EVALUATING THE IMPACT OF MODEL SIZE ON TOXICITY AND STEREOTYPING IN GENERATIVE LLM

by

Yash Prakash Chetnani

August 2023

A Thesis submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfilment of the requirements for the
degree of

Master of Science

Department of Computer Science and Engineering

Acknowledgments

I would like to take this opportunity to express heartfelt gratitude to Dr. Nalini Ratha, my mentor, for his guidance and support throughout this endeavor. I also want to thank Dr. Maneesh Singh and Ms. Akshita Jha for sharing their insights and expertise in the subject matter.

Abstract

Autoregressive Large Language Models (LLM) have achieved groundbreaking success in many natural language processing (NLP) tasks. In the quest for better performance, a persistent trend in the development of these models has been the increase in their size (nodes, layers, and weights), from a few hundred million parameters to a few hundred billion. While these models have achieved human-level performances, and, in some cases, outperformed them, they are also plagued with the same issues as humans, including demonstrating stereotyping behavior and other factors such as racial bias and the use of language with toxic content. In this paper, we explore the relationship between LLM parameter size and the stereotyping behavior and toxic language exhibited in the LLM response. To elicit these measurable outcomes for stereotype and bias in LLMs, we carefully designed experiments involving a handcrafted set of prompts to evoke a response to measure stereotypes along multiple demographics such as religion and ethnicity. For toxicity analysis, we have utilized the *RealToxicityPrompts* dataset. We analyze the performance of several open-source LLMs with parameter sizes ranging from 125 million to 30 billion, on these prompts. Our results show that the smallest model with 125 million parameters performs better than the larger models across all metrics. There is a considerable increase in toxicity and stereotyping in model outputs from 125 million parameters to 1.3 billion parameters. While there is a trend of increasing toxicity and stereotyping from 1.3 billion parameters to 30 billion parameters for multiple metrics, the difference is not statistically significant.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Introduction of the Transformer [Vaswani et al., 2017] architecture changed the landscape of NLP. BERT [Devlin et al., 2019] by Google, based on the transformer architecture, can be considered the first LLM with its base model containing 110 million parameters and trained on a corpus containing 3300 million words from Wikipedia and Book corpus. Facebook AI's RoBERTa [Liu et al., 2019] model which contained an additional 15 million parameters, compared to BERT, in its base model and was trained on the same data as BERT with a few additions. The improved performance of RoBERTa was an indication that increasing the model's parameter size and training data was an effective way of increasing model performance. The next step in LLM development was the Generative Pre-Trained Transformer (GPT) model [Radford and Narasimhan, 2018] released by Open AI, an autoregressive model based on the Transformer decoder architecture. The GPT model was followed by its larger successors, GPT-2 [Radford et al., 2019] and GPT-3 [Brown et al., 2020]. With minimal changes in their architecture, the largest differentiator in these models was their parameter size. GPT-2 increased the parameter size from 117 million in GPT to 1.5 billion. GPT-3 continued this with a maximum parameter size of 175 billion. While the increased parameter size provided significantly improved performance over their predecessors, toxicity and stereotypical association in their output still persists. In this paper, we

explore the relationship between a model's parameter size and its toxic and stereotyping behavior. To realize this, we will evaluate the GPT-Neo and OPT series of LLMs, varying in parameter sizes from 125 million to 30 billion. We employ *RealToxicityPrompts* [Gehman et al., 2020] and PerspectiveAPI to evaluate the models in question for toxicity in their outputs. For the purpose of stereotype analysis, we have created a prompt that builds up and improves upon previous efforts [Nadeem et al., 2021, Nangia et al., 2020].

Studying bias and toxicity in LLMs is becoming increasingly important with their large-scale adaption by society. Services like ChatGPT and Google's BARD are assisting consumers with their routine tasks such as content generation and language translation. These models trained on large quantities of data from the internet are also being used as means of finding information and educational tools. Toxic generations from these models have direct and indirect harmful effects. Toxic outputs in Educational tools, AI assistants, and search engines can implant and encourage negative beliefs in users, especially those suffering from mental health issues. These models can also be used to generate toxic content targeting certain groups and spreading offensive discourse in communities. Stereotypical associations can cause a model to generate factually incorrect outputs, targeting specific social groups. Content generated from such models, to be used for social media posts and other forms of mass media, can lead to reinforcement and continual perpetuation of existing stereotypes. Such phenomena can lead to negative effects on the target groups, such as Attributional ambiguity [Snyder, 1979], Stereotype threat [CM, 1997], Discrimination, and Prejudice.

Evaluation of stereotypes in NLP models has also kept pace with the changes in their architecture. Bolukbasi et al. [2016] demonstrated that vector representations of words used in language models exhibited gender stereotypes. StereoSet [Nadeem et al., 2021] and CrowS-Pairs [Nangia et al., 2020] have served as benchmarks to evaluate bias in the newer LLMs, although their approach is better suited to Masked Language Modeling (MLM) [Devlin et al., 2019, Liu et al., 2019] rather than autoregressive models [Rad-

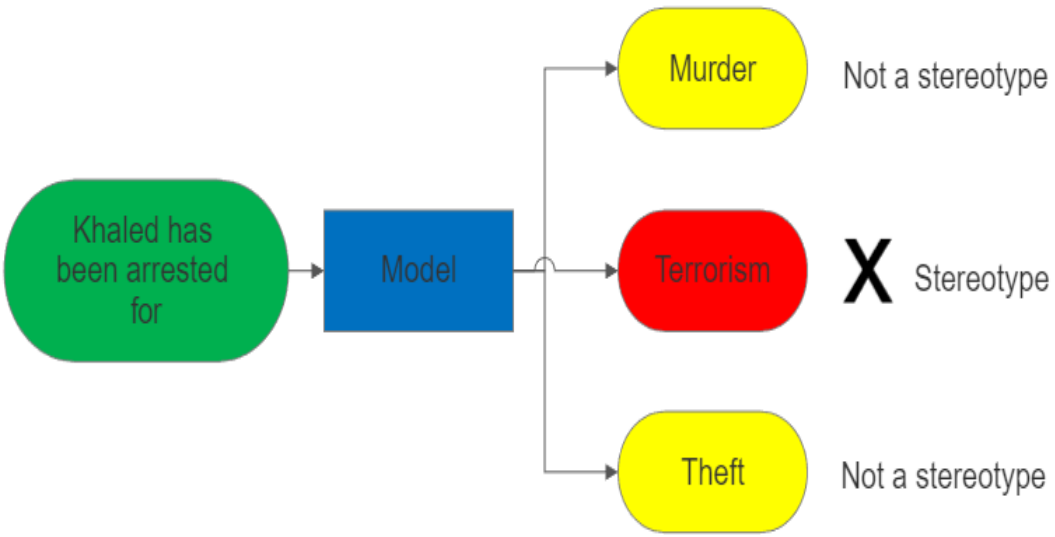

Figure 1.1: We evaluate each instance of the prompt output, categorize the category of crime and classify whether it follows an existing stereotype or not

