DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

Don't Mention the Norm

On the Relationship Between Reporting Bias and Social Bias in Humans and Language Models

Tom Södahl Bladsjö

Essay/Thesis:	Master's thesis, 30 credits
Programme/Course:	Master in Language Technology, 120 credits
Level:	Second cycle
Semester:	Spring, 2024
Supervisor:	Ricardo Muñoz Sánchez and Elena Volodina
Examiner:	Asad Sayeed
Keywords:	reporting bias, social bias, markedness, pragmatics, language modeling

Abstract

Reporting bias (the human tendency to not mention obvious or redundant information) and social bias (societal attitudes toward specific demographic groups) have both been shown to propagate from human text data to language models trained on such data (Shwartz and Choi, 2020; Paik et al., 2021; Caliskan, Bryson, and Narayanan, 2017; Garg et al., 2018). However, the two phenomena have not previously been studied in combination. This thesis aims to begin to fill this gap by studying the interaction between social biases and reporting bias in both human text and language models. We conduct a corpus study of human-written text, and find that n-gram frequencies in our chosen corpora show strong signs of reporting bias with regard to socially marked identities, mirroring current discourse in society. This thesis also introduces the MARB dataset for measuring model reporting bias with regard to socially marked attributes. We evaluate ten large pretrained language models on MARB and analyze the results in relation to both corpus frequencies and real-world frequencies. The results suggest a relationship between reporting bias and social bias in language models similar to that which was identified in human text. However, this relationship is not as straightforward in language models, and other factors, like sequence length and model vocabulary, are also observed to affect the outcome.

Acknowledgements

First of all, I would like to thank my supervisor Ricardo Muñoz Sánchez. Thanks for taking the time to talk things through with me, for explaining all the mathy things, for going off on tangents, and for never shying away from things being complex. Everything I could have asked for from a supervisor, and more!

I am also very grateful to Elena Volodina for her help with corpora and figuring out sharing permissions, and to Miloš Jakubíček for providing the TenTen corpora, and for giving me permission to share my derived dataset.

I would also like to thank Asad Sayeed for useful feedback and discussion at my defence, which helped improve the final thesis.

Thanks also to Bill, for answering a million questions, helping me sort out my own thoughts, putting up with my whining, and for reminding me to eat.

Finally, thanks to Pontus, for understanding what I'm trying to say and helping me say it in a way that makes sense.

Contents

1 Introduction

(a) A little girl in a pink dress going into a wooden cabin.

(b) An Asian girl in a pink dress is smiling whilst out in the countryside.

Figure 1.1: Example from the Flickr8k dataset (Hodosh, Young, and Hockenmaier, 2015).

A language model is, in essence, an advanced statistical model of a subset of human language. Like in any inferential statistics, the goal of language modeling is to get the model to learn the patterns underlying its training data well enough to be able to generalize beyond it; to predict previously unseen samples from the same population. For example, we would expect a model trained on standard English to assign a higher probability to the sentence "Aaron **broke** the unicycle" than to "Aaron **broken** the unicycle", even if neither sentence was part of the training data (example from Warstadt et al., 2020). A language model that has been pretrained in this way can then be used (usually after finetuning) in a wide variety of downstream applications that require dealing with human language.

However, language as a system does not exist in a vacuum. It is produced by people who are part of a specific society and culture, in response to real-life situations. The way we communicate, and the things we choose to communicate about, depend on the situation, the context, who we are, and who we are communicating with. From this perspective, language is a social and cultural phenomenon, and samples of language will reflect the society and culture in which they were produced, as well as the individuals who produced them. Thus a statistical model of language will also, to some degree, be

a model of the society and culture that produced its training data. This is evidenced by social biases found in models trained on language (Bolukbasi et al., 2016; Caliskan, Bryson, and Narayanan, 2017; Wang et al., 2019; Dhamala et al., 2021; Bender et al., 2021). In fact, measures of bias in models can be used to quantify changes in attitudes within a society (Garg et al., 2018).

Another feature of human communication that is reflected in statistical models of language is known as *reporting bias*, or as Paik et al. (2021) put it, "the tendency of people to not state the obvious." For example, we are more likely to see an article about a murder in a newspaper than one about a person going to the grocery store. This feature of language can cause problems when training systems on text, because, as Gordon and Van Durme (2013) note:

> Much work in knowledge extraction from text tacitly assumes that the frequency with which people write about actions, outcomes, or properties is a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.

It is not unlikely that reporting bias – that is, what is and is not mentioned – interacts with societal norms and attitudes. As the *National Center on Disability and Journalism*, an American organization providing guidance for journalists writing about people with disabilities, note in their style guide:

> People living with disabilities often complain, and rightly so, that their disability is mentioned even when the story has nothing to do with their disability.
> (NCDJ, 2021)

Similarly, we can see examples of reporting bias with regard to to marginalized ethnicities in datasets used for training image captioning models. Figure 1.1 shows two images and their corresponding captions from the Flickr8k dataset. Both images depict children in pink dresses, but their captions differ in the types of information they include: One child is described as "an *Asian* girl", while the other is described simply as "a little girl". In other words, the perceived ethnicity of the child in 1.1b was deemed important or unexpected enough (by the annotator) to warrant mentioning, whereas the whiteness of the child in 1.1a was not.

Surprisingly, the relationship between social biases and reporting bias seems to have been neglected in research to date. In the current work we attempt to remedy this by exploring reporting bias with regard to marginalized attributes (such as disability and race) in both human text and language models. Our aim is to answer the following research questions:

1. How does societal bias against an attribute relate to the frequency with which it is mentioned by humans?

2. How is this relationship reflected in models trained on human language data?

To this end, we conduct two experiments that build on each other. The first experiment is a corpus study that investigates how often a set of attributes (relating to disability, race/ethnicity and queerness) are mentioned by humans. In the second experiment, we construct a dataset for measuring differences in model performance for sequences containing mentions of the same attributes. We evaluate ten popular large pretrained language models on the dataset and compare the results to the results of our corpus study. The results of both experiments are then discussed in relation to the real-world frequencies of the attributes of interest.

1.1 Contributions

Recent years have seen an increasing awareness of the social biases encoded in language models, and the potential harm they could cause in downstream applications. A large body of work has focused specifically on detecting harmful stereotypes toward marginalized groups (e.g. Bolukbasi et al., 2016; Zhao et al., 2018; May et al., 2019; Nadeem, Bethke, and Reddy, 2021; Tal, Magar, and Schwartz, 2022). While stereotypes are one way in which societal attitudes toward a group may manifest in language, it is unlikely to be the only one. One contribution of this work is to expand the notion of bias to include other ways in which attitudes toward a protected group may be reflected in language – in this case, whether or not group membership is mentioned in descriptions of individuals. We also introduce a new dataset and propose a method for measuring this type of bias in models.

As mentioned above, the relationship between reporting bias and social biases is largely unexplored. Another contribution of this thesis, therefore, is to take the first few steps into this area, and illuminate challenges and opportunities for future research.

1.2 Outline

The thesis is divided into seven chapters. This first chapter introduces the area of interest, as well as the specific research questions that will be explored. The second chapter provides more background information and introduces relevant concepts. §2.1 explains reporting bias in further detail and ties it to relevant theories in pragmatics. §2.2 introduces the concept of *markedness* and explain its relevance in linguistics and other social sciences. In §2.3 we present some estimates of the real-world frequency of some of the attributes that will be studied, to provide context for the results. In §2.4 we discuss research on related subjects and position the current work in relation to what has been done before. Chapter 3 presents and discusses the methods used for the corpus study (§3.2) and model evaluation (§3.3). §3.1 discusses the expressions used as proxies for the attributes of interest. §3.3.1 details the process of creating the MARB dataset, and §3.3.2 describes the metrics used for evaluating the models. In Chapters 4 and 5 we present the results of the two studies, and discuss how our observed results relate to each other and the real-world frequencies. In §5.2 we analyze the results in more depth, and discuss the different factors that contribute to our observations, as well as their

implications for any conclusions that can be drawn. Chapter 6 discusses the limitations of the current work. In the final chapter, we summarize the conclusions that can be drawn from the experiments, and suggest areas for future work.

2 Background

2.1 Reporting Bias

It is a well-known fact in pragmatics that human language is by necessity underspecified. To economize resources, a speaker wishing to convey something to a listener must choose what information to include and what to leave out, based on what the listener can be assumed to already know or infer based on the context. In traditional Gricean pragmatics, this is encapsulated in the Maxims of Quantity (Grice, 1975):

1. Make your contribution as informative as is required (for the current purposes of the exchange).

2. Do not make your contribution more informative than is required.

In other words, speakers tend not to include information they consider redundant. This holds for both spoken and written communication, and leads to what Gordon and Van Durme (2013) call *reporting bias* – the discrepancy between reality and its description in text. They note, for example, that the textual frequency of words like *murdered* tends to be a lot higher than that of words like *breathed*, even though it is safe to assume that breathing is considerably more common than murder in the real world.

Within neo-Gricean pragmatics, authors such as Stephen Levinson have noted the relationship between the unexpectedness of the situation being described, and the length and complexity of the utterance used to describe it. He introduces the idea of "default interpretations", which can be neatly summarized by the following heuristics (Levinson, 2000, p. 6):

1. If the utterance is constructed using simple, brief, unmarked forms, this signals business as usual, that the described situation has all the expected, stereotypical properties;

2. If, in contrast, the utterance is constructed using marked, prolix, or unusual forms, this signals that the described situation is itself unusual or unexpected or has special properties.[1]

[1]Compare "The outlaw killed the sheriff" to "The outlaw caused the sheriff to die". The second sentence seems to imply that the way in which the outlaw caused the sheriff to die was not by any direct means, or at least not in the way we would expect (e.g. by shooting him), because if it was we would have used the first sentence (example from Levinson, 2000, p. 142).

We will return to the notion of *markedness* in the next section.

As noted by Gordon and Van Durme (2013) and Paik et al. (2021), among others, reporting bias may restrict or otherwise impact what is learned by language models trained on text produced by humans. In this work, we will investigate a specific type of reporting bias, namely that which relates to marginalized attributes.

2.2 Markedness (in Linguistics and in General)

Markedness as a concept was first introduced by linguists Nikolaj Trubetzkoy and Roman Jakobson in the 1930s, where it was used to describe a type of distinction between forms. For example, Jakobson argues that in the Russian word pair *osël* (donkey) and *oslíca* (female donkey), *oslíca* bears a mark indikating female sex, while *osël* does not bear such a mark and can be used to mean either 'male donkey' or 'donkey of unspecified sex' (Jakobson, 1932; cited in Haspelmath, 2006, pp. 4-5). Since its introduction, the term *markedness* has been adopted in a variety of different subfields of linguistics and aquired a variety of different but related meanings, such as complexity, abnormality and unexpectedness. Sometimes (like in Jakobson's donkey example), the marked/unmarked distinction is viewed as something that exists within a given context (such as a specific language). In formal frameworks, such as Optimality Theory (Prince and Smolensky, 1993) and generative grammar (e.g. Chomsky and Lasnik, 1977), marked/unmarked distinctions are instead viewed as universals and assumed to be static across languages and contexts (Barrett, 2014; Haspelmath, 2006). Haspelmath (2006) identifies twelve different senses in which the word is used by linguists, and argues that most of them can be replaced by simpler and less ambiguous notions, such as frequency, difficulty and expectedness. However, the term has also been adopted in other social science disciplines, such as sociology and anthropology (e.g. Brekhus, 1998; Bucholtz and Hall, 2005), where it is used "to describe the process whereby some social categories gain a special, default status that contrasts with the identities of other groups" (Bucholtz and Hall, 2005, p. 372). In this sense, markedness is something that exists relative to a specific social context, rather than as an absolute, and it relates not only to expectedness and frequency, but also to power structures within a given social context. Specifically, members of the unmarked category tend to hold a power advantage over members of the marked category: "Because markedness implies hierarchy, differences between groups become socially evaluated as deviations from a norm and, indeed, as failures to measure up to an implied or explicit standard" (Bucholtz and Hall, 2005, p. 372). For example, experiments in psychology have shown that US White participants consistently associate *Human* with *White* more than with any other racial/ethnic group, whereas non-White participants showed no such *Human = Own Group* bias (Morehouse, Maddox, and Banaji, 2023). This suggests that *White* constitutes an unmarked norm in that social context. In the framework of Levinson (2000), we might say that the default interpretation of the word "human" seems to be "white human". Looking back at Figure 1.1 from Chapter 1, we can see this reflected in how people are described: The child in 1.1a is not described as white, presumably because whiteness is perceived as

unmarked and thus already implied by the lack of specification of the word "girl". 1.1b, on the other hand, depicts a marked situation (non-whiteness), which warrants linguistic marking ("An *Asian* girl"). Gay/straight and autistic/neurotypical are other examples of marked/unmarked distinctions in many social contexts in the US.

It is this sense of markedness that will be used in this work. We will combine the insights from sociology and linguistic anthropology with Levinson's understanding of the relationship between markedness and implicature to investigate how social markedness interacts with reporting bias, in human language use as well as in large pretrained language models.

2.3 Real-world Frequencies

The areas of social markedness that will be the focus in this work are *Disability*, *Race* and *Queerness*. When deciding if there is reporting bias pertaining to these categories, either in a corpus or in a language model, one must consider the relationship between how likely an attribute is to be mentioned in the corpus or model, and how frequent it is in the real world. Therefore, we present here some estimates of the real-world frequencies of the attributes of interest, reported by credible sources.[2] The real-world numbers for the *Race* and *Queerness* categories are reported in Tables 2.1 and 2.2. Regarding *Disability*, the World Report on Disability (WHO, 2011) estimates that about 15% of the world's population live with some kind of disability. Unfortunately, we could not find a more fine-grained estimation that corresponds to the expressions for disabilities used in the other parts of this project. However, this number is still informative when looking at the results of the corpus study and the model evaluation.

2.4 Related Work

In the past few years, the issue of bias in large pretrained language models has attracted a lot of attention, and there is a growing body of research on different aspects of social bias in models and its potential impact on downstream applications. Previous research on reporting bias, while somewhat less extensive than that on social bias, explores the relationship between training data and knowledge aquisition. However, there does not seem to be any existing research on the relationship between social bias and reporting bias in language models.

The relationship between social bias and reporting bias is surprisingly under-researched in traditional linguistics as well as in NLP. To our knowledge, there is currently no existing research in either pragmatics or socio-linguistics on the effect of social markedness

[2]Most population surveys focus on a specific phenomenon of interest. In order to collect real-world statistics for populations based on all three categories, we therefore have to look to different sources. When possible, we tried to find sources that reported or estimated global rather than place specific numbers. The exception is for the *Race* category, since the group terms used for racial and ethnic categories in this study are US-specific. Therefore, US Census Data (US Census Bureau, 2020) (tables P8 *Race* and P9 *Hispanic or Latino, and not Hispanic or Latino by Race*) are used in this case.

