Word Embedding Bias in Large Language Models

Poomrapee Chuthamsatid - Supervised by Dr. Alex Thomo

University of Victoria, Department of Software Engineering, March 2025

BACKGROUND

The rapid development of large language models (LLMs) has expanded natural language processing (NLP) applications, from text generation to chatbots.

- Word embeddings are the core of these systems, converting words into numeric vectors based on their statistics usage patterns in text corpora..
- However, embeddings often reflect societal biases, reinforcing stereotypes [1].
- For instance, they may link professions like nurses to women and engineers to men.

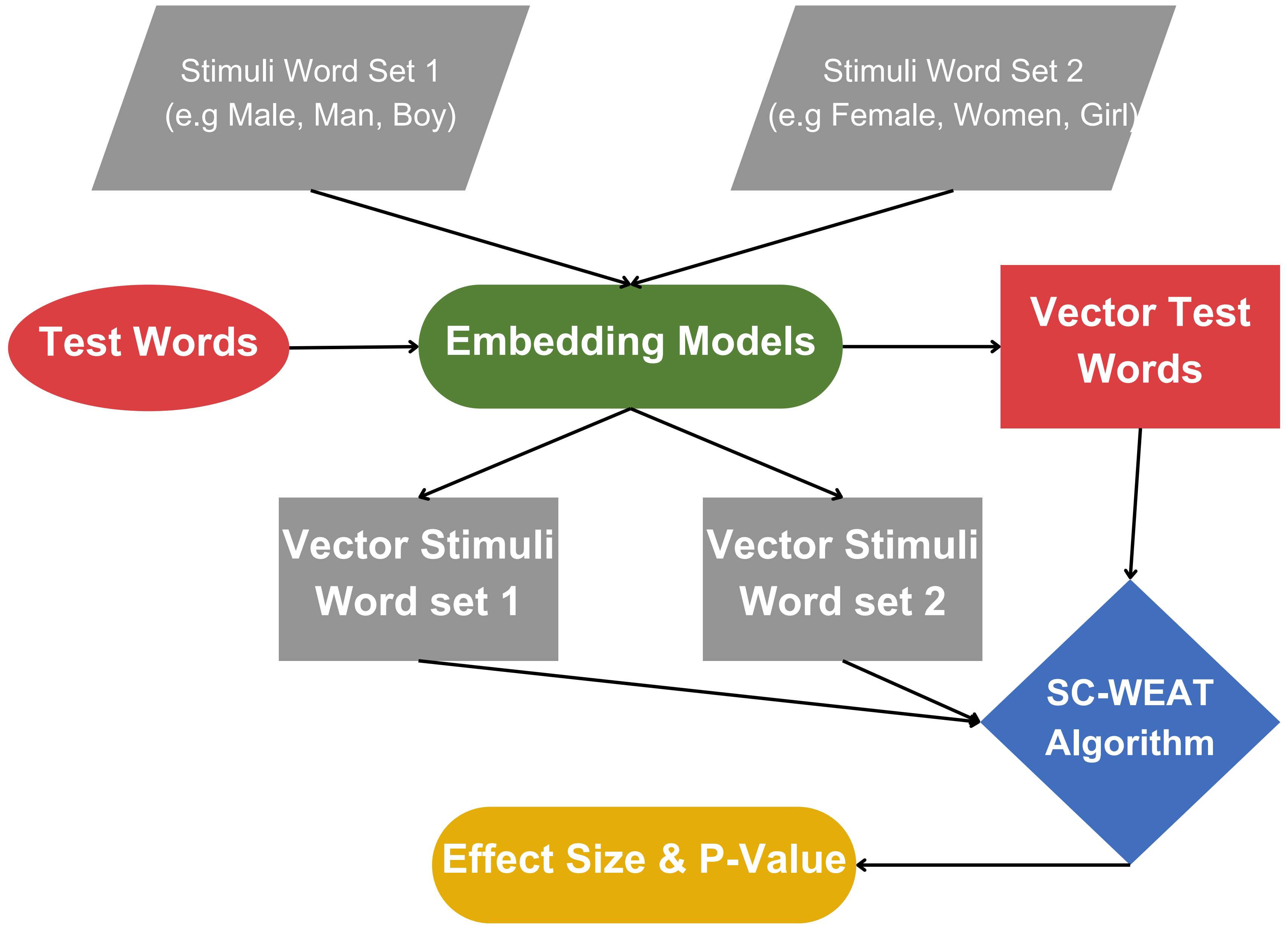
OBJECTIVES

- Analyze gender and race biases in modern LLMs.
  - OpenAI, Cohere, Google, Microsoft, and BGE
- Examine bias in word embeddings and their impact on real-world applications.
  - Tech Industry and Higher Education
- Address biases to ensure fairer and more ethical AI systems.