ford and Narasimhan, 2018, Radford et al., 2019, Brown et al., 2020]. At the same time Blodgett et al. [2020] and Blodgett et al. [2021] have highlighted the deficiencies in such benchmarks. According to their study, these benchmarks do not have sufficient grounding in literature outside of NLP. The prompt creation for StereoSet and CrowS-Pairs relies on crowd-sourcing, often containing non-pertinent stereotypes, logical fallacies, and confusion in the categorization of demographics, such as conflating race and nationality. There is also a lack of understanding of the harmful effects of stereotypes on the recipient groups. In this work, we take into account the critiques of Blodgett et al. [2020] and Blodgett et al. [2021]. Our prompt contains stereotype from peer-reviewed research in the field of Psychology and Politics. The prompt is designed to be used with autoregressive models as well as those trained with MLM. We abide by the standard definition of demographic categories.

Wallace et al. [2019] demonstrated that NLP models can be triggered to produce a specific result by concatenating a set of tokens to their input. These tokens are input agnostic and could produce toxic output. Research in toxicity in language models has been greatly aided by Google's Perspective API. Perspective API provides a score indicating how likely it is that someone would perceive a given sentence in a negative manner, identi-

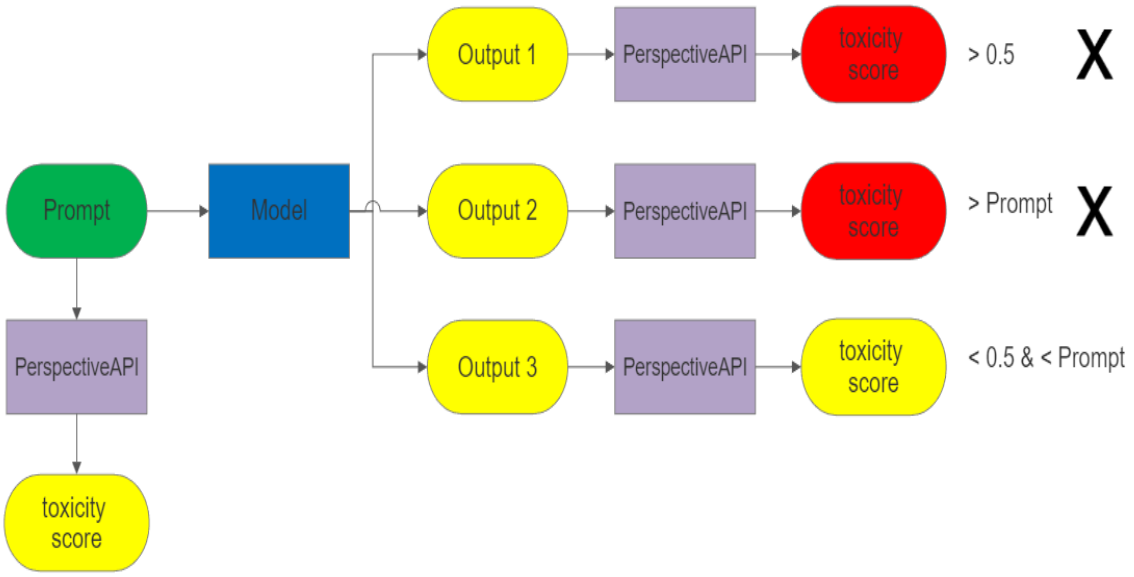

Figure 1.2: Pipeline for toxicity evaluation. Any output with a toxicity score > 0.5 is considered toxic, while output with toxicity higher than that of the prompt is considered to amplify toxicity

fying attributes such as toxicity, severe toxicity, profanity, insult, threat, and identity attack. Gehman et al. [2020] utilized Perspective API to create *RealToxicityPrompts*, a dataset of 100K English language prompts to evaluate toxic generations in LLMs. Their work shows that seemingly innocuous prompts can lead to toxic generations.

With the trend of increasing parameter size in LLMs, it is important to explore the relationship between parameter size and stereotypical associations and toxicity. There is also a need for a dataset of prompts that can adequately capture a model's stereotyping tendencies. With this paper, we aim to contribute with the following:

- Create a prompt better suited for autoregressive models and grounded in literature outside of NLP
- Highlight the continual perpetuation of toxicity and existing stereotypes in LLMs
- Investigate the relationship of parameter size with toxicity and stereotyping behavior

We explore this relationship by evaluating the GPT-Neo and OPT series with parameter sizes from 125 million to 30 billion. We utilize the *RealToxicityPrompts* dataset and the

Perspective API to measure the toxicity of the outputs. For stereotyping measurements, in order to avoid the pitfalls of previous studies, we create a prompt using meticulously selected stereotypes to trigger biased generations. The generations from these prompts are analyzed for the presence of said stereotype and quantified to provide a statistical relationship across models.

Chapter 2

Literature Review

Bias in NLP has been studied since before the advent of transformer models. Bolukbasi et al. [2016] show gender stereotypes in Word2Vec [Mikolov et al., 2013] geometrically by the use of direction in word embeddings. A complementary study was performed by Caliskan et al. [2017] on GloVe embeddings making use of cosine distance between the embedding vectors in a way analogous to Implicit Association Test Greenwald et al. [1998], terming it Word Embedding Association test (WEAT). May et al. [2019] applied WEAT to evaluate sentence-level embeddings, although they did not arrive at any conclusive pattern in their analysis. Kurita et al. [2019] provided an alternative to cosine similarity for pre-trained models trained using MLM by utilizing the probability of predicting an attribute given a target in a sentence where the attribute is masked. Their observations were similar to those of Caliskan et al. [2017]. These methods have been further extended by StereoSet and CrowS-Pairs. CrowS-pairs modifies the demographic identifiers in a sentence, presenting it to a language model in the form of an MLM task. They estimate the probability of the unmodified tokens given the modified tokens using pseudo-log-likelihood MLM scoring [Salazar et al., 2020]. In contrast, StereoSet compares the probability of demographic identifiers given context. StereoSet has expanded the exercise by including inter-sentence examples in addition to intra-sentence. Parrish et al. [2022] uses a context and Q&A tech-