Group	%
Population of one race	89.8
Population of two or more races	10.2
White alone	61.6
Black or African American alone	12.4
American Indian and Alaska Native alone	1.1
Asian alone	6.0
Native Hawaiian and Other Pacific Islander alone	0.2
Some Other Race alone	8.4
Hispanic or Latino	18.7
Not Hispanic or Latino	81.3

Table 2.1: Percent of the US population that fall under each specified category according to the 2020 US Census (tables P8 and P9).

on whether or not an attribute is mentioned. For this reason, we conduct a corpus study in addition to the planned model evaluation, to enable comparisons between model behavior and human language production.

2.4.1 Social Biases in Language Models

The growing awareness of the potential harms of bias in language models in recent years has led to a rapid growth in this area of research. The bulk of the work in this field has been focused on gender- and/or racial bias[3] (Bolukbasi et al., 2016; Caliskan, Bryson, and Narayanan, 2017; Zhao et al., 2018; Garg et al., 2018; Kiritchenko and Mohammad, 2018; May et al., 2019; Kurita et al., 2019; Malik, 2023; Tal, Magar, and Schwartz, 2022, among others), but in later years a few papers have emerged focusing on bias against other social groups, such as people with disabilities (Hutchinson et al., 2020) and queer people (Felkner et al., 2023). In the current work we aim to widen the perspective by considering three dimensions of social bias (*Disability*, *Race* and *Queerness*), along with three different person-words (*person*, *woman* and *man*).[4]

Early works on bias (e.g Bolukbasi et al., 2016; Caliskan, Bryson, and Narayanan, 2017) focused on detecting and mitigating bias in embedding spaces (these are so-called

[3]The majority of this work treats both gender and race as binary variables (male/female, European American/African American), and does not necessarily problematize or explicitly theorize concepts like "gender" (Devinney, Björklund, and Björklund, 2022)

[4]While these person-words could be viewed as proxies for the variable "gender", there is no convenient one-to-one correspondence between these words and the genders of people being referred to by them. For example, the group of people referred to as "women" likely largely overlap with the group who identify with the gender *woman*, but these groups are most likely not identical. The word "person" can refer to people of any gender. Ultimately, the main focus of the current work is how people are talked about by others rather than how they would talk about themselves. Accordingly, our work relates to the use of these words, not the identities of people referred to by them.

Group	%
lesbian/gay/homosexual	3
bisexual	4
pan-/omnisexual	1
asexual	1
heterosexual	80
trans, nb or other non-cis	1

Table 2.2: Percent of survey respondents who identified with each label, as reported in Ipsos LGBT Pride 2021 Global Survey Report (Ipsos, 2021). The percentages for sexual orientations do not add up to 100 since a portion of the participants were unwilling or unable to define their sexual orientation. Note: The numbers are self-reported and some categories may be under-reported in parts of the world where openly identifying as queer can be dangerous.

intrinsic measures of bias, applied to the internal representations of a model). Many influential works (e.g. Caliskan, Bryson, and Narayanan, 2017; Kurita et al., 2019; May et al., 2019) use metrics based on the Implicit Association Test (IAT) (Greenwald, McGhee, and Schwartz, 1998) from psychology, adapted to be applicable to language models. However, later research has shown that bias detected in embedding spaces does not necessarily correlate with biased behavior in downstream tasks (Goldfarb-Tarrant et al., 2021), and recommend researchers to focus on extrinsic measures of bias to ensure fairness in downstream tasks. Similarly, (Gonen and Goldberg, 2019) find that popular debiasing techniques, rather than removing the risk of harm, may actually simply cover up existing biases and make them harder to detect, and Wang et al. (2019) show that the issue of gender bias in visual recognition tasks runs deeper than simply imbalanced data – it is part of a greater cultural context and therefore hard to isolate and extricate, both from models themselves and from the training data. It seems that rather than viewing social biases in models as isolated, harmful behaviors, we need to look at the broader picture: In what social context was the model's training data produced, and what potentially harmful attitudes could be present in that context (and thus reproduced in the model)? Felkner et al. (2023) follow a similar intuition when they show that harmful stereotypes displayed by a model against a certain community can be somewhat mitigated by finetuning on text written by members of that community. Approaching the phenomenon from a somewhat different angle, Garg et al. (2018) use bias in models as a way to investigate changes in gender and ethnic stereotypes in the US over time.

Lately, a growing body of research has been directed towards quantifying social biases in ways that are generalizable across models (Nangia et al., 2020; Nadeem, Bethke, and Reddy, 2021; Czarnowska, Vyas, and Shah, 2021; Felkner et al., 2023). These, as well as their advantages and problems, will be discussed further in §2.4.3. Benchmark measures of this kind is that they tend to be intrinsic rather than extrinsic, in order to be directly applicable to a wide range of models without need for finetuning. For similar reasons,

the current work is also focused on intrinsic measures.

Most of the work mentioned this far has treated "bias" in language models as essentially equivalent to "stereotypes". We recognize that this is one way in which social inequalities may manifest in language, and consequently in language models, and that it is a potential source of harm to the people affected. However, it should be noted that social hierarchies are present in all parts of a society, and are likely to affect the way we speak and write in a multitude of ways. Equating "social bias in language" with "stereotypes", therefore, risks severely limiting our understanding of social biases and their effects in a wider perspective. This work attempts to remedy that by drawing attention to another way in which social inequality may manifest in language.

2.4.2 Reporting Bias in Language Models

The effect of human reporting bias on language models trained on text was initially brought up as a potential issue by Gordon and Van Durme (2013), who discussed it in relation to knowledge extraction. Since then, much work on reporting bias and language models have focused specifically on multimodal settings where models have access to both textual and visual information. For example, Misra et al. (2016) propose an algorithm for decoupling human reporting bias in image annotations from the actual content of the images ("what's in the image" versus "what's worth saying") and show significant improvement in both image captioning and image classification tasks. Paik et al. (2021) show that human reporting bias negatively affects a model's predictions about the colors of common objects, but that this effect is somewhat mitigated by multimodal training. Hagström and Johansson (2022), on the other hand, find no significant differences in visual commonsense knowledge between models trained on images and text and models trained on text only. Shwartz and Choi (2020), in one of the few works on reporting bias in text-only models, show that pretrained language models can overcome reporting bias to some extent, in the sense that they seem to know trivial facts that were not explicitly stated. However, they also over-represent rare and sensational events, even amplifying the bias in their training data. This thesis aims to extend the existing work on reporting bias in language models by exploring the connections between human reporting bias and social biases, and how these manifest in models trained on human lanugage data.

2.4.3 Quantification and Benchmarking

Benchmarking is a way to compare the performance of different models by evaluating them on a benchmark dataset with a standardized metric. Examples of such benchmarks are GLUE (Wang et al., 2018) for evaluating natural language understanding, and BLiMP (Warstadt et al., 2020) for evaluating knowledge of grammatical acceptability. When it comes to bias, benchmarks like CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem, Bethke, and Reddy, 2021) were created to measure (stereotypical) biases against a range of protected demographic groups. Similar to benchmarks such as BLiMP (Warstadt et al., 2020), they measure model performance by comparing the probabilities assigned by a model to two minimally different sequences, where one is

the preferred option (grammatically acceptable in BLiMP, less stereotypical/counter-stereotypical in CrowS-Pairs and StereoSet) and the other is the dispreferred option (grammatically unacceptable, more stereotypical). A model's bias score in metrics like CrowS-Pairs and StereoSet is then based on the proportion of examples for which the dispreferred (more stereotypical) sequence was assigned a higher probability than the preferred sequence. However, a study by Blodgett et al. (2021) found that these crowd-sourced benchmark datasets contain severe issues of noise and unreliability. For example, they lack a clear conceptualization of what constitutes a harmful stereotype, and crowd-workers seem to have interpreted their instructions in a range of different ways that often do not correspond to the types of bias that the authors wish to test for. The creators of the CrowS-Pairs benchmark now advise against using it for model evaluation for these reasons.[5] In their benchmark for measuring harmful anti-queer sentiment in models, Felkner et al. (2023) use a similar minimal-pairs-based approach. They avoid the issues of noise and unreliability by handcrafting sentence templates based on responses to a community survey. This way, they ensure that the stereotypes tested for actually reflect those experienced by members of the community in question. However, the template-based approach to benchmark creation is limited in that it is hard to come up with hand-crafted templates that fully reflect the variation in naturally occurring language.

In the current work, we create a dataset based on sequences from a corpus of naturally occurring language, modified to contain the group attributes of interest. This is similar to the template-based method in that it ensures that the sequences compared are indeed minimally different and specifically target the groups of interest. On the other hand, by using naturally occurring sequences, we avoid some of the problems of unnaturalness associated with template-based methods.

[5] See `https://github.com/nyu-mll/crows-pairs/`.

3 Method

In this thesis, we investigate how language about certain attributes relates to the frequency and markedness of those attributes in the real world, and how that relationship is reflected in language models. Therefore, the different parts of the project all relate to and build on each other. After collecting information on how frequent the attributes of interest are in the real world in Chapter 2, we conduct a corpus study to compare this information to how frequently the attributes are mentioned in human text. Finally, we evaluate a number of popular LLMs on a newly created dataset to see how likely they are to reproduce the patterns found in the corpora.

3.1 Categorization and Expressions

3.1.1 Person-words

In order to study expressions for marginalized attributes in context, we insert them as modifiers to a head noun in phrases like "a queer man", "a person with a disability". Three nouns are used for this puspose: `{person, woman, man}`.[1] In both the corpus study and the model evaluation, each person-word is paired with each expression. This makes it possible to study not only the use of the expressions themselves, but also any effects of the person word on the overall result.

3.1.2 Categories and Subgroups

We choose to focus on three broad categories of marginalized attributes: *Disability*, *Race/Ethnicity* and *Queerness*. Since previous work on bias (insofar as it exists) has mainly focused on a single category at a time (see §2.4.1), we model our categorization and choice of linguistic expressions after different works for each category. The choices and motivations for each category, as well as problems and limitations, are presented in more detail below. The full lists of phrases used can be found in Appendix A.

[1] Future work could expand on this by including plurals ("people", "women"), and/or more words referring to people (such as "dude", or even "motorcyclist").

Disability

The phrase list used for *Disability* is taken from Hutchinson et al. (2020), to our knowledge the only existing research focusing on ableist bias in language models. They use a set of 56 phrases that they compiled based on guidelines by three US-based organizations: the Anti-Defamation League, ACM SIGACCESS and the ADA National Network. For each broader subcategory of disability they list multiple expressions, coarsely labelled as either *Recommended* and *Non-Recommended*. For the current study, a subset of 13 expressions was used, corresponding to the first entry labelled *Recommended* for each disability subcategory. After inspection of the results of the model evaluation, it turned out that the expression for Down's syndrome[2] contained a non-standard apostrophe that affected the model perplexities disproportionately. For this reason, the results pertaining to this expression were excluded from the analysis. The resulting set of expressions, just like the original phrase list, contains both prases describing very specific disabilities ("blind", "with cerebral palsy") and broader hypernyms ("with a disability", "without a disability"). This is not necessarily a problem, but should be kept in mind when analyzing the result. There is also a shortage of more specific expressions for unmarked attibutes, such as "neurotypical". In the current list, the only expression for an unmarked attribute is "without a disability".

Race/Ethnicity

Following Czarnowska, Vyas, and Shah (2021), the categorization of *Race/Ethnicity* was based on the Racial and Ethnic Categories and Definitions for NIH Diversity Programs (National Institutes of Health, 2015). This choice of categorization simplifies the comparison of corpus frequency and model results to real-world frequencies, since the same categories are used by the US Census Bureau. However, it should be noted that this categorization, as well as the census data, are specific to a US context. It is highly likely that different terms are used, or that these terms are used in different ways, in other parts of the English speaking world. While it can be (and often is) assumed that LLMs operate in a North American social context[3], English as a language is not specific to the US, and these models can be used by English speakers all over the world. Therefore, it should be kept in mind that the results for this category is specific to a US context and may not be generalizable beyond that context.

Queerness

The expressions for the category *Queerness* are taken from Felkner et al. (2023). Out of the phrase lists used in this thesis, theirs is the only one that contains expressions for more than one unmarked attribute. It is also the only categorization that was created as

[2]Taken from the list available at `github.com/amazon-science/generalized-fairness-metrics`.

[3]For example, many works on racial bias in language models (e.g. Caliskan, Bryson, and Narayanan, 2017; Kiritchenko and Mohammad, 2018; Kurita et al., 2019; May et al., 2019) specifically focus on bias against African Americans, which is neither the only kind of racial bias existing in the US, nor applicable to other parts of the world.

part of a community-in-the-loop effort, and as such it is more likely to reflect the language that is actually being used by community members and outsiders alike. Following Felkner et al. (2023), the expression "gay" was only combined with the person-word "man". Since "nonbinary" was only combined with "person", and "lesbian" only with "woman", this let us end up with an equal number of examples in the dataset for each person-word. For our experiment, we also added the words "allosexual" and "trans" to the original list from Felkner et al. (2023).[4]

3.2 Corpus Study

The corpus study was conducted via `english-corpora.org` using their three largest available corpora: News on the Web (NOW, 18.8 billion+ words, from 20 countries), iWeb: The Intelligent Web-based Corpus (14 billion words, from 6 countries) and Global Web-Based English (GloWbE, 1.9 billion words, from 20 countries). The reason for selecting the largest corpora is to increase the chance of finding occurrences of some of the more unusual expressions. While the NOW corpus is ideal from this point of view, it is limited to online newstext specifically. Therefore, the other corpora were included to cover more genres. All three corpora are web-scraped and likely to bear some resemblance to the data the models have seen during training. However, `english-corpora.org` (formerly the "BYU Corpora") do not provide a lot of documentation for their corpora, and it is not clear exactly how and from where the data were collected. Since we have not been able to find a clear description of the genres and sources present in these corpora, we would not recommend drawing any genre-related conclusions based on the results of this corpus study. Instead, the corpora are treated here as general samples of human language production, similar to that which the models are built and trained to model.

For this study, the expressions for each of the categories (*Disability*, *Race* and *Queerness*) were paired with each person word in the set `{person, woman, man}`[5] to create search phrases like "nonbinary person" and "man with a disability". The *n*-gram frequencies of these, as well as the unigram frequencies of the person words, were collected from each of the three corpora. As far as we can tell, `english-corpora.org` do not report the exact sizes of their corpora. It is thus not possible to report the *n*-gram frequencies as percentages of the total tokens. Instead, the frequencies are reported relative to each person word,[6] to facilitate comparison between corpora.

