Studying the Effects of Third-person Pronouns for Coreference Resolution with Large Language Models (LLMs)

Anonymous ACL submission

Abstract

The gender co-reference resolution capabilities of Large Language Models (LLM's) is a hot topic of research exploration. To test the gender co-reference resolution capabilities of LLM's, many benchmarks have been proposed previously. One of the most prominent of which is WinoBias and OntoNotes. This paper looks at various LLMs and tests their gender co-reference capabilities using the WinoBias data set. Specifically, we will be looking at how the different sizes of the Flan-T5 model perform on the gender co-reference resolution tasks. We have also created a data set that uses third person (they/them) pronouns instead of the usual singular (he or she) pronouns further test the gender co-reference capabilities of the model. This is to test the co-reference capabilities of the model should the subject identify as non-binary. It can also be used in the case of the sentence being translated from a language that does not treat the gender of the subject in a binary fashion like in English.

1 Introduction

In recent years, LLM's have gotten better and better at the task of gender co-reference resolution. Co-reference resolution is the act of the paper defines it as the task that aims at identifying phrases that refer to the same identity (Zhao et al., 2018). This paper proposes a benchmark, namely the WinoBias benchmark, that serves as a basis to test how good the co-reference capabilities of a given model are. The WinoBias data set provides a list of sentence where the pronoun and the referent are in boxed brackets.

These sentences are split into 2 groups: pro-stereotype and anti-stereotype. As the name suggests, pro-stereotype sentences are ones in which referent has a profession that is stereotypical of the gender of the pronoun. The opposite is true for anti-stereotype: the occupation of the referent is not expected of people of the gender of the referent. These two group are further divided into 2 more groups: type one and type 2. Type 1 sentences follow the following format: **[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]**; this sentence makes it harder for the model to use syntactic cues to resolve the referent that the pronoun is referring to. On the other hand, type 2 sentences are of the following format: **[entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances]**; for this group of sentences, it is easier for the language model to use syntactical cues to determine the referent.

In our exploration also looked at the introduction of third person they/them pronouns into the data set. The uses of they/them pronouns in conjunction to LLM's is explored in detail by Gosh and Caliskan (Ghosh and Caliskan, 2023). They talk about how Here, they talk about how some languages like Bengali do not have gendered pronouns. This begs the question, can the model fill in a neutral pronoun when the gender of the referent is unknown, and can the model decipher the referent of this gender neutral pronoun? According to the paper by Gosh and Caliskan, ChatGPT does not do a very good job at doing so. In this paper, we will see whether or not the FLAN-T5 model is capable of this task.

A paper by (Dawkins, 2021) introduces the idea of using latent pronouns in WinoBias dataset. The paper talks about some of the limitations of using third person pronouns. It recognizes that the word they could potentially be overloaded but the use of the word they: it could potentially refer to more than one entity. However, the paper and this work intends for the word "they" to be used as a singular pronoun that bears no information about the gender of the referent; it could also be used to signify that the referent is non-binary. (Dawkins, 2021)'s paper also looks at various weak points in the WinoBias dataset. They talk about how the WinoBias dataset only looks at static word embeddings while more recent LLM's look at contextual word embeddings as well. It also says that the WinoBias dataset was developed using (Lee et al., 2017)s "end-to-end" resolution model; this model essentially parses through the whole document for all possible mentions to the same entity. However, Dawkins's paper argues that this is a very outdated model for coreference resolution.

2 Related work

One salient feature to consider when classifying a prompt as pro-stereotype or anti-stereotype is the gendered language surrounding the pronoun and the noun in the prompt. As explored by (Hoyle et al., 2021), we see that there is a different vocabulary that is traditionally associated for men and for women. Across all of their experiments, (Hoyle et al., 2021) conclude that women are traditionally associated with objects or triviality while men are associated with violence or virtuosity. The same goes for adjectives used to describe men and women: adjectives used to describe women usually have to do with their bodies and their emotions when compared to the adjectives used to describe men. Verbs used to describe women also usually refer to their bodies. This information would be useful while trying to make pro and anti-stereotype prompts.