DATA SET

<table>
<tr><td><b>Test Word Sets</b></td><td>The most frequent 100,000 words from the GloVe embedding dataset</td></tr>
<tr><td><b>Word Embedding Models</b></td><td>OpenAI, Cohere, Google, Microsoft, and BGE embedding models.</td></tr>
<tr><td><b>Stimuli Words (Attribute Sets)</b></td><td>
<table>
<tr><th>Category</th><th>Stimuli Group</th><th>Stimuli Words</th></tr>
<tr><td rowspan="2">Gender</td><td>Female</td><td>Female, Woman, Girl, Hers, Sister, She, Her, Daughter</td></tr>
<tr><td>Male</td><td>Male, Man, Boy, Brother, He, Him, His, Son</td></tr>
<tr><td rowspan="3">Race</td><td>White</td><td>American, Australian, British, Canadian, White, Caucasian, European, French, German, Italian</td></tr>
<tr><td>Asian</td><td>Asian, Chinese, Japanese, Indonesian, Indian, Korean, Pakistani, Thai, Filipino, Brown</td></tr>
<tr><td>Black</td><td>African, African-American, Black, Congolese, Egyptian, Ethiopian, Haitian, Jamaican, Kenyan, Nigerian</td></tr>
</table>
</td></tr>
<tr><td><b>Big Tech Words</b></td><td>Big Tech companies based on [3]. Such as Google, Amazon, and Facebook</td></tr>
<tr><td><b>Top University Words</b></td><td>The top 50 universities from the 2024 Times Higher Education rankings</td></tr>
</table>

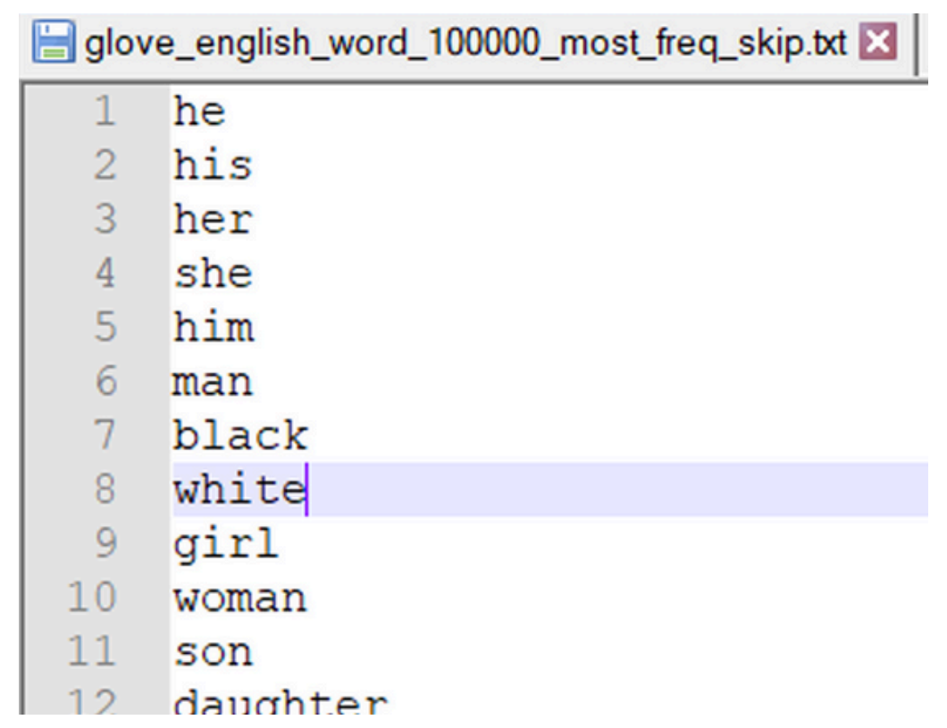
WORK FLOW



SC-WEAT [2]

- Measures bias using cosine similarity between word vectors.

