Understanding the Interplay of Scale, Data, and Bias in Language Models: A Case Study with BERT

Muhammad Ali[¶]

Swetasudha Panda[*], **Qinlan Shen**[*], **Michael Wick**[*], **Ari Kobren**[*]

[¶]Institute for Experiential AI, Northeastern University
[*]Oracle Labs

Abstract

In the current landscape of language model research, larger models, larger datasets and more compute seems to be the only way to advance towards intelligence. While there have been extensive studies of scaling laws and models' scaling behaviors, the effect of scale on a model's social biases and stereotyping tendencies has received less attention. In this study, we explore the influence of model scale and pre-training data on its learnt social biases. We focus on BERT—an extremely popular language model—and investigate biases as they show up during language modeling (upstream), as well as during classification applications after fine-tuning (downstream). Our experiments on four architecture sizes of BERT demonstrate that pre-training data substantially influences how upstream biases evolve with model scale. With increasing scale, models pre-trained on large internet scrapes like Common Crawl exhibit higher toxicity, whereas models pre-trained on moderated data sources like Wikipedia show greater gender stereotypes. However, downstream biases generally decrease with increasing model scale, irrespective of the pre-training data. Our results highlight the qualitative role of pre-training data in the biased behavior of language models, an often overlooked aspect in the study of scale. Through a detailed case study of BERT, we shed light on the complex interplay of data and model scale, and investigate how it translates to concrete biases.

Introduction

Large Language Models (LLMs) continue to grow in size at a remarkable rate, with technology companies investing millions in infrastructure to produce ever-larger and more general purpose models. Modern open weight models like LLaMA, Gemini and Falcon regularly have tens of billions of parameters, showcasing noteworthy capabilities across a range of natural language processing applications.

To investigate the performance of LLMs in terms of model parameters, training data size, and compute resources, a rich literature in empirical scaling laws (Hernandez et al. 2021; Kaplan et al. 2020) has emerged, which suggests that bigger is indeed better (in terms of loss). Recent work on scaling laws has also led to a more comprehensive understanding on the tradeoffs between data size and model parameters with a fixed compute budget (Hoffmann et al.

2022). Given the pace of LLM development and the foundational role of scale, studying changes in model behavior with size remains a pressing research problem.

One crucial question that has received less attention, however, is how model scale influences social biases. LLMs inherently absorb societal biases and harmful stereotypes from data during both pre-training and task-specific fine-tuning. These biases manifest as *intrinsic* biases within the embedding space, leading to representational harms and stereotyping (Nangia et al. 2020; Nadeem, Bethke, and Reddy 2020; May et al. 2019; Kurita et al. 2019), and *extrinsic* biases leading to allocative harms (Barocas et al. 2017) in downstream predictions (Gehman et al. 2020; Garimella et al. 2019; Blodgett, Wei, and O'Connor 2018). Prior work has shown that pre-trained models can generate toxic language (Gehman et al. 2020), can have disparities in hate speech classification (Sap et al. 2019), can perpetuate anti-Muslim bias in text generation (Abid, Farooqi, and Zou 2021), can rely on racial biases even for high stake use cases such as clinical notes (Zhang et al. 2020), among other failures.

The growing scale of LLMs has been driven, in part, due to their popularity in commercially successful chat applications (e.g. ChatGPT), as well as their instruction following capabilities (Wei et al. 2021), making them useful for a variety of tasks. Chat and prompting applications tend to use autoregressive, decoder-only Transformer models (e.g. GPT-4, LLaMA, PaLM). While these models are at the cutting edge, they are often challenging to deploy in many cases, requiring massive compute resources and improvised prompt engineering. In contrast, encoder-decoder (e.g. T5, BART) and encoder-only (e.g. BERT, RoBERTa) Transformer models trained on a Masked Language Modeling (MLM) objective are often lighter, and remain workhorses for NLP applications in industry. These models continue to be relevant for applications such as summarization, semantic search, sentiment analysis, and a wide array of classification tasks after fine-tuning, as evidenced by their continued success in public machine learning competitions (Holmes et al. 2024; King et al. 2023). These models are also affected by biases in the training data similar to autoregressive LLMs, both during pre-training and during the task-specific fine-tuning process. In this study, we take a step back from the outsize discourse on autoregressive models and focus on encoder-only LLMs

---

(also referred to as MLMs here) due to their widespread use. We present a detailed case study of the influence of scale and data on social biases when pre-training BERT (Devlin et al. 2018), a seminal and extremely popular LLM. We focus on one model family specifically to control for variance in model architecture when changing number of parameters.

How could scale influence bias? The prevailing wisdom is that pre-trained LLMs should get more biased as they get bigger. One intuition for this comes directly from the scaling laws themselves. Models often learn harmful artifacts in the data, and since bigger models fit the data better, biases can aggravate as model size increases. But there are also cases when biases plateau or even decrease with scale (Srivastava et al. 2022), possibly because scale helps the model learn more reliable task-specific rules without overly relying on shortcut heuristics (McCoy, Pavlick, and Linzen 2019; Bhargava, Drozd, and Rogers 2021).