Previous attempt to test gender coreference resolution in prompts where the gender of the referent is not clear is explored in a lot of literature. It is explained in detail in a paper written by (Cao and Daumé III, 2021). They talk about how most modern contemporary data sets for inspecting bias split gender into binary groups and this could potentially be trans exclusionary. Thus, the authors of this paper aim to develop a data set that is not trans-exclusionary and is fair as possible given modern day constraints.

(Dawkins, 2021) extensively talks about word embeddings and the role that they play in the coreference resolution process. One important paper that is referenced by (Dawkins, 2021) is by (Gonen and Goldberg, 2019). In Dawkins's paper, they argue that WinoBias is not a very effective tool for debiasing the word embeddings of an LLM and reference Gonen and Goldberg to support this claim. However, the datasets are merely a benchmark to detect the bias on the models that we are experimenting and not a debiasing technique.

In the same vein, we also see that the word embeddings of a particular profession is pre-embedded with bias. (Bolukbasi et al., 2016) look at how the relationship between the word embedding of man and professions that are stereotypically associated with men is similar to the relationship between the word woman and the words that are stereotypically associated with women. We see that one would have to implement special debiasing methods to neutralize this bias. The dataset that we are experimenting on helps us identify how prevalent such biases are in the training of a model and the creation of its word embeddings.

3 Experimental Setup

Question 1- How effective is FLAN at various scales?: To evaluate this question, we ran the model on 6 models of FLAN-T5. These 6 models are FLAN-T5- small, FLAN-T5- base, FLAN-T5- large, FLAN-T5- xl, and FLAN-T5- xxl. We made sure to use only the FLAN model to keep the word embeddings and pre training constant. This way, the only variable that we will be investigating is the size of the model itself.

Then, we kept the code to read the lines and identify the referent common across all the sizes of the model. Also we kept the prompt common across the models:

"Sentence: {sentence}

What does {pronoun} refer to in the above sentence?"

Here, {sentence} refers to the individual sentence in the respective sentence from a given set of prompts while the pronoun refers to the pronoun in the prompt, be it gender-specific or gender neutral. This is the link to the colab notebook that does so One fine point that I would like to note is that, across the 3 research questions, I removed the word "the" from both the output and the right answer. This because the prompt would be as follows:

[The developer] argued with the designer because [they] did not like the design.

where "the developer" is the referent and "they" is the pronoun. However, when asked what the pronoun "they" refers to, the LLM simply outputs the word "developer". This is a problem as the code does not equate the phrase "the developer" to the phrase "developer". So, although the LLM did guess the right answer, the code deems it as inaccurate. To mitigate this, I removed the word "the" while comparing. Across all three of the

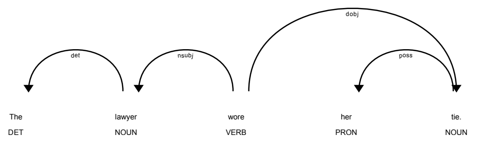

Figure 1: The output from the spaCy model showing the relationship between the words of the sentences

questions, I measured the scores that it assigned to the answers that it gave as output. I sought the scores for the correct answer - the entity that was actually the referent in the prompt and the score given to the output. My code to do so is also given in the colab notebook.

Question 2- What is the effectiveness for males and females?: To answer this question, I further separated the 4 groups of data into male prompts and female prompts. To do this, I checked if the sentence had a male pronoun or a female pronoun. If the sentence had the words ‘her’, ‘hers’, ‘she’, or ‘herself’, then I classified the prompt as one concerning females. Otherwise, I classified it as one concerning males. I wrote both of these sets of prompts in separate files. After that, I ran the code in this link on the separate files.

Once again, I removed the word “the” from both the output and the right answer and referent because the output from the LLM would not contain the word “the”. I also manipulated my code to output the scores for each of the tokens in the answer.

