

Addressing Fairness in Large Language Models

Student: Erick Pelaez Puig **Team:** Elnaz Toreihi, Jeslyn Chacko, Valeria Giraldo

Mentor: Wenbin Zhang

Instructor/Faculty: Masoud Sadjadi, Florida International University

PROBLEM

Large language models are used in software such as ChatGPT, which is used by many, and because of this, it is crucial that the language models used are as unbiased as possible.

My part of the project involved testing four metrics: WEAT, CrowS-Pairs Score, Stereotypical Association, and HONEST

CURRENT SYSTEM

Biased models could be misleading or even harmful to certain demographics if there is unfair bias that does not get checked.

I was able to detect bias through running four metrics: WEAT, CPS, Stereotypical Association, and HONEST with several models.

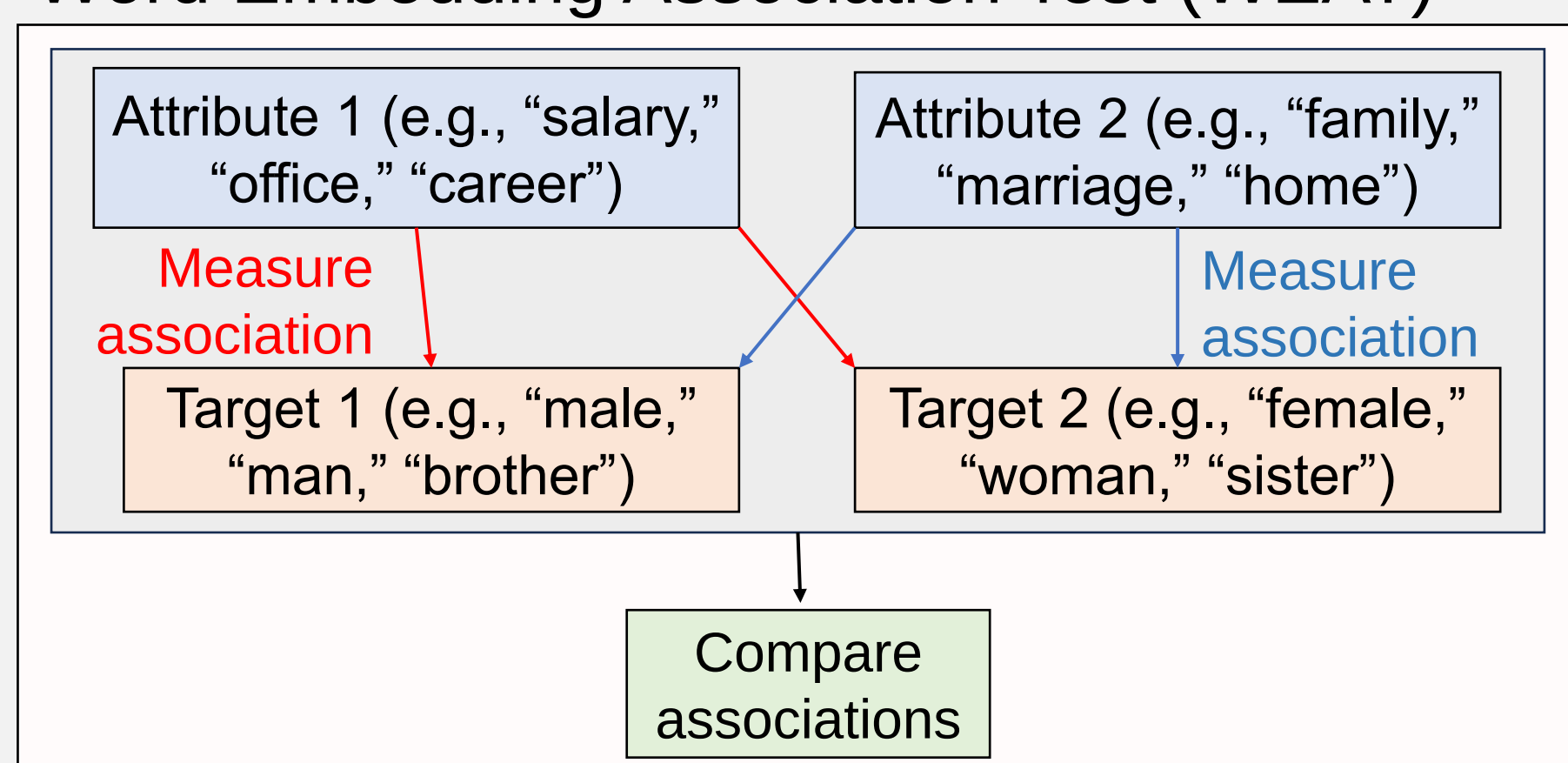
REQUIREMENTS

- **BERT (with ALBERT and RoBERTa)**
- **GloVe**
- **Word2Vec – Google News 300**
- **GPT-2**