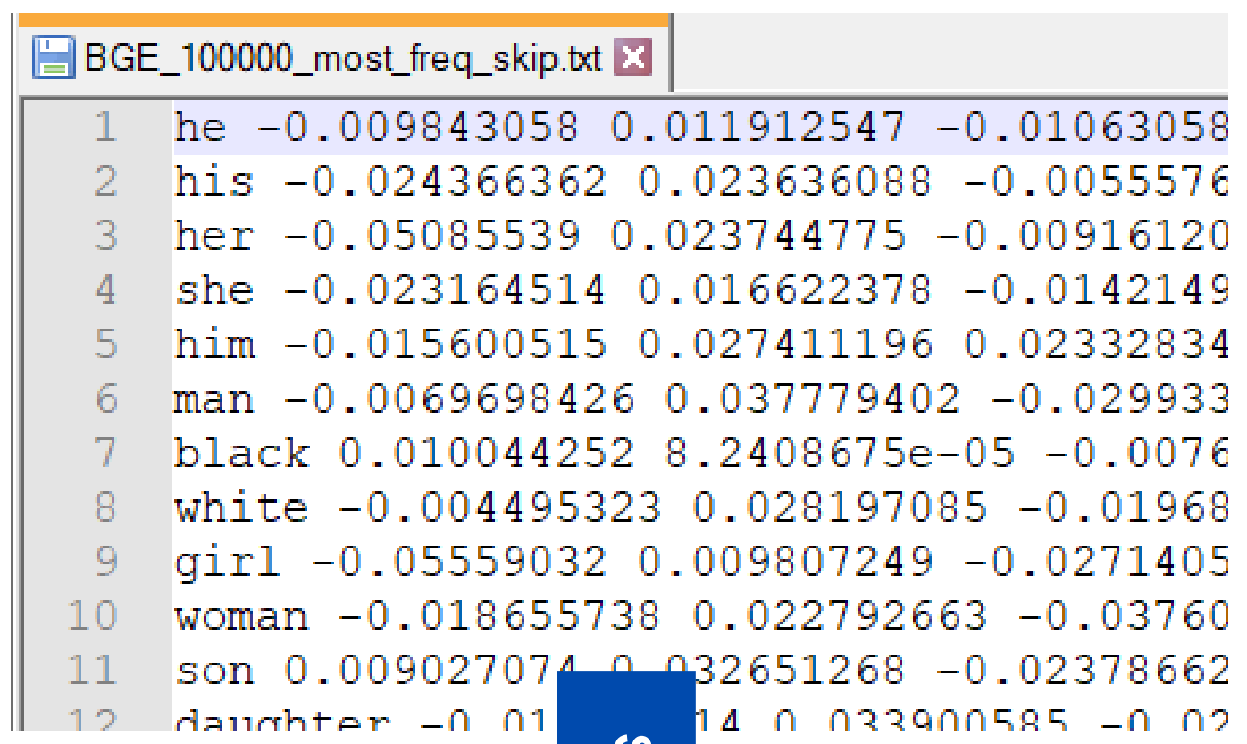
$$ES(\vec{w}, A, B) = \frac{\mathrm{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \mathrm{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\mathrm{std\_dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$



glove_english_word_100000_most_freq_skip.txt

```
he
his
her
she
him
man
black
white
girl
woman
son
daughter
```