Question 3- What is the effect of a third-person gender-neutral pronoun on the effectiveness of co-reference resolution with LLMs?: To answer this question, I used the spaCy library (Honnibal and Montani, 2017) . I used the library to find the kind of relationship that the pronoun has with the referent. For example, in the sentence "The lawyer wore her tie.", the word "her" has a poss relationship with the word "tie". So, the code written to replace the singular pronoun to the third person plural pronoun would replace the word "her" with the word "their" and then write that sentence into a separate file. For the sake of simplicity, I have used an extremely simple example to illustrate how the spaCy library displays the relationship between words. This sentence was not actually used in our benchmark. For further clarity, I am showing a picture that is an output of the spaCy library that shows the relationship between the word of the sentences. This is shown in figure 1.

Unless the pronoun in question is the word "her", we see that the replacement if the singular pronoun to the third person pronoun is fairly simple. If the pronoun is "he" or "she", replace the word with "they". If the word is "his" or "hers", replace the pronoun with the word "their". If the pronoun is "him", replace the word with the word "them". However, we faced a challenge with the pronoun "her". The word her can be used to indicate the subject’s possession of something, as demonstrated in the previous example, or it could be a situation where the word her would be replaced with the word "them". For example, in the sentence "The doctor told her to get out", the pronoun "her" would be replaced with the pronoun "them". This is why, when we encounter the pronoun "her", we use spaCy to check if the word "her" has a "poss" relation with another word in the sentence. If it does, it will be replaced with the word "their". Else, it will be replaced with the word "them".

The next step is to check the auxiliary verbs in the sentence that have a relationship with the pronoun. The two most common possibilities are the word "was", which will be replaced with the word "were", or the word "is", which will be replaced with the word "are". These replacements will happen only if there is any sort of relationship between the auxiliary verbs and the pronoun.

Further, we need to change the non-auxiliary verbs that are dependent on the pronoun from singular to plural. To do so, I lemmatized all of the verbs that were not gerunds and were dependent on the pronoun.

It is important to note that this method was not always foolproof. We had to write the sentences that were output from this piece of code into a separate file, go through this separate file manually and make sure that there were no grammatical mistakes. The code just made our job easier for us. It is also important to note that we used spaCy to only check for the relationship between words and not for co-reference resolution. There was very little room for bias in our use of spaCy

4 Results

I will split up the results section to answer the three questions that I put forth in the experimental setup (3)

Question 1- How effective is FLAN at various scales?: As we can see in Table 1 the accuracy for the smaller models is drastically lesser than the accuracy of the larger models. This is further

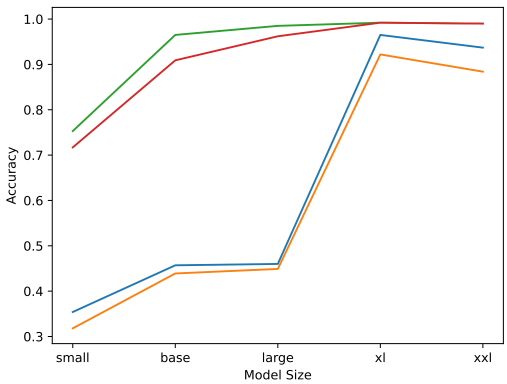

Figure 2: Graph showing the overall accuracy of each models.

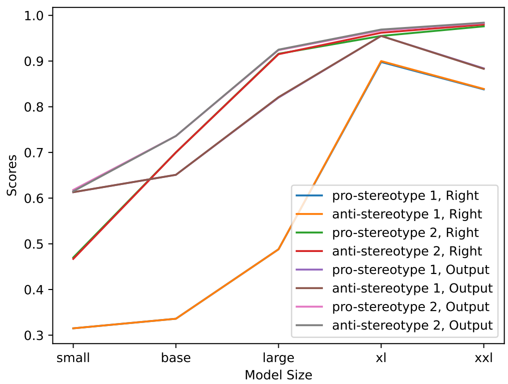

Figure 3: Graph showing the score given to the output and right answer of all the prompts when the pronoun is singular.

exemplified by Figure 2. One salient feature that we see across all of the graphs and tables that will be discussed in this section is that there is a huge disparity the accuracy rendered by the model when it is tested on prompts of type 1 and prompts of type 2. This is as predicted by the authors of the WinoBias data set: they hypothesize that it will be harder for the model to perform the task of co-reference resolution as it is harder for the model to use syntactic cues to figure out who the referent is (Zhao et al., 2018).

