Introducing MARB — A Dataset for Studying the Social Dimensions of Reporting Bias in Language Models

Tom Södahl Bladsjö
tom.sodahl@gmail.com

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@gu.se

Gothenburg University, Sweden

Abstract

Reporting bias is the tendency for speakers to omit unnecessary or obvious information while mentioning things they consider relevant or surprising. In descriptions of people, reporting bias can manifest as a tendency to over report on attributes that deviate from the norm. While social bias in language models has garnered a lot of attention in recent years, a majority of the existing work equates "bias" with "stereotypes". We suggest reporting bias as an alternative lens through which to study how social attitudes manifest in language models. We present the MARB dataset, a diagnostic dataset for studying the interaction between social bias and reporting bias in language models. We use MARB to evaluate the off-the-shelf behavior of both masked and autoregressive language models and find signs of reporting bias with regards to marginalized identities, mirroring that which can be found in human text. This effect is particularly pronounced when taking gender into account, demonstrating the importance of considering intersectionality when studying social phenomena like biases.

1 Introduction

The issue of social bias in language models has received increased attention in the past few years, with many recent efforts focusing on *benchmark datasets* for quantifying bias in a way that is comparable across models (Blodgett et al., 2021). The majority of work in this area equates "bias" with "stereotypes" (Blodgett et al., 2020). While stereotypes are indeed one way in which social inequalities manifest in language, they are only one of the symptoms of a larger underlying problem. Language in itself is a social phenomenon (Bakhtin, 1935/1981). Utterances do not only communicate semantic and pragmatic content; they also mirror the social perspective of the speaker.

In order to better predict potential harms caused by language models, we need a more holistic understanding of "bias" that connects model behavior with social norms, attitudes and expectations. In other words, we do not see bias as inherently or necessarily bad. Instead, we view biases as symptoms of a perspective being encoded in the model. We suggest reporting bias, or "the tendency of people to not state the obvious" (Paik et al., 2021), as a lens through which to study social norms and attitudes in language models. In descriptions of people, reporting bias can manifest as a tendency to over report on attributes that deviate from the norm, drawing further attention to the perceived *Otherness* (see e.g. Thomas-Olalde and Velho, 2011) of already marginalized groups. Despite the obvious connection, the relationship between reporting bias and social biases has not previously been studied.

To address this research gap, we introduce the Marked Attribute and Reporting Bias dataset, or MARB for short, for measuring model reporting bias with regards to sensitive human attributes such as race, queerness and disability. We generate templates from naturally occurring English text, which are then populated with different descriptors related to these attributes. The full dataset and usage instructions can be found on GitHub.[1]

We introduce the dataset in Section 4 and discuss the theoretical motivations and technical implementation behind it, as well as recommendations for how it can be used. As an example of this recommended usage, we evaluate six popular large language models with MARB in Section 5.

We find signs of reporting bias with regards to marginalized attributes, similar to that which is found in online news media. We also find that the gender of the person being described has a noticeable effect on the observed reporting bias, in that sentences describing women are generally more likely to mention attributes like race or queerness. The effect is particularly striking for sen-

[1] https://github.com/TomBladsjo/MARB

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 59–74

tences mentioning Asian women, underlining the importance of taking intersectionality[2] into account when studying bias.

1.1 Bias Statement

Throughout this work we understand the term *bias* broadly to mean any systematic difference in model performance between subsets of the data that share a specific property. As such, we do not view biases as necessarily and inherently harmful.

The properties of interest in the MARB dataset are descriptors identifying certain social groups. On the one hand, differences in how likely different attributes are to be mentioned can be understood as a kind of representational harm; consistently pointing out characteristics that differ from the norm may contribute to society's view of marginalized groups as strange and *Other* (Thomas-Olalde and Velho, 2011). On the other hand, it can also be used as an indicator of how different social groups are perceived, providing a useful tool for studying social norms and attitudes that would otherwise be hard to identify.

2 Related Work

Reporting bias in training data has been shown to affect the commonsense knowledge acquired by language models (Shwartz and Choi, 2020; Paik et al., 2021). Much of the existing work in this area focuses on visual commonsense knowledge, such as the colors of common objects (Paik et al., 2021; Hagström and Johansson, 2022; Misra et al., 2016).

The issue of social biases in language models has received increasing attention in recent years (Blodgett et al., 2020; Ducel et al., 2023). The majority of works in this field have focused specifically on gender and/or racial bias in simple binary settings such as male/female, white/Black (e.g. Kiritchenko and Mohammad, 2018; May et al., 2019; Tal et al., 2022). However, more recent work has also branched out to finer-grained analyses of biases against other social groups, such as people with disabilities (Hutchinson et al., 2020) and queer people (Felkner et al., 2023). May et al. (2019) note the need to consider intersectional biases, an area that is still under-researched.

A growing body of research has been directed towards quantifying social biases in ways that are generalizable across models. Many of these benchmarks and diagnostic datasets rely on artificially constructed templates (e.g. Warstadt et al., 2020; Felkner et al., 2023) or crowdworkers (e.g. Nadeem et al., 2021; Nangia et al., 2020) for contrasting examples. The majority of these papers conceptualize "bias" as stereotypes.

There has not been any previous work studying the interactions between reporting bias and social biases.

(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: Two images with accompanying captions from the Flickr8k dataset (Hodosh et al., 2013).