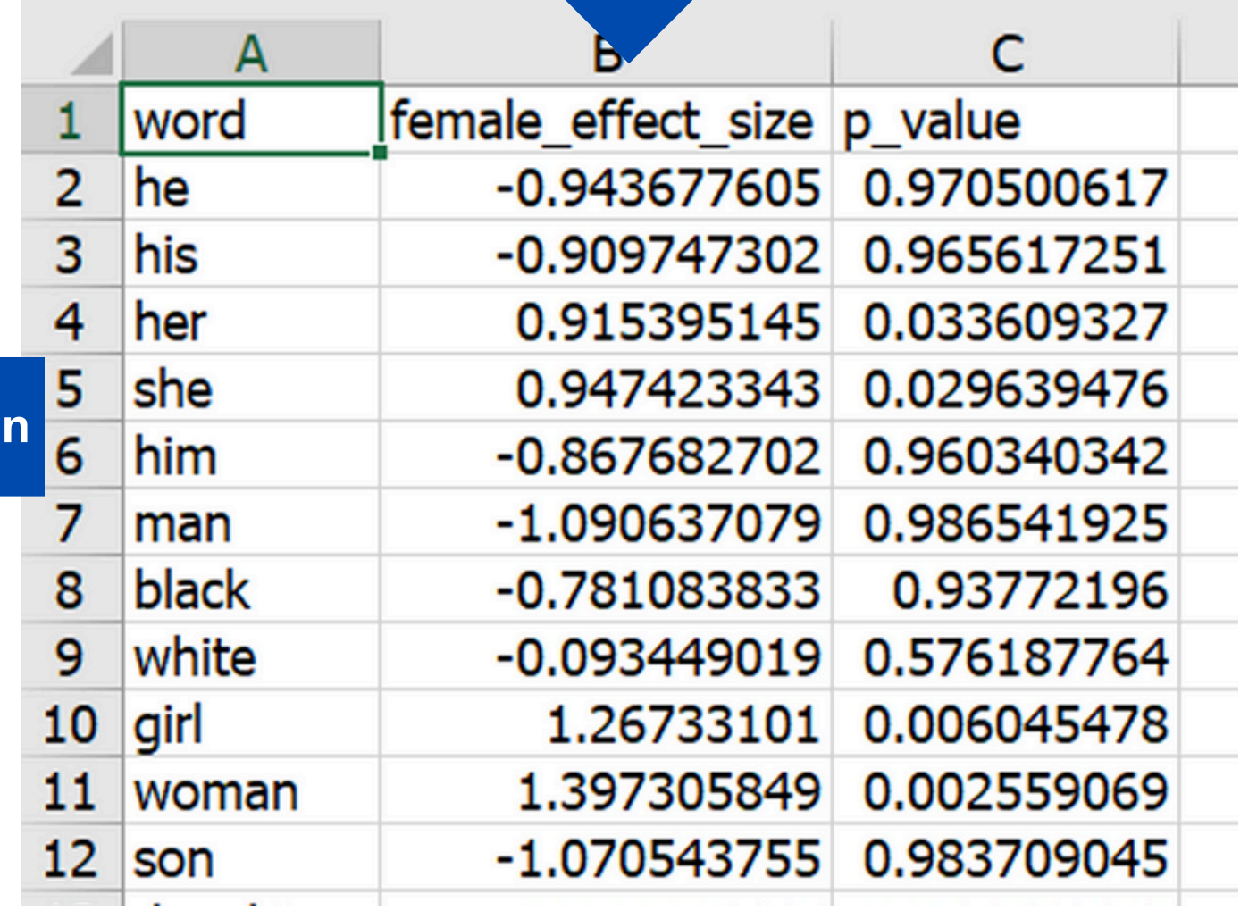
Embedding Models



BGE_100000_most_freq_skip.txt

```
he -0.009843058 0.011912547 -0.01063058
his -0.024366362 0.023636088 -0.0055576
her -0.05085539 0.023744775 -0.00916120
she -0.023164514 0.016622378 -0.0142149
him -0.015600515 0.027411196 0.02332834
man -0.0069698426 0.037779402 -0.029933
black 0.010044252 8.2408675e-05 -0.0076
white -0.004495323 0.028197085 -0.01968
girl -0.05559032 0.009807249 -0.0271405
woman -0.018655738 0.022792663 -0.03760
son 0.009027074 0.032651268 -0.02378662
daughter -0.01[illegible]14 0.033900585 -0.02
```