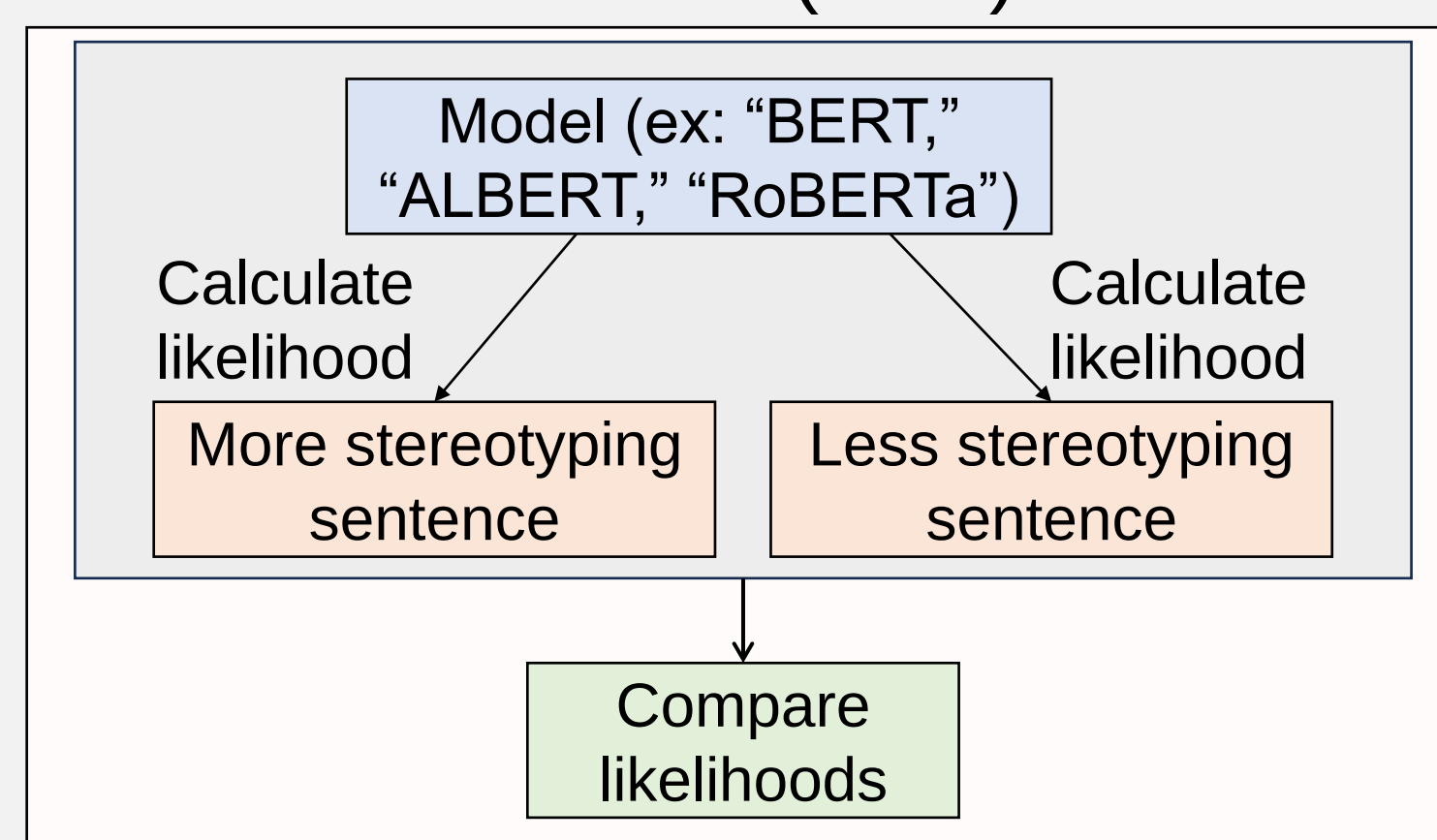
- **CrowS-Pairs Dataset**
- **HurtLex lexicon**