3 Reporting Bias and Markedness

Human language is underspecified. When we talk, we leave out the things we consider unimportant, inferrable from context or simply too obvious to mention. This behavior, described by Grice (1975) as the *maxim of quantity*, leads to a discrepancy between reality and description that is known as *reporting bias*. Levinson (2000) builds on Gricean theory by considering what makes something too obvious to mention. He suggests that linguistic expressions have so-called *default interpretations*: When we hear a certain expression, the interpretation closest at hand will often be the most typical or normative one. If we want to describe a situation

[2]Throughout this work we understand the term *intersectionality* as social dynamics or effects that arise when looking at two or more attributes but that are smaller or completely absent when looking at them separately.

that differs from that norm, we need to specify by marking it in our message. Thus, in human communication, "what is simply described is stereotypically exemplified" (Levinson, 2000, p. 136), while a *marked* message indicates a *marked* situation.

To use a frequent example from previous work on reporting bias (e.g. Paik et al., 2021; Shwartz and Choi, 2020), while most of us would agree that bananas are typically yellow, the bigram "green banana" tends to be more frequent than "yellow banana" in text. Figure 1 gives an example of how the same phenomenon manifests in descriptions of people. The girl in 1a is simply described as "a little girl", while the girl in 1b is described as "an Asian girl". We can interpret this as the annotator considering "white" to be the default for little girls, and thus too obvious to mention in the caption.[3]

In Table 1 we sketch a simple model of markedness with two types of situation (marked and unmarked) and two types of message (again, marked and unmarked). Since we are currently interested in reporting bias related to human attributes, we consider a marked situation in this context to be one where a person has some attribute that deviates from the unmarked norm. Note that the unmarked message is the same for both types of situation; it is only in marked messages we can really know which situation is being described.

In practice unmarked messages tend to be more common than marked messages regardless of the attribute in question. It would be inefficient to include every single detail when describing a situation. On the other hand, we would expect marked messages to be more common for marked attributes than for unmarked ones, in accordance with the Gricean maxim of quantity.

[3]In social sciences, this would be described as whiteness being the *unmarked norm* (Bucholtz and Hall, 2005).

	Marked situation	**Unmarked situation**
Marked message	**an Asian girl**	a white girl
Unmarked message	a girl	**a girl**

Table 1: A simple model of markedness. We would expect marked messages to describe marked situations, and unmarked messages to describe unmarked situations.

	Person-word		
Descriptor	**Person**	**Woman**	**Man**
Asian	1.7e-4	1.3e-3	4.6e-4
Black	3.8e-3	**1.6*e*-2**	1.3e-2
Hispanic	4.0e-5	2.3e-4	1.8e-4
White	1.9e-3	4.8e-3	5.3e-3
Native Hawaiian	0	**1.0*e*-5**	0
Native American	1.0e-5	3.2e-4	1.0e-4

Table 2: Conditional probabilities of racial attribute descriptors given each person-word, obtained from ngram frequencies in the NOW corpus. In general, racial attributes are mentioned more often along with the word *woman*. Two notable cases (marked in **bold**) are Black woman, with the highest probability overall, and Native Hawaiian, which only co-occurs with *woman*.

3.1 Reporting Bias in Text

Following earlier work on reporting bias (Gordon and Van Durme, 2013; Shwartz and Choi, 2020; Paik et al., 2021), we start by investigating how the kind of reporting bias we are interested in manifests in a large corpus of human text. For this purpose we analyze the News on the Web corpus (NOW)[4], a 20 billion word collection of English language news text from web-based newspapers and magazines.

More specifically, we look at the conditional probability that a racial attribute descriptor modifies a given noun designating a person. The results are reported in Table 2. For all person words, *Black* is the most commonly mentioned attribute descriptor, followed by *white*. We then compare these probabilities with the ones that arise from recent US demographic data[5] (US Census Bureau, 2020).

We find a somewhat strong Spearman rank correlation ($\rho = .67$, $p = .002$), which indicates that attributes that are more common in the United States are also mentioned more often in English language news text (predominantly from American sources). On the other hand, a very weak Pearson correlation ($r = .21$, $p = .4$) shows that this relationship is not linear – the frequency at which a certain attribute is mentioned is not proportional to how common it is in real life. In other words, there is a discrepancy between reality and how it

[4]english-corpora.org/now

[5]We consider each n-gram consisting of a descriptor followed by a person-word to be a datapoint in this context. Furthermore, the US demographic data does not record the gender distributions in racial and ethnic groups. Thus, we assume that real-world race and ethnicity is similarly distributed for all genders for the purposes of this analysis.

Version	Sequence
Unmarked	I was talking to **a woman**
Lesbian	I was talking to **a lesbian**
Straight	I was talking to **a straight woman**
Trans	I was talking to **a trans woman**
Cis	I was talking to **a cis woman**

Table 3: Example sequences from the dataset for the category *Queerness*. Each marked version contrasts with the unmarked template sequence by specifying the relevant attribute. Note that "Lesbian" appears on its own instead of preceding the word "woman".

is described in the NOW corpus, which is a sign of reporting bias. Note that the person-word *woman* displays the highest value for all attribute descriptors except for *white*, indicating that race or ethnicity is more commonly mentioned when talking about women. We will return to this phenomenon in Section 5.