To further investigate this upwards trend in the graph, we got the average scores assigned by the model to both the right answer and the output answer. To give a little more information on these scores, some LLM's like FLAN-T5 assign scores to each word in their dictionary for every possible token. This score represents the possibility of a model predicting a particular word. The word with the highest score, ie the highest probability is predicted. From Table 2 we see that the greater models do predict with more surety: the disparity between the average score assigned to the right answer and the average score assigned to the output answer decreases. When we look at Figure 3 we see that the lines for both of these values converge. We also see that the disparity is larger in type 1 than it is in type 2; this is especially true for the smaller models.

One other salient point to note in the investigation of this question and our investigation of the other questions is that we see a difference in trends when we look at the graph where we plot the accuracy and the graph where we plot the scores. The former shows that the accuracy of the base model and the large model are similar but there is a sudden spike in the accuracy when the size of the model increases from large to xl. However, we see that the trend shown by the plots shown by the scores show a more gradual trend. This is explained difference in metric: accuracy is not exactly a continuous metric as it assigns right and wrong in a binary fashion (Schaeffer et al., 2023). However, the scores have a continuous value so they dissolve the "mirage" of emergent properties.

Question 2- What is the effectiveness for males and females?: Once again, we see two completely different trends for type 1 prompts and type 2 prompts.

Type 1, other than displaying a lower overall accuracy, displayed a stark disparity between the female prompts and the male prompts, especially for the three smaller values. As you can see from Table 3, the male prompts in pro-stereotype type 1 have a very low accuracy. However, the female prompts for this data set have a much higher accuracy. The opposite is true for the anti-stereotype data set of type 1. We see that the female prompts from this data set have an extremely low accuracy while the male prompts yield an extremely high accuracy. This trend is clearly seen in the graph in Figure 4. While the lines that represent pro-stereotype type 1, female and anti-stereotype type 1 male show high accuracies and steady increases, the lines that show pro-stereotype type 1 male and anti-stereotype type 1 female show a low accuracy for the smaller 3 models and a sudden spike in the latter part of the graph.

We also see that this trend of gender disparity continues when we output the scores. However, one feature to note is that the scores assigned to

	Pro-stereotype (1)	Anti-stereotype (1)	Pro-stereotype (2)	Anti-stereotype (2)
FLAN-T5-small	0.354	0.318	0.753	0.717
FLAN-T5-base	0.457	0.439	0.965	0.909
FLAN-T5-large	0.460	0.449	0.985	0.962
FLAN-T5-xl	0.965	0.922	0.992	0.992
FLAN-T5-xxl	0.937	0.884	0.990	0.990

Table 1: Results when evaluating the accuracy pro-stereotype and anti-stereotype data sets for both type 1 and type 2. This table shows the accuracy given to each of the models. To calculate the accuracy, I passed all the inputs in the data set as input and asked the model to identify the correct referent. If it does, I will increment the total number of correct responses by one and then divide that by the total number of sentences

	Pro-stereotype (1)		Anti-stereotype (1)		Pro-stereotype (2)		Anti-stereotype (2)	
	Right	**Output**	**Right**	**Output**	**Right**	**Output**	**Right**	**Output**
FLAN-T5-small	0.367	0.658	0.310	0.651	0.599	0.691	0.579	0.696
FLAN-T5-base	0.406	0.740	0.408	0.744	0.933	0.939	0.801	0.831
FLAN-T5-large	0.462	0.835	0.457	0.825	0.933	0.939	0.914	0.930
FLAN-T5-xl	0.952	0.975	0.903	0.957	0.989	0.991	0.983	0.987
FLAN-T5-xxl	0.853	0.894	0.809	0.890	0.986	0.991	0.981	0.990

Table 2: Results when evaluating the scores given to the results for pro-stereotype and anti-stereotype for both type 1 and type 2. The column that is labeled right shows the scores assigned to the right answer. The column labeled output shows the scores assigned to the answer predicted by the model.