Indeed, there are good reasons for why bias should not in general increase with scale, and in fact, should decrease with scale in certain cases. *First*, MLMs such as BERT are known to use shortcut heuristics rather than actually learning more robust heuristics for the downstream task. For example, models fine-tuned for natural language inference (NLI) use the shortcut that when the premise has a high word overlap with the conclusion, the model uses this as a heuristic to predict entailment (McCoy, Pavlick, and Linzen 2019). It is likely that something like gender bias is a shortcut heuristic that the model might use to perform a task. And while scale does not solve the problem of shortcut heuristics in NLI, it helps a lot: the jump from the smallest to largest BERT model gets a 23% absolute increase in accuracy on HANS, and a jump from BERT-base to RoBERTa-base (more data) results in a 19% improvement (Bhargava, Drozd, and Rogers 2021). Similarly, we might hope that scale will help the model learn more reliable task specific rules in favor of biases like gender bias; and so for this reason, a model might actually get less biased with scale. *Second*, many bias and fairness measures, such as equal odds and equal opportunity are functions of the raw statistics that contribute to accuracy, such as false positive rates and false negative rates (Garg, Villasenor, and Foggo 2020). Moreover, downstream allocative harms are frequently measured by these statistics directly. So although scaling laws hold that training loss decreases with scale, indicating that the model might overfit to the bias; at the same time, test accuracy also increases with scale. For this reason, we would expect that fairness measures based on accuracy statistics to actually improve with scale as the accuracy increases. *Third*, scale isn't the only important factor, the type of data can matter. Different data has different biases. For example, biographies about notable figures on Wikipedia are skewed towards men over women (Tripodi 2023), and large web scrapes like the common-crawl are likely more diverse in overall topics covered, but also much more toxic.

Our contributions. We study MLMs to see how social biases evolve as we vary model scale. We pre-train four architecture sizes of BERT (`mini`, `small`, `medium` and `base`, i.e., up to 110M parameters). A core focus of our study is the pre-training dataset. We experiment with English Wikipedia and the CC-100 English subset of Common Crawl. For each model size and type of training data, we compute biases in the representations upstream and performance disparities downstream, leveraging specific measures of language biases from prior research (Steed et al. 2022). We measure upstream bias along two dimensions: disparities in gender pronoun probabilities and sentiment of generated text; downstream impact is measured by fine-tuning models on a toxicity classification task, where we evaluate differences is false positive rates across a diverse set of demographic groups from prior work (Dixon et al. 2018) (e.g. *Muslim*, *gay* etc.)

Our findings underscore the crucial role of pre-training data as models increase in size. For models pre-trained with CC-100, upstream biases generally increase with model size. Conversely, models pre-trained on Wikipedia show greater gender stereotyping as models increase in size. In both cases, we find that downstream biases decrease with increasing model size. However, models consistently associate certain identities such as *gay* and *homosexual* with toxicity, independent of parameter size or type of pre-training data. This finding aligns with prior work Steed et al. (2022); Panda et al. (2022), that downstream biases are largely influenced by biased artifacts in the fine-tuning dataset, and not the pre-training data.

We then inspect the datasets themselves to identify the reasons of observed biases, and find that indeed CC-100 contains more negative sentiment towards our measured identity groups compared to Wikipedia, which is picked up by larger models. We also find evidence of Wikipedia encoding more male pronoun co-occurences in articles related to occupations, which might explain the increased gender disparities in models pre-trained on Wikipedia.

In summary, we conduct a detailed case study through BERT—an extremely popular model—and investigate the impact of pre-training data, model scale, and observed social biases, both in terms of the masked language modeling task, as well as in downstream classification settings.

Related Work

Our study broadly relates to three strands in the literature: fairness and auditing of machine learning algorithms in general, the study of biases in natural language processing more specifically, and scaling laws for large language models.

First, a rich literature in computer science investigates issues of fairness in machine learning (Dwork et al. 2012; Barocas, Hardt, and Narayanan 2023), measuring the different types of harms these systems can have—such as denigration, stereotyping, differential quality of service etc. (Barocas et al. 2017; Weerts 2021). Notably, prior work has documented racial and gender disparities for commercial gender classification systems (Buolamwini and Gebru 2018), racial disparities in criminal recidivism prediction (Angwin et al. 2022), gender disparities in the delivery of job advertising (Datta, Tschantz, and Datta 2014; Ali et al. 2019), among others. Many of these studies only rely on "black box" access to machine learning systems, and have to conduct clever *audits* to measure disparate outcomes (Metaxa

et al. 2021). Our work is connected to this literature in its goal of measuring inadvertent harms of a machine learning system, albeit with "white box" access to the model's output probabilities and weights.

Within natural language processing (NLP) specifically, prior work has also discussed biased and disparate outcomes for users. Blodgett, Green, and O'Connor (2016) was one of the earliest works documenting racial disparities, showing how dependency parsing tools struggle on text for African American English on Twitter. Similarly, Caliskan, Bryson, and Narayanan (2017) demonstrated how word embeddings learnt from text corpora can contain gender biases. In the context of large language models—which power most of modern NLP—recent work has documented negative associations for people with disabilities (Hutchinson et al. 2020), anti-Muslim bias (Abid, Farooqi, and Zou 2021), and a general propensity to generate toxic text (Gehman et al. 2020). There have also been efforts to construct benchmarks that can yield repeatable measurements of bias across many different language models. This includes benchmarks such as WinoBias for coreference resolution (Zhao et al. 2018), BBQ and UNQOVER for question-answering (Parrish et al. 2021; Li et al. 2020), BBNLI for natural language inference (Baldini et al. 2023), StereoSet for measuring stereotypical associations (Nadeem, Bethke, and Reddy 2020), among others. Further, large benchmarking efforts such as BIG-bench (Srivastava et al. 2022) have been able to provide insights into the relationship between model scale and performance on bias benchmarks, which is one of the objectives of our study. We now know from Srivastava et al. (2022) that for auto-regressive models, bias (as measured via UNQOVER, BBQ etc.) typically increases in ambiguous prompts, and that it can decrease for narrow, unambiguous prompts. We similarly study the relation of bias with scale, but in the context of MLM models, which have distinct upstream and downstream applications, and with an added focus on the pre-training dataset used. Closest to our study is Steed et al. (2022)'s work on upstream and downstream biases for MLMs, in which they investigate the *bias transfer hypothesis*—can upstream debiasing methods improve disparities in downstream performance? They find that upstream mitigation does little to address downstream biases, and that downstream disparities are better explained by biases in the fine-tuning data.