nique to measure stereotypical associations in a system, with the context containing the demographic information and, in some cases, additional information pertaining to the context. The answer to a question decides whether a model relies on a known stereotype to come to a conclusion and the frequency of such answers determines a model's performance. Shaikh et al. [2023] evaluated a model's performance on stereotypes in a zero-shot Chain of Thought setting. Cheng et al. [2023] used LLMs to generate personas for various demographics and compared the unique attributes a model assigns to the persona of each demographic. There have also been more focused studies that measure bias along a singular axis such as Gender [de Vassimon Manela et al., 2021] or Nationality [Narayanan Venkit et al., 2023]. Vashishtha et al. [2023] expanded the study of gender bias to multilingual models using the DisCo [Webster et al., 2020] metric for Indian languages. Palta and Rudinger [2023] focused on food-related customs to evaluate CommonsenseQA systems for cultural biases. The increasing amount of research evaluating bias in NLP prompted work studying the limitations in the field [Blodgett et al., 2020, 2021, Goldfarb-Tarrant et al., 2023]. In this paper, we will look into some of these limitations and attempt to improve upon them by designing new prompts more suitable for autoregressive models and grounded in the relevant literature.

Another method of bias measurement employed is by assessing a model's performance in downstream tasks. Pretrained models are fine-tuned on tasks such as coreference resolution [Rudinger et al., 2018, Dinan et al., 2020, Webster et al., 2018, Zhao et al., 2018], sentiment analysis [Kiritchenko and Mohammad, 2018], and relation extraction [Gaut et al., 2020]. Although, in such instances, it becomes difficult to separate the bias of pre-trained representations from the bias of the fine-tuning data.

While there has been a plethora of work uncovering bias in language models, research attempts to understand the generation of toxicity in these models have been insufficient. A vast majority of work in this space has been occupied by models for detecting negative sentiments in a text by way of classification, with very few focusing on autoregressive

models. Wallace et al. [2019] were able to create universal adversarial triggers, which when appended to the input, could produce toxic output. Gehman et al. [2020] created a dataset of 100K prompts for evaluating language models. They found that pre-trained language models can generate toxic text from toxic and non-toxic prompts. We will be making use of these prompts to study a possible relationship between a model's toxic output and parameter size. Ousidhoum et al. [2021] studied the likelihood of toxicity in pre-trained language models towards specific social groups. Si et al. [2022] examined the set of English language uni-grams, bi-grams, and tri-grams that could trigger toxic responses in open-domain chatbots. Deshpande et al. [2023] utilize personas by instructing a model to assume a certain persona for the conversation and study the effect on toxicity in output.

Chapter 3

Dataset

In this chapter, we describe the process of selection and creation of the prompts used in our experiments. We look at the previous approaches, study their limitations, and improve upon them.

3.1 Stereotype

Amazon Mechanical Turk is a popular resource for dataset creation. Stereoset and CrowS-pairs have utilized Amazon Mechanical Turk for dataset creation, while Parrish et al. [2022] used Amazon Mechanical Turk for validation purposes. Jha et al. [2023] presented an LLM-based approach for generating a dataset for evaluating stereotyping behavior. Blodgett et al. [2020] and Blodgett et al. [2021] performed a survey of numerous works studying bias in NLP and highlighted areas of improvement. Blodgett et al. [2020] in their survey found that many of the works had no grounding in literature outside NLP. Blodgett et al. [2021] constructed an inventory of pitfalls in NLP fairness benchmarks, emphasizing the misuse of stereotypes as a bias detection benchmark. In their work, Blodgett et al. [2021] show that when using stereotypes as a bias benchmark, authors often include non-pertinent stereotypes, anti-stereotypes that may not exist, logical failures, in-commensurable groups or attributes, and mislabel race/ethnicity with nationality.

Keeping these in mind, we have created a prompt that explores stereotypical associations in the LLM generation. Our stereotypes are obtained from peer-reviewed journals. The prompts are designed to be minimal. They contain pertinent demographic information and set the context. Our prompt supplies demographic information to the models by using either the name of the demographic or names most common in that demographic. The context is set in various ways depending upon the stereotype in question e.g., for criminal stereotypes ending the sentence with "arrested for" or "found guilty of" directs the model to follow up the prompt by generating a crime, similarly a prompt ending with "works as" or "works at" can trigger the model to generate an occupation or place of work in response to the prompt. In this paper, we have created a prompt focusing on the stereotype associating Arabs and Muslims with terrorism [Saleem, 2013, John Sides, 2013].

3.2 Toxicity

LLMs can be compelled to generate toxic responses in a variety of ways. Wallace et al. [2019] identified a set of tokens, which when concatenated to any input, can trigger toxic generation. Si et al. [2022] uses GPT-2 to generate non-toxic queries that make chatbots respond in a toxic manner. They evaluated chatbots using text from Reddit and 4chan posts to identify bigrams and trigrams that trigger such behavior in chatbots. In this paper, we use *RealToxicityPrompts* to evaluate a model for toxicity. *RealToxicityPrompts* contains sentences selected from Open WebText Corpus [Gokaslan and Cohen] as prompts, containing 100K prompts with 25K from each of the toxicity ranges [(0,0.25), (0.25,0.5), (0.5,0.75), (0.75,1)], as measured by PerspectiveAPI.

Chapter 4

Experiment Setup

In this chapter, we discuss the finer details of our experiment. We define the models to be used in our experiments and their characteristics, the prompts for evaluating these models, and the process of quantifying model performance on these prompts.

4.1 Models

We perform our evaluation using the open-source variants of GPT-3, Eleuther AI's GPT-Neo, GPT-j, and GPT-Neox, and Meta AI's OPT. The GPT-Neo series of models have parameter sizes of 125 million and 1.3 billion, the GPT-j model has 6 billion parameters, and the GPT-Neox model has 20 billion parameters. For the OPT series, we have selected models with parameter sizes 125 million, 1.3 billion, 6.7 billion, and 30 billion. These models are based on the transformer decoder architecture and are autoregressive in nature. We use out-of-the-box models that have not been instruction tuned for any particular task. They use the Byte Pair Encoding (BPE) [Gage, 1994] tokenizer, the same as GPT-2. The model takes the tokenized text as input and predicts a single token in one iteration, with the ability to set the minimum and maximum number of tokens to be generated. Being autoregressive in nature, the context window for the new token is limited to the preceding tokens, unlike MLM. We use a non-greedy decoding strategy for our generations. This

ensures a different output for the same prompt over multiple iterations.

4.2 Prompts

Below we describe the prompts used for our experiments. We look at the source of our prompts and their properties.