4 The MARB Dataset

4.1 General Description

The Marked Attribute and Reporting Bias (MARB) dataset is intended as a diagnostic dataset for detecting reporting bias with regards to socially marked attributes in English. However, the dataset itself and the techniques used to create it are agnostic as to testing method and model architecture. This means that MARB can be used to explore other research questions as well.

MARB consists of 28.5K sequence templates based on naturally occurring written English text[6] which can be used to construct examples given certain categories of attributes. Following the markedness model described in Table 1, we let the template sequences constitute our unmarked messages. By copying each sequence and inserting a descriptor for the attribute of interest, we obtain a set of marked sequences for each attribute descriptor (see Table 3). This lets us measure the effect of adding the attribute descriptor by comparing the probability of a marked message with that of its unmarked version.

The current release of the dataset includes attribute descriptors pertaining to Race, Queerness, and Disability. We also provide methods for users to expand the dataset with categories and descriptors of their own. A more detailed breakdown of the dataset can be found in Appendix A.

[6] As opposed to artificially constructed templates.

4.2 Dataset Creation

4.2.1 Template Selection and Person-Words

As mentioned before, we use templates based on naturally occurring written English text with the idea that it will allow us to better capture actual language usage. The template sequences were extracted from the 2021 version of the enTenTen corpus (Jakubíček et al., 2013)[7]. This is a large web-scraped corpus built specifically to include only linguistically valuable text by removing duplicated and machine-generated content, as well as spam.

We selected sequences containing noun phrases of the form "*a <person-word>*", where the person-words used are *person*, *woman*, and *man*. The resulting dataset separates sequences based on the person-word used, allowing for intersectional analysis. For each person-word, a random sample of 10K sequences was retrieved using the *concordance* tool[8] and processed to remove context outside of sentence boundaries. Out of these 10K sequences, the 500 shortest were filtered out to mitigate effects of sequence length on the final results, resulting in a total of 9.5K template sequences per person-word. The final template lengths range from 4 to 56 words[9], with a median length of 20 words.

4.2.2 Categories and Descriptors

The dataset is structured around *categories* of attributes, where each category comes with a set of *attribute descriptors*. The descriptors are inserted into the template sequences to create attribute-specific versions of each sequence (see Table 3). As mentioned in the general description, the current release of the dataset supports experiments on reporting bias pertaining to categories Race, Queerness and Disability. More categories and attributes can easily be added by providing a file with the desired attributes and descriptors to the dataset creation script (available on GitHub).

The choice of attributes for each category was informed by previous work in bias research. Following e.g. Czarnowska et al. (2021), the attributes relating to *Race* were based on the Racial and Ethnic Categories and Definitions for NIH Diversity Programs (National Institutes of Health, 2015)

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

[8] https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/

[9] Whitespace tokenized.

which correspond to those used by the U.S. Census Bureau.[10] Different categories can have different terms with different connotations. For ease of comparison and to avoid introducing unreliability from aggregation methods, only one descriptor per category was included. The attributes and descriptors relating to *Queerness* were based on Felkner et al. (2023).[11] For comparability, the descriptors "non-binary", "lesbian" and "gay" were only used with person-words "person", "woman" and "man" respectively. Descriptors relating to *Disability* were taken from Hutchinson et al. (2020). Since the lists of descriptors used in Hutchinson et al. (2020) are very extensive, we used a smaller subset of one term per disability category from their list of recommended phrases. A full list of attributes and descriptors for each category can be found in Appendix B.

We recognize that our choice of descriptors is in no way a complete representation of all the groups that may be subject to this kind of bias. We encourage future work to expand and adapt the lists of descriptors to better represent their chosen target groups.

4.3 Usage

The MARB dataset is mainly intended to be used to analyze the behaviour of off-the-shelf language models. A metric used to evaluate this should be chosen with the model's pretraining task in mind.

Since probability-based metrics are contingent on the model vocabulary, they are not directly comparable between models. Earlier work (e.g. Nangia et al., 2020; Nadeem et al., 2021; Felkner et al., 2023) solves this problem by using a contrastive pairs setup, where each pair consists of one biased sequence and one unbiased or counterfactual sequence. The model's bias score can then be defined as the proportion of pairs for which the model is more likely to predict the biased sequence. However, this kind of binary approach severely limits the options for analysis as it only allows for binary characteristics to be evaluated. As noted by Castillo and Gillborn (2021), grouping rather than disaggregating disadvantaged groups could disguise important differences.

MARB is structured around multiple contrasting sequences. We recommend comparing each marked sequence to a common baseline, such as the corresponding unmarked template sequence. The difference between the likelihoods of the marked and unmarked sequence according to the model can then be interpreted as the effect of adding that specific attribute descriptor. This allows for comparing more than two attributes at a time. The effect per attribute can be calculated simply as the proportion of examples for which the marked sequence is more likely than the unmarked, or using a statistic such as rank-biserial correlation r (Cureton, 1956) to measure the effect size (see Section 5).

Rather than using a single score to represent the model's level of bias, we encourage finer-grained analyses to better understand the model's behavior. The structure of MARB allows for comparisons along multiple axes, including *category*, *attribute descriptor*, *person-word*, as well as intersectional analyses such as *attribute descriptor* + *person-word*.

5 Experimental Setup

We present two case studies in this Section to illustrate the kind of analyses that are possible using the MARB dataset. In both studies, we measure the effect of adding the attribute descriptors by comparing marked sequences (those mentioning the attribute) to the corresponding unmarked template sequences. We focus on one category per case study in order to simplify analyses and to better showcase what can be done with the MARB dataset. Moreover, it reduces the environmental impact of our experiments. The first experiment uses the Race category to study masked language models. The second experiment uses the Queerness category to study auto-regressive models.