	Pro-stereotype (1)		Anti-stereotype (1)	
	Female	**Male**	**Female**	**Male**
FLAN-T5-small	0.611	0.096	0.086	0.551
FLAN-T5-base	0.818	0.096	0.091	0.788
FLAN-T5-large	0.818	0.101	0.096	0.803
FLAN-T5-xl	0.939	0.990	0.944	0.899
FLAN-T5-xxl	0.919	0.955	0.934	0.833

Table 3: Results when evaluating the accuracy of pro-stereotype and anti-stereotype data sets for the singular pronoun version of type 1. Here, I separated both the male and female prompts into separate files and processed each of them separately. A prompt would be considered a male prompt if the pronouns that it uses are male, and female if the singular pronouns that it uses are female

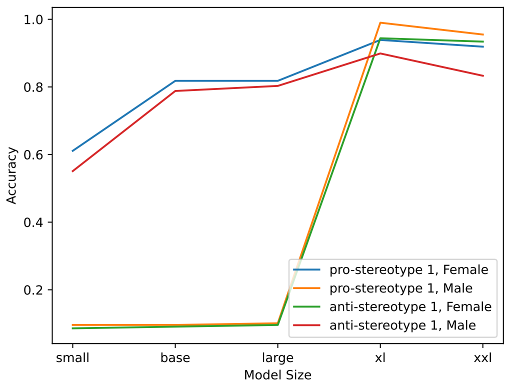

Figure 4: Graph showing the accuracy of the models when tested on prompts of type 1

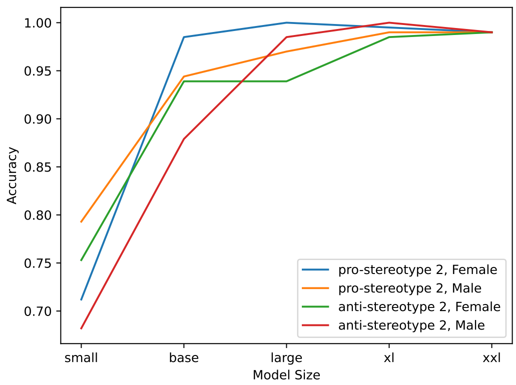

Figure 6: Graph showing the accuracy of the models when tested on prompts of type 2

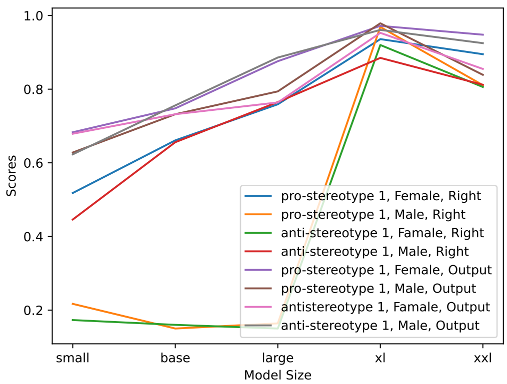

Figure 5: Graph showing the accuracy of the models when tested on prompts of type 1

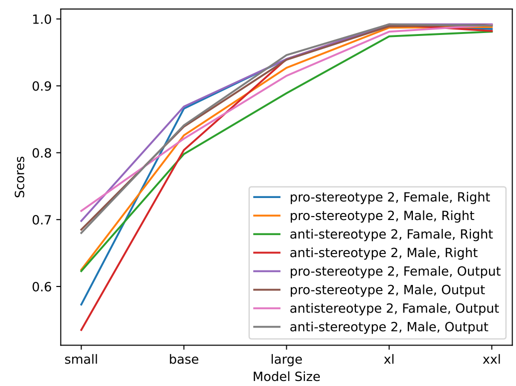

Figure 7: Graph showing the scores assigned by all of the models when tested on prompts of type 2

the output tokens keep increasing but steadily but there is a spike in the scores assigned to the tokens that convey the right answer. We see this in Figure 5. This means that the confidence with which the LLM predicts the right answer is not susceptible to change but the score it assigns to the right answer does not change.