SC-WEAT Algorithm



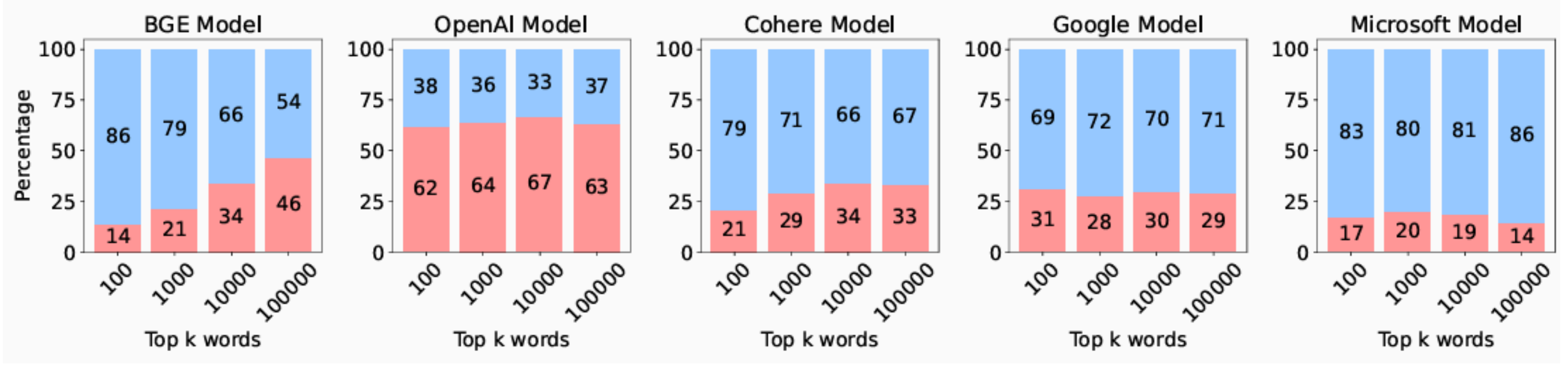
	A	B	C
1	word	female_effect_size	p_value
2	he	-0.943677605	0.970500617
3	his	-0.909747302	0.965617251
4	her	0.915395145	0.033609327
5	she	0.947423343	0.029639476
6	him	-0.867682702	0.960340342
7	man	-1.090637079	0.986541925
8	black	-0.781083833	0.93772196
9	white	-0.093449019	0.576187764
10	girl	1.26733101	0.006045478
11	woman	1.397305849	0.002559069
12	son	-1.070543755	0.983709045