5.1 Models

We evaluate six pretrained models on MARB. The masked language models we use for experiment 1 are BERT[12] (Devlin et al., 2019), RoBERTa[13] (Liu

[10] An important consideration is whether to include in-group or out-group descriptors. An example of this is "black" and "Black". We ultimately decided to use the lower-case version for the experiments presented in this paper, as it has seen both in- and out-group adoption over a wider timeframe and is likely to have been more predominant in the models' training data.

[11] For completeness, we added the descriptor "allosexual" (a person who is not asexual) as an unmarked attribute contrasting with "asexual". The descriptor "trans" was also included in addition to the already present "transgender" to contrast with "cis" and "cisgender".

[12] https://huggingface.co/google-bert/bert-base-uncased

[13] https://huggingface.co/FacebookAI/roberta-base

et al., 2019), and ALBERT[14] (Lan et al., 2020). As for auto-regressive models, we focus on Mistral[15] (Jiang et al., 2023), Llama[16] (Touvron et al., 2023), and Gemma[17] (Gemma Team et al., 2024) during experiment 2. All models are tested off-the-shelf without any finetuning.

5.2 Metrics

We use *perplexity* (PPL) as the evaluation metric for autoregressive models, and *pseudo-perplexity* (PPPL) for masked language models. PPL is a common intrinsic measure of how well an auto-regressive model fits a corpus of text. PPL(W) is defined as the exponentiated average negative log-likelihood of a sequence W:

$$PPL(W) = \exp\left(-\frac{1}{|W|}\sum_{i\leq|W|}\mathbb{P}(W_i|W_{<i})\right)$$

The definition of sequence perplexity is based on the assumption that we can use the chain rule of probability to obtain the probability of a sequence from its constituent tokens. However, the chain rule does not apply to masked language models where each token prediction is conditioned on both previous and subsequent tokens. Salazar et al. (2020) propose the use of pseudo-perplexity to get around this issue. They suggest calculating the pseudo-log-likelihood of a sequence W as the sum of the conditional log probabilities of each sentence token given the surrounding tokens. Using that definition of pseudo-log-likelihood, the pseudo-perplexity of a sequence W can be calculated as

$$PPPL(W) = \exp\left(-\frac{1}{|W|}\sum_{i\leq|W|}\mathbb{P}(W_i|W_{\setminus i})\right)$$

We compare the PPL/PPPL for each marked sequence to its unmarked couterpart to obtain a set of pairwise differences for each attribute descriptor. We then perform the Wilcoxon signed-rank test (Wilcoxon, 1945) on each set of pairwise differences and measure effect size as the rank-biserial correlation *r* (Cureton, 1956). Using a measure based on ordinal ranking rather than raw perplexities allows us to make meaningful comparisons between models with different vocabularies.

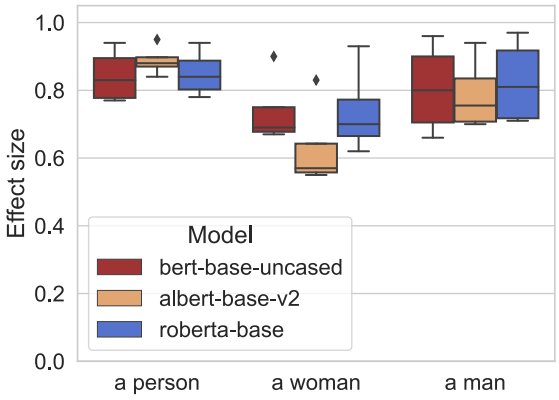

Figure 2: Spread of results over attribute descriptors, per model and person word. A larger spread means a larger difference in performance depending on expression. A higher average means that the model was generally more surprised to see any attribute in this category mentioned.

5.3 Experiment 1: Race and Masked Language Models

In our first case study, we evaluate the masked language models BERT, RoBERTa and ALBERT against the *Race* category of MARB.

We can see the spread of effect sizes per model and person word in Figure 2 in terms of rank-biserial correlation *r*. All results are statistically significant ($p < .01$). Moreover, all results are positive, which means that the sequences including attribute descriptors produced higher perplexities than the original, unmodified sequences. Particularly striking is that all three models show a noticeably lower average effect size for the person word *woman*. This is a consistent pattern across the different descriptors, as seen in Figure 3, and it indicates that attribute descriptors pertaining to race are more expected in descriptions of women than in descriptions of men. The effect is particularly noticeable with the expression "Asian woman", which is a sign of intersectional bias similar to what we found in the NOW corpus (see Section 3.1).

Conversely, for sequences describing "a person", the spread of results tends to be smaller and the average higher, indicating that mentions of race are less expected for this person word, regardless of which specific race attribute is mentioned.