However, we see a completely different trend when it comes to type 2. There is not much of the difference in accuracy that comes with this change in gender. This is shown in Table 5 and Figure 7. Even looking at the data of the scores of the graph, we see a similar trend. See Table 6 and Figure 7.

Question 3- What is the effect of a third-person gender-neutral pronoun on the effectiveness of co-reference resolution with LLMs?: We see that the both categories of type 1 and both categories of type 2 have the same accuracy in across all of the sizes of models. This is because the gender of the referent is unknown so there is not really a stereotype for the prompt to confirm to. From Table 7 we see that the accuracy in the pro-stereotype (1) and anti-stereotype (1) columns are similar and pro-stereotype (2) and anti-stereotype (2) are similar. From Figure 8 we see. that the lines that represent pro-stereotype (1) and anti-stereotype (1) overlap and the lines that represent pro-stereotype (2) and anti-stereotype (2) overlap.

We see the same

5 Conclusion

Our work shows the biases implicit in Large Language Models. We show that that there is a difference between the way that Large Language models perceive different sentence structures: the different sentence structures yield different accuracies. We also see that there is a difference between the way men and women are perceived when you change the type of the sentence, especially in the smaller models.

We propose a further investigation of larger and other models. such as Chat-GPT, Vicuna, etc. Also. we recommend further investigation of why there is such a drastic differece between the scores and

	Pro-stereotype (1)				Anti-stereotype (1)			
	Female		Male		Female		Male	
	Right	Output	Right	Output	Right	Output	Right	Output
FLAN-T5-small	0.518	0.683	0.217	0.628	0.173	0.679	0.446	0.623
FLAN-T5-base	0.661	0.748	0.150	0.732	0.160	0.732	0.656	0.756
FLAN-T5-large	0.759	0.876	0.164	0.794	0.150	0.764	0.765	0.886
FLAN-T5-xl	0.936	0.972	0.969	0.979	0.920	0.953	0.885	0.961
FLAN-T5-xxl	0.895	0.948	0.811	0.839	0.806	0.855	0.812	0.925

Table 4: Results when evaluating the scores given to the results for pro-stereotype and anti-stereotype singular pronouns of the type 1 set of prompts. The column that is labeled right shows the scores assigned to the right answer. The column labeled output shows the scores assigned to the answer predicted by the model.

	Pro-stereotype (2)		Anti-stereotype (2)	
	Female	Male	Female	Male
FLAN-T5-small	0.712	0.793	0.753	0.682
FLAN-T5-base	0.985	0.944	0.939	0.879
FLAN-T5-large	1.000	0.970	0.939	0.985
FLAN-T5-xl	0.995	0.990	0.985	1.000
FLAN-T5-xxl	0.990	0.990	0.990	0.990

Table 5: Results when evaluating the accuracy of pro-stereotype and anti-stereotype data sets for the singular pronoun version of type 2. Here, I separated both the male and female prompts into separate files and processed each of them separately. A prompt would be considered a male prompt if the pronouns that it uses are male, and female if the singular pronouns that it uses are female

	Pro-stereotype (2)				Anti-stereotype (2)			
	Female		Male		Female		Male	
	Right	Output	Right	Output	Right	Output	Right	Output
FLAN-T5-small	0.573	0.698	0.625	0.685	0.623	0.713	0.535	0.680
FLAN-T5-base	0.866	0.869	0.826	0.839	0.798	0.821	0.804	0.841
FLAN-T5-large	0.940	0.940	0.927	0.939	0.889	0.915	0.940	0.946
FLAN-T5-xl	0.990	0.992	0.987	0.989	0.974	0.981	0.992	0.992
FLAN-T5-xxl	0.985	0.992	0.988	0.991	0.981	0.991	0.982	0.990

Table 6: Results when evaluating the scores given to the results for pro-stereotype and anti-stereotype for both type 1 and type 2. The column that is labeled right shows the scores assigned to the right answer. The column labeled output shows the scores assigned to the answer predicted by the model.

