, et al. 2023. Tokenizer choice for llm training: Negligible or crucial? *arXiv preprint arXiv:2310.08754*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi, and Sheema Usmani. 2021. Gender bias in online language translators: Visualization, human perception, and bias/accuracy tradeoffs. *IEEE Internet Computing*, 25(5):53–63.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

*Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. ArXiv:1608.07187 [cs].

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Kate Crawford. 2017. The trouble with bias. http://youtube.com/watch?v=fMym_BKWQzk. Talk given at NeurIPS December 2017.

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.

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.

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. *arXiv preprint*. ArXiv:2309.00770 [cs].

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Johannes Graën, Tannon Kew, Anastassia Shaitarova, and Martin Volk. 2019. Modelling large parallel corpora: The zurich parallel corpus collection. In *Proceedings of the 7th Workshop on Challenges in the Management of Large Corpora (CMLC)*, pages 1–8. Leibniz-Institut für Deutsche Sprache.

J. Graën, D. Batinic, and M. Volk. 2014. Cleaning the Europarl corpus for linguistic applications. In *Konvens 2014*. Stiftung Universität Hildesheim.

Najeh Hajlaoui, David Kolovratnik, Jaakko Vaeyrynen, Ralf Steinberger, and Dániel Varga. 2014. DCEP - Digital corpus of the European parliament. In *Proc. LREC 2014 (Language Resources and Evaluation Conference). Reykjavik, Iceland*, pages 3164–3171.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Stefan Höfler and Michael Piotrowski. 2011. Building corpora for the philological study of Swiss legal texts. *Journal for Language Technology and Computational Linguistics*, 26(2):77–89.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2023. Kobbq: Korean bias benchmark for question answering. *arXiv preprint arXiv:2307.16778*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit, volume 5*, pages 79—86. Asia-Pacific Association for Machine Translation (AAMT).

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017a. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017b. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing Biases and the Impact of Multilingual Training across Multiple Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280, Singapore. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*.

Yang Liu. 2024. Quantifying stereotypes in language. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1223–1240, St. Julian's, Malta. Association for Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

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. 2023a. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Roberto Navigli, Simone Conia, and Björn Ross. 2023b. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):1–21.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

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.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 385–388, New York, NY, USA. Association for Computing Machinery.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

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.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Luca Soldaini and Kyle Lo. 2023. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI. ODC-By, `https://github.com/allenai/pes2o`.

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

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models. *Preprint*, arXiv:2010.06032.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Xiaoxian Yang, Zhifeng Wang, Qi Wang, Ke Wei, Kaiqi Zhang, and Jiangang Shi. 2024. Large language models for automated q&a involving legal documents: a survey on algorithms, frameworks and applications. *International Journal of Web Information Systems*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. 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 5276–5284, Hong Kong, China. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

## A  BBQ Bias Scores

Here we present the bias scores for the BBQ dataset covering the nine demographic attributes. Figure 5 we show the bias scores for the monolingual and multilingual LLMs that we trained and in Figure 6 we show the scores for the open-source models.

## B  Intrinsic Evaluation of the LLMs

The training losses of all six mono- and multilingual models are depicted in Figure 7 in the Appendix. Additionally, in Figure 8, we illustrate that during training, all models consistently decrease to a perplexity of approximately $10 \pm 2.5$ on a holdout validation set, with slight variations observed depending on the language. As all models use different tokenizers, the training loss and the validation perplexity are not directly comparable to each other. Also, the curated corpora, and therefore the training- and validation sets differ slightly depending on the language. Nonetheless, all models show a consistent improvement during training.

## C  Datasets

Our web documents in the corpora consist of Oscars[3] (Abadji et al., 2021), that were generated by the ungoliant pipeline[4] based on 20 Common Crawl WET Archives (2014-42, 2015-14, 2015-48, 2016-22, 2016-43, 2017-13, 2017-47, 2018-30, 2018-47, 2019-22, 2020-24, 2020-45, 2021-31, 2021-49, 2022-27, 2022-40, 2022-49, 2023-06, and 2023-14).

The curated datasets consist of *The Pile* (Gao et al., 2020), *RedPajama* (Computer, 2023), and single datasets that do not belong to a collection. From the Pile subcorpora, we selected: Phil Archive, PMC Abstracts, PMC Extracts, Open-WebText, NIH Exporter, and Free Law Opinions V2. From RedPajama we use Books and StackExchange.