Analysis & Virtualization

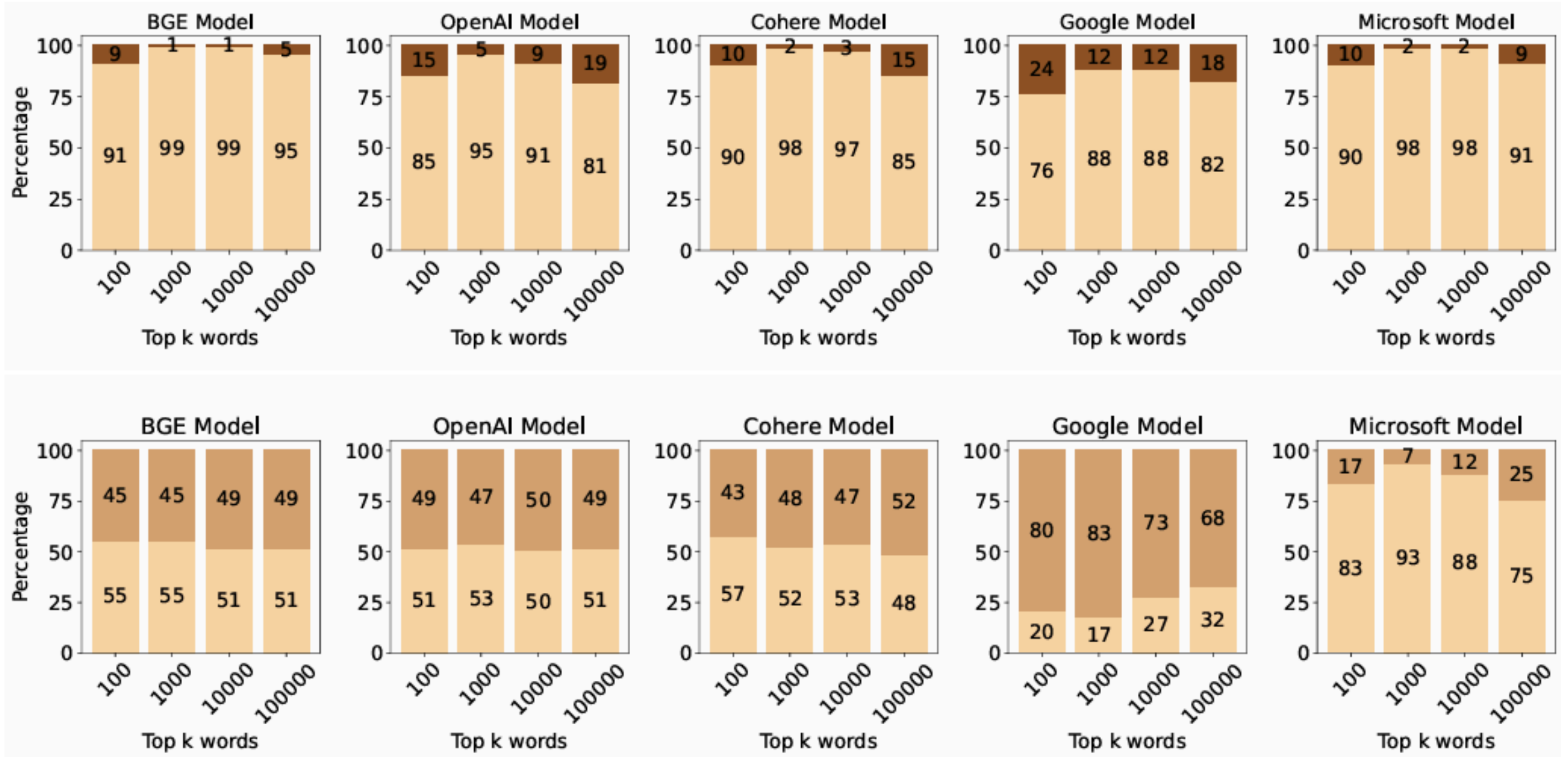
Analysis

- Bias Analysis by Frequency Range and Effect Size
- Semantic Categories of Gender- and Race-Associated Words
- Bias in Big Tech and Higher Education Contexts

