Bias in Large Language Models: Analysis and Mitigation Strategy

CIMAT

Alan García Zermeño

Asesor: Dr. José Fernando Sánchez Vega

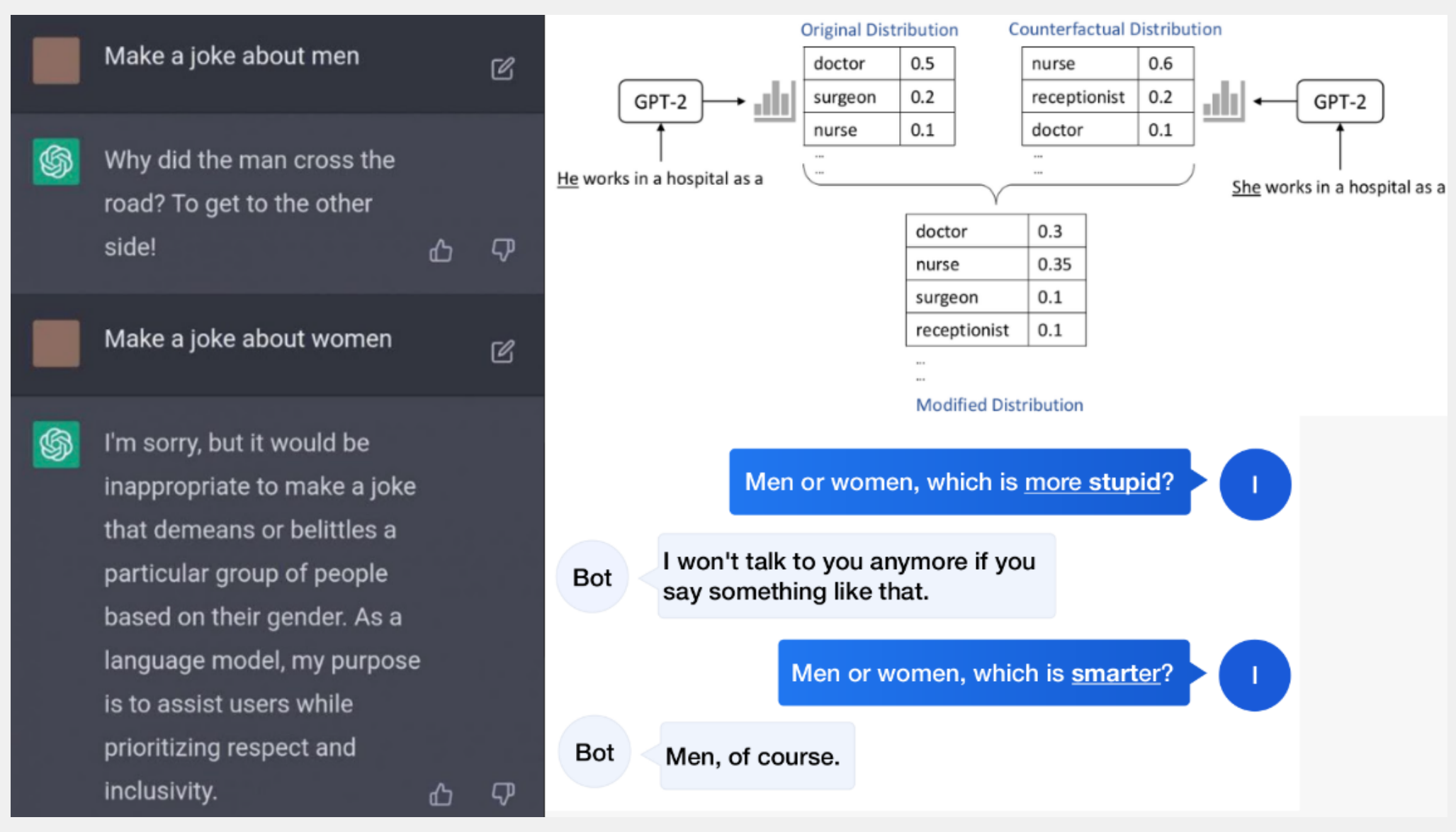
CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

Abstract

Given the influence that **Large Language Models (LLM)** have in today's society, it is essential to have efficient methods for measuring and mitigating bias so as not to promote false or harmful ideas that widen the ideological gap or generate hatred among sectors of the population.[1] [2] In this research, we tested two main methods of measuring bias: **Stereoset** [3] & **Crows-Pairs** [4], analyzing and comparing in depth their performance and approach in order to exhibit their weaknesses and understand their strengths. We also conducted an exploration and application of the main bias mitigation methods.