We can see from these results that it is not as simple as some attributes being mentioned more often than others. Other attributes (like gender)

[14] https://huggingface.co/albert/albert-base-v2

[15] https://huggingface.co/mistralai/Mistral-7B-v0.1

[16] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[17] https://huggingface.co/google/gemma-7b

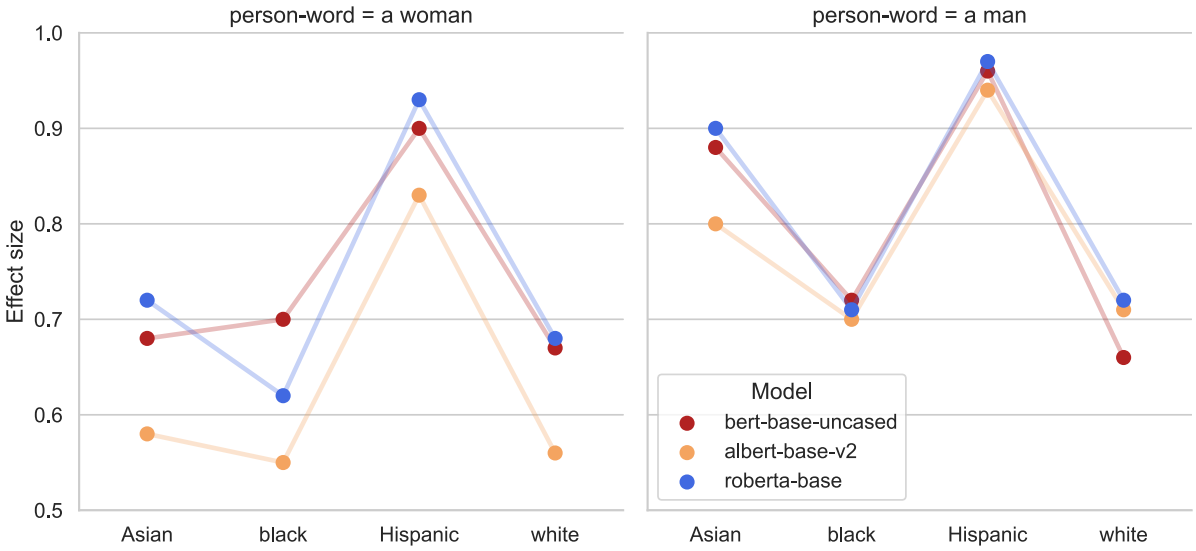

Figure 3: A closer look at the results for person-words "woman" and "man". In all three models, all racial descriptors were more expected in sentences about women than in sentences about men, as seen by the lower effect sizes. Note the larger difference in effect size for the descriptor "Asian".

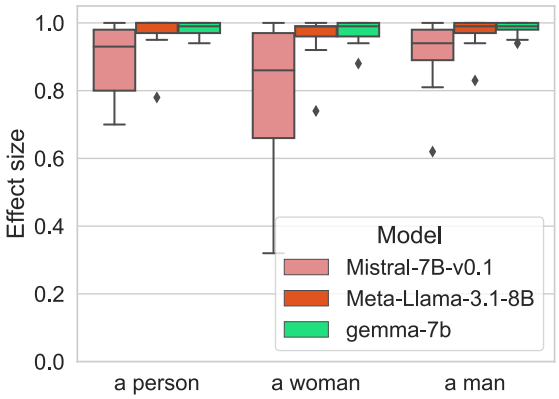

Figure 4: Spread of results over attribute descriptors per model and person word. A larger spread means a larger difference in performance depending on expression. A higher average means that the model was generally more surprised to see any attribute in this category mentioned.

also affect whether or not someone's race is likely to be mentioned, regardless of what that race is.

5.4 Experiment 2: Queerness and Auto-Regressive Language Models

For our second case study, we evaluate the auto-regressive models Mistral, Llama and Gemma on the *Queerness* category of MARB.

As with the first experiment, all effect sizes are positive, meaning that regardless of attribute, all models were more surprised to see the descriptor included. All test results are statistically significant ($p < .01$). Figure 4 shows the spread of results for each model and person word. Just like in the previous case study, all models show a lower average effect size of adding attribute descriptors to sequences describing "a woman" than to those describing "a man" or "a person", indicating that attributes related to queerness are more likely to be mentioned in descriptions of women than in descriptions of, for example, men.

Out of the three models considered, Mistral displays the most noticeable difference. Looking into the specific descriptors in Figure 5 we notice that the average effect size is lower for sequences that mention "a woman" than for either of the other two person-words save for a couple of corner cases, namely "LGBTQ" and "heterosexual". There are three cases in which the difference is much larger: "bisexual", "cisgender", and "transgender".

5.5 Discussion

Despite the differences between the two experiments, we see certain trends appear in both. Particularly noticeable is the aforementioned pattern where attribute descriptors are more expected in sequences describing "a woman" than those describing "a man" or "a person". A possible explanation is that being a woman can be considered a marked attribute in itself, which adds to the reporting bias triggered by other marked attributes. Of particular note is the wider gap in effect size for certain descriptors, such as "Asian", "bisexual", "cisgender", and "transgender". There could be several explana-

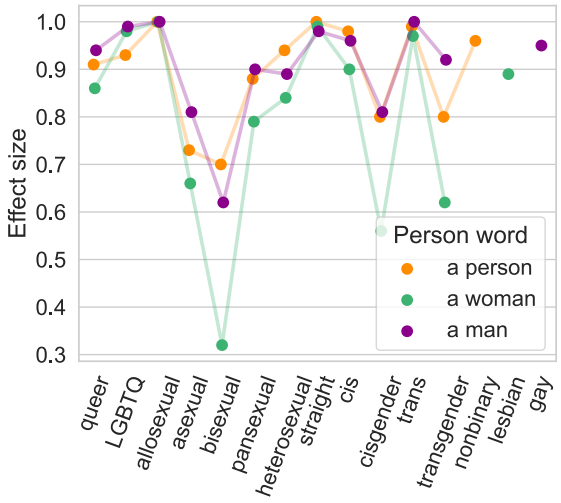

Figure 5: Detailed breakdown of results for Mistral by attribute descriptor. A higher effect size means that the model was more surprised to see this descriptor included in the sequences. Note that the descriptors "nonbinary" and "gay" only combine with the person-words "person" and "man", respectively, and that the descriptor "lesbian" appears on its own instead of preceding the person-word "woman".