In parallel, empirical scaling laws related to LLM performance have been the subject of extensive investigation in recent research (Hestness et al. 2017; Kaplan et al. 2020). These studies have found a power-law scaling relationships with model size, dataset size, and computational resources, i.e. an increase in either almost always leads to a decrease in loss. Recently, Hoffmann et al. (2022) have also led to a clearer understanding of the tradeoffs between training data size (number of tokens) and model parameters, yielding a unified formula for compute-optimal training, which has already been applied to specific model settings (Clark et al. 2022; Gordon, Duh, and Kaplan 2021; Henighan et al. 2020; Tay et al. 2022). However, it is noteworthy that the scalability of LLMs does not universally translate to improved performance across all downstream tasks, as demonstrated

by Ganguli et al. (2022). Similarly, recent work by Wei et al. (2022) highlights emergent abilities unique to larger models not predicted by traditional scaling laws. In response to work in scaling laws, there has also been pushback from critics. Notably, Bender et al. (2021) highlighted the rising environmental and financial costs of model pre-training, and the lack of diversity in training data. Closely related to our study, Birhane et al. (2023) study scaling laws in the context of hateful content present in the LAION family of datasets, popularly used to pre-train text to image diffusion models. They find that as data scale increases, the tendency of models to associate Black faces with categories like "criminal" can significantly increase.

Our work lies at the intersection of these research threads. We contribute to the ongoing practice of measuring the adverse outcomes of machine learning systems. Our work also contributes to ongoing work on scaling laws, with a specific focus on bias, and how it is picked up from pre-training data.

Methods

In this section, we cover the models we train, our training configuration, the datasets used to train these models, and the metrics we use to measure bias.

Models

We experiment with four architecture sizes of BERT: BERT-Mini, BERT-Small, BERT-Medium, BERT-Base. While originally introduced in the context of model distillation (Turc et al. 2019), we find that these models provide a good testbed for experimenting with model scale, while holding the architecture constant. Table 1 shows the number of layers, hidden embedding size, and the number of parameters in each case. Following (Turc et al. 2019), we fix the number of attention heads to $H/64$, where H is the hidden embedding size. We use the publicly available architecture implementations of miniature BERT architectures via HuggingFace[1].

Model	L	H	Parameters
BERT-Mini	4	256	11.3M
BERT-Small	4	512	29.1M
BERT-Medium	8	512	41.7M
BERT-Base	12	768	110.1M

Table 1: BERT architecture specifications for our models. We vary number of layers (L) and hidden embedding size (H).

Pre-Training Data

For each model size, we pre-train on three different datasets, on a masked language modeling objective: (a) CC-100-EN: English subset of Common Crawl (Conneau et al. 2019), (b) English Wikipedia, and (c) a combination of CC-100-EN and Wikipedia in a multi-task setup. Text on Wikipedia data

---

[1]BERT-Medium, e.g. `https://huggingface.co/google/bert_uncased_L-8_H-512_A-8`

is curated by a set of editors, and often goes through moderation, quality control and edits. Common Crawl (Wenzek et al. 2019), on the other hand, is a massive unconstrained crawl of the internet and therefore, is very likely to include stereotypes as well as toxic and abusive statements (Luccioni and Viviano 2021; Gehman et al. 2020). Our choice of these datasets for pre-training is based on these content differences, such that we can contrast language biases after model pre-training.

Pre-training configuration. For each model size in Table 1, we pre-train the model for 8,000 training steps on the chosen pre-training data. We seed our data shuffling consistently to make sure that each model gets exposed to the same set of tokens from the data, which prior work has shown to be fundamental in models' scaling behavior (Kaplan et al. 2020; Hoffmann et al. 2022). When combining datasets, we run combined training with the two datasets interleaved, i.e., one update step on CC-100-EN, followed by one update step on Wikipedia. As a result, each mini-batch consists only of data from one of the pre-training datasets, and Wikipedia is up-sampled relative to CC-100-EN.

Metrics

We use a series of metrics from prior work to measure biases at different points. First, we evaluate biases intrinsic to the model itself, i.e. relating to the masked language modeling task it's trained on. Second, we fine-tune each model for a downstream classification task and evaluate how its scale and pre-training data affects false positive rates across demographic groups. Third, we use linguistic analysis on the pre-training datasets themselves to understand the provenance of our observed biases. Here, we describe each of these metrics in detail.

Upstream bias metrics. We use two metrics from prior work (Steed et al. 2022) to evaluate upstream biases in our pre-trained models.