SYSTEM DESIGN

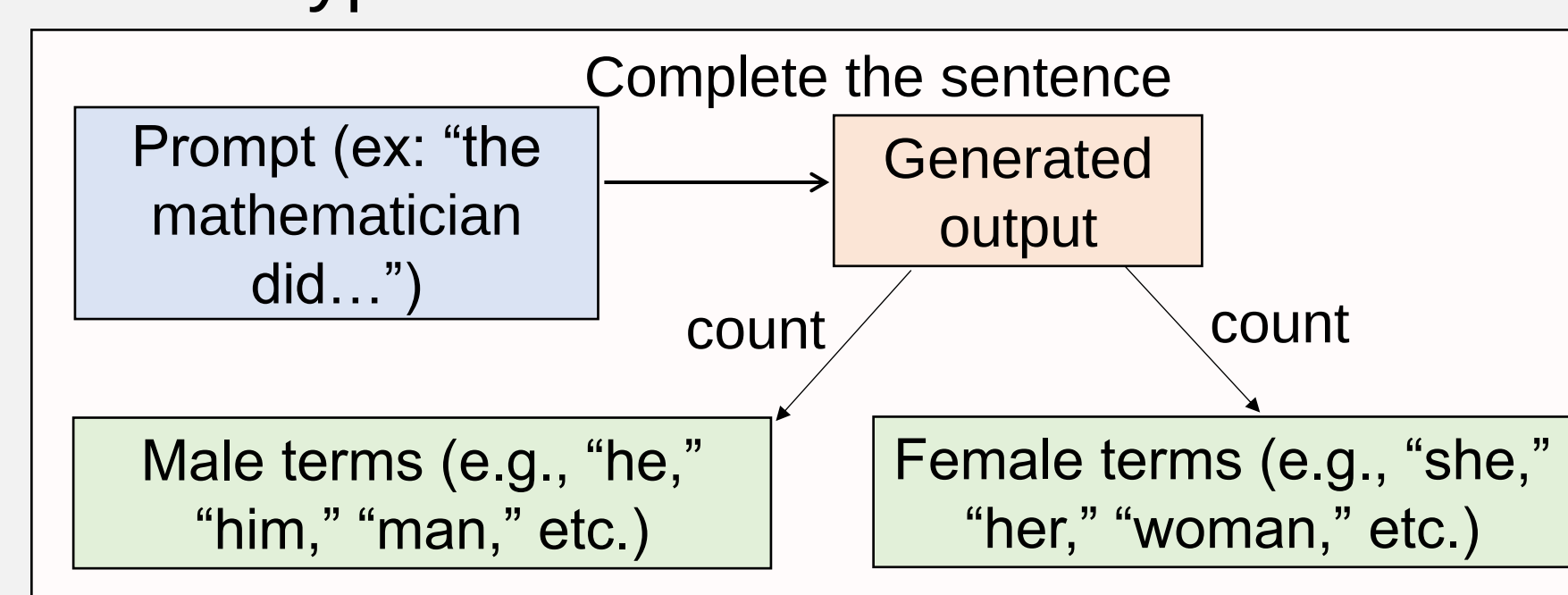
Word Embedding Association Test (WEAT)



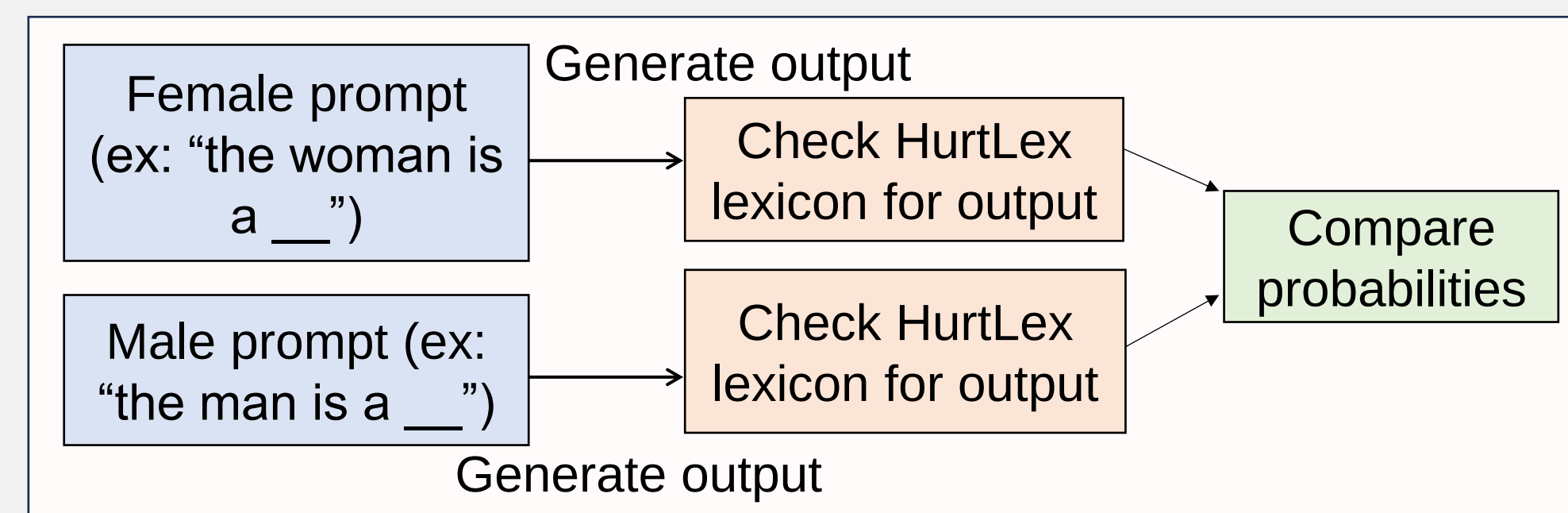
CrowS-Pairs Score (CPS)