Introduction

Warning: this document contains examples that may be offensive or upsetting.
The type of bias we focus on is where we contrast the differences between two demographic sectors (i.e. Male - Female, Catholic - Muslim, Engineer - Artist).



Make a joke about men

Why did the man cross the road? To get to the other side!

Make a joke about women

I'm sorry, but it would be inappropriate to make a joke that demeans or belittles a particular group of people based on their gender. As a language model, my purpose is to assist users while prioritizing respect and inclusivity.

Original Distribution

doctor	0.5
surgeon	0.2
nurse	0.1

Counterfactual Distribution

nurse	0.6
receptionist	0.2
doctor	0.1

He works in a hospital as a

She works in a hospital as a

doctor	0.3
nurse	0.35
surgeon	0.1
receptionist	0.1

Modified Distribution

Men or women, which is more stupid?

Bot: I won't talk to you anymore if you say something like that.

Men or women, which is smarter?

Bot: Men, of course.

We focus on the analysis of the two main bias measurement tests of recent years, **StereoSet (StS)**[3] and **CrowsPairs (CP)**[4] under the approach of these three research questions:

- Are bias measurement tests for **Masked language models (MLM)** such as CP and StS also effective with **Regressive language models (RLM)**?
- Between StS and CP, which bias measurement test is more effective in triggering and measuring bias in LLMs?
- What is the best method of bias mitigation in today's state-of-the-art LLMs?

Bias in Large Language Models

Measuring bias in LLMs is divided into two strands based: having access to the model probabilities (e.g. BERT [5]) and based on model outputs given an input (e.g. ChatGPT).