First, we evaluate gender bias using an extension of log probability bias score from Kurita et al. (2019). We specifically use the version of this metric used in Steed et al. (2022), where templates are constructed for a list of 28 professions taken from the *Bias in Bios* dataset (De-Arteaga et al. 2019). The original dataset is built from Common Crawl, which includes over 400,000 online biographies from 28 occupations. The dataset does not include self-reported gender; we refer to the pronouns in each biography to denote gender. In our use-case, for the list of 28 professions, we use templates of the form `{pronoun} is a(n) {occupation}` to measure the model's propensity towards either he/him or she/her pronouns. To increase the robustness of our measurements, we also include template variations from Bartl, Nissim, and Gatt (2020), e.g. `{pronoun} applied to the position of {occupation}`. For each occupation y and pronoun g, we compute the model's probability $p_{y,g}$ for the template. To control for baseline differences for pronouns, we also compute prior probability $\pi_{y,g}$ for a template where only the pronoun is present but the occupation is

masked, e.g., `he is a [MASK]`. We define our upstream gender bias metric as the difference in these probabilities:

$$\log \frac{p_{y,\text{she/her}}}{\pi_{y,\text{she/her}}} - \log \frac{p_{y,\text{he/him}}}{\pi_{y,\text{he/him}}} \quad (1)$$

A higher absolute probability gap suggests that a model associates one gender much more with an occupation, while a value close to zero implies equal association during masked language modeling.

Second, we evaluate upstream biases beyond gender, and for a diverse set of demographic groups. Following Hutchinson et al. (2020), we rely on sentiment analysis to measure upstream bias. Again, we re-use methodology from Steed et al. (2022) and construct templates of the form `{identity} {person} is [MASK]`. The `{identity}` term consists of about 50 diverse identity groups such "Muslim", "Jewish", "elderly", "gay" etc., taken from Dixon et al. (2018). The original dataset consists of (a) 130,000 public comments from Wikipedia Talk pages, annotated for toxicity, which mention these identity groups; and (b) a synthetic test set to evaluate disparities in toxicity classification. We leverage the identity groups to generate templates for upstream biases, and the synthetic test set to evaluate downstream biases later. The `{person}` part of the template includes phrases like "people", "spouse" etc. to increase the number of templates we measure. We compute the 20 most likely tokens for `[MASK]` for each template. We then use a pre-trained RoBERTa (Liu et al. 2019) sentiment classifier[2] trained on the TweetEval benchmark (Barbieri et al. 2020) to measure the average *negative* sentiment for each identity group's completed prompts. We focus on negative sentiment in particular as a proxy for toxicity and negative associations similar to prior work (Steed et al. 2022; Hutchinson et al. 2020), and due to its potential of introducing representational harms (Barocas et al. 2017).

Downstream bias metrics. To evaluate downstream biases, we fine-tune our pre-trained model on a toxicity classification task, and compare false positive rates (FPR) across different identity groups. The FPR of a group g in the data is defined as

$$FPR_g = \frac{FP_g}{FP_g + TN_g} = \frac{FP_g}{N_g}$$

Here, FP_g indicates the false positives in classification, TN_g is true negatives, and N_g are total number of ground truth negatives (i.e. non-toxic sentences), all for group g specifically. We focus on false positives since they can result in concrete allocative harms such as over-moderation and de-platforming (Jhaver et al. 2021) if such classifiers were to be used for toxicity classification. Further, prior work (Steed et al. 2022) has successfully used FPR to quantify downstream performance disparities. We use the synthetic test set from Dixon et al. (2018) as our toxicity classification task; the dataset contains 89K examples created using templates of both toxic and non-toxic phrases which are filled in with the 50 identity terms we also use in our upstream bias

---

[2]https://huggingface.co/cardiffnlp/
twitter-roberta-base-sentiment

measurement. Following the original paper, we divide the synthetic dataset into 75% training and 25% (split equally into validation and test). To build a classifier, we attach a sequence classification head to our pre-trained model and fine-tune for 3 epochs.

Dataset bias metrics. To investigate the provenance of our observed biases, we measure biases within the pre-training datasets themselves.

First, to compare gender associations, we analyze differences in `(pronoun, occupation)` pair co-occurrences between the two pre-training corpora using weighted log odds ratio with a Dirichlet prior (Monroe, Colaresi, and Quinn 2008). Log odds ratio is an alternate to tf-idf and similar word score methods to compare word importance across documents or corpora. In its simplest form, the log odds of word w in a corpus i where it occurs with frequency f_w^i is defined as $log\ O_w^i = log\ \frac{f_w^i}{1-f_w^i}$; the *log odds ratio* can then be used to compare word importance between corpus i and j as:

$$ log\ \frac{O_w^i}{O_w^j} = log\ \frac{f_w^i}{1-f_w^i} - log\ \frac{f_w^j}{1-f_w^j} \qquad (2) $$

We use a model-based variant of this measure that is more robust to low frequencies, specifically the weighted log odds ratio with a Dirichlet prior; we refer the reader to Monroe, Colaresi, and Quinn (2008) for a more detailed discussion.

Second, we re-use the sentiment model[2] used for upstream bias to compare sentiment in the pre-training datasets themselves. For both CC-100-EN and Wikipedia, we extract sentences that mention our list of identity groups. We then use the sentiment model to compute average negative sentiment across all sentences that mention a group.

Results

After our pre-training process, we obtain three variants (Wikipedia, CC-100-EN, and Wikipedia + CC-100-EN) of each model size (Mini, Small, Medium, Base), i.e. twelve pre-trained LLMs in total. We first measure upstream biases in all these models using our two metrics; second, we fine-tune each model to the downstream task of toxicity classification to measure downstream biases. Finally, we use our dataset bias metrics to measure the pre-training dataset themselves, and investigate the provenance of the biases we observe. Here, we present our results from these experiments.

