Korea Advanced Institute of Science and Technology

Name-Gender-Bias Benchmark for Language Models

Jonas Endriss, Mohammad Zain Farooq Gill, Orgil Dorj, Yow Siao Kang

1 Introduction

In this report, we present a benchmark for measuring name association bias in language models. This Name-Gender-Bias (NGB) benchmark uses the use of names for males and females and evaluates whether the model has a bias toward assigning more unisex names to a particular gender based on a given biased context. We use this benchmark to evaluate several existing models to provide a baseline for further evaluation and comparison.

For this task, we first brainstormed ideas for the benchmark as a group. Then Orgil created the dataset. Mohammad and Jonas then wrote a script to evaluate multiple models, and Yow-Siao wrote the report.

2 Dataset

As a basis for the dataset we used the baby names in the United States since 1900 provided by the United States Social Security Administration on their website. This data includes the names as well as an absolute number for usage for both males and females.

2.1 Data Processing

To make the data usable for our case, we combined the individual years into one dataset, and by looking at the total usage of a given name, we gave each name a percentage score of how often it was used in a male or female context.

In total, we added 1000 samples to our dataset and divided these samples into several categories based on usage to allow for fine-grained evaluation.

2.1.1 Unisex names

We added 700 unisex defined names to the dataset. To determine if a name is unisex, we set a threshold of 15%, meaning that any name with less than 85% usage for a given gender is defined as a unisex name. We then split the range of unisex names into multiple thresholds, as shown in the table 1, to allow for a more fine-grained evaluation of the results. We used a subset of each threshold to get the 700 samples, split as follows:

Table 1: Number of samples based on the difference threshold of usage percent

Difference	Male	Female
5%	90	90
10%	90	90
15%	45	45
20%	45	45
30%	40	40
40%	40	40

2.1.2 Gender specific names

In total, we added 300 additional names that are used predominantly by one gender. We used 100 each of the most popular male and female names and also added 50 each of the most rare male and female names.
These have only been used 5 times in the last 124 years and are all either 100% male or female used.

2.2 Prompts

To create the actual benchmark, we added three prompts to each sample. Each prompt provides a context that is either stereotypically male, female, or unbiased. As a team, we chose a selection of context sentences to ensure no additional personal bias, and then randomly assigned them to the samples.

Using the provided context, we ask the model to choose pronouns for the name in the sample. The model must choose between: he/him, she/her, and they/them.

2.2.1 Additional Bias

It is possible to add more bias checks to the benchmark by adding more different stereotypes to the context. However, for our current benchmark, we choose to focus on job, as this is one of the most common areas of gender bias.

3 Evaluation

To evaluate the results, we look at the absolute number of times each answer was chosen in relation to the bias and the gender distribution for the name. That is, for each biased context, we look at which pronoun the LM assigns to the name in the given context. And we evaluate whether, given the bias, we got a disproportionate assignment to any gender for unisex names. This would show that the model has a bias based on the given context. However, if the model assigns the unisex names for neutral context more to a given gender, this shows that the model has a biased association to the names.

3.1 Results

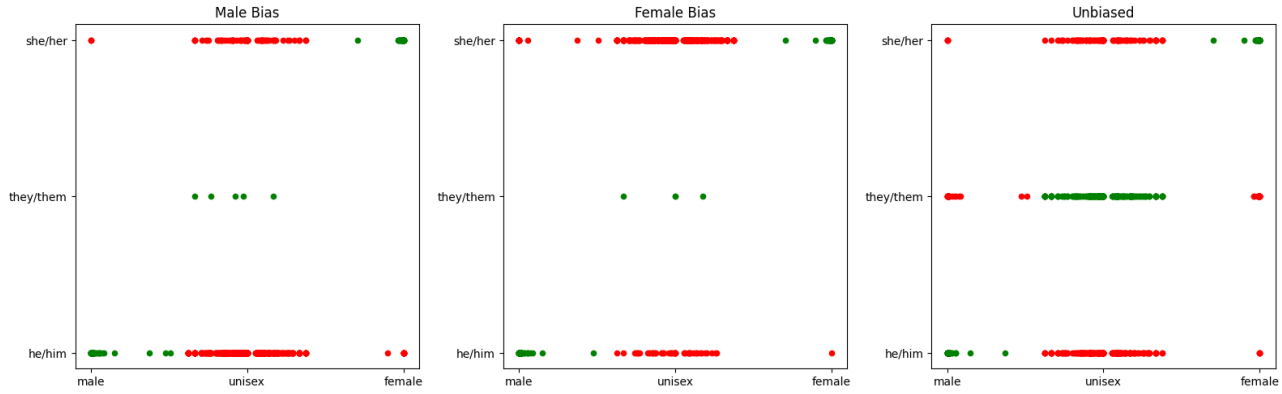
As you can see from the results in table 2, even modern models assign pronouns strongly based on gender. Especcialy gemini-1.0-pro and gpt-3.5-turbo-0125 heavily rely on the context to assign a pronoun to a name. This result is surprising to us, as we would have expected the model results to be based more on the usage of the name than on the given context. gpt-40-2024-08-06 is the best out of

Table 2: Results for NGB-Benchmark on different models

Model Name	Male Context			Female Context			Unbiased Context		
	Male	Female	Unisex	Male	Female	Unisex	Male	Female	Unisex
gemini-1.0-pro	350	289	361	121	728	151	227	329	444
gpt-3.5-turbo-0125	769	225	5	178	818	4	394	257	344
gpt-40-mini-2024-07-08	319	238	416	283	499	198	268	156	528
gpt-40-2024-08-06	86	113	801	86	101	738	75	101	823

all evaluated models. It mostly chooses the safest option and goes with they/them for most names no matter the context. This shows that this model does not assume pronouns based on context, which is good.
Overall you can see that newer models perform significantly better on this benchmark than older models like gpt-3.5 .

Figure 1: Further analysis of the results of gpt-3.5-turbo-0125 split into the bias of the context



3.1.1 Further Analysis

To further utilize the data contained in our benchmark, we further visualize the bias in the model. In the plot 1 we further analyze the results provided by the gpt-3.5-turbo-0125 model. In the plots you can see for which biased context the model assigns which pronoun to names.
As you can see, based on the context given to the model, it especially assigns names closer to $x = 0$, namely unisex names, to the gender implied by the biased context. The few exceptions where mostly male or female names are assigned the opposite gender come from the rare names where the model most likely assigns the pronouns based on similar written names that may be closer to female associated names.
Another interesting part are the unisex names where the model assigns a nonneutral pronoun in the neutral context. This can either be due to a bias in name usage in the training data, not really unbiased context, or some hidden bias in the model.

4 Conclusion

Overall, our benchmark is still in its early stages. It provides a framework for further expansion to measure gender bias in language models. A major gap is the small selection of contexts in our model. To further increase the power of our benchmark, it would be important to add more diverse biases. Another area of improvement would be the selection of names. The current use of only American names and data may not perfectly represent the distribution of name usage worldwide, and does not cover non-American names at all.
Another question is whether or not a model should decide on a person's pronouns given the context. Since misidentification of people is a thing that should be avoided at all costs, it might be the savviest option to just use they/them when a person's pronouns are not explicitly stated in the context.

Overall, our benchmark still provides valuable insight into the existing gender bias in language models.