The remaining datasets are:

1. The Wikimedia dump of 2023-09-01[5]

2. All the News V2.0[6] is a corpus of newspaper articles crawled from over 26 different publications from January 2016 to April 1, 2020.

3. CoStEP[7] is a cleaned-up and corrected version of the EuroParl corpus(Graën et al., 2014)(Koehn, 2005)

4. DCEP[8] is a companion corpus to CoStEP, containing documents published by the European Parliament. (Hajlaoui et al., 2014)

5. Dissertations[9] is a collection of dissertations from the Deutsche Nationalbibliothek.

6. MAREC/IREC[10]: The MAtrixware REsearch Collection / The Information retrieval facility Research Collection is a patent corpus of over 19 million documents from the EP, WO, US, and JP patent offices.

7. Medi-Notice[11] is part of the Zurich Parallel Corpus Collection. It is a multilingual corpus compiled from information leaflets for medications and pharmaceutical products published by the Swiss Agency for Therapeutic Products.(Graën et al., 2019)

8. Swiss Policy[12] contains documents of the Swiss Legislation Corpus (Höfler and Piotrowski, 2011)

9. OpenSubtitles 2018[13,14] is a collection of translated movie subtitles.

10. The peS2o (Soldaini and Lo, 2023) dataset is a collection of 40M creative open-access academic papers, cleaned, filtered, and formatted for pre-training of language models (Lison and Tiedemann, 2016)

11. The EUR-Lex dataset[15] is a multilingual collection of case laws, decisions, directives, recommendations, regulations, and proposals of the European Union.

---