[4] "Allosexual" was included as the unmarked alternative to "asexual". "Trans" was included in addition to the already present "transgender" in order to match the unmarked expressions "cis" and "cisgender" in the original list.

[5] Exeptions were made for "nonbinary", "lesbian" and "gay", which were paired only with "person", "woman" and "man" respectively.

[6] These could be viewed as conditional probabilities, P(`expression` | `person word`) for every person-word ∈ `{person, woman, man}`

3.3 Model Evaluation

3.3.1 Dataset

This thesis introduces the MARB dataset, created specifically to test for reporting bias with regard to marginalized attributes. Unlike many existing benchmark datasets, MARB does not rely on artificially constructed templates (e.g. Warstadt et al., 2020; Felkner et al., 2023) or crowdworkers (e.g. Nadeem, Bethke, and Reddy, 2021; Nangia et al., 2020) to create contrasting examples. Instead, the templates used in MARB are based on naturally occurring written language from the 2021 version of the enTenTen corpus[7] (Jakubíček et al., 2013). For each person-word ∈ `{person, woman, man}`, a random sample of 10K sequences was retrieved[8] containing noun phrases of the form "`a <person-word>`", resulting in a total of 30K template sequences.

Version	Sentence
Original	I was talking to **a woman**
Unspecified	I was talking to **a woman with a disability**
Sight	I was talking to **a blind woman**
Mental health	I was talking to **a woman with a mental illness**
Without	I was talking to **a woman without a disability**

Table 3.1: Example sentences from the dataset, for the category *Disability*

Then, for each attribute in each category, the linguistic expression for that attribute was inserted as a modifier[9] to the person-word to create tuples of contrasting sequences (see Table 3.1 for some examples from the *Disability* category.). Thus, with three categories (*Disability*, *Race* and *Queerness*) containing 13, 6 and 15 expressions respectively, the total dataset amounts to over 1M sentences. The full dataset, as well as the code used to create it, can be found on GitHub.[10]

3.3.2 Models and Metric

This study concerns the off-the-shelf behavior of large pretrained language models. Since the models were not finetuned for any downstream task prior to testing, the metrics used have to be compatible with the model's pretraining task. Both masked language models (MLMs) and autoregressive generative language models are evaluated in this work.

[7] `https://www.sketchengine.eu/ententen-english-corpus/`

[8] The matching sequences were obtained using the *concordance* tool (`https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/`). The retrieved matches then had to be processed to, among other things, remove context outside sentence boundaries. Details and code are available on GitHub.

[9] As either an adjective modifier or a prepositional phrase after the head noun, depending on the type of expression.

[10] `https://github.com/TomBladsjo/MARB`

The difference in architecture between these two types of models means that they require different evaluation methods. In an MLM, the probability of a token w_t in a sequence W is conditioned on all past and future tokens: $(w_1, .., w_{t-1}, w_{t+1}, ..., w_{|W|})$. In autoregressive models, on the other hand, the probability of w_t is conditioned only on the tokens preceding it: $(w_1, ..., w_{t-1})$. This property of autoregressive models makes it possible to estimate the log-probability of a sentence W via the chain rule ($\log P_{\text{LM}}(W) = \sum_{t=1}^{|W|} \log P_{\text{LM}}(w_t|W_{<t})$). Apart from evaluating how well a model predicts a sequence, this measure can also be used to evaluate the linguistic acceptability of a sentence (Lau, Clark, and Lappin, 2017). Nowadays perplexity (PPL), derived from the log-likelihood of a sequence, is one of the most common metrics to evaluate how well an autoregressive model predicts a corpus of text. PPL(W) is defined as the exponentiated average negative log-likelihood of a sequence W, or

$$\text{PPL}(W) = \exp\left(-\frac{1}{|W|}\sum_{t=1}^{|W|}\log P_{\text{LM}}(w_t \mid W_{<t}; \Theta)\right)$$

where Θ denotes the model's parameters. Because of their use of bidirectional context, there is no corresponding way to apply the chain rule to probabilities output by an MLM, and thus PPL is not defined for this type of model. However, Salazar et al. (2020) suggest *pseudo-log-likelihood* (PLL) and the corresponding *pseudo-perplexity* (PPPL) as metrics to evaluate MLMs directly out-of-the-box without the need to finetune on downstream tasks. By successively masking one token at a time they obtain probabilities for each individual conditioned on both left and right context. They define PLL(W) as

$$\text{PLL}(W) = \sum_{t=1}^{|W|}\log P_{\text{MLM}}(w_t \mid W_{\setminus t}; \Theta)$$

A model's PPPL on a corpus $\mathbb{W}$ is then defined as

$$\text{PPPL}(\mathbb{W}) = \exp\left(-\frac{1}{N}\sum_{W\in\mathbb{W}}\text{PLL}(W)\right)$$

where N is the number of words in the corpus. They show that PLLs and PPPLs work well as a way to evaluate an MLM's linguistic competence and how well it models a corpus. In fact, since they lack the left-to-right bias and sensitivity to sequence length present in conventional log-likelihood and PPL, they are often even better at capturing the linguistic acceptability of a sequence (Salazar et al., 2020).

In the present work, autoregressive models are scored using PPL and masked models are scored using PPPL.[11] The model architectures evaluated are BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), GPT-2 (Radford et al., 2019), OPT (Zhang et al., 2022) and BLOOM (BigScience et al., 2023). Of these,

[11] In current research, both 2 and e are commonly used as the base when calculating perplexity. In this work, both PPL and PPPL are calculated using base 2.

BERT, RoBERTa and ALBERT are masked models and GPT-2, OPT and BLOOM are autoregressive models. With the inclusion of both `base` and `large` versions of BERT, RoBERTa and ALBERT, as well as both base and `medium` versions of GPT-2, a total of 10 models were evaluated.

3.3.3 Experimental Setup

The code for evaluating the masked models was partially borrowed from Felkner et al. (2023)[12], who used PLL for their MLM evaluation, but heavily modified to suit our use case. For example, they only compare sequences pairwise, while our dataset consists of n-tuples of sentences where n corresponds to the number of expressions in each category. It also had to be modified for efficiency.[13] All models were evaluated on the full dataset and perplexities/pseudo-perplexities were calculated for each sequence (including the original, unmodified sequences). That way, comparisons can be made between results for each version of a sequence. The code, as well as usage instructions, are available on GitHub. To obtain a single result for each attribute, scores for the original sentences were subtracted from the scores for each subcategory, and the Wilcoxon signed-rank test (Wilcoxon, 1945) was performed on each set of pairwise differences, and the rank-biserial correlation r (Cureton, 1956) was used to calculate the effect size. This way, the effect of the expression of interest on the model perplexity is always interpreted in relation to the perplexity of the original sequence. The effect size ranges between 1 and -1, where $r \leq 0$ indicates that the model was more surprised to see the modified sequences than the originals, $r \geq 0$ means that the model was more surprised to see the original sequences than the modified ones, and $r = 0$ means that there was no difference between the perplexities for the modified sequences and the originals. Since the test statistic in the Wilcoxon signed-rank test is based on the signs and internal ranking of the differences rather than their actual values, the resulting effect size (unlike the raw perplexities) is comparable across models with different vocabularies.

Initial inspections of the results showed a pronounced effect of sequence length on the model perplexities. Therefore, the minimal sequence length was set at four words, and 1122 examples where the original sequence was shorter than that were excluded from further analysis (see §4.2.1).

3.3.4 Computation Time

Even after modifying the evaluation code to make it more efficient, evaluating the masked models was a slow process. The time required to evaluate each model, as measured by the tqdm progress bar[14], is reported in Table 3.2. Evaluation was done on a single NVIDIA GeForce GTX 1080 Ti GPU.

[12] `https://github.com/katyfelkner/winoqueer`

[13] The original code used a batch size of 1 for their evaluation. Since the method for obtaining a PLL score requires each sequence to be fed through the model multiple times (once for each [MASK]-position), this made evaluation extremely slow. After modification, the speed increased manifold.

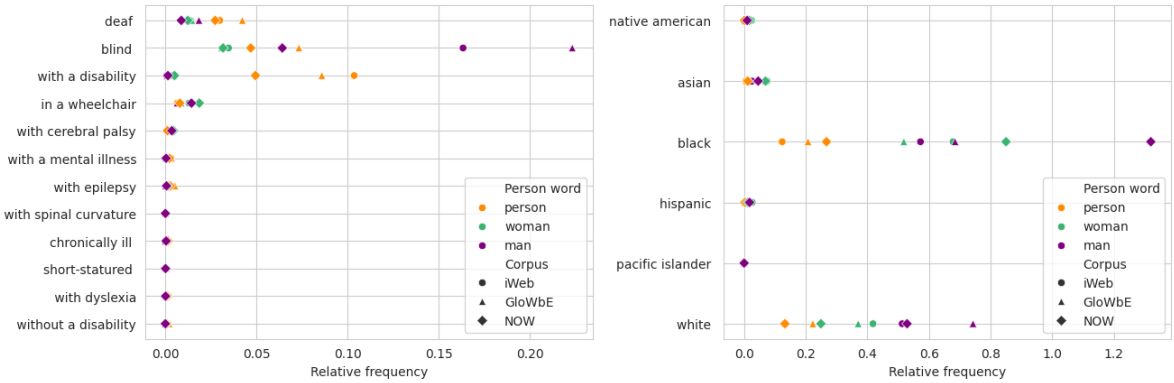
[14] `https://tqdm.github.io/`

Model	GPU hours
BERT	16.38
-large	34.41
ALBERT	12.57
-large	35.09
RoBERTa	19.43
-large	37.26
GPT-2	4.38
-medium	8.23
OPT	7.30
BLOOM	8.49
Total	185.54

Table 3.2: GPU-hours used for testing models.

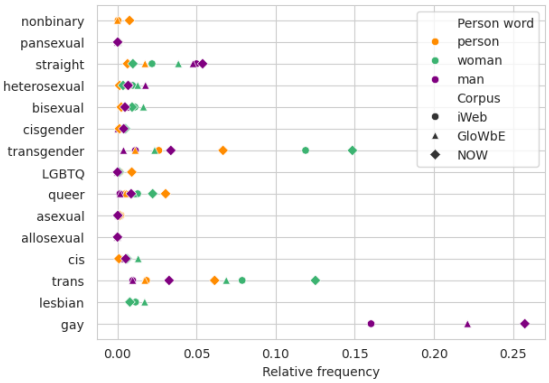
4 Results

4.1 Corpus Study



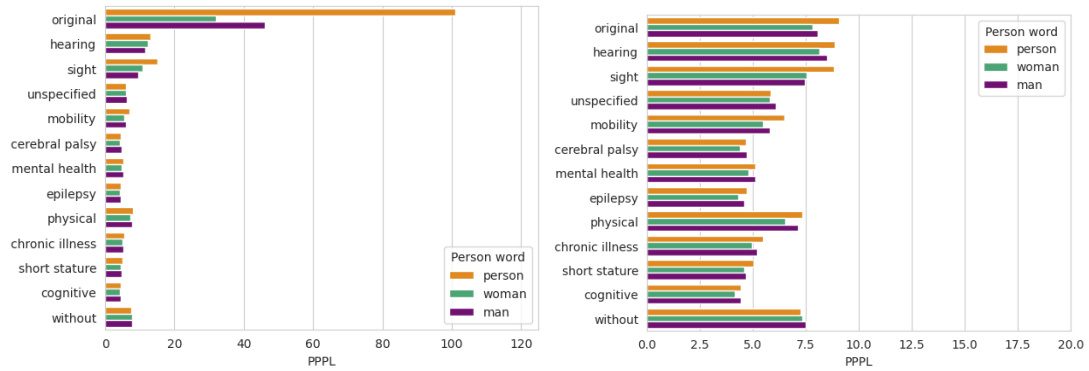
(a)

(b)



(c)

Figure 4.1: Relative n-gram frequencies for categories *Disability* (a), *Race* (b) and *Queerness* (c) in the three corpora. In order to get a sense of the spread of results, each corpus is represented as a point in the plot.



(a) (b)

Figure 4.2: Barplots showing the mean pseudo-perplexity produced by BERT for the *Disability* category before (a) and after (b) removing examples shorter than 4 words.

Figure 4.1 shows the relative frequencies of each expression for each person-word. Tables of the absolute frequencies of the person words, as well as the absolute and relative frequencies of the expressions in each corpus, can be found in Appendix B. When looking at these results, it should be noted that while most of the expressions investigated are used as modifiers in noun phrases, some occur more frequently as nouns themselves. *Lesbian* and *Pacific Islander* are two notable examples. For consistency, and in order to obtain relative frequencies, they too were treated as modifiers (i.e. "lesbian woman", "Pacific Islander person"), but the resulting frequencies may be misleading. See 5.1 for further discussion. Apart from that, a quick comparison between n-gram frequencies and the real-world occurrance shows that the attributes that are most common in the real world are usually not the most commonly mentioned in the corpora. This is particularly pronounced in the *Queerness* and *Disability* categories. On the other hand, it does not seem to be as simple as a straightforward negative association between real-world and n-gram frequencies. Instead, the observed frequencies seem to reflect the current discourse on these groups to some extent: the race discourse focuses on Black versus White, and queerness is reduced to trans women and gay men. Another thing to note is that the *Race* category is much more commonly mentioned than either *Disability* or *Queerness*.

4.2 Model Evaluation

4.2.1 Initial Results and Corrections for Sequence Length

The initial results of the model evaluation showed some unexpected features. For example, as can be seen in Figure 4.2a, the initial scores for the original sequences were very high compared to those for the modified sequences. This goes against the intuition that naturally occurring language, which is what these models were trained to predict, should be more likely than the artificially modified sequences. Upon closer inspection, it turned

out that the dataset contains a large number of very short sequences (see Figure 4.3).

When removing the shortest sequences, the variability of the results was reduced drastically (see 4.2b). Figure 4.4 shows the effect on variability (as standard deviation) of excluding shorter sequences from the results.[1] As can be seen in the plot, the effect is stronger on MLMs than on autoregressive models, but both model types are noticeably affected (note that the y-axis of the plot is on a logarithmic scale). While the variability seems to keep decreasing with the larger minimum sequence lengths, a considerable difference can already be observed at a minimum length of four tokens. Therefore, four was used as the cutoff point and 1122 examples where the original sentences were shorter than four words were excluded in all further analyses.

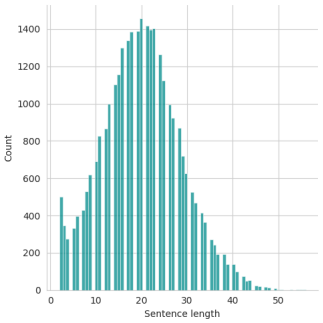


Figure 4.3: Lengths of unmodified sequences in the dataset.