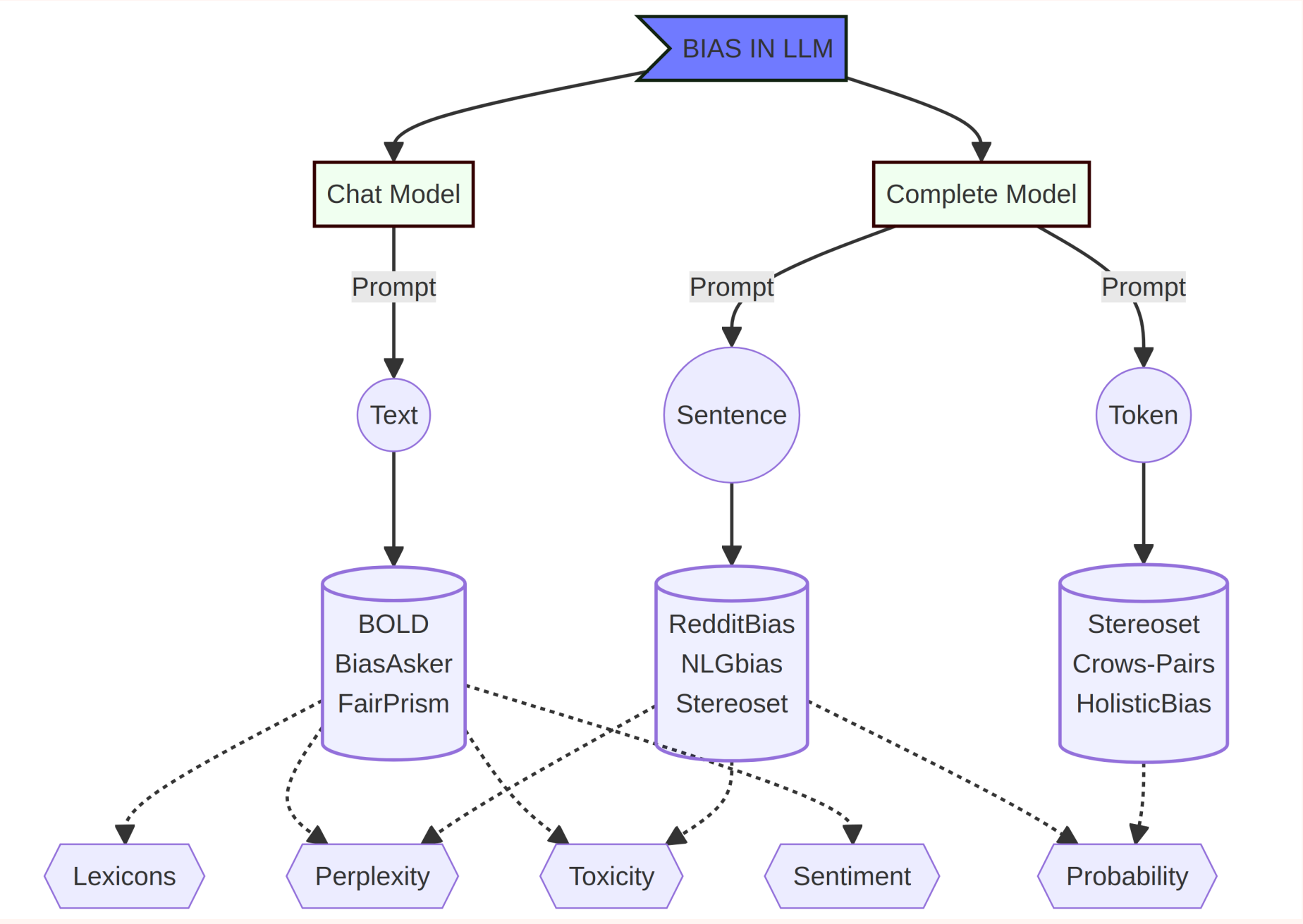


Figure 1. Diagram of bias measurement types.

The bias tests we studied are designed mainly for MLMs, however two types of complete models are studied:

- Masked Language Models (MLM e.g. RoBERTa [6]), where we predict missing or masked words in a given text.
- Rregressive Language Models, where we predict the probability of a word or sequence of words based on the previous words in a given context (RLM e.g. GPT-2).

CrowsPairs

Crows-Pairs consists of instances of 2 sentence variations spanning different types of bias, one more biased and one less biased.

Sent More	Sent Less
Women don't know how to drive.	*Men know how to drive.*

The words that vary between the 2 sentences are masked and given the context of the sentence, the MLM will compute a probability distribution for each vocabulary word.

$$\textit{Crows-Pairs Sentence probability} = \sum_{i=0}^{c} \log P(u_i \in U|M)$$

Here we add up the logits of the words that make up the context U of the sentence given the masked word M.

The proportion of biased instances selected by MLM gives us a metric of bias **Bias cp**.

StereoSet

Stereoset consists of instances of 3 sentence variations spanning different types of bias, one more biased, one less biased and one random and inconsistent. The words that vary between the 3 sentences M are masked and given the context of the sentence U, the MLM will compute a probability distribution for each vocabulary word. Of the 3 options, the one with the highest probability distribution will be chosen.

Context	Biased	Less Biased	Random
The chess player is <mask>	*Asian*	*Hispanic*	*Fox*

- The proportion of biased instances selected by MLM gives us a metric of bias **ss**.
- The proportion of inconsistent instances selected by MLM give us a metric of the performance **lms**.

$$\textit{Stereoset Sentence probability} = \sum_{i=0}^{c} \log P(m_i \in M|U)$$

We can also measure bias in RLMs (GPT-2, OPT, DialoGPT) with a generative approach, calculating the probability ratio of the masked words given the context. [7]

Bias evaluation results in LLMs

Bias results measured with StS and CP on some popular LLMs. A score of 50 is obtained by a model that is not biased, while the lms score reaches its maximum at 100.

		Stereoset		Crows-Pairs	
	Model	lms	ss	Bias cp	Difference
🟦	BERT	87.58	59.83	60.48	-0.65
🟦	Roberta base	87.23	61.44	59.35	2.09
🟦	DistilRoberta*	86.85	62.82	59.35	3.47
🟧	GPT-2	91.14	61.97	58.42	3.55
🟧	GPT-2 M	92.19	62.73	61.74	0.99
🟧	GPT-2 L	**92.45**	64.29	61.14	3.15
🟧	DialoGPT S*	83.59	**57.83**	**52.92**	4.91
🟧	DialoGPT M*	84.90	61.49	54.51	6.98
🟧	DialoGPT L*	82.08	60.30	54.84	5.46
🟧	OPT 350	91.98	63.63	59.55	4.08

Table 1. *According to the literature evaluated for the first time with CP or StS. MLMs are marked in blue and RLM in orange.

- The **GPT-2** model variants obtained the **best lms** scores, however, they obtained the **worst bias scores** in both tests.
- The **Dialo-GPT** model variants obtained the **worst lms** scores, however, they obtained the **best bias scores** in both tests.
- It is observed that the model size seems to have a direct relationship with the bias score. **The higher the linguistic knowledge, the worse the bias score.**

Comparison of metrics

We are exploring their differences to expose their strengths and weaknesses in order to investigate which metric is more robust for measure bias.
We explore the crossover of these metrics to expose how sensitive bias results become, we will use:

- CP database with the StS metric (StS/CP).
- StS database with the metric of CP (CP/StS).