**Figure 5**

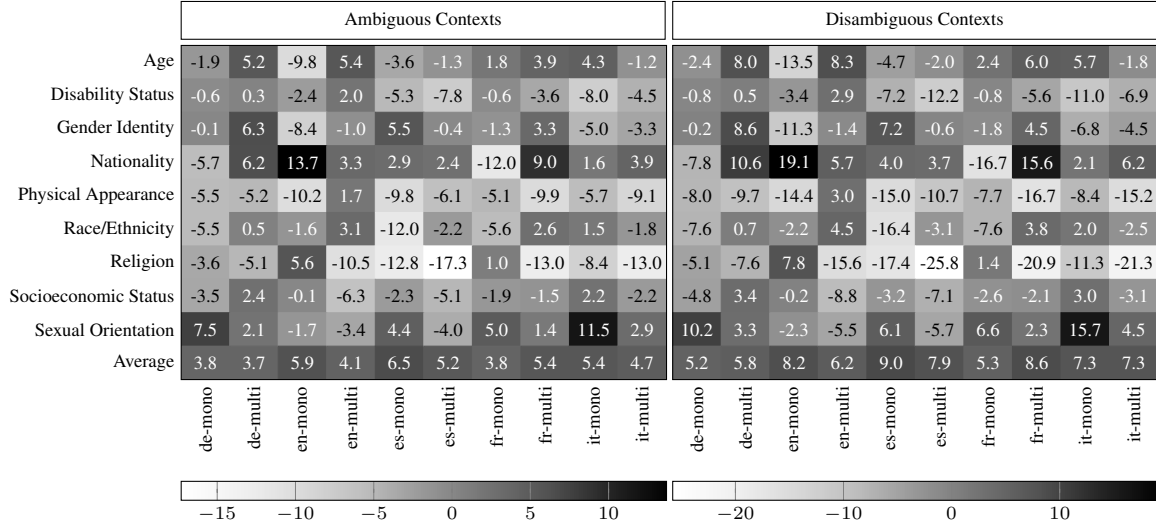| | Ambiguous Contexts | | | | | | | | | | Disambiguous Contexts | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | de-mono | de-multi | en-mono | en-multi | es-mono | es-multi | fr-mono | fr-multi | it-mono | it-multi | de-mono | de-multi | en-mono | en-multi | es-mono | es-multi | fr-mono | fr-multi | it-mono | it-multi |
| Age | -1.9 | 5.2 | -9.8 | 5.4 | -3.6 | -1.3 | 1.8 | 3.9 | 4.3 | -1.2 | -2.4 | 8.0 | -13.5 | 8.3 | -4.7 | -2.0 | 2.4 | 6.0 | 5.7 | -1.8 |
| Disability Status | -0.6 | 0.3 | -2.4 | 2.0 | -5.3 | -7.8 | -0.6 | -3.6 | -8.0 | -4.5 | -0.8 | 0.5 | -3.4 | 2.9 | -7.2 | -12.2 | -0.8 | -5.6 | -11.0 | -6.9 |
| Gender Identity | -0.1 | 6.3 | -8.4 | -1.0 | 5.5 | -0.4 | -1.3 | 3.3 | -5.0 | -3.3 | -0.2 | 8.6 | -11.3 | -1.4 | 7.2 | -0.6 | -1.8 | 4.5 | -6.8 | -4.5 |
| Nationality | -5.7 | 6.2 | 13.7 | 3.3 | 2.9 | 2.4 | -12.0 | 9.0 | 1.6 | 3.9 | -7.8 | 10.6 | 19.1 | 5.7 | 4.0 | 3.7 | -16.7 | 15.6 | 2.1 | 6.2 |
| Physical Appearance | -5.5 | -5.2 | -10.2 | 1.7 | -9.8 | -6.1 | -5.1 | -9.9 | -5.7 | -9.1 | -8.0 | -9.7 | -14.4 | 3.0 | -15.0 | -10.7 | -7.7 | -16.7 | -8.4 | -15.2 |
| Race/Ethnicity | -5.5 | 0.5 | -1.6 | 3.1 | -12.0 | -2.2 | -5.6 | 2.6 | 1.5 | -1.8 | -7.6 | 0.7 | -2.2 | 4.5 | -16.4 | -3.1 | -7.6 | 3.8 | 2.0 | -2.5 |
| Religion | -3.6 | -5.1 | 5.6 | -10.5 | -12.8 | -17.3 | 1.0 | -13.0 | -8.4 | -13.0 | -5.1 | -7.6 | 7.8 | -15.6 | -17.4 | -25.8 | 1.4 | -20.9 | -11.3 | -21.3 |
| Socioeconomic Status | -3.5 | 2.4 | -0.1 | -6.3 | -2.3 | -5.1 | -1.9 | -1.5 | 2.2 | -2.2 | -4.8 | 3.4 | -0.2 | -8.8 | -3.2 | -7.1 | -2.6 | -2.1 | 3.0 | -3.1 |
| Sexual Orientation | 7.5 | 2.1 | -1.7 | -3.4 | 4.4 | -4.0 | 5.0 | 1.4 | 11.5 | 2.9 | 10.2 | 3.3 | -2.3 | -5.5 | 6.1 | -5.7 | 6.6 | 2.3 | 15.7 | 4.5 |
| Average | 3.8 | 3.7 | 5.9 | 4.1 | 6.5 | 5.2 | 3.8 | 5.4 | 5.4 | 4.7 | 5.2 | 5.8 | 8.2 | 6.2 | 9.0 | 7.9 | 5.3 | 8.6 | 7.3 | 7.3 |