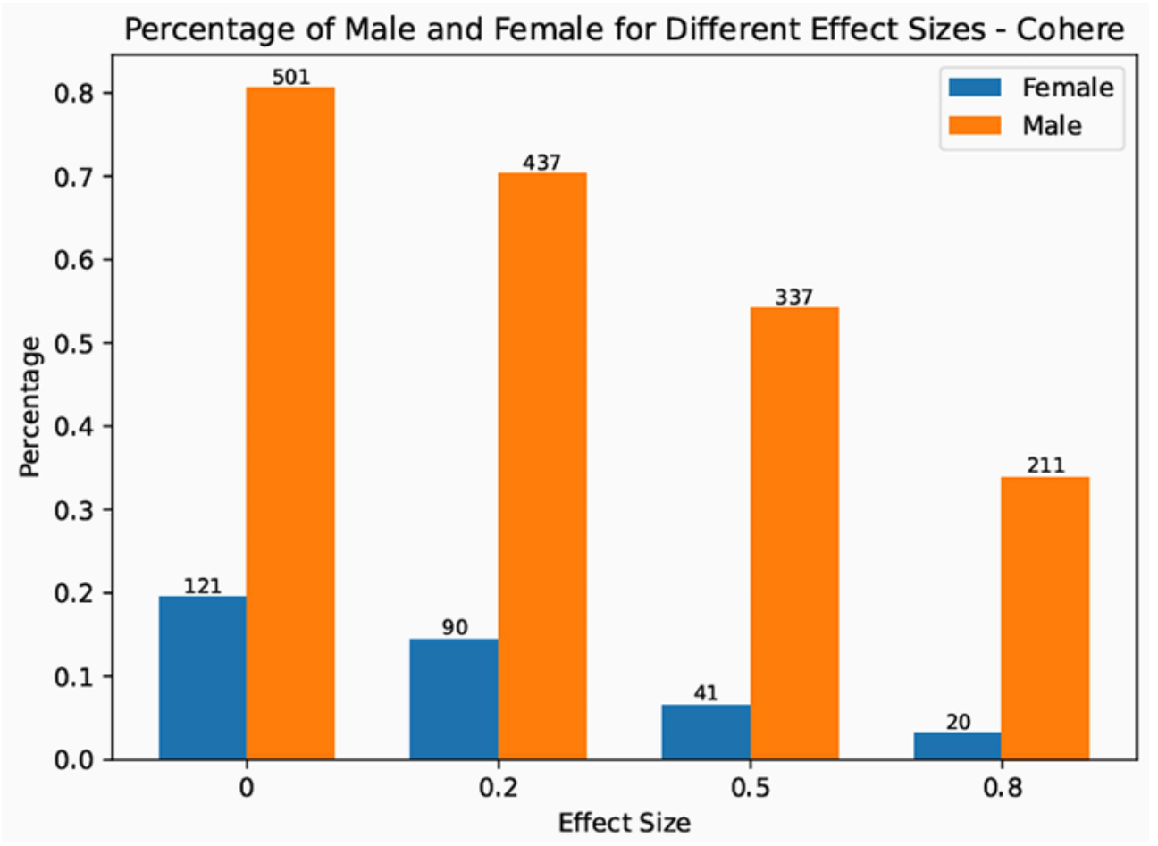
RESULTS



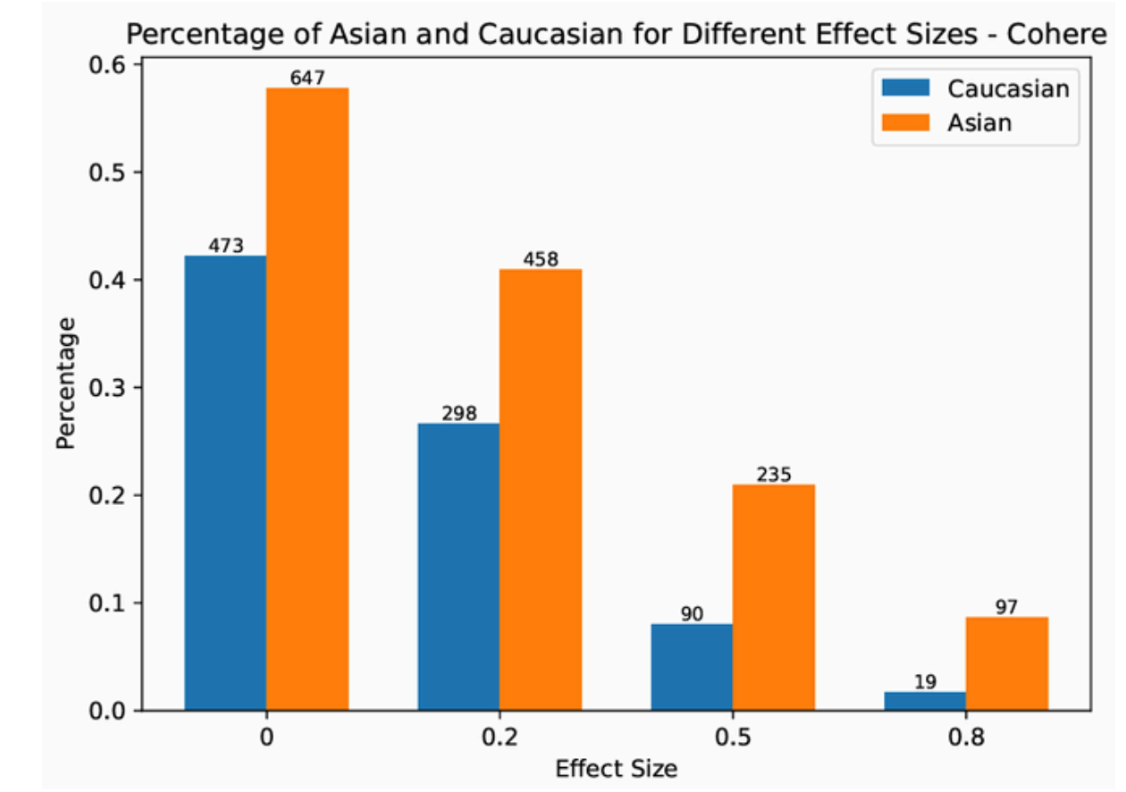
Gender Association of Top Words: Male is light blue, female is pink



Race Association of Top Words.1st row: White (lighter color) vs Black (darker color); 2nd row: White (lighter color) vs Asian (darker color)



Big Tech Association by Gender



Top University Association by Race

CONCLUSION

- **Male** group association dominates in most models
- **Black** group consistently underrepresented
- **Male** / **Asian** groups dominate in Big Tech
- **Male** / **Caucasian** groups dominate in Higher Education

REFERENCES

[1] Eric Michael Smith et al. ""I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset". In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 9180–9211.

[2] Aylin Caliskan et al. "Gender bias in word embeddings: A comprehensive anal ysis of frequency, syntax, and semantics". In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022, pp. 156–170.

[3] Mohamed Abdalla and Moustafa Abdalla. "The Grey Hoodie Project: Big to bacco, big tech, and the threat on academic integrity". In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021, pp. 287–297.

This research was supported by the Jamie Cassels Undergraduate Research Awards