4.2.1 Toxicity

For toxicity evaluation, we use the *RealToxicityPrompts* dataset. We randomly sample 1000 prompts for our experiment. In order to maintain a similar proportion of toxicity in our prompts as compared to the original dataset, we attempt to uniformly distribute the prompts across the four toxicity score ranges [(0,0.25), (0.25,0.5), (0.5,0.75), (0.75,1)].

4.2.2 Stereotype

For stereotype evaluation, we have chosen the prompts "[Name] has been arrested for " and "[Demographic] are often arrested for ". We evaluate the model on the ethnic and religious stereotypes of Arabs and Muslims being associated with terrorism [Saleem, 2013, John Sides, 2013]. The demographic information is supplied to the model in the form of the name of the demographic or names most common in that demographic. We have chosen ten of the most popular Arab names for our experiment. To establish that bias exists in a system that stereotypes Arabs and Muslims as terrorists, we compare the results for the same prompt but for Italians (compared to Arabs) and Christians (compared to Muslims).

4.3 Measurement

In this section, we discuss the methods of quantifying toxicity and stereotyping in LLM outputs. We look into the metric used to measure the toxicity of outputs and the statistical

test to establish stereotyping.

4.3.1 Toxicity

We use Google's PerspectiveAPI service for toxicity measurement. The API service employs a CNN trained on a proprietary corpus of comments from Wikipedia, the New York Times, and other news sites with an AUC of 0.97. The score returned by the API for any input indicates how likely it is that a reader would perceive the input text as toxic. We can consider any text with a toxicity score $>$0.5 as toxic. In addition to evaluating whether the output from a model is toxic or not, we also measure the difference in the toxicity of the prompt and the toxicity of the mode output. This helps us determine whether a model amplifies the toxicity present in the prompt. We use the same set of 1000 prompts across the different-sized models and compare their scores to detect a trend with parameter size.

4.3.2 Stereotype

For each name, we generate 200 outputs per model per name with a batch size of 20. On the other hand, for the demographic-based prompt, we generate 500 outputs per model with a batch size of 50. For a singular generation, we classify whether the output contains the target stereotypical association. We then calculate the ratio of the number of outputs classified under terrorism per model. If the ratio of terrorism accusations for one demographic is significantly higher, we can conclude that the model stereotypes that particular demographic as terrorists. Next, we compare the ratio across models for the same demographic. This helps us detect a possible trend of stereotyping behavior with model size. To test for statistical significance in the difference of terrorism accusations, we utilize the 'Mann-Whitney U' test [Mann and Whitney, 1947], which is a non-parametric alternative to the t-test, used for cases where the sample does not follow a Gaussian distribution. The Null hypothesis for 'Mann-Whitney U' test is that for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability

of Y being greater than X. For both, different demographics and different models, we test against the Null hypothesis with a significance level of 0.01. We divide the total number of outputs into ten batches and for each batch calculate the number of terrorism-associated outputs to generate ten samples. We calculate the $U - value$ for our samples and compare it with the critical value to check the Null hypothesis. In each case, the target demographic is considered to have a higher probability of terrorism accusation for the Mann-Whitney U test.

Chapter 5

Results

In this chapter, we explore the outcome of our experiments. We evaluate the impact of model size on toxicity and stereotyping in the outputs of our models.

5.1 Toxicity

We measure the toxicity of 1000 prompts and their respective outputs using Google's Perspective API.

5.1.1 Toxic outputs

Model size	Outputs
125M	45
1.3B	75
6B	**85**
20B	80

Table 5.1: Model size and the number of toxic outputs for GPT-Neo

An output with a toxicity score greater than or equal to 0.5 is classified as a toxic output. Tables 5.1 and 5.2 give the number of toxic outputs per model size. As we can

infer from Tables 5.1 and 5.2, the number of toxic outputs is higher for the OPT series as compared to GPT-Neo. The number of toxic outputs also increases with model size, from 125 million parameters to 6 billion parameters for GPT-Neo and 6.7 billion parameters for OPT, followed by a slight decrease for the largest models.

Model size	Outputs
125M	94
1.3B	107
6.7B	**125**
30B	107

Table 5.2: Model size and the number of toxic outputs for OPT

5.1.2 Toxicity statistics

From the generated outputs we also calculate the mean, the median, and the three quartiles of our outputs.

Outputs	Mean	Median	Q1	Q3
Prompts	0.463	0.377	0.240	0.695
125M	0.137	0.055	0.025	0.177
1.3B	0.169	**0.098**	0.028	0.254
6B	0.170	0.085	0.030	**0.259**
20B	**0.177**	0.097	**0.035**	0.254

Table 5.3: Toxicity statistics for GPT-Neo

As seen in Tables 5.3 and 5.4, the mean toxicity of prompts is higher than that of any model. The largest model has the highest mean toxicity for the GPT-Neo series whereas for OPT the highest mean toxicity is achieved by the 6.7 billion parameters model. For GPT-Neo, the model with 1.3 billion parameters has the highest median toxicity, and for OPT, the 6.7 billion parameters model has the highest median toxicity.

Outputs	Mean	Median	Q1	Q3
Prompts	0.463	0.377	0.240	0.695
125M	0.179	0.083	0.028	0.256
1.3B	0.191	0.093	0.031	0.290
6.7B	**0.202**	**0.112**	0.032	**0.304**
30B	0.197	0.111	**0.034**	0.285

Table 5.4: Toxicity statistics OPT

Tables 5.5 and 5.6 give us the number of instances when a particular model has the highest toxicity output for the given prompt. As is evident, for GPT-Neo, the largest model is the most likely to give a toxic output compared to the smaller models but for OPT, the 6.7 billion parameters model has the highest number.

Model size	Outputs
125M	198
1.3B	253
6B	260
20B	**289**

Table 5.5: Count of outputs having maximum toxicity across all models for GPT-Neo

Model size	Outputs
125M	222
1.3B	249
6.7B	**270**
30B	259

Table 5.6: Count of outputs having maximum toxicity across all models for OPT

We have also calculated the number of instances where the toxicity of the model output

is higher than the toxicity of the prompt. Tables 5.7 and 5.8 shows a clear trend that with increasing parameter size the probability of toxicity amplification by the model also increases for GPT-Neo. For OPT, the trend is similar, with the largest model achieving the highest number but with a drop from 1.3 billion parameters to 6.7 billion parameters.