Figure 5: Heat map of BBQ biases using our monolingual and multilingual models. The left side shows bias for the ambiguous contexts, while the right shows bias scores for the disambiguous contexts.

**Figure 6**

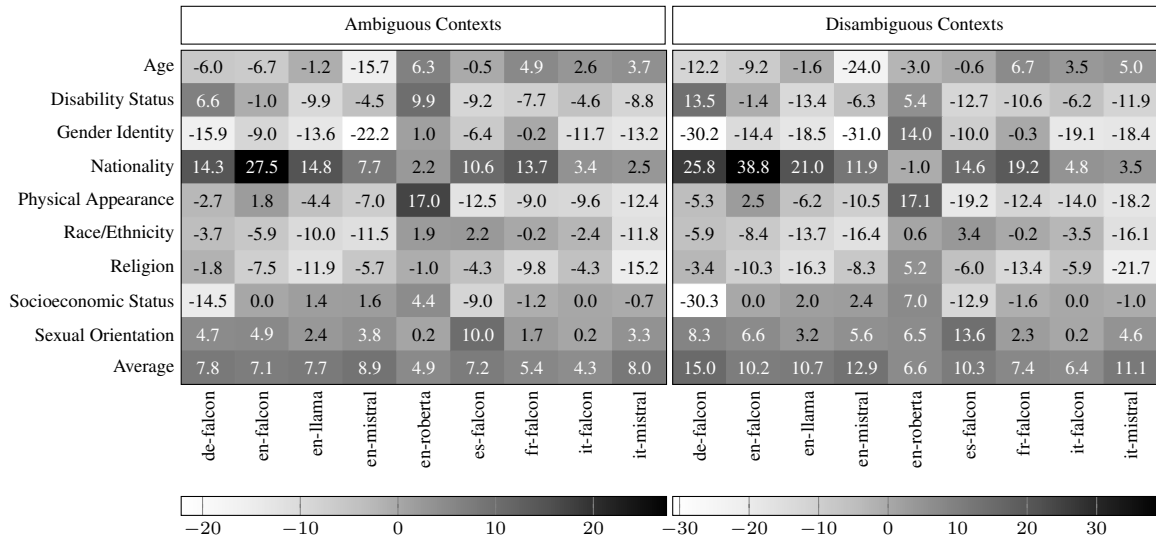| | Ambiguous Contexts | | | | | | | | | Disambiguous Contexts | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | de-falcon | en-falcon | en-llama | en-mistral | en-roberta | es-falcon | fr-falcon | it-falcon | it-mistral | de-falcon | en-falcon | en-llama | en-mistral | en-roberta | es-falcon | fr-falcon | it-falcon | it-mistral |
| Age | -6.0 | -6.7 | -1.2 | -15.7 | 6.3 | -0.5 | 4.9 | 2.6 | 3.7 | -12.2 | -9.2 | -1.6 | -24.0 | -3.0 | -0.6 | 6.7 | 3.5 | 5.0 |
| Disability Status | 6.6 | -1.0 | -9.9 | -4.5 | 9.9 | -9.2 | -7.7 | -4.6 | -8.8 | 13.5 | -1.4 | -13.4 | -6.3 | 5.4 | -12.7 | -10.6 | -6.2 | -11.9 |
| Gender Identity | -15.9 | -9.0 | -13.6 | -22.2 | 1.0 | -6.4 | -0.2 | -11.7 | -13.2 | -30.2 | -14.4 | -18.5 | -31.0 | 14.0 | -10.0 | -0.3 | -19.1 | -18.4 |
| Nationality | 14.3 | 27.5 | 14.8 | 7.7 | 2.2 | 10.6 | 13.7 | 3.4 | 2.5 | 25.8 | 38.8 | 21.0 | 11.9 | -1.0 | 14.6 | 19.2 | 4.8 | 3.5 |
| Physical Appearance | -2.7 | 1.8 | -4.4 | -7.0 | 17.0 | -12.5 | -9.0 | -9.6 | -12.4 | -5.3 | 2.5 | -6.2 | -10.5 | 17.1 | -19.2 | -12.4 | -14.0 | -18.2 |
| Race/Ethnicity | -3.7 | -5.9 | -10.0 | -11.5 | 1.9 | 2.2 | -0.2 | -2.4 | -11.8 | -5.9 | -8.4 | -13.7 | -16.4 | 0.6 | 3.4 | -0.2 | -3.5 | -16.1 |
| Religion | -1.8 | -7.5 | -11.9 | -5.7 | -1.0 | -4.3 | -9.8 | -4.3 | -15.2 | -3.4 | -10.3 | -16.3 | -8.3 | 5.2 | -6.0 | -13.4 | -5.9 | -21.7 |
| Socioeconomic Status | -14.5 | 0.0 | 1.4 | 1.6 | 4.4 | -9.0 | -1.2 | 0.0 | -0.7 | -30.3 | 0.0 | 2.0 | 2.4 | 7.0 | -12.9 | -1.6 | 0.0 | -1.0 |
| Sexual Orientation | 4.7 | 4.9 | 2.4 | 3.8 | 0.2 | 10.0 | 1.7 | 0.2 | 3.3 | 8.3 | 6.6 | 3.2 | 5.6 | 6.5 | 13.6 | 2.3 | 0.2 | 4.6 |
| Average | 7.8 | 7.1 | 7.7 | 8.9 | 4.9 | 7.2 | 5.4 | 4.3 | 8.0 | 15.0 | 10.2 | 10.7 | 12.9 | 6.6 | 10.3 | 7.4 | 6.4 | 11.1 |