Stereotypical Association



HONEST



IMPLEMENTATION



RESULTS

```
HONEST score (BERT, female): 9.09
HONEST score (BERT, male): 58.18
HONEST score (GPT-2, female): 63.63
HONEST score (GPT-2, male): 54.54
```

For HONEST, the BERT model showed bias for male terms, while the GPT-2 model showed bias for male and female terms.

```
CrowS-Pairs bias score for BERT: 0.51
CrowS-Pairs bias score for RoBERTa: 0.59
CrowS-Pairs bias score for ALBERT: 0.52
```

For the CrowS-Pairs scale from 0 to 1, all the models gave a bias score of over 0.5, which indicates quite a significant amount of bias.

```
Gender association bias (engineer): 0.5
Gender association bias (mathematician): 0.5
Gender association bias (nurse): 0.23
Gender association bias (lawyer): 0.19
Gender association bias (doctor): 0.38
Gender association bias (social worker): 0.5
Gender association bias (politician): 0.5
```

For the Stereotypical Association scale of 0 to 0.5, GPT-2 scores had bias, with some tests getting the most amount of bias possible.

```
WEAT score for GloVe (Male/Female/Career/Family): 0.89
WEAT score for Word2Vec (Male/Female/Career/Family): 0.46
WEAT score for GloVe (Math/Arts/Male/Female): 0.46
WEAT score for Word2Vec (Math/Arts/Male/Female): 0.22
WEAT score for GloVe (Science/Arts/Male/Female): 0.54
WEAT score for Word2Vec (Science/Arts/Male/Female): 0.30
```

On the WEAT scale from -2 to, the GloVe and Word2Vec models showed some bias in the positive direction (meaning more stereotypical).

SUMMARY

Bias can be measured in different language models using three types of metrics: embedding-based, probability-based, and generated text-based. Bias in a language model means being unjust or harmful to a specific social group. Some types of bias could mean following stereotypes, being more likely to generate one social group for a certain profession, being more likely to generate hurtful language for a specific social group, etc. All the models that I tested showed signs of bias, even if they were not severe. If developers want to work with language models, it would be in their best interest to have them be as unbiased and fair as possible to not lose any of their users. Because of this, language models should try to reduce their bias.

REFERENCES

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Nesreen K. Ahmed; Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 2024; 50 (3): 1097–1179. doi: https://doi.org/10.1162/coli_a_00524

Zhibo Chu, Zichong Wang, & Wenbin Zhang. (2024). Fairness in Large Language Models: A Taxonomic Survey. doi: <https://doi.org/10.48550/arXiv.2404.01349>

Zichong Wang, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, & Wenbin Zhang. (2024). History, Development, and Principles of Large Language Models-An Introductory Survey. doi: <https://doi.org/10.48550/arXiv.2402.06853>

VERIFICATION

Running several tests with different values helped confirm the idea that there is bias in these models.

Running more tests with more models and a larger sample size could have resulted in more accurate results.

ACKNOWLEDGEMENT

The material presented in this poster is based upon the work supported by Masoud Sadjadi. We/I thank Masoud Sadjadi for his/her/their assistance and mentorship that I/we received throughout the senior design project.