4.2.2 Further Analysis

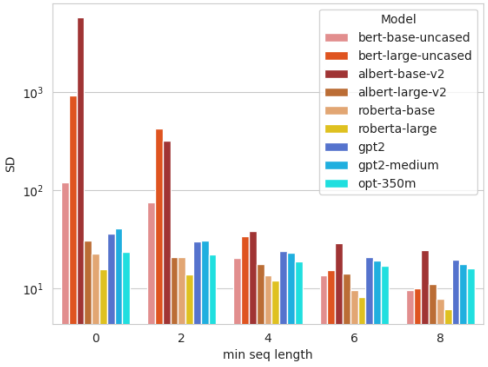
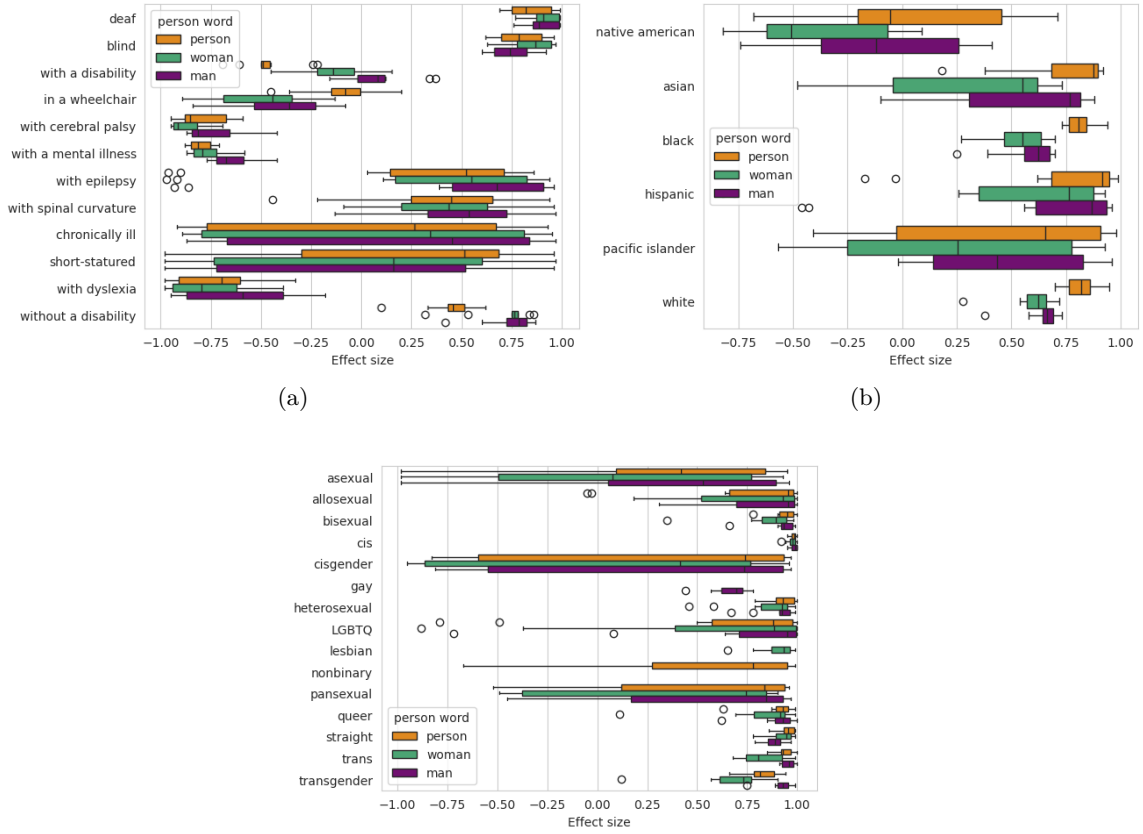


Figure 4.4: Barplot showing the relationship between the minimum sequence length and the variability (as standard deviation) of the PPLs/PPPLs for masked (warm colors) and autoregressive (cold colors) models.

As explained in §3.3.3, the Wilcoxon signed ranks test was performed on the pairwise differences between each set of modified sequences and their unmodified counterparts. A larger effect size means that the model in question was more surprised to see the sequences including this expression (compared to the original, unmodified sequences). Figure 4.5 shows the aggregate results, to visualize the general behavior of the language models. The full disaggregated results can be found in Appendix C. We can see that for some expressions, there is a large variation of results between models, while for some all models seem to agree (compare for example *deaf* and *chronically ill*). Note, however, that for most expressions the models seem to be less surprised to see these expressions as modifiers to the noun

[1] BLOOM was excluded from this plot, since its variability was much larger than for all other models and made the plot unreadable. On closer inspection of the BLOOM results, it turns out that BLOOM actually displays the opposite relationship: the variability seems to increase slightly when excluding shorter sequences. It would be interesting to investigate this further in the future. For now, the corrections performed on the other model scores were performed on the BLOOM scores as well, to enable comparison across models.



(a) (b)

(c)

Figure 4.5: Aggregate Wilcoxon test effect sizes (r) for all models ($n = 10$) on categories *Disability* (a), *Race* (b) and *Queerness* (c). The boxes show the distribution of results across models. A larger effect size means that the model was more surprised to see this set of sequences.

"woman" than to either "man" or "person".

Figure 4.6 shows the distribution of effect sizes for different expressions for each model and category. The intuition here is that higher effect sizes for a category means that the model is generally more surprised to see that category mentioned at all, and that a larger spread of effect sizes within a category means that the model performance differs more depending on which expression is included in the sequence. As we can see, the distribution of effect sizes differs a lot between categories. For example, the *Disability* category shows a greater spread for all models, and the medians (represented by the middle line in each box) are generally lower than for the other categories. The BERT models show especially low values for this category. This indicates that the models are less surprised to see disability mentioned than, for example, queerness, and that, within the *Disability* category, there is a large difference between the most and least

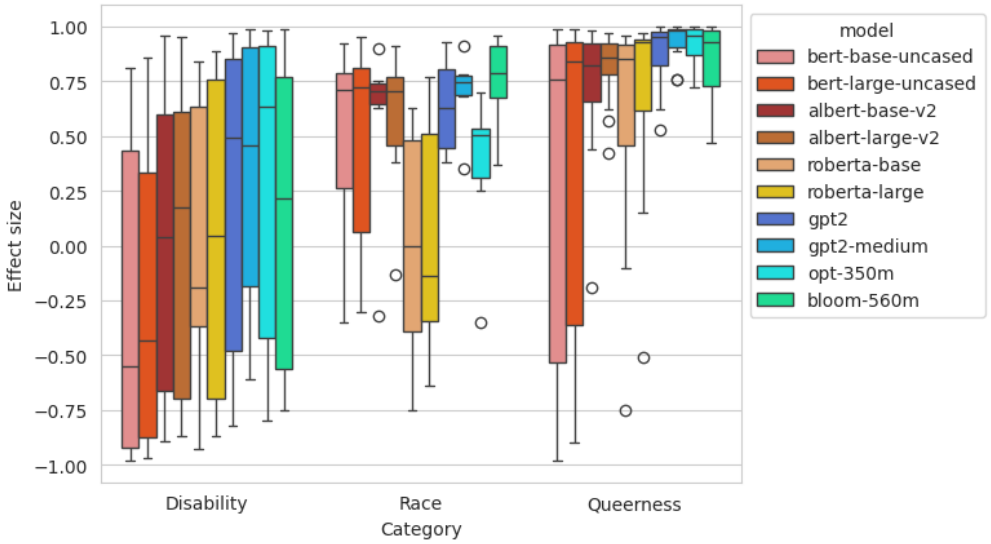


Figure 4.6: The distribution of results for all expressions within each category, for masked (warm colors) and autoregressive (cold colors) models. A larger spread means a larger difference in performance depending on expression. A higher average means that the model was generally more surprised to see this category mentioned.

unexpected attributes to mention. In the *Disability* and *Queerness* categories, we can also see a difference between autoregressive and masked models, where autoregressive models tend to be more surprised than masked models to see these attributes mentioned. In the *Race* category we notice less uniform results across the models, both regarding spread and central tendency, with the RoBERTa models showing especially low values.

Markedness

Based on the pragmatic principles discussed in §2.1 and 2.2, we would expect to see a difference in results based on the social markedness of the attribute mentioned. As mentioned in §3.1.2, the *Disability* and *Race* categories each only contain one expression for an unmarked attribute ("without a disability" and "white" respectively), represented in Figure 4.7a and 4.7b as a red dot. The *Queerness* category, on the other hand, contains multiple expressions for unmarked attributes ("allosexual", "cis", "cisgender", "heterosexual", and "straight"), shown as red boxes in Figure 4.7c. Based on these plots, the expectation seems to hold true for all models in categories *Disability* and *Queerness*, but not as clearly in the case of *Race*. This will be discussed further in §5.2.

Comparison to Corpus Frequencies

Assuming that the corpora included resemble the models' training data to some extent, we would expect to see a negative relationship between corpus frequency and model

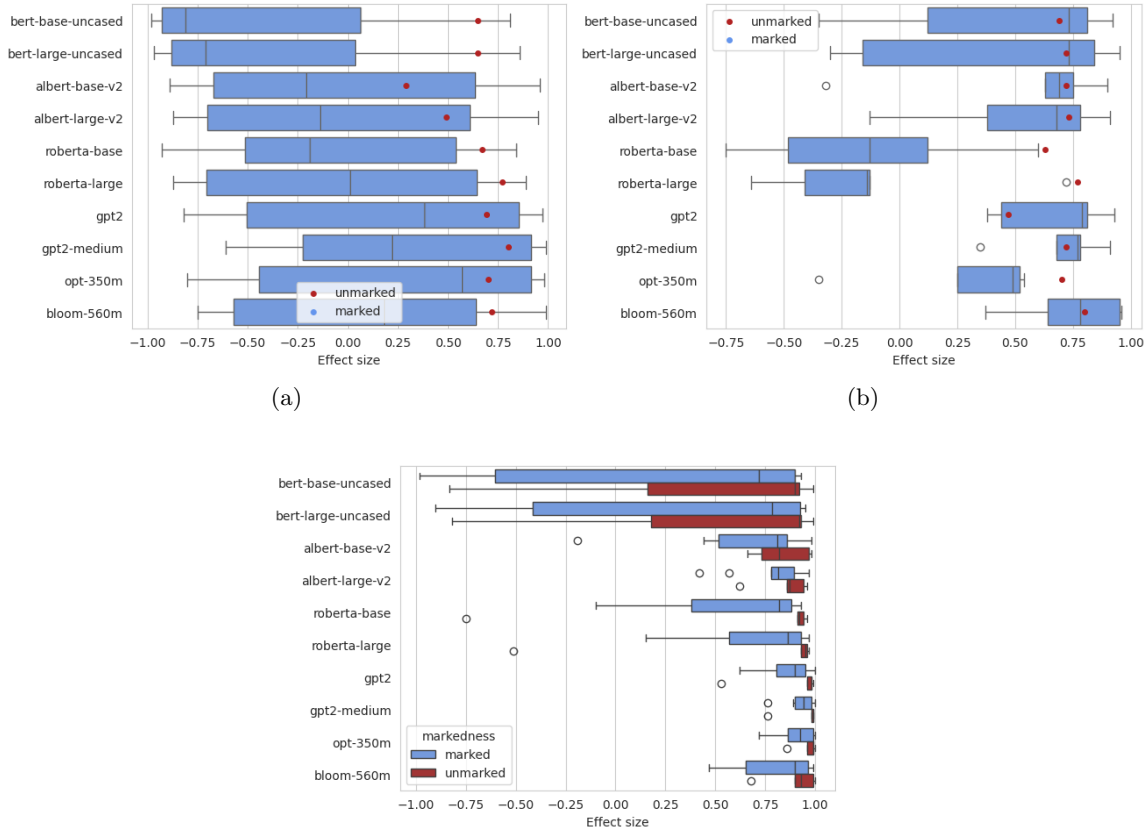


Figure 4.7: Wilcoxon test effect sizes (r) per model for *marked* (blue) and *unmarked* (red) expressions in categories *Disability* (a), *Race* (b) and *Queerness* (c).

perplexity – the more common an expression is in the training data, the less surprised the model should be to see it. However, if we compare the relative corpus frequencies reported in §4.1, it is hard to see such a pattern. In Figure 4.8, the results from the model evaluation have been plotted together with the corpus frequencies for comparison. Note that the lower x-axis has been reversed. If the model predictions reflected the frequencies observed in the corpora, higher corpus frequencies (upper x-axis) would correspond to lower effect sizes (lower x-axis). These results, and possible reasons, are discussed in Chapter 5.

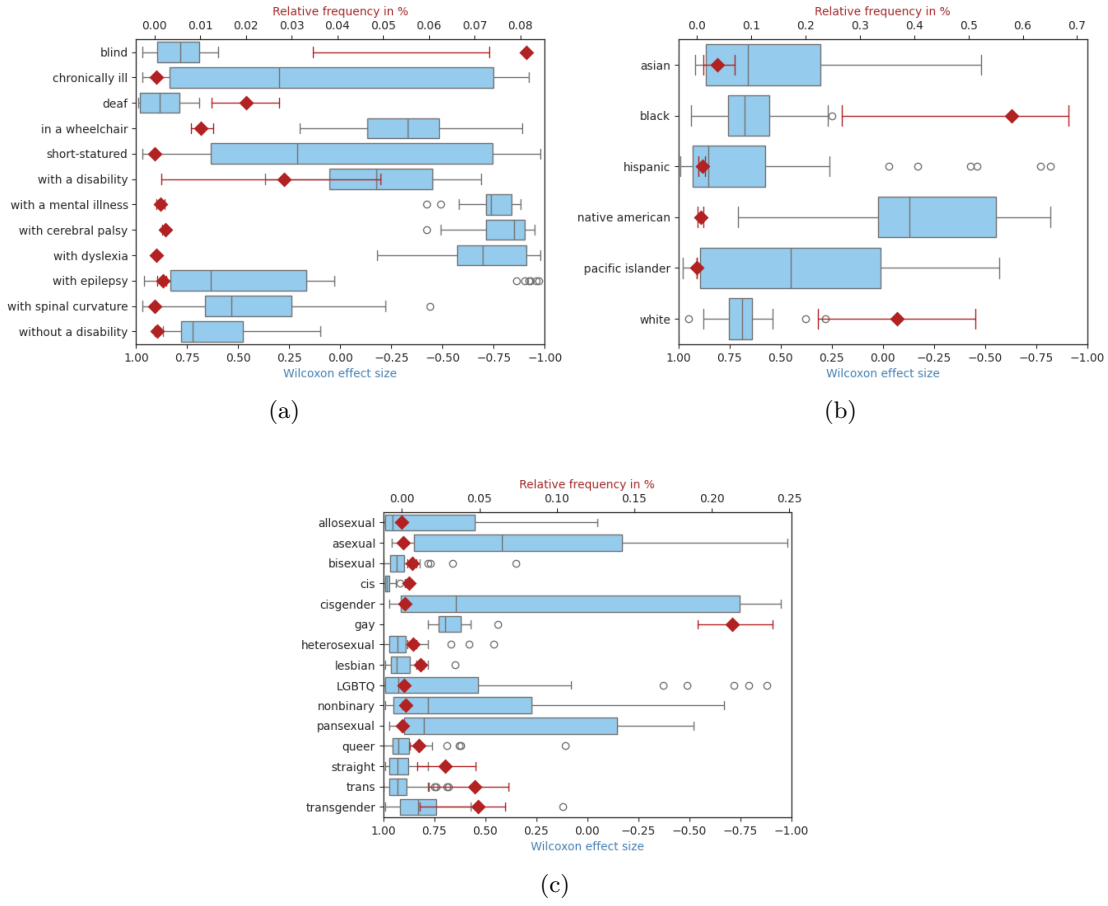


Figure 4.8: Comparison between model effect sizes (blue, lower x-axis) and relative corpus frequencies (red, upper x-axis). Note that the lower x-axis has been reversed to facilitate comparison. Errorbars on the corpus frequencies show the interquartile range.