Figure 6: Heat map of BBQ biases using open source models. The left side shows bias for the ambiguous contexts, while the right shows bias scores for the disambiguous contexts.
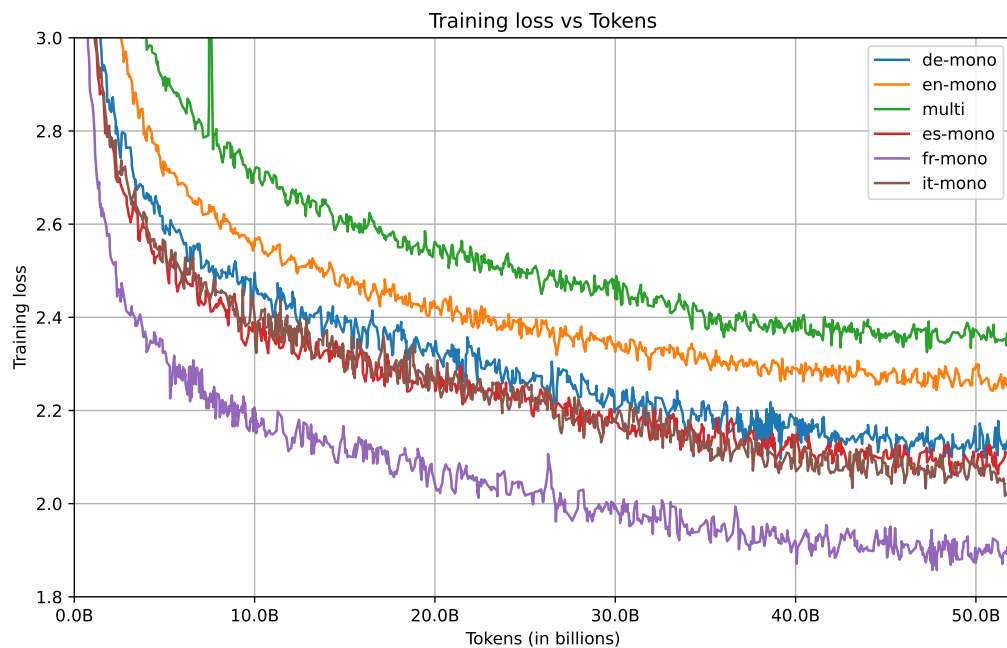
Figure 7: The plot shows the training loss per tokens for the monolingual and multilingual models.
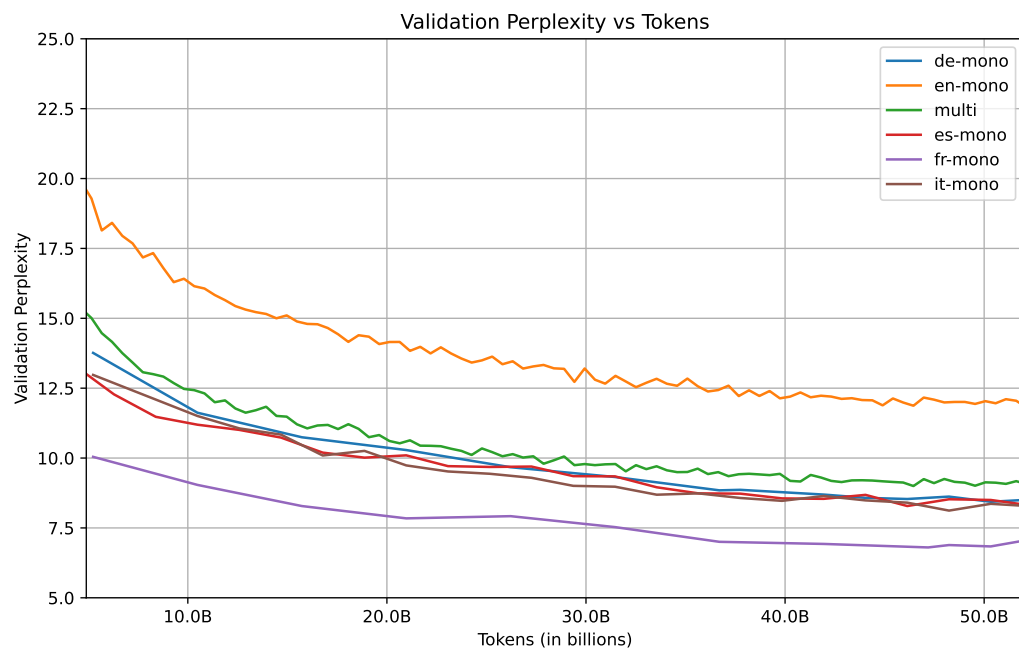


Figure 8: The plot shows the validation perplexity per tokens for the monolingual and multilingual models.