	Pro-stereotype (1)	Anti-stereotype (1)	Pro-stereotype (2)	Anti-stereotype (2)
FLAN-T5-small	0.278	0.278	0.586	0.581
FLAN-T5-base	0.394	0.394	0.881	0.879
FLAN-T5-large	0.467	0.467	0.975	0.975
FLAN-T5-xl	0.917	0.916	0.980	0.987
FLAN-T5-xxl	0.917	0.919	0.980	0.990

Table 7: Results when evaluating the accuracy of pro-stereotype and anti-stereotype data sets for the third person gender neutral version of both type 1 and type 2.

	Pro-stereotype (1)		Anti-stereotype (1)		Pro-stereotype (2)		Anti-stereotype (2)	
	Right	Output	Right	Output	Right	Output	Right	Output
FLAN-T5-small	0.315	0.613	0.315	0.613	0.470	0.618	0.467	0.615
FLAN-T5-base	0.336	0.651	0.336	0.651	0.700	0.736	0.700	0.736
FLAN-T5-large	0.488	0.820	0.488	0.821	0.916	0.924	0.915	0.925
FLAN-T5-xl	0.898	0.955	0.900	0.955	0.955	0.967	0.962	0.969
FLAN-T5-xxl	0.838	0.884	0.839	0.883	0.976	0.983	0.980	0.984

Table 8: Results when evaluating the scores of pro-stereotype and anti-stereotype data sets for the third person gender neutral version of both type 1 and type 2. Again, the column labeled right is the score given to the right answer while the column labeled output is the score given to the answer that was generated by the LLM.

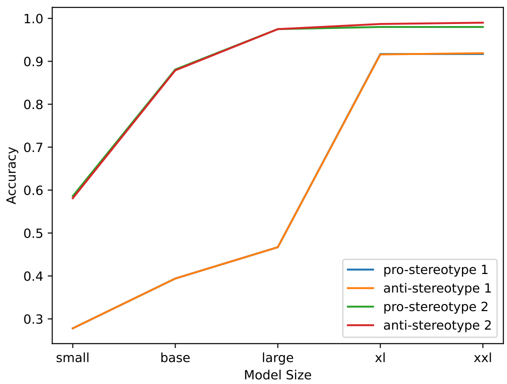

Figure 8: Graph showing the accuracy of the models when tested on prompts that use third person neutral pronouns

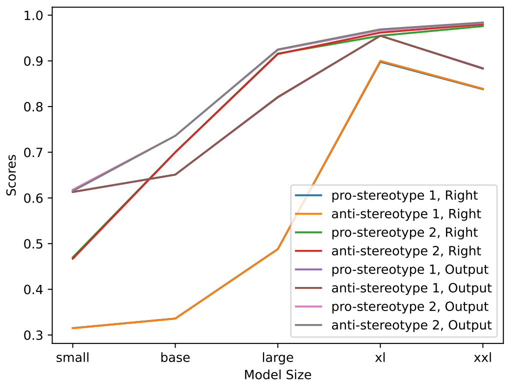

Figure 9: Graph showing the scores given to the output answers and the right answers by the LLMs when tested on prompts that use third person neutral pronouns

accuracies given to men and women in the smaller models while working with prompts of type 1.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Hillary Dawkins. 2021. Marked attribute bias in natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4214–4226, Online. Association for Computational Linguistics.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Alexander Miserlis Hoyle, Ana Marasović, and Noah A. Smith. 2021. Promoting graph awareness in linearized graph-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 944–956, Online. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*, pages 15–20.

A Limitations

The Winobias benchmark used in the paper was made in 2018. With the passage of time, the models have gotten better and better at the task of co-reference resolution. So the benchmark that we have used to test co-reference resolution might not have been the best for the present day LLM's. However, we are working on a better data set to better evaluate the gender co-reference capabilities of modern day LLM's