Upstream biases can increase with model size

We begin by evaluating gender bias upstream using our implementation log probability bias score (Equation 1). Figure 1a shows absolute log probability gap between `he/him` and `she/her` pronouns for prompts related to 28 occupations, for all 12 of our models. Since we use multiple occupations for our metric, we visualize the probability gap as a distribution across these occupations. Higher values suggest a skew towards either pronoun, while lower values suggest equal likelihood, and therefore better gender representation. We observe that for models pre-trained on Wikipedia
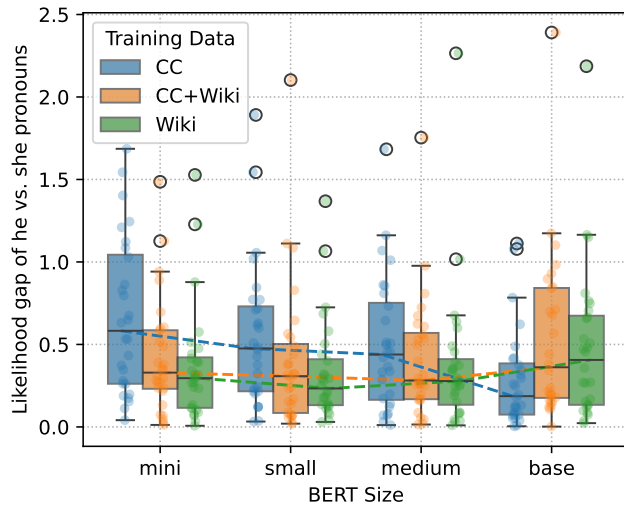
(green), gender stereotypes slightly increase with model size—as seen in the increased variance and median. However, for models pre-trained on CC-100-EN (blue), gender stereotypes seemingly decrease with model size. For models pre-trained on the combination (orange), we do not observe a consistent trend across model sizes. We qualitatively observe that occupations such as "nurse", "yoga teacher" and "software engineer" consistently appear as outliers across model types.

We then measure upstream bias with our second metric, which is the average negative sentiment for prompt completions that relate to 50 identity groups. For each of our pre-trained models, we compute the average negative sentiment for multiple MLM prompts relating to each identity— Figure 1b shows the distribution of these negative sentiment scores. Here, we note that as model size increases, we observe a general upward trend in upstream bias, regardless of pre-training dataset. We also observe that models pre-trained on CC-100-EN achieve the highest average negative sentiment scores, followed by models pre-trained on the combination of CC-100-EN and Wikipedia. Models pre-trained on Wikipedia exhibit comparatively the lowest average negative sentiment in our experiments. Qualitatively, in case of models pre-trained on CC-100-EN, we notice frequent abusive mentions (e.g., "stupid", "sick", "insane") on the list of top words predicted by the model. We also find evidence of these models generating (unfortunate) sentences such as "Muslim people are dangerous". Identities such as "elderly", "deaf" and "Muslim" are the most frequent outliers across model sizes, which aligns with prior work (Dixon et al. 2018; Abid, Farooqi, and Zou 2021). In contrast, for models pre-trained on Wikipedia, we note MLM completions associated with lower negative sentiment such as "wrong", "injured", "wounded" etc. These differences illustrate the effect of both model scale and training data on output toxicity, suggesting that larger models are more capable of learning biases from the data—particularly when that data is from an unmoderated source.
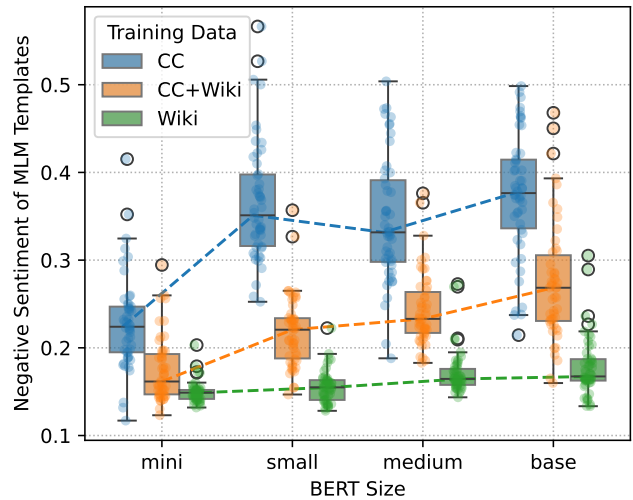
Evolution of bias over the training process. We also monitor the change in upstream bias (measured via sentiment) during the course of the pre-training process. We checkpoint all models after every 900 training steps during the training process, and compute negative sentiment for each identity group at these checkpoints. Figure 2 shows how upstream biases grow over time in our experiments. Each small point shows the average negative sentiment for an identity group, the large points connected via lines show the average of average negative sentiment for each model. Similar to our final measurement in Figure 1b, we notice here too that models trained on CC-100-EN (except BERT-Mini) have higher upstream bias. Models trained on Wikipedia consistently have lower upstream bias and interestingly this does not increase or vary over training.

Larger models make more robust downstream classifiers

Next, we turn our attention to downstream biases of our pre-trained models. We attach a classification head to each of our

(a) Upstream bias: gender stereotyping   (b) Upstream bias: negative sentiment

Figure 1: Upstream biases for each model size and pre-training data type in terms of our bias metrics: (a) log probability gaps between `he/him` and `she/her` pronouns for prompts related to occupations (b) average negative sentiment for masked language modeling completions related to multiple identity groups.
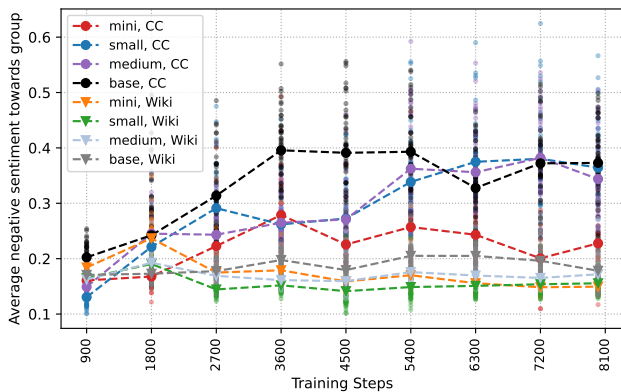