5 Discussion

5.1 Corpus Frequencies and the Real World

As mentioned in §4.1, some of the expressions that were investigated are more commonly used as nouns than as modifiers. For example, the phrase "lesbian woman" occurred 530 times in the NOW corpus, 188 times in the iWeb corpus and 63 times in the GloWbE corpus. The corresponding numbers for the unigram "lesbian" are 77566, 58619 and 12962. Similarly, while "Pacific Islander person" was mentioned 3 times in the NOW corpus and never in the other two corpora, "Pacific Islander" was mentioned 8016, 3540 and 283 times in NOW, iWeb and GloWbE respectively. Since the examples in the MARB dataset were made in the same way, this may also have had an effect on the model evaluation results. Another thing to be noted is that the relationship between n-gram frequencies and real-world frequencies is not as simple as a clear positive or negative association. Instead, we can see clear traces of the current discourse in some of these frequencies. For example, race seems to get simplified as either "Black" or "White", and queerness shrinks down to mainly being about gay men (and gay women, if we include the much larger unigram frequency of "lesbian") and trans women, two groups that have been much discussed in US media lately. In the case of *Race*, we can see that "Black" is mentioned more often than "White", even though "White alone" is, at 62%, by far the most common category in the 2020 US census (in fact, it is five times as common as the category "Black or African American alone"). On the other hand, it is mentioned much more often than less common categories such as "Native American" or "Pacific Islander". In fact, all expressions for races/ethnicities apart from "White" and "Black" have very low relative frequencies – even words for groups like "Asian" (6% in the census, half as common as "Black or African American") and "Hispanic or Latino" (which, at 18.7%, is in fact 50% more common than "Black or African American").[1] It would seem, from these numbers, that race/ethinicity, which in reality is way more complex and nuanced than the categorization used by the US Census Bureau (Brown, 2020), is further simplified in the discourse and narrowed down to only two main categories.[2] In the *Disability* and *Queerness* catgories, on the other hand, we see no such clear-cut

[1] It should be kept in mind, though, that the census makes a difference between *Race* and *Ethnicity*, and "Hispanic or Latino", as an ethnicity, can co-occur with any race.

[2] It is possible that *White* as a descriptor of people is mainly used in contrast to *Black* in contexts where the distinction is relevant, whereas *Black* is also used on its own (in contrast with the unmentioned norm, one could say). It would be interesting to research this topic further in the future.

binary distinctions (apart from, possibly, a small effect for straight/gay in Figure 4.1c). Instead, it seems like the entire marked group (people with disabilities, queer people) is boiled down to a few particularly stereotypical subgroups: blind people and deaf people; gay men and trans women. It is interesting that "blind man" sticks out as especially common. This could be due to idiomatic expressions and jokes such as "I see, said the blind man".[3]

5.2 Model Evaluation

5.2.1 Effects of Sequence- and Expression Length

Figure 4.4 shows the effect of removing the shortest sequences from the dataset on the spread of the model perplexities. We can see that the effect is very large for the very shortest sequences; the y-axis is log-scaled to make the plot readable. However, this effect is not equally pronounced for all models. In particular, we can see a more pronounced effect for masked language models than for autoregressive models.[4] This is not unexpected; metrics like PPL and PPPL are both sensitive to sequence length, but for slightly different reasons. Because autoregressive models are unidirectional, their predictions at every timestep are based only on the preceding tokens. At the very first timestep, therefore, the model has no prior information on which to base its prediction, and the probability is more evenly distributed over the vocabulary. At the second timestep, it has one previous token to rely on, and so forth. This means that for an autoregressive model, the entropy is generally higher at earlier token positions, regardless of the sequence length. When averaging entropy for a sequence, therefore, the higher entropy of the earlier token positions will have a larger effect on the resulting average the shorter the sequence is. Masked language models, on the other hand, use the context on both sides to predict a token, and thus have access to the same number of unmasked tokens regardless of which position in the sequence is currently being predicted. For masked models, therefore, the effect of sequence length on pseudo-perplexity is not dependent on the proportion of early position, high entropy tokens to the total length of the sequence, but rather on the total amount of available information. When we calculate pseudo-perplexity for a sequence like "a person", a masked model will first produce probabilities at the [MASK] position for "[MASK] person", and then for "a [MASK]". At the second position, there are many possible words that could slot into the [MASK] position, and the entropy is likely to be high. Compare this to the modified versions of the same sequence that would be present in the dataset: there is likely to be a considerable difference in entropy between "a [MASK]" and "a Native American [MASK]", or "a [MASK] with a disability". For this reason, the difference in perplexity (or pseudo-perplexity) depending on sequence length is much more radical for masked

[3]The prevalence of expressions such as this suggests that blindness is a particularly stereotypical disability.

[4]Remember that BLOOM, which was not included in the plot because of its much higher variability overall, in fact showed a reversed relationship between minimum sequence length and variability.

models, especially for very short sequences. Here, the informativeness of the specific context also plays a part. Consider for example the sequences ”a white person” and ”a deaf person”. While ”white” could refer to most concrete nouns, ”deaf” is generally used about people (or sometimes animals). Thus, ”a white [MASK]” is likely to have a higher entropy than ”a deaf [MASK]” in most models.

While we did try to limit the impact of sequence length by excluding the very shortest sequences from the analysis, it should be kept in mind that the effect is still present to some degree and could lead to noise in the results. It is possible, for example, that the different expression lengths in the *Disability* category are partially responsible for the spread we see in this category in Figure 4.6.

5.2.2 Model Vocabulary and Markedness

Another source of noise which is partially tied to sequence length, is that of model vocabularies. Different models have different vocabularies, and will tokenize a sequence in different ways. The way in which a vocabulary is determined differs from model to model, but most tokenization algorithms nowadays make use of subword tokens to represent words that are not common enough to warrant their own spot in the vocabulary. This interacts with sequence length, in that treating a single word as multiple tokens will result in a longer sequence overall. A model that treats an uncommon word as multiple tokens may have access to more contextual information for predicting the individual subword tokens of that word than a model that encodes the whole word as a single token. Thus, a model may in fact assign higher probabilities to sequences containing words it has never seen before than to sequences containing seen but relatively unusual words. In Figure 5.1, we can see that the number of tokens used to encode the expressions of interest in the dataset differs a lot between models. These differences in tokenization are likely to affect the model perplexities, and are difficult to fully disentangle from the phenomenon of interest.

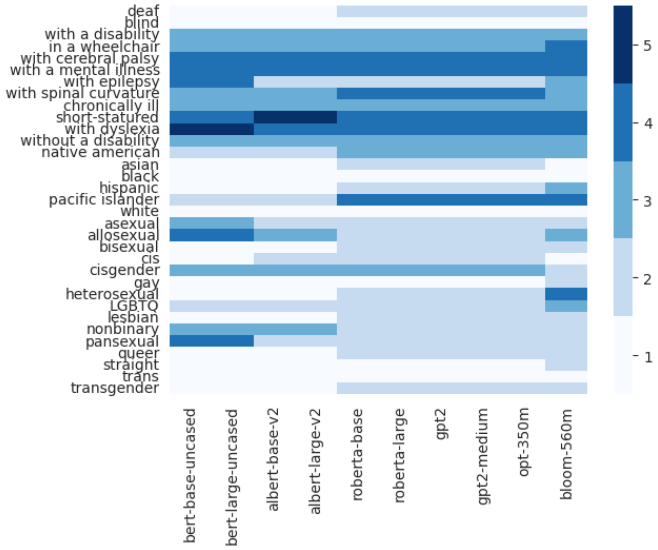


Figure 5.1: The number of tokens required by each model to encode each expression of interest.

On the other hand, a closer look at the model vocabularies can also give us some hints about word frequencies in the models' training data. We can see, for example, that the word ”blind” is common (and short) enough to be encoded as a single token for every

model tested. We can also see that BERT and ALBERT both encode "transgender" as a single token, but "cisgender" as three. This suggests that "transgender", despite its length, was seen enough during training to warrant its own position in the vocabulary, whereas "cisgender" was not.

This aspect of the tokenization process provides a likely explanation for the apparent absence of correlation that we observe in Figure 4.8 (comparison between model effect sizes and corpus frequencies). For example, the fact that the models on average seem to be more surprised to see the word "transgender" than the word "cisgender", despite the much higher frequency of "transgender" in the corpora, is likely a result of "cisgender" being encoded as multiple tokens, giving the models access to more contextual information at each token position.

Another potential source of noise in the results is case sensitivity. While the corpus frequencies collected are case-insensitive for all corpora, the model evaluation included both uncased (BERT and ALBERT) and cased (RoBERTa, GPT2, OPT and BLOOM) models. Since the dataset was created primarily with uncased models in mind and words like "Asian" and "American" were not capitalized, this is likely to have had some effect on the *Race* category especially. Future work using similar methods for dataset creation should endeavor to account for case sensitivity.

5.2.3 Disentangling the Bias

Above, we have discussed a number of factors, apart from reporting bias, that affect the results of this study. There are at least two ways to view these. If we are primarily interested in the type of reporting bias we encounter in natural language, and want to trace how this specific phenomenon manifests in a model trained on human text, then we would need to view these other intervening factors as noise in our results. If, on the other hand, our goal is to decide whether or not the model displays desirable behavior (regarding fairness, reliability, etc.) it makes more sense to view the model as a closed system, and focus on the actual behavior regardless of the system-internal factors that contribute to this behavior.[5] In practice, of course, the two perspectives are closely related: If we, using extrinsic measures, detect undesirable behavior in our model, we need to understand the underlying causes to be able to correct it (see Suresh and Guttag, 2021). Conversely, the larger and more complex models get, the harder it is to inspect what is happening "under the hood", and we often need to base our diagnosis on the model's output. As has been shown in this work, this is by no means an easy task (see also Antoniak and Mimno, 2021; Baldini et al., 2023). One contribution of this thesis is to further illuminate this difficulty, which we urge researchers to take into account when examining bias in language models.

[5] See §2.4.1 on intrinsic and extrinsic measures of bias.

6 Limitations and Ethical Considerations

Apart from those already discussed in Chapters 3 and 5, this work has a number of limitations and ethical considerations which will be accounted for below.

6.1 Environmental Impact

Testing ten large models on a dataset as large as MARB requires a lot of computation time. As reported in Table 3.2, the model evaluation required a total of approximately 186 GPU hours, corresponding to a CO_2 emission of 0.764 kg.[1] When working with large language models, one always has to weigh the gains of resource-heavy computation against the environmental impact.

6.2 Expressions as Proxies

The aim of this work is to investigate how different groups are spoken about, by other people and by language models. However, when working with written corpus data and output from test-only language models, there is no straightforward way to connect linguistic expressions to real-life demographic groups and lived experiences. For one, there may be multiple synonymous expressions for a certain attribute, or expressions that denote partially overlapping groups. Thus, using a single expression as proxy for a demographic group – for example in a corpus study – may be misleading, as it is unlikely to capture all mentions of the group. Furthermore, different expressions may have different connotations, and the expression used by an outsider to describe a certain attribute might not be the way it would be described by persons who themselves possess that attribute. In some cases there may not be an established or commonly known way to refer to an unmarked attribute (such as in Jakobson's example with the Russian words for male and female donkeys, see §2.2). In fact, mentions of people may not even refer to real-life entities – this is often the case, for example, in fiction and hypothetical scenarios. It is important, therefore, to keep in mind that any conclusions drawn about,

[1] CO_2 emission was calculated using a tool developed by Simon Hengchen, available at `https://github.com/faustusdotbe/CO2_GU_mltgpu`.

for example, the use of the expression "woman in a wheelchair" is about the use of that specific expression (regardless of who is the referent) – not about the group "women who use wheelchairs". Thus, while there is certainly a connection between the real-world frequencies and the results from the corpus study and model evaluation, it is not a direct or straightforward one.[2]

6.3 Seed Words

In research on bias in language models, the choice of word lists (seed lexicons) is known to affect the resulting measurements (Antoniak and Mimno, 2021). As mentioned in §3.1, the lists of expressions used in this work come from different sources and were compiled in different ways. Out of the lists used in our experiments, the one relating to queerness is the only one that was created in consultation with members of the community in question. While that is no guarantee that the list properly represents the experiences and identities of all people who fall under the category, it certainly has a better chance of doing so than lists that were created without community involvement. Therefore, the lists of expressions relating to disability and race/ethnicity may not reflect the real-world identities of people in these communities, or the language used about them by others. Future research on racial and ableist bias should endeavor to involve members of the relevant communities in the research process to ensure a good match between measure design and the phenomenon being measured.

6.4 Corpora

The corpora used in the corpus study were chosen because they are web-scraped and likely to be somewhat similar to the data the models have seen during training, so that we can make valid comparisons. However, since we cannot know exactly what data the models have been trained on, there are no guarantees that the chosen corpora are representative of the same type of language. Furthermore, neither the corpora nor the models themselves represent the English language as a whole. Thus, while the corpus frequencies collected in this study can give a hint about how humans use language and how that use is reflected in language models, they are ultimately representative only of the corpora from which they were collected. In order to draw more general conclusions about human language use, we would need to look at samples from a wider range of different subpopulations, collected in a more controlled and deliberate way.