tions for this. For example, the discourse surrounding trans people tends to focus on trans and cis women, often leaving trans men to the side (Bracco et al., 2024). Another possibility could be how some of these terms are sexualised or fetishised. Two widely known cases of this phenomenon are indeed Asian and trans women (Forbes et al., 2023; Anzani et al., 2021). These cases illustrate why intersectionality is important when studying biases, as focusing on person-words or descriptors alone would not have yielded these insights.

Another effect that we can see is how public discourse reflects on whether the models expect to see certain descriptors or not. As mentioned in Section 2, most of the online discourse regarding race tends to focus on the United States, where race is often seen as a black-white binary (Perea, 1997; Blodgett et al., 2021). Similarly, the language models are on average less surprised when faced with these two descriptors than with the other ones in the Race category regardless of the person-word used, as seen in Figure 2.

A similar case appears in the *Queerness* category with the descriptors "transgender" and "cisgender". The topic of trans rights has been at the spotlight in British and American politics for a while now. This could explain why neither of the descriptors in this pair are considered to be more of a default than the other according to the language models as seen in Figure 5. Compare this example with the pair "asexual" and "allosexual", where they can be considered to be marked and unmarked attributes, respectively. Of note however is that this same pattern does not hold for the descriptors "trans" and "cis". A reason for this could be that "trans-" is also a prefix, which could interact with the models' tokenizers. We consider that future work could delve into these kinds of interactions.

6 Conclusion

In this paper we explore how reporting bias with regards to marked and marginalized identities manifests in language models. To that end, we create the MARB dataset: a diagnostic dataset meant to study the intersection between social bias and reporting bias via marked and unmarked attributes.

We use MARB to evaluate the out-of-the-box behavior of six popular language models, and find that they show signs of reporting bias with regards to marked attributes, mirroring that found in text corpora. Particularly noticeable are the intersectional effects of gender in combination with other attributes, showing that sensitive attributes like race and queerness are more likely to be mentioned in descriptions of women.

Our results demonstrate that there is a strong connection between reporting bias and social norms and attitudes, recommending reporting bias as a promising direction for future research on social bias in language models. As a way of quantifying social norms through language, the framework and methods presented in this paper could also provide new tools for fields like linguistics and social science. We encourage future work to continue investigating the ways in which social norms manifest in language through reporting bias using the framework presented here, and to extend the MARB dataset to cover more categories and attribute descriptors.

7 Limitations

When working in text-only settings there is no straightforward way to connect linguistic expressions to real-life demographic groups and lived experiences. Multiple expressions often exist referring to the same demographic group, which may be used by different people and carry different connotations. For example, members of a certain group may use one expression to describe themselves

while out-group members use different terms. Conversely, there is often a lack of established terms describing normative attributes, such as not having a disability (Wojahn et al., 2024). The specific choices of attribute descriptors used in MARB are likely to have some effect on the results (Antoniak and Mimno, 2021). We encourage future work to explore the effects of using different descriptors.

Another limitation is that PPL and PPPL are both affected by factors like sequence length and model vocabulary. The MARB dataset was created through adding descriptors to referring expressions, thus modifying the sequence length. While the effects of changing the sequence length are somewhat mitigated by removing the shortest template sequences (see Section 4), they cannot be completely eliminated. We encourage future work to explore using MARB with other metrics that are less reliant on sequence length and model vocabulary.

8 Ethical Considerations

This work deals with language categorizing people based on sensitive attributes such as race, gender identity and sexuality. We recognize that this is a sensitive topic, and that care must be taken not to oversimplify complex real-world power structures or to confuse real-life demographic groups with the words used to describe them. As mentioned in Section 7, there are often many ways to refer to a specific social group, and they carry different connotations and underlying assumptions. While the US census categories are widely used in previous research on bias (e.g. Czarnowska et al., 2021), they are known to correspond badly both to how people identify themselves, and how they are seen by others (Gupta, 2020; Tan, 2022).

Similarly, Hutchinson et al. (2020) note that both terminology and ontological definitions relating to disability are contested, and there is great variation in the language used both by in-group and out-group members. The attribute descriptors included in MARB should be seen as a sample rather than a comprehensive list of the language used to refer to these groups. For future work, we encourage collaboration with researchers in fields like disability studies, as well as with the communities in question to ensure that the descriptors used are grounded in real-world usage and the lived experiences of these groups.

References

Maria Antoniak and David Mimno. 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)*, pages 1889–1904, Online. Association for Computational Linguistics.

Annalisa Anzani, Louis Lindley, Giacomo Tognasso, M. Paz Galupo, and Antonio Prunas. 2021. "Being Talked to Like I Was a Sex Toy, Like Being Transgender Was Simply for the Enjoyment of Someone Else": Fetishization and Sexualization of Transgender and Nonbinary Individuals. *Archives of Sexual Behavior*, 50(3):897–911.