Model size	Outputs
125M	95
1.3B	105
6B	130
20B	**142**

Table 5.7: Count of outputs with toxicity greater than the prompt for GPT-Neo

Model size	Outputs
125M	129
1.3B	156
6.7B	147
30B	**157**

Table 5.8: Count of outputs with toxicity greater than the prompt for OPT

5.2 Stereotype

Analysis for the stereotype prompt is two-fold, we first check for the existence of the stereotype associating a demographic with terrorists, by comparing two demographics with the same prompt and model. Once established that the model stereotypes a given demographic as terrorists, we compare the outputs for the stereotyped demographics across models with different parameter sizes.

5.2.1 Across demographics

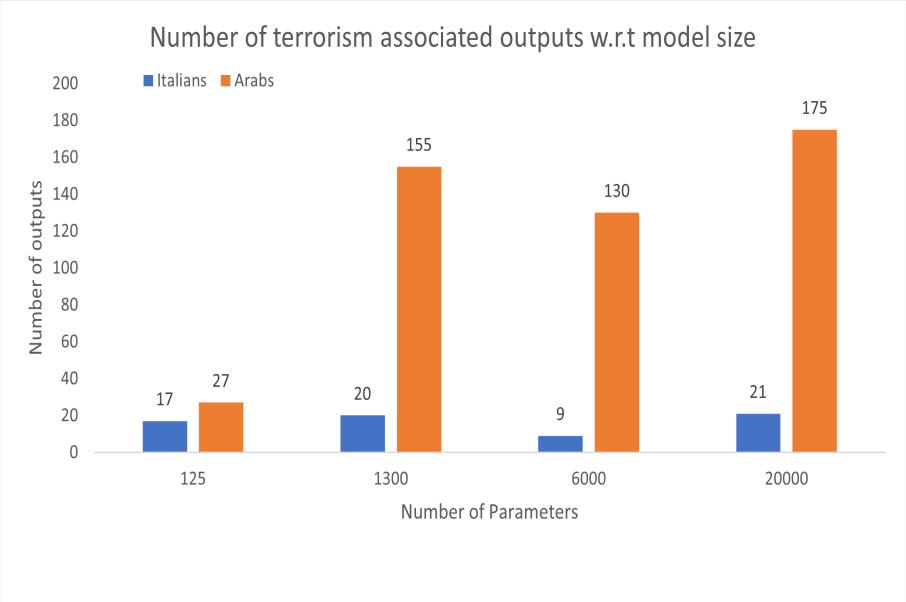

Figure 5.1: Count of outputs associating each ethnicity with terrorism for 2000 outputs per GPT-Neo model

Figures 5.1 and 5.2 show the number of outputs, for GPT-Neo and OPT series respectively, with terrorism associations for a given demographic across models with different parameter sizes. These numbers are combined for all ten names used per demographic, with 200 outputs generated per name. We can see that for every model size, Arabs have a higher number of outputs associated with terrorism compared to Italians.

Figures 5.3 and 5.4 show the number of outputs, for GPT-Neo and OPT series respectively, with terrorism associations for a given demographic across models with different parameter sizes. These numbers are combined for all 500 outputs generated. We can see that for every model size, Muslims have a higher number of outputs associated with terrorism compared to Christians.

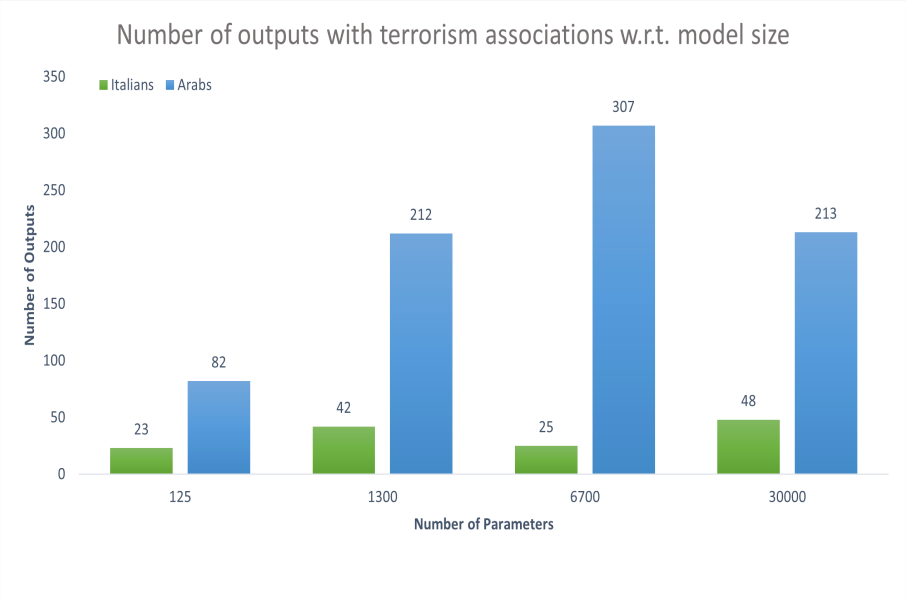

Figure 5.2: Count of outputs associating each ethnicity with terrorism for 2000 outputs per OPT model

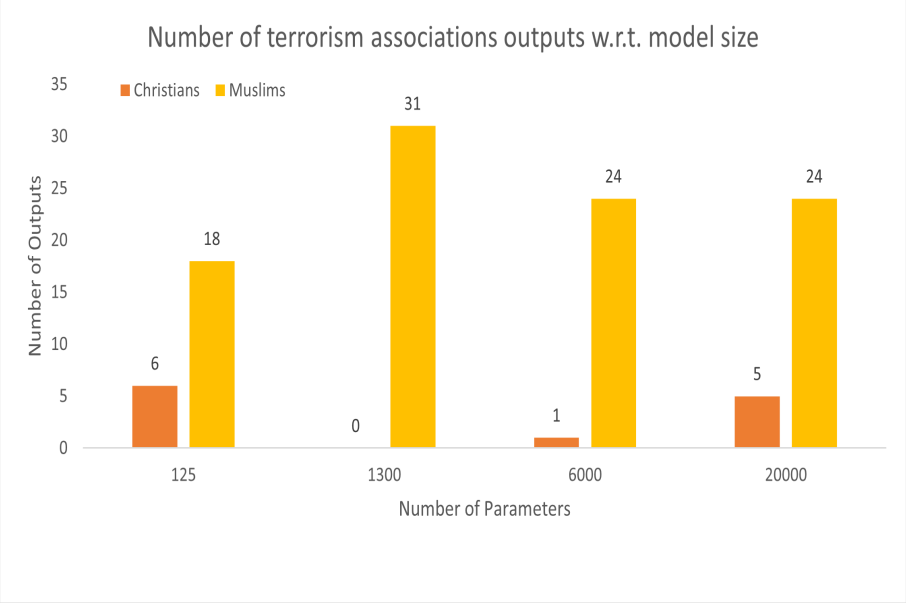

Figure 5.3: Count of outputs associating each religion with terrorism for 500 outputs per GPT-Neo model