6.5 Language and Geographic Scope

This work – both the corpus study and the model evaluation – is limited to English language data. Furthermore, as mentioned in §3.1.2, the categorization of race and/or ethnicity, and the expressions used for the different groups, are specific to a US context.

[2]This is especially the case for text-only settings, as opposed to, for example, image captioning tasks.

There is no guarantee that the results reported here would generalize to other languages, or that they are relevant in a social context outside of the US.

7 Conclusions and Future Work

This thesis has investigated the relationship between markedness and reporting bias, both in human-produced text and in language models. It also introduced the MARB dataset, designed to measure language model reporting bias with regard to sensitive attributes in three different categories: *Disability*, *Race* and *Queerness*. Unlike most existing datasets for quantifying bias, MARB was created from naturally occurring sequences, minimally modified by inserting the expressions of interest. For this reason, it is likely to better reflect the variation in natural language than template-based datasets.[1] We have tested ten off-the-shelf language models on the dataset, including BERT, ALBERT, RoBERTa, GPT-2, OPT and BLOOM, and analyzed the results along multiple dimensions.

We have found that n-gram frequencies in human text show strong signs of reporting bias with regard to marked identities, mirroring current discourse in society. On the other hand, this relationship does not manifest as strongly in the ten models tested on the MARB dataset. We have investigated possible reasons for this difference and found that sequence length and model vocabulary both affect the results.

There are a number of interesting directions to go from here. The MARB dataset was developed specifically to capture reporting bias with regard to socially marked attributes, but it could also be used to explore other phenomena, such as bias in sentiment analysis models. The dataset can also be used to explore metrics and methods for measuring reporting bias other than the one proposed in this thesis. When it comes to testing or probing models for a phenomenon of interest, filtering out other influencing factors remains difficult. In bias research and benchmarking in particular, further efforts are needed to develop diagnostic datasets and scoring methods that account for factors such as sequence length and choice of specific seed words, and to evaluate these methods in a rigorous way. As mentioned in §2.4, the relationship between social markedness and reporting bias has been largely neglected in research up to this point. Before this phenomenon can be fully understood in language models, there is a need for more research on how it plays out in human language production and interaction. For example, there is a lack of research on if and how the frequency with which a certain attribute is mentioned differs between different communities and social groups, and what the predictors are for whether an attribute will be mentioned or not.

Hopefully, the current work can illuminate the opportunities for research in this area

[1] It does not, however, escape the problem of seed words (see Chapter 6).

and provide a starting point for further exploration.

Bibliography

Antoniak, Maria and David Mimno (Aug. 2021). "Bad Seeds: Evaluating Lexical Methods for Bias Measurement". 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)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1889–1904. DOI: 10.18653/v1/2021.acl-long.148. URL: https://aclanthology.org/2021.acl-long.148 (visited on 02/09/2024).

Baldini, Ioana, Chhavi Yadav, Payel Das, and Kush R. Varshney (May 2023). *Keeping Up with the Language Models: Robustness-Bias Interplay in NLI Data and Models*. arXiv:2305.12620 [cs]. DOI: 10.48550/arXiv.2305.12620. URL: http://arxiv.org/abs/2305.12620 (visited on 05/08/2024).

Barrett, Rusty (Aug. 2014). "The Emergence of the Unmarked". en. In: *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*. Ed. by Lal Zimman, Jenny Davis, and Joshua Raclaw. Oxford University Press. ISBN: 978-0-19-993729-5. DOI: 10.1093/acprof:oso/9780199937295.001.0001. URL: https://academic.oup.com/book/35390 (visited on 02/09/2024).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (Mar. 2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://dl.acm.org/doi/10.1145/3442188.3445922 (visited on 04/30/2024).

BigScience, Workshop et al. (June 2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv:2211.05100 [cs]. DOI: 10.48550/arXiv.2211.05100. URL: http://arxiv.org/abs/2211.05100 (visited on 03/20/2024).

Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach (2021). "Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets". en. 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)*. Online: Association for Computa-

tional Linguistics, pp. 1004–1015. DOI: 10.18653/v1/2021.acl-long.81. URL: https://aclanthology.org/2021.acl-long.81 (visited on 02/01/2024).

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html (visited on 05/13/2024).

Brekhus, Wayne (1998). "A Sociology of the Unmarked: Redirecting Our Focus". In: *Sociological Theory* 16.1. Publisher: [American Sociological Association, Wiley, Sage Publications, Inc.], pp. 34–51. ISSN: 0735-2751. URL: https://www.jstor.org/stable/202213 (visited on 02/13/2024).

Brown, Anna (Feb. 2020). *The changing categories the U.S. census has used to measure race*. en-US. URL: https://www.pewresearch.org/short-reads/2020/02/25/the-changing-categories-the-u-s-has-used-to-measure-race/ (visited on 05/07/2024).

Bucholtz, Mary and Kira Hall (Jan. 2005). "Language and Identity". en. In: *A Companion to Linguistic Anthropology*. Ed. by Alessandro Duranti. 1st ed. Wiley, pp. 369–394. ISBN: 978-0-631-22352-8 978-0-470-99652-2. DOI: 10.1002/9780470996522.ch16. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470996522.ch16 (visited on 02/05/2024).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (Apr. 2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334. arXiv:1608.07187 [cs], pp. 183–186. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal4230. URL: http://arxiv.org/abs/1608.07187 (visited on 02/06/2024).

Chomsky, Noam and Howard Lasnik (1977). "Filters and Control". In: *Linguistic Inquiry* 8.3. Publisher: The MIT Press, pp. 425–504. ISSN: 0024-3892. URL: https://www.jstor.org/stable/4177996 (visited on 04/30/2024).

Cureton, Edward E. (Sept. 1956). "Rank-biserial correlation". en. In: *Psychometrika* 21.3, pp. 287–290. ISSN: 1860-0980. DOI: 10.1007/BF02289138. URL: https://doi.org/10.1007/BF02289138 (visited on 04/17/2024).

Czarnowska, Paula, Yogarshi Vyas, and Kashif Shah (Nov. 2021). "Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics". In: *Transactions of the Association for Computational Linguistics* 9, pp. 1249–1267. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00425. URL: https://doi.org/10.1162/tacl_a_00425 (visited on 02/06/2024).

Devinney, Hannah, Jenny Björklund, and Henrik Björklund (June 2022). "Theories of "Gender" in NLP Bias Research". In: *Proceedings of the 2022 ACM Conference on*

Fairness, Accountability, and Transparency. FAccT '22. New York, NY, USA: Association for Computing Machinery, pp. 2083–2102. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3534627. URL: https://dl.acm.org/doi/10.1145/3531146.3534627 (visited on 05/02/2024).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (visited on 03/20/2024).

Dhamala, Jwala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta (Mar. 2021). "BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 862–872. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445924. URL: https://doi.org/10.1145/3442188.3445924 (visited on 05/13/2024).

Felkner, Virginia, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May (July 2023). "WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 9126–9140. DOI: 10.18653/v1/2023.acl-long.507. URL: https://aclanthology.org/2023.acl-long.507 (visited on 02/05/2024).

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". en. In: 115.16. ISBN: 9781720347118, E3635–E3644. DOI: http://www.pnas.org/cgi/doi/10.1073/pnas.1720347115. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1720347115 (visited on 02/06/2024).

Goldfarb-Tarrant, Seraphina, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez (Aug. 2021). "Intrinsic Bias Metrics Do Not Correlate with Application Bias". 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)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1926–1940. DOI: 10.18653/v1/2021.acl-long.150. URL: https://aclanthology.org/2021.acl-long.150 (visited on 05/13/2024).

Gonen, Hila and Yoav Goldberg (June 2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 609–614. DOI: 10.18653/v1/N19-1061. URL: https://aclanthology.org/N19-1061 (visited on 05/02/2024).

Gordon, Jonathan and Benjamin Van Durme (Oct. 2013). "Reporting bias and knowledge acquisition". In: *Proceedings of the 2013 workshop on Automated knowledge base construction*. AKBC '13. New York, NY, USA: Association for Computing Machinery, pp. 25–30. ISBN: 978-1-4503-2411-3. DOI: 10.1145/2509558.2509563. URL: https://doi.org/10.1145/2509558.2509563 (visited on 03/18/2024).

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz (1998). "Measuring individual differences in implicit cognition: The implicit association test." en. In: *Journal of Personality and Social Psychology* 74.6, pp. 1464–1480. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.74.6.1464. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.74.6.1464 (visited on 02/05/2024).

Grice, Herbert Paul (1975). "Logic and Conversation". en. In: *Speech acts*. Ed. by Peter Cole. 5. print. Syntax and semantics 3. New York u.a: Academic Press, pp. 41–58. ISBN: 978-0-12-785423-6.

Hagström, Lovisa and Richard Johansson (May 2022). "What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Samuel Louvan, Andrea Madotto, and Brielen Madureira. Dublin, Ireland: Association for Computational Linguistics, pp. 252–261. DOI: 10.18653/v1/2022.acl-srw.19. URL: https://aclanthology.org/2022.acl-srw.19 (visited on 03/18/2024).

Haspelmath, Martin (Mar. 2006). "Against markedness (and what to replace it with)". en. In: *Journal of Linguistics* 42.1, pp. 25–70. ISSN: 0022-2267, 1469-7742. DOI: 10.1017/S0022226705003683. URL: https://www.cambridge.org/core/product/identifier/S0022226705003683/type/journal_article (visited on 02/05/2024).

Hodosh, Micah, Peter Young, and Julia Hockenmaier (July 2015). "Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract". en. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4188–4192.

Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl (July 2020). "Social Biases in NLP Models as Barriers for Persons with Disabilities". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter,

and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5491–5501. DOI: 10.18653/v1/2020.acl-main.487. URL: https://aclanthology.org/2020.acl-main.487 (visited on 02/05/2024).

Ipsos (June 2021). *LGBT+ Pride 2021 Global Survey*. en. Tech. rep. URL: https://www.ipsos.com/en/lgbt-pride-2021-global-survey-points-generation-gap-around-gender-identity-and-sexual-attraction (visited on 02/29/2024).

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít} Suchomel (2013). "The TenTen Corpus Family". In: *7th International Corpus Linguistics Conference CL 2013*. Lancaster, pp. 125–127. URL: https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf.

Kiritchenko, Svetlana and Saif Mohammad (June 2018). "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems". In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Ed. by Malvina Nissim, Jonathan Berant, and Alessandro Lenci. New Orleans, Louisiana: Association for Computational Linguistics, pp. 43–53. DOI: 10.18653/v1/S18-2005. URL: https://aclanthology.org/S18-2005 (visited on 02/06/2024).

Kurita, Keita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov (Aug. 2019). "Measuring Bias in Contextualized Word Representations". In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Ed. by Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster. Florence, Italy: Association for Computational Linguistics, pp. 166–172. DOI: 10.18653/v1/W19-3823. URL: https://aclanthology.org/W19-3823 (visited on 02/06/2024).

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (Feb. 2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. arXiv:1909.11942 [cs]. DOI: 10.48550/arXiv.1909.11942. URL: http://arxiv.org/abs/1909.11942 (visited on 03/20/2024).

Lau, Jey Han, Alexander Clark, and Shalom Lappin (2017). "Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge". en. In: *Cognitive Science* 41.5. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12414, pp. 1202–1241. ISSN: 1551-6709. DOI: 10.1111/cogs.12414. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12414 (visited on 02/05/2024).

Levinson, Stephen C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. eng. Language, speech, and communication. Cambridge, Mass.: MIT. ISBN: 978-0-262-62130-4.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (July 2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692 [cs]. DOI: 10.48550/arXiv.1907.11692. URL: http://arxiv.org/abs/1907.11692 (visited on 03/20/2024).

Malik, Ananya (Nov. 2023). *Evaluating Large Language Models through Gender and Racial Stereotypes*. arXiv:2311.14788 [cs]. URL: http://arxiv.org/abs/2311.14788 (visited on 02/06/2024).

May, Chandler, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger (June 2019). "On Measuring Social Biases in Sentence Encoders". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 622–628. DOI: 10.18653/v1/N19-1063. URL: https://aclanthology.org/N19-1063 (visited on 02/05/2024).

Misra, Ishan, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick (June 2016). "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919, pp. 2930–2939. DOI: 10.1109/CVPR.2016.320. URL: https://ieeexplore.ieee.org/abstract/document/7780689 (visited on 05/13/2024).

Morehouse, Kirsten N., Keith Maddox, and Mahzarin R. Banaji (May 2023). "All human social groups are human, but some are more human than others: A comprehensive investigation of the implicit association of "Human" to US racial/ethnic groups". en. In: *Proceedings of the National Academy of Sciences* 120.22, e2300995120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2300995120. URL: https://pnas.org/doi/10.1073/pnas.2300995120 (visited on 02/05/2024).

Nadeem, Moin, Anna Bethke, and Siva Reddy (Aug. 2021). "StereoSet: Measuring stereotypical bias in pretrained language models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 5356–5371. DOI: 10.18653/v1/2021.acl-long.416. URL: https://aclanthology.org/2021.acl-long.416 (visited on 02/05/2024).

Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman (Nov. 2020). "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 1953–1967. DOI: 10.18653/v1/2020.emnlp-main.154. URL: https://aclanthology.org/2020.emnlp-main.154 (visited on 02/05/2024).

National Institutes of Health (Apr. 2015). *NOT-OD-15-089: Racial and Ethnic Categories and Definitions for NIH Diversity Programs and for Other Reporting Purposes*. Notice Number: NOT-OD-15-089. URL: https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html (visited on 02/16/2024).

NCDJ (Aug. 2021). *Disability Language Style Guide | National Center on Disability and Journalism*. en-US. Walter Cronkite School of Journalism and Mass Communication, Arizona State University. URL: https://ncdj.org/style-guide/ (visited on 02/15/2024).

Paik, Cory, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann (Nov. 2021). "The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 823–835. DOI: 10.18653/v1/2021.emnlp-main.63. URL: https://aclanthology.org/2021.emnlp-main.63 (visited on 03/18/2024).

Prince, Alan and Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. en. Tech. rep. ROA Version, 8/2002. University of Colorado at Boulder: Rutgers University Center for Cognitive Science and Computer Science Department.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners". en. In: Published by OpenAI.

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff (July 2020). "Masked Language Model Scoring". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 2699–2712. DOI: 10.18653/v1/2020.acl-main.240. URL: https://aclanthology.org/2020.acl-main.240 (visited on 02/05/2024).

Shwartz, Vered and Yejin Choi (Dec. 2020). "Do Neural Language Models Overcome Reporting Bias?" In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6863–6870. DOI: 10.18653/v1/2020.coling-main.605. URL: https://aclanthology.org/2020.coling-main.605 (visited on 02/05/2024).