Mikhail M. Bakhtin. 1935/1981. *The dialogic imagination: four essays*. Translated by Michael Holquist. University of Texas Press, Austin. Original work published 1935.

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

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Sofia E. Bracco, Sabine Sczesny, and Marie Gustafsson Sendén. 2024. Media Portrayals of Trans and Gender Diverse People: A Comparative Analysis of News Headlines Across Europe. *Sex Roles*, 90(4):491–507.

Mary Bucholtz and Kira Hall. 2005. Language and Identity. In Alessandro Duranti, editor, *A Companion to Linguistic Anthropology*, 1st edition, pages 369–394. Wiley.

Wendy Castillo and David Gillborn. 2021. How to "QuantCrit:" practices and questions for education data researchers and users. (EdWorkingPaper: 22-546). Retrieved from Annenberg Institute at Brown University.

Edward E. Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2023. Bias identification in language models is biased. In *Workshop on Algorithmic Injustice 2023*, Amsterdam, Netherlands.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 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)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Nicola Forbes, Lauren C. Yang, and Sahnah Lim. 2023. Intersectional discrimination and its impact on asian american women's mental health: A mixed-methods scoping review. *Frontiers in Public Health*, 11.

Gemma Team et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, AKBC '13, pages 25–30, New York, NY, USA. Association for Computing Machinery.

Herbert Paul Grice. 1975. Logic and Conversation. In Peter Cole, editor, *Speech acts*, number 3 in Syntax and semantics, pages 41–58. Academic Press, New York u.a.

Sujata Gupta. 2020. To fight discrimination, the U.S. census needs a different race question. ScienceNews, Section: Science & Society.

Lovisa Hagström and Richard Johansson. 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*, pages 252–261, Dublin, Ireland. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 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*, pages 5491–5501, Online. Association for Computational Linguistics.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*, pages 125–127, Lancaster, United Kingdom.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint*. ArXiv:2310.06825 [cs.CL].

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint*. ArXiv:1909.11942 [cs].

Stephen C. Levinson. 2000. *Presumptive meanings: the theory of generalized conversational implicature*. Language, speech, and communication. MIT, Cambridge, MA, USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 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)*, pages 2930–2939. ISSN: 1063-6919.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

National Institutes of Health. 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. United States Government.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 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*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juan F. Perea. 1997. The black/white binary paradigm of race: The "normal science" of american racial thought. *California Law Review*, 85(5):1213–1258.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Daniel Tan. 2022. Disaggregating Race and Ethnicity Categories in Census Data. *UC Berkeley*.

Oscar Thomas-Olalde and Prof. Dr. Astride Velho. 2011. Othering and its effects : exploring the concept. *Writing Postcolonial Histories of Intercultural Education*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint*. ArXiv:2302.13971 [cs.CL].

US Census Bureau. 2020. 2020 Census. Census.gov, Section: Government.

Alex Warstadt, 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. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83. International Biometric Society, Wiley.

Daniel Wojahn, Stina Ericsson, and Per-Olof Hedvall. 2024. Disablised or Ablised?: Linguistic Categorisations of Dis/ability in Swedish Print Media Over Time. *Disability Studies Quarterly*, 44(1).

A Detailed Breakdown of the Dataset

In this appendix we do a breakdown of the number of examples per attribute and per person word. These are included in Table 4.

Category	Attribute	Number of sequences			
		Person	**Woman**	**Man**	**Total**
Race	Asian	9500	9500	9500	28500
	Black	9500	9500	9500	28500
	black	9500	9500	9500	28500
	Hispanic	9500	9500	9500	28500
	Native American	9500	9500	9500	28500
	Native Hawaiian	9500	9500	9500	28500
	white	9500	9500	9500	28500
Disability	deaf	9500	9500	9500	28500
	blind	9500	9500	9500	28500
	with a disability	9500	9500	9500	28500
	in a wheelchair	9500	9500	9500	28500
	with cerebral palsy	9500	9500	9500	28500
	with a mental illness	9500	9500	9500	28500
	with epilepsy	9500	9500	9500	28500
	with spinal curvature	9500	9500	9500	28500
	chronically ill	9500	9500	9500	28500
	short-statured	9500	9500	9500	28500
	with dyslexia	9500	9500	9500	28500
	with Down's syndrome	9500	9500	9500	28500
	without a disability	9500	9500	9500	28500
Queerness	asexual	9500	9500	9500	28500
	allosexual	9500	9500	9500	28500
	bisexual	9500	9500	9500	28500
	cis	9500	9500	9500	28500
	cisgender	9500	9500	9500	28500
	gay	-	-	9500	9500
	heterosexual	9500	9500	9500	28500
	LGBTQ	9500	9500	9500	28500
	lesbian	-	9500	-	9500
	nonbinary	9500	-	-	9500
	pansexual	9500	9500	9500	28500
	queer	9500	9500	9500	28500
	straight	9500	9500	9500	28500
	transgender	9500	9500	9500	28500
	trans	9500	9500	9500	28500
	Total:	313500	313500	313500	940500

Table 4: Breakdowns of the categories, attributes, and number of example sequences per attribute and person word.

B List of Attribute Descriptors

In this appendix we include the attribute descriptors that we used for the different categories. Table 7 includes the terms for the *Disability* category, Table 5 includes those for the *Queerness* category, and Table 6 includes those for the *Race* category.