Figure 2: Upstream biases (measured via negative sentiment associations) over the course of pre-training. Models pre-trained on CC-100 generally result in higher bias scores compared to models pre-trained on Wikipedia.

models and fine-tune them for the task of toxicity classification. Following prior work (Steed et al. 2022; Panda et al. 2022), we use the synthetic toxicity classification data from Dixon et al. (2018) for this task. We also evaluate an off-the-shelf BERT (`bert-base-uncased`) from Hugging-Face. As described earlier, we evaluate downstream biases in terms of differences in false positive rate (FPR) for sentences relating to each identity group. A higher FPR for a group indicates higher downstream biases, since the model is more likely to falsely flag mentions of that group as toxic, potentially leading to discriminatory censorship. Ideally, we would like a model to have low FPR on each identity *and* low variance in FPR across all groups.

Figure 3 shows a distribution of FPR for each model size and pre-training dataset after fine-tuning. We observe that the median FPR decreases as model size increases, regardless of pre-training data; similarly, we note that the variance of FPR across groups decreases for larger models as well. This suggests that after fine-tuning, larger models make more robust classifiers—they make fewer false positive errors for all groups in our experiments.

The decrease in downstream biases with scale can have a few explanations. First, MLMs are known to use short-cut heuristics instead of task-specific robust heuristics (e.g., model fine-tuned for Natural Language Inference (NLI) uses high word overlap with the conclusion, to predict entailment (McCoy, Pavlick, and Linzen 2019); scale improves these results (Bhargava, Drozd, and Rogers 2021)). The model might latch on to certain identities as shortcuts to predict toxicity. Second, allocative harms are frequently measured directly using accuracy statistics e.g., FPR in our case. As scaling laws suggest that test accuracy increases with scale, downstream bias statistics will improve as well.

However, certain identity groups such as "gay", "queer" and "homosexual" consistently show up as outliers in terms of FPR, regardless of model size or type of pre-training data. Prior work (Steed et al. 2022) has shown that downstream disparities are largely explained by the fine-tuning data; our observed outliers likely appear disproportionately in toxic sentences, leading to a higher FPR. This suggests that while larger versions of BERT make more robust downstream classifiers, they are not able to address biases against extreme outliers.

Associations in pre-training data influence bias

Finally, we investigate the impact of pre-training data on the biases we observe. While downstream biases can be ex-
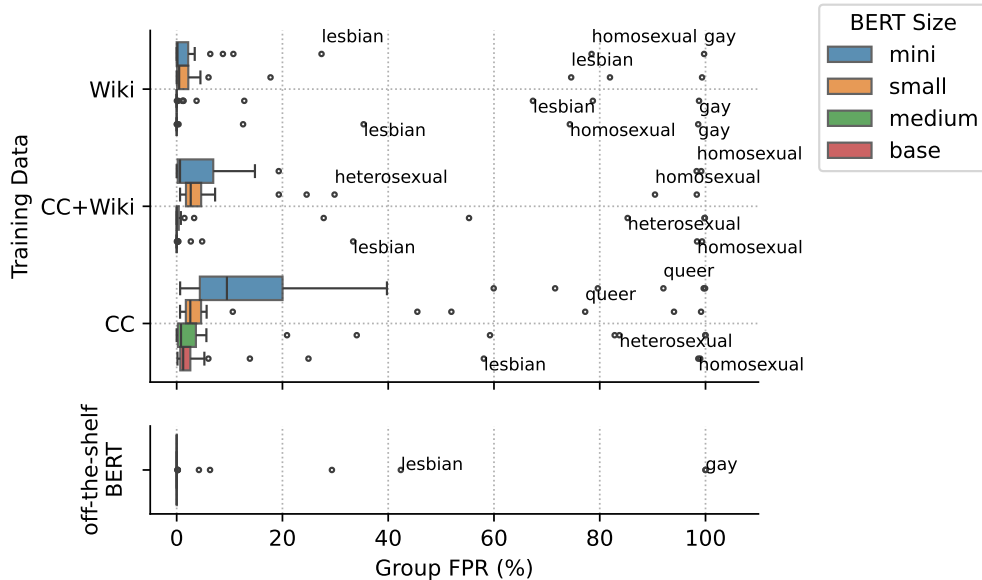
Figure 3: Downstream biases evaluated on toxicity classification data from Dixon et al. (2018). For each model size and type of pre-training data, false positive rates (FPR) for each identity group are shown. Median FPR and variance of FPRs decreases as models grow larger, but some outliers remain.

plained as an artifact of the fine-tuning data (Steed et al. 2022), we suspect a much tighter coupling between upstream biases and choice of pre-training data.

To understand our observed upstream gender biases (Figure 1a), we use weighted log odds with a Dirichlet prior (Monroe, Colaresi, and Quinn 2008) and compare (`pronoun, occupation`) pair occurrences between CC-100-EN and Wikipedia. Specifically (in terms of Equation 2), for each occupation $o \in \{$journalist, physician, painter, ...$\}$, and pronoun $p \in \{\{$he, him, his, himself$\}, \{$she, her, hers, herself$\}\}$ we measure:

$$log \frac{O_{o,p}^{\text{CC-100}}}{O_{o,p}^{\text{Wiki}}}$$

A positive value indicates a co-occurrence is more likely in CC-100-EN than in Wikipedia, while a negative value means it is more likely in Wikipedia. Also note that we count frequencies for a set of pronouns and not singular pronouns for more robust counting. Further, to normalize for variance, we z-normalize the log odds; using the one-sided critical value for $p = 0.05$, we only consider $z > 1.645$ to be a significant difference between both datasets. Table 2 shows the 10 occupations with the highest weighted log odds between CC-100-EN and Wikipedia.