Suresh, Harini and John Guttag (Nov. 2021). "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '21. New York, NY, USA: Association for Computing Machinery, pp. 1–9. ISBN: 978-1-4503-8553-4. DOI: 10.1145/3465416.3483305. URL: https://dl.acm.org/doi/10.1145/3465416.3483305 (visited on 05/02/2024).

Tal, Yarden, Inbal Magar, and Roy Schwartz (July 2022). "Fewer Errors, but More Stereotypes? The Effect of Model Size on Gender Bias". In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Chris-

tian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen. Seattle, Washington: Association for Computational Linguistics, pp. 112–120. DOI: 10.18653/v1/2022.gebnlp-1.13. URL: https://aclanthology.org/2022.gebnlp-1.13 (visited on 02/06/2024).

US Census Bureau (2020). *2020 Census*. Section: Government. URL: https://www.census.gov/2020census (visited on 03/15/2024).

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (Nov. 2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446 (visited on 05/13/2024).

Wang, Tianlu, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez (Oct. 2019). "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504, pp. 5309–5318. DOI: 10.1109/ICCV.2019.00541. URL: https://ieeexplore.ieee.org/document/9008527 (visited on 05/02/2024).

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (2020). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English". In: *Transactions of the Association for Computational Linguistics* 8. Ed. by Mark Johnson, Brian Roark, and Ani Nenkova. Place: Cambridge, MA Publisher: MIT Press, pp. 377–392. DOI: 10.1162/tacl_a_00321. URL: https://aclanthology.org/2020.tacl-1.25 (visited on 03/18/2024).

WHO (2011). *World Report on Disability*. en. Tech. rep. World Health Organization and the World Bank.

Wilcoxon, Frank (1945). "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6. Publisher: [International Biometric Society, Wiley], pp. 80–83. ISSN: 0099-4987. DOI: 10.2307/3001968. URL: https://www.jstor.org/stable/3001968 (visited on 04/17/2024).

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (June 2022). *OPT: Open Pre-trained Transformer Language Models*. arXiv:2205.01068 [cs]. DOI: 10.48550/arXiv.2205.01068. URL: http://arxiv.org/abs/2205.01068 (visited on 03/20/2024).

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (June 2018). "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)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 15–20. DOI: 10.18653/v1/N18-2003. URL: https://aclanthology.org/N18-2003 (visited on 05/02/2024).

Appendices

A Lists of Expressions

Disability
deaf
blind
with a disability
in a wheelchair
with cerebral palsy
with a mental illness
with epilepsy
with spinal curvature
chronically ill
short-statured
with dyslexia
with Down's syndrome
without a disability

Table A.1: List of expressions for category *Disability*

Race
native american
asian
black
hispanic
pacific islander
white

Table A.2: List of expressions for category *Race*

Queerness
allosexual
asexual
bisexual
cis
cisgender
gay
heterosexual
LGBTQ
lesbian
nonbinary
pansexual
queer
straight
trans
transgender

Table A.3: List of expressions for category *Queerness*

B *N*-gram frequency tables

expression	NOW	iWeb	GloWbE
person	4733987	4303004	842958
woman	3549911	1649501	366023
man	6766790	4124411	915096

Table B.1: Frequencies of the different person-words in each corpus.

		NOW		iWeb		GloWbE	
		count	conditional	count	conditional	count	conditional
blind	man	4335	0.0641	6730	0.1632	2042	0.2231
	person	2216	0.0468	2762	0.0642	617	0.0732
	woman	1120	0.0316	571	0.0346	113	0.0309
chronically ill	man	24	0.0004	5	0.0001	0	0.0000
	person	72	0.0015	0	0.0000	7	0.0008
	woman	18	0.0005	4	0.0002	0	0.0000
deaf	man	588	0.0087	511	0.0124	168	0.0184
	person	1292	0.0273	1281	0.0298	356	0.0422
	woman	436	0.0123	230	0.0139	53	0.0145
in a wheelchair	man	970	0.0143	335	0.0081	59	0.0064
	person	372	0.0079	278	0.0065	70	0.0083
	woman	660	0.0186	213	0.0129	32	0.0087
short-statured	man	7	0.0001	4	0.0001	1	0.0001
	person	3	0.0001	7	0.0002	1	0.0001
	woman	4	0.0001	0	0.0000	0	0.0000
with Down's syndrome	man	60	0.0009	12	0.0003	10	0.0011
	woman	48	0.0014	17	0.0010	3	0.0008
	person	50	0.0011	20	0.0005	12	0.0014
with a disability	man	98	0.0014	29	0.0007	10	0.0011
	person	2339	0.0494	4454	0.1035	723	0.0858
	woman	179	0.0050	66	0.0040	17	0.0046
with a mental illness	man	30	0.0004	11	0.0003	2	0.0002
	person	103	0.0022	134	0.0031	28	0.0033
	woman	11	0.0003	6	0.0004	3	0.0008
with cerebral palsy	man	236	0.0035	92	0.0022	16	0.0017
	person	52	0.0011	67	0.0016	17	0.0020
	woman	160	0.0045	33	0.0020	8	0.0022
with dyslexia	man	8	0.0001	3	0.0001	2	0.0002
	person	41	0.0009	69	0.0016	4	0.0005
	woman	3	0.0001	1	0.0001	0	0.0000
with epilepsy	man	39	0.0006	16	0.0004	2	0.0002
	person	121	0.0026	134	0.0031	44	0.0052
	woman	41	0.0012	18	0.0011	6	0.0016
with spinal curvature	man	0	0.0000	0	0.0000	0	0.0000
	person	0	0.0000	0	0.0000	0	0.0000
	woman	0	0.0000	0	0.0000	0	0.0000
without a disability	man	0	0.0000	0	0.0000	0	0.0000
	person	31	0.0007	64	0.0015	18	0.0021
	woman	2	0.0001	0	0.0000	0	0.0000

Table B.2: Absolute and relative n-gram frequencies for category Disability.

		NOW		iWeb		GloWbE	
		count	conditional	count	conditional	count	conditional
asian	man	3090	0.0457	1074	0.0260	190	0.0208
	person	793	0.0117	242	0.0056	67	0.0079
	woman	4692	0.0693	1236	0.0749	255	0.0697
black	man	89318	1.3199	23606	0.5723	6266	0.6847
	person	18085	0.2673	5309	0.1234	1748	0.2074
	woman	57512	0.8499	11184	0.6780	1897	0.5183
hispanic	man	1207	0.0178	584	0.0142	63	0.0069
	person	182	0.0027	100	0.0023	12	0.0014
	woman	813	0.0120	437	0.0265	40	0.0109
native american	man	669	0.0099	162	0.0039	12	0.0013
	person	64	0.0009	25	0.0006	7	0.0008
	woman	1144	0.0169	404	0.0245	40	0.0109
pacific islander	man	8	0.0001	0	0.0000	1	0.0001
	person	3	0.0000	0	0.0000	0	0.0000
	woman	12	0.0002	0	0.0000	3	0.0008
	ϵ	8016	-	3540	-	283	-
white	man	35788	0.5289	21148	0.5128	6797	0.7428
	person	8967	0.1325	5699	0.1324	1881	0.2231
	woman	16862	0.2492	6896	0.4181	1355	0.3702

Table B.3: Absolute and relative n-gram frequencies for category Race.

		NOW		iWeb		GloWbE	
		count	conditional	count	conditional	count	conditional
LGBTQ	man	3	0.0000	0	0.0000	0	0.0000
	person	608	0.0090	67	0.0016	4	0.0005
	woman	67	0.0010	2	0.0001	0	0.0000
allosexual	man	0	0.0000	0	0.0000	0	0.0000
	person	3	0.0000	13	0.0003	0	0.0000
	woman	0	0.0000	0	0.0000	0	0.0000
asexual	man	12	0.0002	17	0.0004	9	0.0010
	person	91	0.0013	94	0.0022	4	0.0005
	woman	20	0.0003	19	0.0012	2	0.0005
bisexual	man	315	0.0047	151	0.0037	52	0.0057
	person	165	0.0024	83	0.0019	41	0.0049
	woman	622	0.0092	181	0.0110	60	0.0164
cis	man	349	0.0052	76	0.0018	29	0.0032
	person	64	0.0009	61	0.0014	20	0.0024
	woman	416	0.0061	90	0.0055	48	0.0131
cisgender	man	264	0.0039	49	0.0012	3	0.0003
	person	84	0.0012	34	0.0008	2	0.0002
	woman	335	0.0050	58	0.0035	2	0.0005
gay	man	17424	0.2575	6611	0.1603	2025	0.2213
heterosexual	man	453	0.0067	306	0.0074	162	0.0177
	person	89	0.0013	71	0.0017	51	0.0061
	woman	239	0.0035	156	0.0095	47	0.0128
lesbian	woman	530	0.0078	188	0.0114	63	0.0172
	ϵ	77566	-	58619	-	12962	-
nonbinary	person	516	0.0076	15	0.0003	0	0.0000
pansexual	man	8	0.0001	1	0.0000	0	0.0000
	person	30	0.0004	7	0.0002	2	0.0002
	woman	19	0.0003	4	0.0002	2	0.0005
queer	man	580	0.0086	63	0.0015	16	0.0017
	person	2058	0.0304	209	0.0049	41	0.0049
	woman	1503	0.0222	210	0.0127	38	0.0104
straight	man	3638	0.0538	2058	0.0499	437	0.0478
	person	429	0.0063	284	0.0066	147	0.0174
	woman	659	0.0097	360	0.0218	141	0.0385
trans	man	2203	0.0326	396	0.0096	88	0.0096
	person	4157	0.0614	787	0.0183	146	0.0173
	woman	8463	0.1251	1300	0.0788	252	0.0688
transgender	man	2279	0.0337	467	0.0113	35	0.0038
	person	4512	0.0667	1121	0.0261	94	0.0112
	woman	10051	0.1485	1962	0.1189	86	0.0235

Table B.4: Absolute and relative n-gram frequencies for category Queerness.

C Full Result Tables

model	person word	hearing	sight	unspecified	mobility	cerebral palsy	mental health	epilepsy	physical	chronic illness	short stature	cognitive	down's syndrome	without
bert-base-uncased	person	0.69*	0.69*	-0.61*	-0.15*	-0.95*	-0.87*	-0.96*	0.3*	-0.92*	-0.98*	-0.98*	0.98*	0.43*
	woman	0.87*	0.77*	-0.03	-0.4*	-0.94*	-0.83*	-0.97*	0.33*	-0.89*	-0.98*	-0.98*	0.99*	0.76*
	man	0.87*	0.65*	-0.03	-0.33*	-0.85*	-0.72*	-0.93*	0.45*	-0.87*	-0.98*	-0.95*	0.99*	0.72*
	total	0.81*	0.71*	-0.23*	-0.29*	-0.91*	-0.81*	-0.95*	0.36*	-0.9*	-0.98*	-0.97*	0.99*	0.65*
bert-large-uncased	person	0.78*	0.72*	-0.5*	-0.0	-0.87*	-0.79*	-0.9*	0.23*	-0.79*	-0.96*	-0.94*	0.99*	0.43*
	woman	0.91*	0.87*	-0.06*	-0.28*	-0.92*	-0.73*	-0.92*	0.18*	-0.82*	-0.97*	-0.95*	1.0*	0.77*
	man	0.88*	0.74*	0.09*	-0.23*	-0.82*	-0.6*	-0.86*	0.29*	-0.74*	-0.97*	-0.88*	1.0*	0.73*
	total	0.86*	0.78*	-0.16*	-0.16*	-0.87*	-0.71*	-0.89*	0.23*	-0.78*	-0.97*	-0.93*	1.0*	0.65*
albert-base-v2	person	0.69*	0.74*	-0.49*	-0.36*	-0.88*	-0.85*	0.49*	0.61*	-0.71*	0.96*	-0.7*	-0.22*	0.1*
	woman	0.77*	0.69*	-0.22*	-0.72*	-0.94*	-0.87*	0.34*	0.54*	-0.75*	0.97*	-0.8*	-0.43*	0.32*
	man	0.76*	0.7*	0.07*	-0.58*	-0.86*	-0.77*	0.64*	0.53*	-0.49*	0.96*	-0.57*	-0.07*	0.42*
	total	0.74*	0.71*	-0.21*	-0.56*	-0.89*	-0.83*	0.5*	0.56*	-0.65*	0.96*	-0.69*	-0.23*	0.29*
albert-large-v2	person	0.74*	0.62*	-0.45*	-0.14*	-0.86*	-0.74*	0.5*	0.57*	-0.81*	0.95*	-0.69*	-0.65*	0.33*
	woman	0.81*	0.63*	-0.09*	-0.46*	-0.93*	-0.72*	0.47*	0.6*	-0.81*	0.96*	-0.79*	-0.73*	0.53*
	man	0.78*	0.6*	0.12*	-0.39*	-0.83*	-0.63*	0.65*	0.64*	-0.73*	0.93*	-0.61*	-0.58*	0.6*
	total	0.78*	0.61*	-0.14*	-0.33*	-0.87*	-0.7*	0.55*	0.61*	-0.78*	0.95*	-0.7*	-0.65*	0.49*
roberta-base	person	0.78*	0.63*	-0.49*	-0.02	-0.93*	-0.85*	0.55*	-0.44*	0.3*	0.09*	-0.92*	-0.93*	0.45*
	woman	0.88*	0.8*	-0.19*	-0.33*	-0.95*	-0.86*	0.63*	-0.09*	0.46*	-0.33*	-0.96*	-0.96*	0.78*
	man	0.85*	0.61*	0.1*	-0.23*	-0.87*	-0.72*	0.7*	-0.13*	0.61*	-0.34*	-0.91*	-0.86*	0.77*
	total	0.84*	0.68*	-0.19*	-0.19*	-0.91*	-0.81*	0.62*	-0.22*	0.46*	-0.19*	-0.93*	-0.92*	0.67*
roberta-large	person	0.86*	0.83*	-0.22*	0.2*	-0.85*	-0.73*	0.68*	-0.22*	0.38*	-0.43*	-0.87*	-0.92*	0.61*
	woman	0.91*	0.87*	0.11*	-0.13*	-0.91*	-0.75*	0.75*	-0.08*	0.57*	-0.87*	-0.91*	-0.95*	0.84*
	man	0.89*	0.75*	0.34*	-0.08*	-0.81*	-0.58*	0.81*	-0.01	0.66*	-0.85*	-0.84*	-0.82*	0.83*
	total	0.89*	0.82*	0.08*	0.01	-0.86*	-0.69*	0.75*	-0.1*	0.54*	-0.72*	-0.87*	-0.9*	0.77*
gpt2	person	0.94*	0.89*	-0.69*	-0.05*	-0.75*	-0.88*	0.72*	0.33*	0.77*	0.7*	-0.64*	0.97*	0.5*
	woman	0.98*	0.93*	-0.45*	-0.72*	-0.88*	-0.84*	0.85*	0.26*	0.89*	0.61*	-0.58*	0.96*	0.76*
	man	0.97*	0.74*	-0.16*	-0.55*	-0.7*	-0.73*	0.94*	0.54*	0.9*	0.53*	-0.45*	0.98*	0.8*
	total	0.97*	0.86*	-0.44*	-0.45*	-0.78*	-0.82*	0.85*	0.38*	0.86*	0.61*	-0.56*	0.97*	0.69*
gpt2-medium	person	0.98*	0.95*	-0.24*	0.18*	-0.59*	-0.71*	0.84*	0.68*	0.86*	0.43*	-0.33*	0.97*	0.62*
	woman	0.99*	0.97*	0.15*	-0.42*	-0.75*	-0.58*	0.9*	0.64*	0.95*	0.16*	-0.39*	0.97*	0.86*
	man	0.99*	0.92*	0.37*	-0.21*	-0.49*	-0.42*	0.95*	0.75*	0.96*	0.06*	-0.18*	0.98*	0.87*
	total	0.99*	0.95*	0.1*	-0.15*	-0.61*	-0.57*	0.9*	0.69*	0.93*	0.22*	-0.3*	0.98*	0.8*
opt-350m	person	0.95*	0.9*	-0.49*	-0.11*	-0.62*	-0.84*	0.86*	0.67*	0.93*	0.64*	-0.5*	-0.81*	0.47*
	woman	0.99*	0.97*	-0.3*	-0.59*	-0.8*	-0.83*	0.94*	0.69*	0.95*	0.58*	-0.6*	-0.9*	0.78*
	man	0.99*	0.85*	-0.05*	-0.49*	-0.64*	-0.72*	0.96*	0.8*	0.97*	0.49*	-0.35*	-0.78*	0.81*
	total	0.98*	0.91*	-0.28*	-0.4*	-0.69*	-0.8*	0.92*	0.72*	0.95*	0.57*	-0.49*	-0.83*	0.7*
bloom-560m	person	0.99*	0.96*	-0.47*	-0.45*	-0.65*	-0.79*	0.03*	0.94*	0.23*	0.6*	-0.59*	0.42*	0.52*
	woman	0.99*	0.95*	-0.21*	-0.89*	-0.69*	-0.66*	0.11*	0.96*	0.23*	0.16*	-0.68*	0.48*	0.75*
	man	0.99*	0.86*	0.01	-0.84*	-0.42*	-0.49*	0.39*	0.97*	0.3*	0.26*	-0.37*	0.71*	0.84*
	total	0.99*	0.93*	-0.22*	-0.75*	-0.59*	-0.65*	0.18*	0.95*	0.25*	0.35*	-0.55*	0.55*	0.72*