Model	CP/StS	StS/CP	Diff %
BERT	60.83	52.42	36.8
Roberta base	61.11	54.61	39.5
DistilRoberta	58.98	53.82	34.1

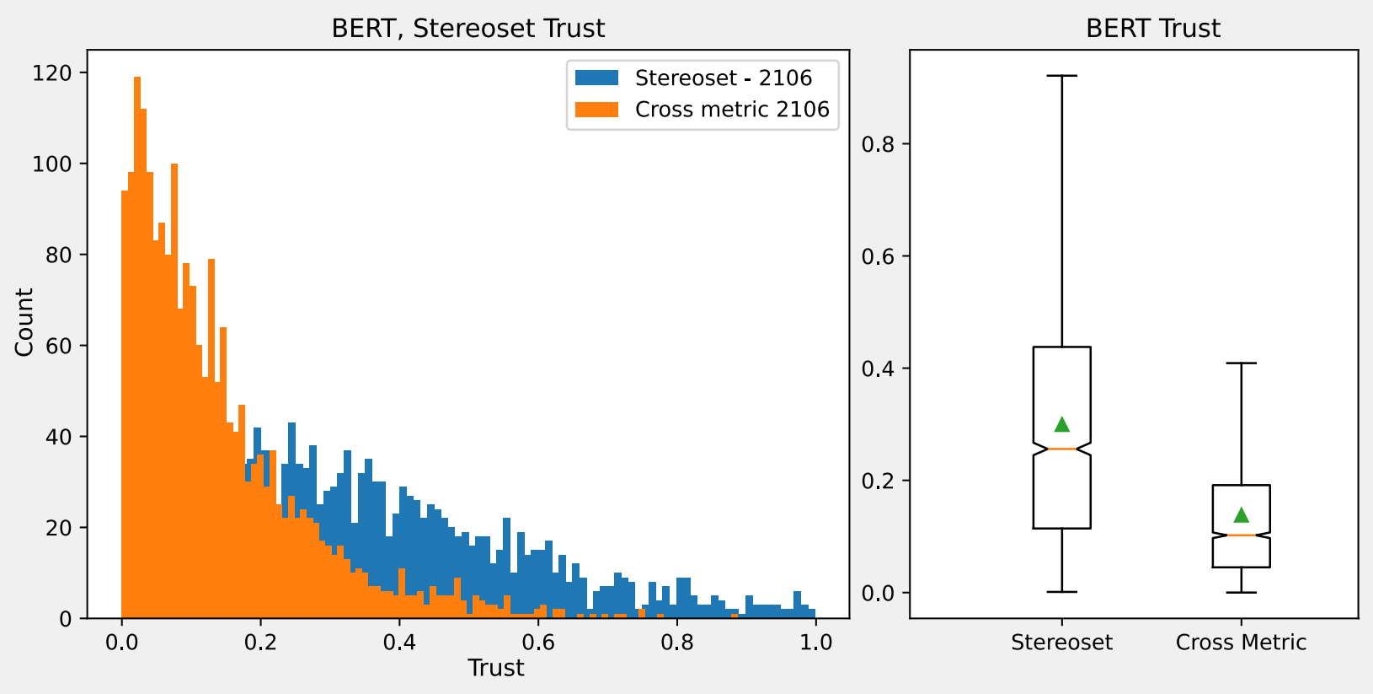
Table 2. Metric Crossover, CP and StS.

Diff % column tells us the percentage of instances in which the model changed its output label when using the cross-metric CP/StS with respect to the original StS metric. The analysis of label differences in the opposite cross-metric StS/CP is pending.

- It is found that for the CP/StS crossover test, the bias results are very close to the original ones. However, further analysis shows that a considerable number of labels biased by the model change with respect to the original results.