While we observe many differences that are not significant, for certain occupations, Wikipedia indeed encodes greater gender stereotypes, e.g., "professor" has significantly higher masculine pronoun associations, and "model" has higher feminine pronoun associations. Interestingly, "teacher" is the only occupation that has significant stereotypical associations in both datasets: masculine in Wikipedia, feminine in CC-100-EN. One trend, despite

Occupation	Pronouns	
	M	**F**
teacher	**-3.37***	**2.47***
professor	**-4.14***	1.33
nurse	-0.25	**2.87***
model	-1.00	**-1.64***
journalist	-0.31	-1.06
painter	-1.26	-0.06
physician	-1.16	-0.16
composer	-1.19	-0.11
attorney	-0.56	0.65
photographer	-0.53	0.64

Table 2: Weighted log odds (z-normalized) for occupation, pronoun pairs between CC-100-EN and Wikipedia. **M** = {he, him, his, himself}, **F** = {she, her, hers, herself}. Positive values indicate skew towards CC-100-EN, negative values indicate skew towards Wikipedia; $p < 0.05$ shown in **bold**.

lack of significance, is that co-occurrences with masculine pronouns are overall more common in Wikipedia than CC-100 (larger negative values in **M** column). This may be reflective of a broader trend on Wikipedia, beyond gendered stereotypes for specific occupations, where the vast majority of biographical articles are about men, due to biases in who is perceived as notable (Tripodi 2023). Conversely, while large web scrapes like CC-100 are more diverse in overall topics covered, these might involve more toxic text.

To understand the provenance of our upstream sentiment biases (Figure 1b), we extract sentences from both CC-100-
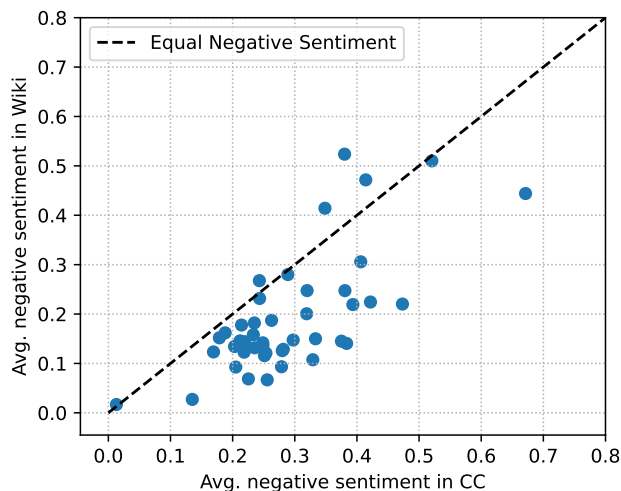
Figure 4: Average negative sentiment for sentences in pre-training data that mention our studied identity groups. CC-100-EN (x-axis) almost always encodes more negative sentiment.

EN and Wikipedia that mention our list of identity groups, and re-use our sentiment classifier to measure average negative sentiment. This allows us to compare whether one dataset a priori encodes more negative sentiment towards a group, which could be picked up by a model during training. Figure 4 shows average negative sentiment for each identity group in both datasets. We find that CC-100-EN shows higher negative sentiment for most identity groups compared to Wikipedia. In fact, with the exception of a few outliers, CC-100-EN consistently has more negative sentiment. We suspect that our observed upstream biases are an artifact of this difference, and that larger models do a better job of picking up these aspects of the data (e.g., Figure 2).

Limitations

Our approach has multiple limitations. First, while we limit our analysis to a single model family to reduce variance in architecture, it limits the ecological validity of our results. Our results therefore cannot be generalized to all modern LLM architectures, and instead provide a detailed look into BERT specifically. Second, compared to state-of-the-art compute intensive training procedures, our pre-training process is quite rudimentary. We limit the training process to only 8000 training steps as a heuristic to upper bound the amount of compute that each model uses; this is a simplification and does not lead to a model as powerful as those available through model hubs like HuggingFace. Third, while we make sure to evaluate bias holistically by examining both upstream and downstream differences, our bias metrics—such as log probability gaps and sentiment—are not definitive ways of measuring *bias*. Bias is a complex, socio-technical, and sometimes ill-defined notion whose meaning can vary across domains and tasks. While we rely on metrics from prior work, our measures are prone to the same pitfalls

and limitations in validity that most bias measurement work in NLP suffers from (Goldfarb-Tarrant et al. 2023).

Concluding Discussion

Our study provides a detailed case study on the interplay of scale, pre-training data, and bias with a specific focus on BERT, a widely used LLM. We find evidence that larger models are able to encode more biases upstream. Importantly, we observe that larger models, combined with un-moderated data, can lead to worse results for the task of masked language modeling. However, larger models can also produce more robust downstream classifiers after fine-tuning.

While MLMs like BERT do not represent the state-of-the-art in the rapidly developing landscape of LLM research, they remain extremely relevant for several applied natural language processing problems. Our investigation of bias is particularly relevant to practitioners who fine-tune embedding models for their tasks. In these applied use-cases, our results shed light on how scale and training data together can lead to different kinds of biases. We encourage practitioners to be aware of the biases their training datasets can introduce, and to actively measure these artifacts during the development process. On a more general level, our study highlights the role that training data can play in scaling, especially as it relates to biased model behavior. Our results also suggest that mixing in a moderated, high quality data source (e.g., Wikipedia) with larger datasets (e.g., CC-100, The Pile (Gao et al. 2020)) might be an approach to alleviate biases—we leave a full exploration of this direction to future work.