12. Bundestag - Plenarprotokolle[16] comprises transcripts of sessions of the German Bundestag.

13. Bundestag - Drucksachen[17] contains all bills that are negotiated in the Bundestag.

14. Bundesgerichtshof - Entscheidungen[18] is a collection of decisions of the German Federal Court.

15. German legal cases contain German court decisions and the corresponding citation network(Ostendorff et al., 2020).

---

[16] https://www.bundestag.de/dokumente/protokolle/plenarprotokolle
[17] https://www.bundestag.de/drucksachen
[18] https://www.bundesgerichtshof.de/DE/Entscheidungen/entscheidungen_node.html

| Source | French | Spanish | Italian | German | English |
|---|---|---|---|---|---|
| OSCAR | 67,015,753,339 | 82,837,352,642 | 33,071,482,584 | 75,706,524,323 | 839,963,018,551 |
| wm_wikisource | 12,988,728 | 37,410,708 | 29,544,756 | 2,692,741 | 367,439,571 |
| wm_wikipedia | 857,581,175 | 741,118,908 | 541,125,604 | 954,833,450 | 2,564,847,030 |
| wm_wikibooks | 7,815,084 | 6,663,686 | 12,404,472 | 6,887,881 | 49,415,989 |
| wm_wikinews | 975,592 | 3,185,339 | 1,140,250 | 2,286,078 | 6,365,015 |
| wm_wikivoyage | 2,565,645 | 4,385,308 | 5,185,341 | 8,509,482 | 19,080,823 |
| pile_openwebtext2 | 104,372,804 | 114,879,971 | 49,069,122 | 89,603,385 | 10,146,045,156 |
| pile_pmc_extracts | 7,907,869 | 6,286,202 | 235,112 | 6,718,264 | 12,140,605,892 |
| pile_pmc_abstracts | 80,031 | 112,119 | 5,504,671 | 87,948 | 3,111,690,781 |
| pile_nih_exporter | - | - | - | - | 303,366,349 |
| pile_v2_philarchive | 10,340,245 | 30,992,077 | 14,778,488 | 8,523,507 | 328,042,520 |
| pile_v2_freelaw | - | - | - | - | 10,401,621,085 |
| rp_book | 292,138,590 | 237,135,131 | 68,968,376 | 66,016,756 | 16,444,915,334 |
| rp_stackexchange | 488,250 | 46,343,855 | 254,003 | 530,997 | 7,522,581,967 |
| marec_irec | 1,431,629,251 | 29,607,774 | 11,569 | 2,135,066,541 | 7,524,414,926 |
| dcep | 93,782,213 | 90,816,394 | 84,386,513 | 75,058,889 | 98,615,360 |
| pes2o | 1,099,711 | 165,370 | 43,128 | 172,599 | 42,203,308,709 |
| allthenews | 107,250 | 1,724,157 | 36,697 | 24,150 | 1,394,745,801 |
| dissertations | 5,765,763 | 12,711,847 | 5,504,671 | 802,610,026 | 3,222,585,878 |
| opensubtitles2018 | 46,811,431 | 46,811,431 | 29,675,610 | 23,502,394 | 84,686,545 |
| medi_notice | 25,105,375 | - | 6,840,687 | 19,659,873 | - |
| swiss_policy | 177,783,858 | - | 31,041,467 | 352,783,813 | - |
| costep | 41,337,687 | 41,667,792 | 38,395,535 | 36,017,291 | 41,435,877 |
| eurlex | 917,636,855 | 81,5163,256 | 856,298,092 | 782,332,455 | 862,491,674 |
| bt_plenarprotokolle | - | - | - | 226,030,395 | - |
| bt_drucksachen | - | - | - | 929,440,378 | - |
| bgh_entscheidungen | - | - | - | 100,384,663 | - |
| german_legal_cases | - | - | - | 749,409,675 | - |

Table 5: Amount of words per dataset for the monolingual models.

| Hyperparameter | Value |
| --- | --- |
| seq_length | 2048 |
| gr_clip_mode | p2_norm |
| gr_clip_thres. | 1.0 |
| num_tokens | 57B |
| learning_rate | 6e-5 |
| betas | [0.9, 0.95] |
| eps | 1e-8 |
| weight_decay | 1e-1 |
| precision | BF_16 |
| vocab_size_mono | 32,768 |
| vocab_size_multi | 100,352 |
| n_layer | 32 |
| n_head_qkv | 32 |
| ffn_hidden | 6656 |
| n_embd | 2560 |
| dropout | false |
| epsilon | 1e-5 |
| linear_biases | false |
| activation_function | swiglu |

Table 6: Hyperparamters of the mono- and multilingual 2.6B parameter models.