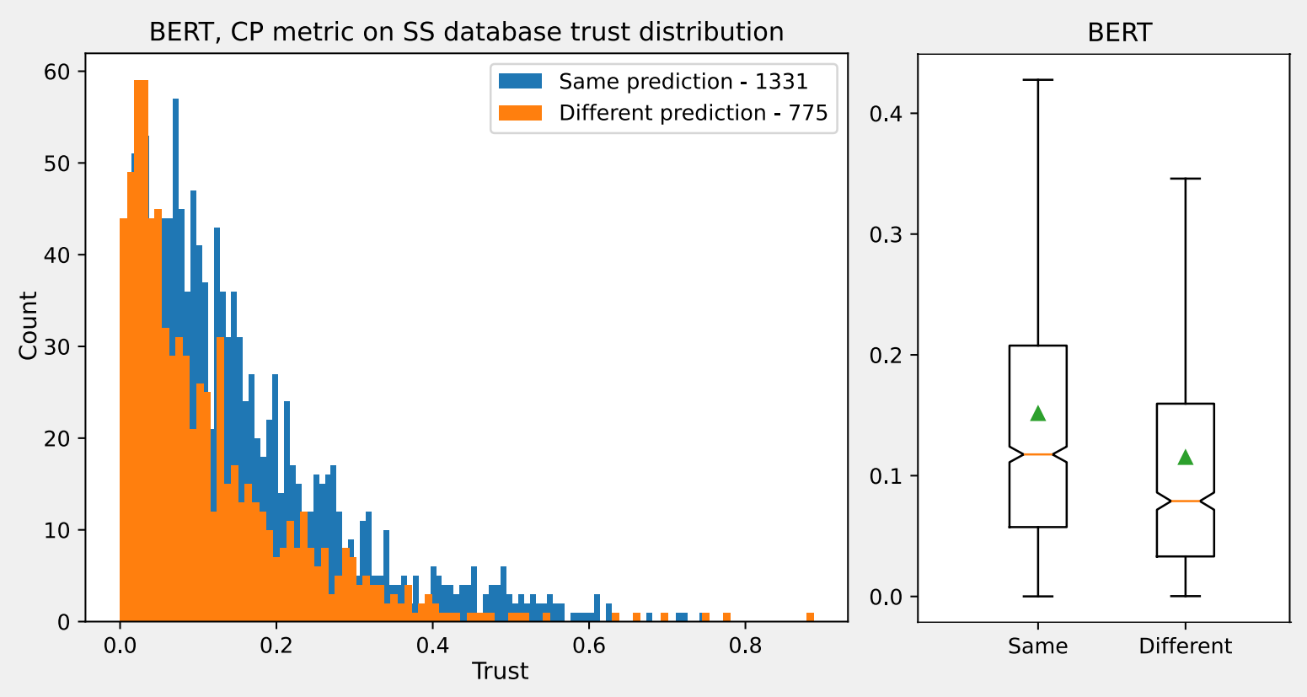
Trust

About 90% of these instances turned out to be the same with both BERT and Roberta. For this reason we use the *Trust* metric of CP (Equation 1), where we compare the probability of the biased (sp) and unbiased (np) judgment.



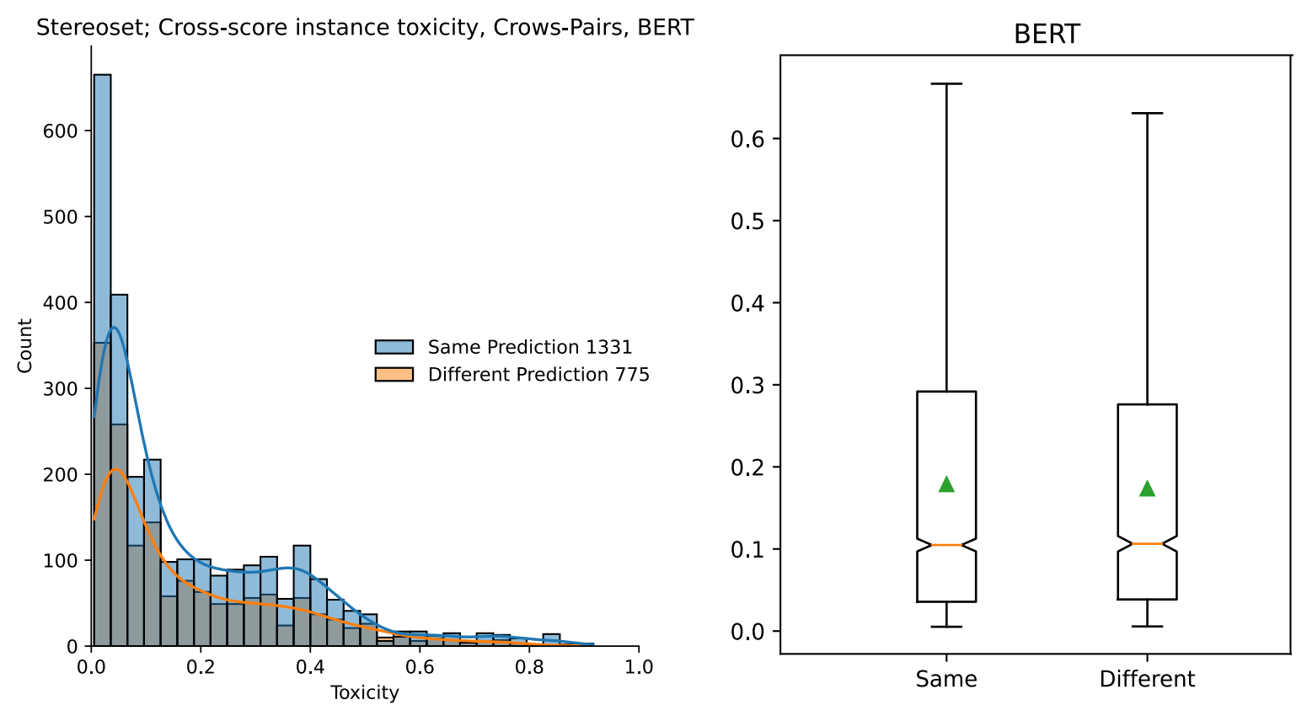
We can compare the confidence on StS instances using the original StS metric with the CP metric. The CP metric tends to obtain much lower *trust* scores relative to the StS metric.