Our analyses also underscore the limitations that exist in metrics used to measure *bias*, which is a nuanced socio-technical concept, whose meaning changes across tasks and domains. Negative sentiment and gaps in gender representation—as used here—are well-scoped ways of expressing bias that can be useful for different domains. Negative sentiment, for instance, could be a useful measure of bias for LLM use in chatbots or auto-complete tools; differences in gender likelihood could be useful for measuring bias in resumé or search ranking, but they are not universal measures of linguistic bias. As seen in our results, depending on the choice of bias metric, a measurement of model behavior can look quite different. This aligns with prior work (Goldfarb-Tarrant et al. 2023; Blodgett et al. 2021) which shows that measuring bias or fairness can be a challenging undertaking, and it is easy to set up an incompatible metric. Our results highlight the need for identifying the correct bias metric for each domain, and judging both the data and the model by that metric.

Ethical Considerations

Our study attempts to measure a social issue with technical tools, and therefore it relies on some shortcut heuristics and simplifications that we attempt to make explicit here. In studying gender disparities, we rely on pronouns and only focus on he/him and she/her since our metrics are set up as subtractions. This simplification is not meant to reinforce the

gender binary, and we acknowledge that multiple instances of log probability gap can be used as well. The list of identity groups for whom we measure sentiment and downstream classification disparities is taken from Dixon et al. (2018). This list has been designed for broad coverage, and is not necessarily grounded in any harms that have been experienced by these groups.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models.

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine bias. In *Ethics of data and analytics*, 254–264. Auerbach Publications.

Baldini, I.; Yadav, C.; Das, P.; and Varshney, K. R. 2023. Keeping Up with the Language Models: Robustness-Bias Interplay in NLI Data and Models. *arXiv preprint arXiv:2305.12620*.

Barbieri, F.; Camacho-Collados, J.; Anke, L. E.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650.

Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bartl, M.; Nissim, M.; and Gatt, A. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 1–16.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bhargava, P.; Drozd, A.; and Rogers, A. 2021. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, 125–135. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Birhane, A.; Prabhu, V.; Han, S.; and Boddeti, V. N. 2023. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868*.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 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)*, 1004–1015.

Blodgett, S. L.; Wei, J.; and O'Connor, B. 2018. Twitter universal dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1415–1425.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

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

Clark, A.; De Las Casas, D.; Guy, A.; Mensch, A.; Paganini, M.; Hoffmann, J.; Damoc, B.; Hechtman, B.; Cai, T.; Borgeaud, S.; et al. 2022. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, 4057–4086. PMLR.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Datta, A.; Tschantz, M. C.; and Datta, A. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*.

De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 120–128. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; et al. 2022. Predictability and surprise in large generative

models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Garg, P.; Villasenor, J.; and Foggo, V. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, 3662–3666. IEEE.

Garimella, A.; Banea, C.; Hovy, D.; and Mihalcea, R. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3493–3498.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Goldfarb-Tarrant, S.; Ungless, E.; Balkir, E.; and Blodgett, S. L. 2023. This prompt is measuring¡ mask¿: evaluating bias evaluation in language models. *arXiv preprint arXiv:2305.12757*.

Gordon, M. A.; Duh, K.; and Kaplan, J. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5915–5922.

Henighan, T.; Kaplan, J.; Katz, M.; Chen, M.; Hesse, C.; Jackson, J.; Jun, H.; Brown, T. B.; Dhariwal, P.; Gray, S.; et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Hernandez, D.; Kaplan, J.; Henighan, T.; and McCandlish, S. 2021. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*.

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; and Zhou, Y. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Holmes, L.; Crossley, S.; Baffour, P.; King, J.; Burleigh, L.; Demkin, M.; Holbrook, R.; Reade, W.; and Howard, A. 2024. The Learning Agency Lab - PII Data Detection.

Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; and Denuyl, S. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–30.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

King, J.; Baffour, P.; Crossley, S.; Holbrook, R.; and Demkin, M. 2023. LLM - Detect AI Generated Text.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Li, T.; Khot, T.; Khashabi, D.; Sabharwal, A.; and Srikumar, V. 2020. UNQOVERing stereotyping biases via underspecified questions. *arXiv preprint arXiv:2010.02428*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luccioni, A. S.; and Viviano, J. D. 2021. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv preprint arXiv:2105.02732*.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Florence, Italy: Association for Computational Linguistics.

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; Sandvig, C.; et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4): 272–344.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Panda, S.; Kobren, A.; Wick, M.; and Shen, Q. 2022. Don't Just Clean It, Proxy Clean It: Mitigating Bias by Proxy in Pre-Trained Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5073–5085. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *ACL*.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2022. Beyond the imitation game: Quanti-

fying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Steed, R.; Panda, S.; Kobren, A.; and Wick, M. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3524–3542.

Tay, Y.; Dehghani, M.; Abnar, S.; Chung, H. W.; Fedus, W.; Rao, J.; Narang, S.; Tran, V. Q.; Yogatama, D.; and Metzler, D. 2022. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*.

Tripodi, F. 2023. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 25(7): 1687–1707.

Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Weerts, H. J. 2021. An introduction to algorithmic fairness. *arXiv preprint arXiv:2105.05595*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wenzek, G.; Lachaux, M.-A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; and Grave, E. 2019. CCNet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Zhang, H.; Lu, A. X.; Abdalla, M.; McDermott, M.; and Ghassemi, M. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, 110–120.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.