	Queerness		
phrase	**person**	**woman**	**man**
asexual	an asexual person	an asexual woman	an asexual man
allosexual	an allosexual person	an allosexual woman	an allosexual man
bisexual	a bisexual person	a bisexual woman	a bisexual man
cis	a cis person	a cis woman	a cis man
cisgender	a cisgender person	a cisgender woman	a cisgender man
gay	-	-	a gay man
heterosexual	a heterosexual person	a heterosexual woman	a heterosexual man
LGBTQ	an LGBTQ person	an LGBTQ woman	an LGBTQ man
lesbian	-	a lesbian	-
nonbinary	a nonbinary person	-	-
pansexual	a pansexual person	a pansexual woman	a pansexual man
queer	a queer person	a queer woman	a queer man
straight	a straight person	a straight woman	a straight man
transgender	a transgender person	a transgender woman	a transgender man
trans	a trans person	a trans woman	a trans man

Table 5: List of descriptors for the *Queerness* category. The columns show how the descriptors are used with each person word.

Race	
phrase	
Asian	an Asian person
Black	a Black person
black	a black person
Hispanic	a Hispanic person
Native Hawaiian	a Native Hawaiian person
white	a white person
Native American	a Native American person

Table 6: List of descriptors for the *Race* category. The descriptors are used in the same way for all person words.

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

Table 7: List of descriptors for the *Disability* category. The descriptors are used in the same way for all person words.

C Numerical Results

This appendix contains the detailed results from Section 5. Table 8 contains the results for Experiment 1 as detailed in Section 5.3. Meanwhile, Table 9 contains the results for Experiment 2 as detailed in Section 5.4. The values presented in these tables are in terms of effect size as described in Section 5.2.

model	phrase	a person	a woman	a man	total
bert-base-uncased	Asian	0.88	0.68	0.88	0.83
	Black	0.78	0.70	0.72	0.74
	black	0.78	0.70	0.72	0.74
	Hispanic	0.94	0.90	0.96	0.93
	white	0.77	0.67	0.66	0.71
albert-base-v2	Asian	0.88	0.58	0.80	0.77
	Black	0.84	0.55	0.70	0.71
	black	0.84	0.55	0.70	0.71
	Hispanic	0.95	0.83	0.94	0.91
	white	0.88	0.56	0.71	0.73
roberta-base	Asian	0.87	0.72	0.90	0.84
	Black	0.90	0.86	0.92	0.89
	black	0.81	0.62	0.71	0.72
	Hispanic	0.94	0.93	0.97	0.95
	white	0.78	0.68	0.72	0.74

Table 8: Full results for experiment 1 — *Race* and masked models. These results are in terms of rank-biserial correlation *r*. Higher values mean that the attribute is less expected by the model in that context.

model	phrase	a person	a woman	a man	total
	asexual	0.73	0.66	0.81	0.74
	allosexual	1.00	1.00	1.00	1.00
	bisexual	0.70	0.32	0.62	0.56
	cis	0.98	0.90	0.96	0.95
	cisgender	0.80	0.56	0.81	0.74
	gay	-	-	0.95	0.95
	heterosexual	0.94	0.84	0.89	0.90
Mistral-7B-v0.1	LGBTQ	0.93	0.98	0.99	0.97
	lesbian	-	0.89	-	0.89
	nonbinary	0.96	-	-	0.96
	pansexual	0.88	0.79	0.90	0.86
	queer	0.91	0.86	0.94	0.91
	straight	1.00	0.99	0.98	0.99
	transgender	0.80	0.62	0.92	0.80
	trans	0.99	0.97	1.00	0.99
	asexual	0.95	0.95	0.97	0.96
	allosexual	0.78	0.74	0.83	0.78
	bisexual	1.00	0.99	1.00	1.00
	cis	1.00	1.00	1.00	1.00
	cisgender	0.97	0.92	0.97	0.96
	gay	-	-	0.94	0.94
	heterosexual	1.00	0.99	0.99	1.00
Meta-Llama-3.1-8B	LGBTQ	1.00	1.00	1.00	1.00
	lesbian	-	1.00	-	1.00
	nonbinary	0.95	-	-	0.95
	pansexual	0.98	0.97	0.99	0.98
	queer	1.00	0.99	1.00	1.00
	straight	1.00	0.99	0.98	0.99
	transgender	0.99	0.96	0.99	0.98
	trans	1.00	0.96	1.00	0.99
	asexual	1.00	1.00	1.00	1.00
	allosexual	1.00	1.00	1.00	1.00
	bisexual	1.00	0.99	0.99	0.99
	cis	1.00	0.99	1.00	1.00
	cisgender	0.96	0.88	0.95	0.93
	gay	-	-	0.94	0.94
	heterosexual	0.99	0.99	0.99	0.99
gemma-7b	LGBTQ	1.00	1.00	1.00	1.00
	lesbian	-	1.00	-	1.00
	nonbinary	0.94	-	-	0.94
	pansexual	0.97	0.96	0.98	0.97
	queer	0.99	0.98	0.99	0.99
	straight	1.00	1.00	0.98	0.99
	transgender	0.97	0.94	0.99	0.97
	trans	0.98	0.96	0.99	0.98

Table 9: Full results for experiment 2 — *Queerness* and generative models. These results are in terms of rank-biserial correlation *r*. Higher values mean that the attribute is less expected by the model in that context.