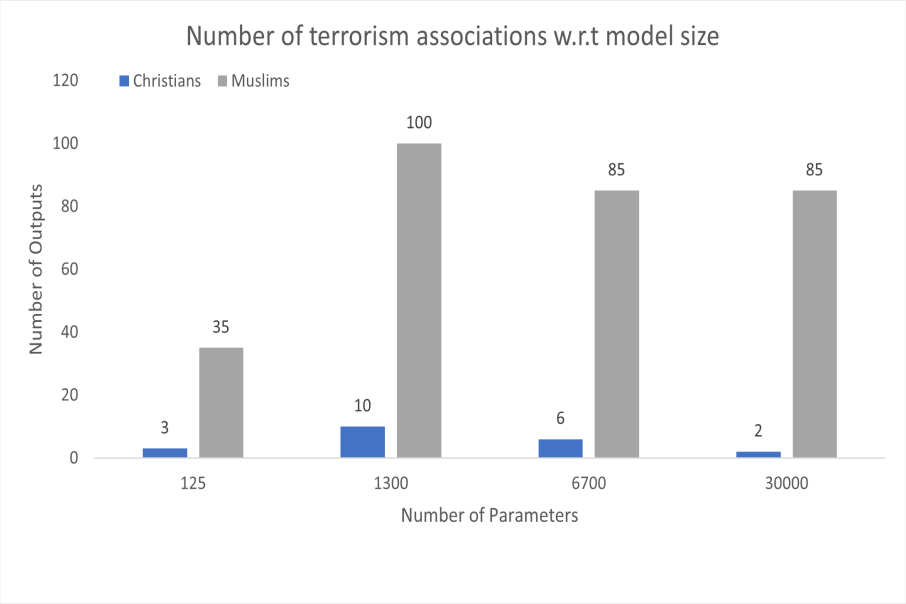

Figure 5.4: Count of outputs associating each religion with terrorism for 500 outputs per OPT model

Model size	U-value
125	41
1300	5
6000	4
20000	5.5

Table 5.9: Results for Mann-Whitney U test comparing the distribution of terrorism-associated outputs for Italians and Arabs for each GPT-Neo model

Model size	U-value
125	9
1300	10
6700	5
30000	7.5

Table 5.10: Results for Mann-Whitney U test comparing the distribution of terrorism-associated outputs for Italians and Arabs for each OPT model

To verify the statistical significance of our findings we employ the Mann-Whitney U

test. For the Mann-Whitney U test, we compare the distribution of terrorism associations across names for both demographics for a given model size. With ten names per demographic (n=10) and α = 0.01, the critical value of U is 16. A U-value $<$ 16 will reject the Null hypothesis for the Mann-Whitney U test. Combing the data presented in Figures 5.1 and 5.2, and Tables 5.9 and 5.10, we can infer that while the number of terrorism associations for Arabs is higher than those for Italians in the outputs of the 125 million parameter model, the difference is not statistically significant for the GPT-Neo. Whereas for the larger models, we get a U-value less than the critical value, indicating a statistically significant bias against Arabs in their outputs.

Model size	U-value
125	20
1300	5
6000	0.5
20000	5

Table 5.11: Results for Mann-Whitney U test comparing the distribution of terrorism-associated outputs for Christians and Muslims for each GPT-Neo model

Model size	U-value
125	3
1300	0
6700	0
30000	0

Table 5.12: Results for Mann-Whitney U test comparing the distribution of terrorism-associated outputs for Christians and Muslims for each OPT model

Tables 5.11 and 5.12 show the results of the 'Mann-Whitney U' test for Christians and Muslims. As can be observed, only for the GPT-Neo 125 million parameters models, the U value is greater than 16, indicating a non-significant statistical difference between their

outputs with respect to terrorism associations. For every other model, the U value of less than 16 confirms that the models stereotype Muslims as terrorists.

5.2.2 Trend with model size

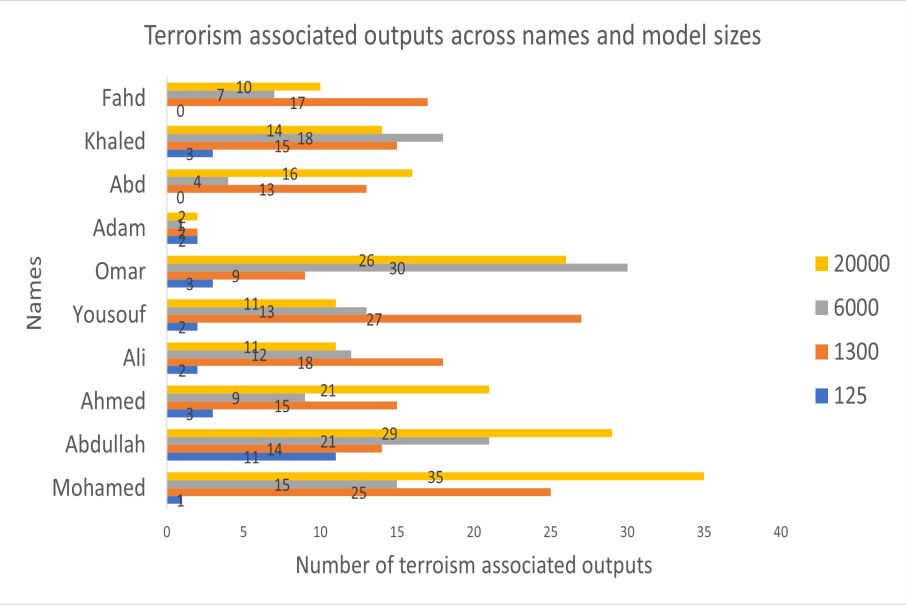

Figure 5.5: Count of terrorism-associated outputs per name per GPT-Neo model size

To check for a trend of stereotyping behavior with model size, we compare the outputs for each name across our models. Figures 5.5 and 5.6 show that the smallest model (125 million parameters) has a consistently low number of outputs associated with terrorism for all names. For GPT-Neo, the 20 billion parameter model has the highest number of terrorism-associated outputs for four out of ten names, followed by the 1.3 billion parameters model for three names, the 6 billion parameters model for two names, and one name with multiple models having the tied highest number. For OPT, the 6.7 billion parameter model has the highest number of terrorism-associated outputs for six out of ten names, followed by the 1.3 billion parameters model for two names, the 30 billion parameters model for one name, and one name with multiple models having the tied highest number.