Table C.1: Wilcoxon test effect sizes (r) for category *Disability*. Asterisk indicates $p < 0.01$.

model	person word	native american	asian	black	hispanic	pacific islander	white
bert-base-uncased	person	-0.24*	0.87*	0.78*	0.93*	-0.18*	0.76*
	woman	-0.63*	0.67*	0.69*	0.89*	0.06*	0.66*
	man	-0.17*	0.86*	0.7*	0.94*	0.45*	0.64*
	total	-0.35*	0.81*	0.73*	0.92*	0.12*	0.69*
bert-large-uncased	person	-0.04*	0.87*	0.79*	0.96*	-0.41*	0.78*
	woman	-0.46*	0.73*	0.7*	0.93*	-0.47*	0.69*
	man	0.01	0.88*	0.67*	0.96*	-0.02	0.67*
	total	-0.16*	0.84*	0.73*	0.95*	-0.3*	0.72*
albert-base-v2	person	-0.09*	0.88*	0.83*	0.94*	0.79*	0.88*
	woman	-0.56*	0.55*	0.53*	0.81*	0.46*	0.54*
	man	-0.31*	0.77*	0.68*	0.93*	0.61*	0.69*
	total	-0.32*	0.75*	0.69*	0.9*	0.63*	0.72*
albert-large-v2	person	-0.06*	0.92*	0.83*	0.95*	0.6*	0.86*
	woman	-0.26*	0.55*	0.57*	0.84*	0.02	0.64*
	man	-0.07*	0.81*	0.59*	0.93*	0.41*	0.66*
	total	-0.13*	0.78*	0.68*	0.91*	0.38*	0.73*
roberta-base	person	-0.68*	0.38*	0.75*	-0.17*	-0.06*	0.73*
	woman	-0.82*	-0.24*	0.45*	-0.82*	-0.34*	0.56*
	man	-0.74*	0.18*	0.55*	-0.46*	0.01	0.58*
	total	-0.75*	0.12*	0.6*	-0.48*	-0.13*	0.63*
roberta-large	person	-0.58*	0.18*	0.85*	-0.03*	0.08*	0.86*
	woman	-0.8*	-0.48*	0.63*	-0.77*	-0.57*	0.72*
	man	-0.54*	-0.1*	0.64*	-0.43*	0.05*	0.73*
	total	-0.64*	-0.13*	0.72*	-0.41*	-0.14*	0.77*
gpt2	person	0.62*	0.89*	0.73*	0.87*	0.95*	0.7*
	woman	0.09*	0.63*	0.27*	0.72*	0.9*	0.28*
	man	0.41*	0.82*	0.25*	0.81*	0.94*	0.38*
	total	0.38*	0.79*	0.44*	0.81*	0.93*	0.47*
gpt2-medium	person	0.63*	0.9*	0.86*	0.9*	0.95*	0.83*
	woman	0.03*	0.59*	0.51*	0.63*	0.88*	0.6*
	man	0.35*	0.77*	0.61*	0.77*	0.9*	0.69*
	total	0.35*	0.77*	0.68*	0.78*	0.91*	0.72*
opt-350m	person	-0.05*	0.62*	0.76*	0.62*	0.71*	0.81*
	woman	-0.6*	-0.15*	0.36*	0.26*	0.45*	0.61*
	man	-0.39*	0.19*	0.39*	0.56*	0.42*	0.65*
	total	-0.35*	0.25*	0.52*	0.49*	0.54*	0.7*
bloom-560m	person	0.71*	0.9*	0.94*	0.99*	0.98*	0.95*
	woman	0.0	0.28*	0.64*	0.9*	0.93*	0.66*
	man	0.34*	0.65*	0.68*	0.95*	0.96*	0.71*
	total	0.37*	0.64*	0.78*	0.95*	0.96*	0.8*

Table C.2: Wilcoxon test effect sizes (r) for category *Race*. Asterisk indicates $p < 0.01$.

model	person word	Asexual	Allosexual	Bisexual	Cis	Cisgender	Gay	Heterosexual	LGBTQ	Lesbian	NB	Pansexual	Queer	Straight	Trans	Transgender
bert-base-uncased	person	-0.98*	-0.03	0.9*	0.98*	-0.78*	nan	0.91*	-0.79*	nan	-0.67*	-0.36*	0.94*	0.94*	0.93*	0.67*
	woman	-0.98*	0.18*	0.9*	0.99*	-0.9*	nan	0.92*	-0.88*	0.93*	nan	-0.43*	0.93*	0.88*	0.74*	0.71*
	man	-0.98*	0.31*	0.94*	0.99*	-0.81*	0.68*	0.93*	-0.72*	nan	nan	-0.42*	0.93*	0.87*	0.92*	0.89*
	total	-0.98*	0.16*	0.91*	0.99*	-0.83*	0.68*	0.92*	-0.8*	0.93*	-0.67*	-0.4*	0.93*	0.9*	0.87*	0.76*
bert-large-uncased	person	-0.85*	-0.05*	0.96*	0.99*	-0.83*	nan	0.93*	-0.49*	nan	-0.47*	-0.52*	0.92*	0.93*	0.93*	0.79*
	woman	-0.93*	0.27*	0.95*	1.0*	-0.87*	nan	0.93*	-0.37*	0.94*	nan	-0.49*	0.94*	0.93*	0.83*	0.78*
	man	-0.92*	0.31*	0.96*	1.0*	-0.77*	0.73*	0.92*	0.08*	nan	nan	-0.45*	0.93*	0.9*	0.94*	0.92*
	total	-0.9*	0.18*	0.95*	0.99*	-0.82*	0.73*	0.93*	-0.25*	0.94*	-0.47*	-0.48*	0.93*	0.92*	0.91*	0.84*
albert-base-v2	person	0.04*	0.72*	0.91*	0.98*	0.97*	nan	0.79*	0.5*	nan	0.98*	0.87*	0.87*	0.86*	0.98*	0.82*
	woman	-0.59*	0.52*	0.77*	0.98*	0.96*	nan	0.58*	0.25*	0.65*	nan	0.76*	0.69*	0.78*	0.97*	0.57*
	man	-0.05*	0.71*	0.92*	0.98*	0.97*	0.44*	0.78*	0.64*	nan	nan	0.86*	0.88*	0.81*	0.98*	0.94*
	total	-0.19*	0.66*	0.87*	0.98*	0.97*	0.44*	0.73*	0.47*	0.65*	0.98*	0.83*	0.82*	0.82*	0.98*	0.8*
albert-large-v2	person	0.51*	0.64*	0.94*	0.96*	0.95*	nan	0.89*	0.88*	nan	0.97*	0.8*	0.89*	0.92*	0.97*	0.81*
	woman	0.12*	0.52*	0.82*	0.96*	0.92*	nan	0.79*	0.82*	0.78*	nan	0.73*	0.76*	0.85*	0.94*	0.58*
	man	0.58*	0.69*	0.92*	0.97*	0.96*	0.57*	0.91*	0.93*	nan	nan	0.83*	0.85*	0.79*	0.98*	0.9*
	total	0.42*	0.62*	0.9*	0.96*	0.94*	0.57*	0.87*	0.88*	0.78*	0.97*	0.79*	0.84*	0.86*	0.96*	0.78*
roberta-base	person	0.33*	0.95*	0.91*	0.95*	-0.67*	nan	0.93*	0.88*	nan	0.19*	0.05*	0.91*	0.95*	0.85*	0.78*
	woman	0.03*	0.98*	0.84*	0.92*	-0.95*	nan	0.9*	0.95*	0.85*	nan	-0.49*	0.85*	0.93*	0.69*	0.72*
	man	0.48*	0.97*	0.9*	0.95*	-0.61*	0.62*	0.92*	0.97*	nan	nan	0.12*	0.91*	0.85*	0.91*	0.91*
	total	0.3*	0.96*	0.89*	0.94*	-0.75*	0.62*	0.92*	0.93*	0.85*	0.19*	-0.1*	0.89*	0.91*	0.83*	0.81*
roberta-large	person	0.25*	0.96*	0.96*	0.97*	-0.37*	nan	0.97*	0.94*	nan	0.52*	0.34*	0.94*	0.96*	0.9*	0.86*
	woman	-0.21*	0.97*	0.89*	0.94*	-0.84*	nan	0.93*	0.97*	0.93*	nan	-0.21*	0.91*	0.96*	0.75*	0.75*
	man	0.36*	0.97*	0.93*	0.96*	-0.35*	0.71*	0.95*	0.98*	nan	nan	0.32*	0.94*	0.88*	0.92*	0.94*
	total	0.15*	0.97*	0.93*	0.96*	-0.51*	0.71*	0.95*	0.97*	0.93*	0.52*	0.17*	0.93*	0.93*	0.87*	0.86*
gpt2	person	0.81*	0.99*	0.99*	0.99*	0.66*	nan	0.99*	0.99*	nan	0.7*	0.94*	0.96*	0.99*	0.92*	0.89*
	woman	0.69*	0.99*	0.94*	0.97*	0.25*	nan	0.96*	1.0*	0.95*	nan	0.86*	0.92*	0.97*	0.68*	0.74*
	man	0.85*	0.99*	0.98*	0.99*	0.63*	0.62*	0.97*	1.0*	nan	nan	0.94*	0.97*	0.9*	0.95*	0.96*
	total	0.79*	0.99*	0.97*	0.98*	0.53*	0.62*	0.98*	1.0*	0.95*	0.7*	0.92*	0.95*	0.96*	0.86*	0.88*
gpt2-medium	person	0.86*	0.99*	1.0*	1.0*	0.82*	nan	1.0*	1.0*	nan	0.89*	0.95*	0.98*	0.99*	0.96*	0.94*
	woman	0.89*	0.99*	0.97*	0.99*	0.58*	nan	0.99*	1.0*	0.98*	nan	0.86*	0.97*	0.98*	0.88*	0.84*
	man	0.93*	0.99*	0.99*	0.99*	0.84*	0.76*	0.99*	1.0*	nan	nan	0.94*	0.99*	0.95*	0.98*	0.98*
	total	0.89*	0.99*	0.99*	0.99*	0.76*	0.76*	0.99*	1.0*	0.98*	0.89*	0.92*	0.98*	0.98*	0.95*	0.93*
opt-350m	person	0.85*	1.0*	0.99*	0.99*	0.89*	nan	1.0*	0.99*	nan	0.86*	0.93*	0.99*	0.98*	0.92*	0.94*
	woman	0.8*	1.0*	0.98*	0.97*	0.77*	nan	0.99*	1.0*	0.99*	nan	0.8*	0.99*	0.97*	0.78*	0.9*
	man	0.91*	1.0*	0.99*	1.0*	0.9*	0.72*	0.98*	1.0*	nan	nan	0.89*	1.0*	0.92*	0.97*	0.99*
	total	0.85*	1.0*	0.99*	0.99*	0.86*	0.72*	0.99*	1.0*	0.99*	0.86*	0.88*	0.99*	0.96*	0.9*	0.95*
bloom-560m	person	0.95*	0.96*	0.78*	1.0*	0.95*	nan	0.84*	0.79*	nan	0.99*	0.96*	0.63*	0.99*	1.0*	0.66*
	woman	0.93*	0.89*	0.35*	1.0*	0.76*	nan	0.46*	0.81*	0.97*	nan	0.9*	0.11*	0.99*	0.99*	0.12*
	man	0.96*	0.94*	0.66*	1.0*	0.94*	0.78*	0.67*	0.92*	nan	nan	0.97*	0.62*	0.97*	1.0*	0.75*
	total	0.95*	0.93*	0.61*	1.0*	0.9*	0.78*	0.68*	0.85*	0.97*	0.99*	0.95*	0.47*	0.99*	0.99*	0.54*

Table C.3: Wilcoxon test effect sizes (r) for category *Queerness*. Asterisk indicates $p < 0.01$.