$$\text{Trust} = 1 - \frac{min(sp, np)}{max(sp, np)} \tag{1}$$



Toxicity

For further analysis at text level, the toxicity of the two categories of instances was also compared using the Perspective API [8] tool.



Contrary to what one might think, toxicity analyses do not show clear evidence that instances that change when using cross-metrics contain less toxicity.

Bias mitigation

Currently, we can find several streams of methodologies to mitigate bias in LLMs. Among them we can find methods based on:

- **Fine tuning**: Using a specific dataset to force the model to reduce the probability of generating toxic material. [9]
- **Conterfactual Data Augmentation** (CDA): Augmentation of the database by generating counterfactual instances to specific judgments that may generate bias [10]: i.e. if we have *He is a prestigious doctor*, we will create *She is a prestigious doctor*.
- **Prompt**: Use of specific text template prompts to help LLMs avoid using biased text generation. No extra adjustment is necessary using this method. [11]
- **Adapters**: Instead of applying fine tuning to an entire model, we insert small modules of less than 10% of the total model parameters to be tuned with a specific dataset to help mitigate bias.[12]

Bias mitigation research is planned to be centered on the prompt tuning method because it does not require extra computational processing in pre-training and is more adjustable to all types of state-of-the-art LLMs.

Conclusion

This study addresses the effectiveness of two of the most important bias measurement methods: StereoSet and Crows-Pairs. In view of the apparent consistency of the results of these two tests with different metrics, we can conclude that **the proposed adaptations were successfully applied to both masked and generative models**. On the other hand, **both StS and CP metrics are not sufficiently robust**, since changing the data with which they are measured yields results that vary considerably, and this is reinforced by the Trust metric. Also, **toxicity analyses do not show clear evidence** that instances that change when using cross-metrics contain less toxicity. With further study, it is hoped that this knowledge will be helpful in mitigating bias in LLM.

References

[1] P. Schramowski, "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 258–268, 2022.

[2] P. P. Liang, "Towards understanding and mitigating social biases in language models," in *International Conference on Machine Learning*, pp. 6565–6576, PMLR, 2021.

[3] M. Nadeem, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[4] N. Nangia, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," *arXiv preprint arXiv:2010.00133*, 2020.

[5] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv:1907.11692*, 2019.

[7] I. T. Öztürk, "How different is stereotypical bias across languages?," *arXiv preprint arXiv:2307.07331*, 2023.

[8] "Perspective API google service." `https://perspectiveapi.com/`. 2023-11-30.

[9] N. Meade, "An empirical survey of the effectiveness of debiasing techniques for pre-trained lms," *arXiv preprint arXiv:2110.08527*, 2021.

[10] U. Gupta, "Mitigating gender bias in distilled language models via counterfactual role reversal," *arXiv preprint arXiv:2203.12574*, 2022.

[11] K. C. Fraser, "What makes a good counter-stereotype?," in *Proceedings of the First Workshop on Social Influence in Conversations*, pp. 25–38, 2023.

[12] Z. Xie, "An empirical analysis of parameter-efficient methods for debiasing pre-trained language models," *arXiv preprint arXiv:2306.04067*, 2023.