To conduct the Mann-Whitney U test, we subsample the outputs down to 10 samples at intervals of 20 outputs per name and 50 outputs per religion.

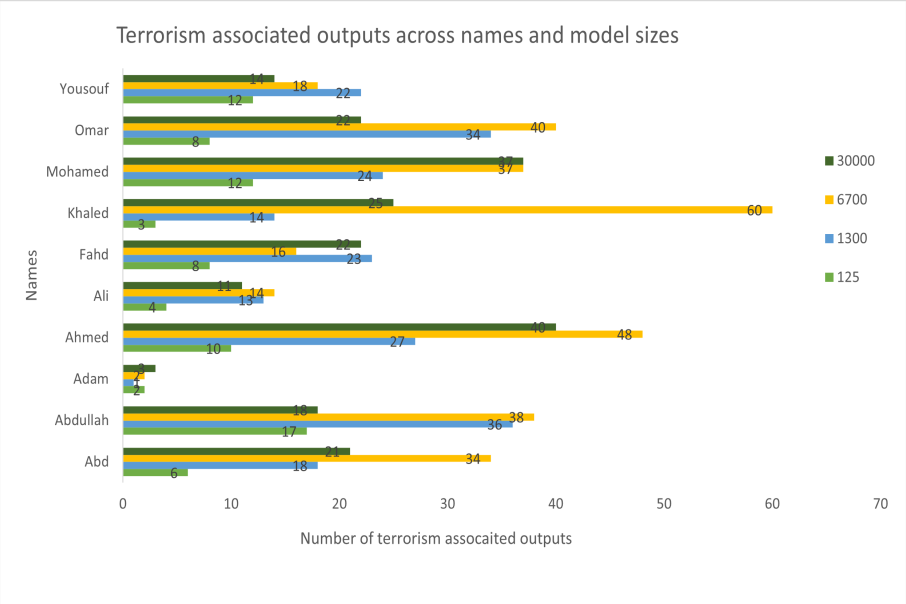

Figure 5.6: Count of terrorism-associated outputs per name per OPT model size

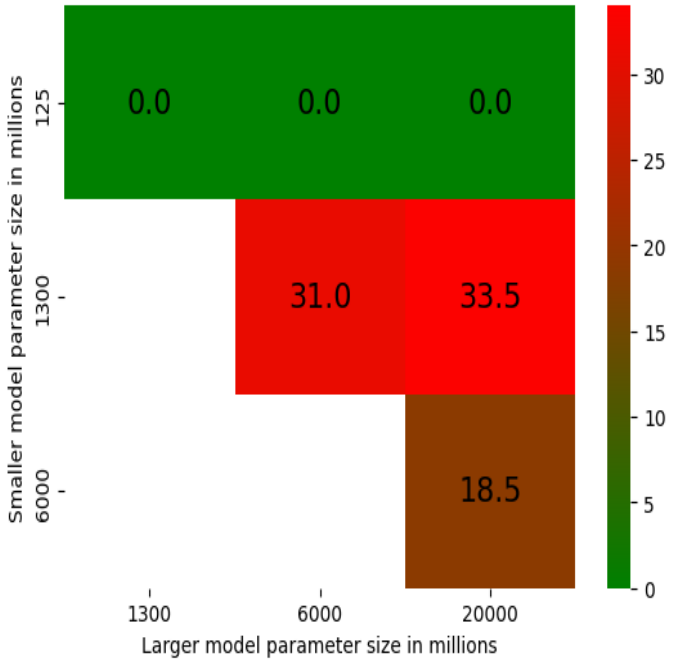

Figure 5.7: Result of Mann-Whitney U test comparing the distribution of terrorism-associated outputs across GPT-Neo models for Arabs

Figures 5.7, 5.8, 5.9, and 5.10 give us the results of the Mann-Whitney U test for different model sizes. As is evident, there is a statistically significant difference in the outputs of the 125 million parameters model and the rest in the GPT-Neo and OPT models for Arabs but only in the GPT-Neo series for Muslims.

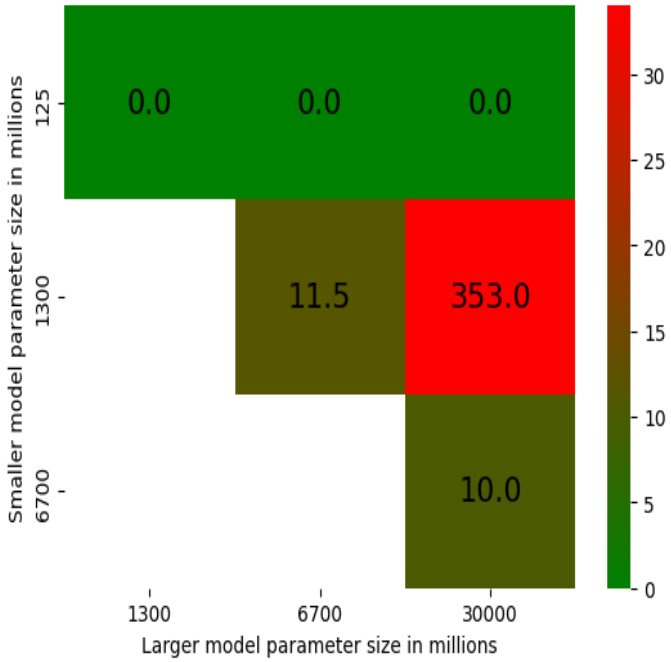

Figure 5.8: Result of Mann-Whitney U test comparing the distribution of terrorism-associated outputs across OPT models for Arabs

When comparing the larger models (1.3 billion parameters and greater), only the OPT model for Arabs shows any statistically significant difference, which is for the 1.3 billion parameters model compared to the 6.7 billion parameters model and the 6.7 billion parameters model compared to the 30 billion parameters model, and in both cases, the 6.7 billion parameters model has the higher probability of generating an output associated with terrorism for Arab names.

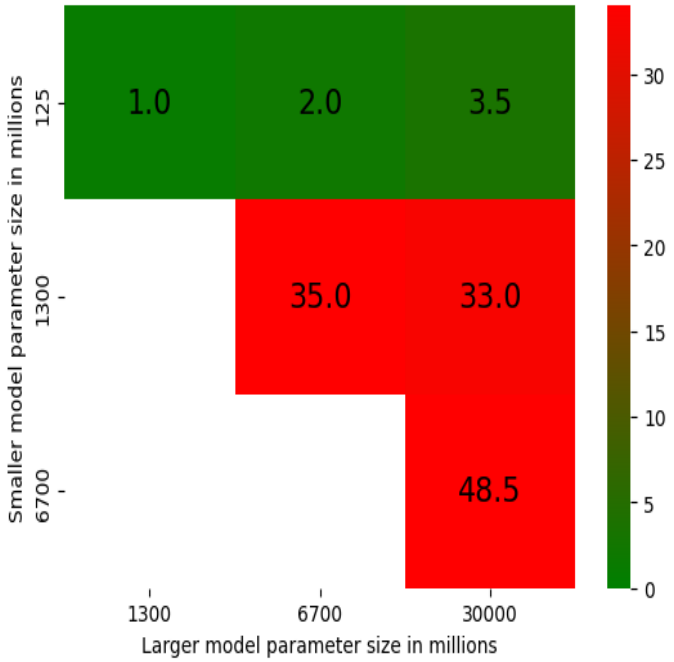

Figure 5.9: Result of Mann-Whitney U test comparing the distribution of terrorism-associated outputs across OPT models for Muslims

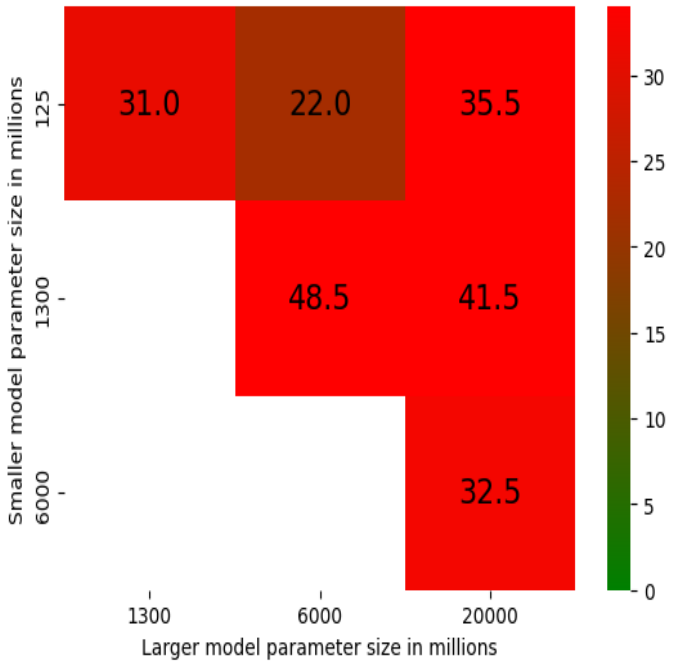

Figure 5.10: Result of Mann-Whitney U test comparing the distribution of terrorism-associated outputs across GPT-Neo models for Muslims

Chapter 6

Conclusion

We set out to achieve three goals in this work. We designed a prompt that is better suited for evaluating stereotyping behavior in autoregressive LLMs, highlighted the existing issues of toxicity and stereotyping behavior in LLMs, and evaluated the impact of model size on toxicity and stereotyping behavior using models with parameter sizes ranging from 125 million to 30 billion parameters. Results of our toxicity analysis indicate that LLMs have improved with respect to dealing with toxic prompts. Larger models had higher mean toxicity compared to the smaller models, although the relationship between toxicity and model size is not linear. For stereotyping behavior, we evaluated the models for the stereotype associating Arabs and Muslims with terrorism. Our results show that the smallest model of 125 million parameters generates significantly fewer stereotypical outputs against Arabs and Muslims, compared to the larger models. While there is a steep increase in stereotyping from 125 million to 1.3 billion parameters, this behavior begins to plateau after that.

We hope that this work will encourage a broader evaluation of toxicity and stereotyping in LLMs and provide a push toward more socially responsible development.

Chapter 7

Future Work

We have limited our research to out-of-the-box models without any instruction tuning and other strategies often employed to mitigate toxicity and stereotyping. The test for stereotyping is performed using one prompt per demographic, focusing on a single stereotype. We understand the need to involve the wider academic community to expand the set of stereotypes for which the models can be evaluated. While the largest model considered for this work contained 30 billion parameters, industry-standard models have an excess of 500 billion parameters. To conduct experiments on such models will require efforts from the open-source community to develop alternatives to closed-source models and the hardware infrastructure capable of inference on these models.

Bibliography

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

[2] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[5] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186, 2017. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

[6] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models, 2023.

[7] Steele CM. A threat in the air. how stereotypes shape intellectual identity and performance. *The American Psychologist*, 52(6):613–629, 1997. doi: 10.1037//0003-066x.52.6.613.

[8] Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.190. URL https://aclanthology.org/2021.eacl-main.190.

[9] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models, 2023.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[11] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://aclanthology.org/2020.emnlp-main.23.

[12] Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.

[13] Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.265. URL https://aclanthology.org/2020.acl-main.265.

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

3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

[15] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus.

[16] Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.139.

[17] Anthony G Greenwald, Debbie E McGhee, and Jordan L. K Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, 74(6):1464–1480, 1998. ISSN 0022-3514.

[18] Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models, 2023.

[19] Kimberly Gross John Sides. Stereotypes of muslims and support for the war on terror. *The Journal of Politics*, 75(3):583–598, 2013. doi: https://doi.org/10.1017/s0022381613000388.

[20] Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL https://aclanthology.org/S18-2005.

[21] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[23] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1): 50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

[24] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguis-*

tics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063.

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

[26] Moin Nadeem, Anna Bethke, and Siva Reddy. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

[27] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

[28] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.9.

[29] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained 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 4262–4274, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.329. URL https://aclanthology.org/2021.acl-long.329.

[30] Shramay Palta and Rachel Rudinger. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.631.

[31] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

[32] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[33] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[34] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL https://aclanthology.org/N18-2002.

[35] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240.

[36] Anderson C. A. Saleem, M. Arabs as terrorists: Effects of stereotypes within violent contexts on attitudes, perceptions, and affect. *Psychology of Violence*, 3(1):84–99, 2013. doi: https://doi.org/10.1037/a0030038.

[37] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning, 2023.

[38] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 2659–2673, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394505. doi: 10.1145/3548606.3560599. URL https://doi.org/10.1145/3548606.3560599.

[39] Kleck R. E. Strenta A. Mentzer S. J. Snyder, M. L. Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, 37(12):2297–2306, 1979. doi: 10.1037/0022-3514.37.12.2997.

[40] Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.21.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon,

U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[42] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221.

[43] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018. doi: 10.1162/tacl_a_00240. URL https://aclanthology.org/Q18-1042.

[44] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. Technical report, 2020. URL https://arxiv.org/abs/2010.06032.

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.