Word Embedding Bias in Large Language Models

Poomrapee Chuthamsatid, Shera Potka, and Alex Thomo

University of Victoria, British Columbia, Canada

Abstract. This paper extends prior research on bias in word embeddings by addressing significant limitations in previous studies, such as Caliskan et al. (2017, 2022), which focused primarily on older models like GloVe and FastText and examined mainly gender bias. In contrast, our work investigates biases in modern large language models (LLMs), including OpenAI and Google embeddings, and expands the scope to both gender- and race-associated biases. We analyze biases across different word frequency ranges, using SC-WEAT tests, clustering, and t-SNE visualizations to uncover deeper insights into thematic clusters. Additionally, we explore how these biases are related to real-world sectors like the tech industry and higher education. By broadening the scope and applying more contemporary models, our research provides a more comprehensive understanding of bias in LLMs compared to earlier studies.

Keywords: Bias · Word Embeddings · Large Language Models

1 Introduction

The rapid development of Large Language Models (LLMs) has significantly expanded the use of Natural Language Processing (NLP) across diverse applications, ranging from text generation to intelligent chatbots. At the core of these applications are word embeddings, which convert words into numeric vectors based on co-occurrence statistics in large text corpora. Despite their power, word embeddings tend to inherit biases from the human-generated texts they are trained on, often reflecting demographic factors such as race, gender, and other social identities. These inherited biases introduce unintended prejudices into NLP models, leading to unfair outcomes in various applications [17].

Understanding and addressing these biases is crucial for the development of fair and unbiased LLMs. Mitigating stereotypes in NLP can help create technologies that promote fairness across a wide range of applications. Previous research has consistently shown that word embeddings reflect societal biases, whether intentionally or unintentionally [3, 4].

In a seminal study, Caliskan et al. (2017) introduced the Word Embedding Association Test (WEAT), demonstrating how pretrained GloVe embeddings replicate human biases related to gender, race, and age. The study revealed that words associated with women were more likely to be linked to caregiving roles, while words associated with men were tied to careers and competence. These biases, embedded in NLP systems, risk perpetuating harmful stereotypes with real-world consequences in domains like hiring, education, and healthcare.

Building on this foundation, Caliskan et al in [5] expanded the analysis by examining not just semantic associations between words, but also the frequency, syntax, and broader categories of biased words. By analyzing factors like word frequency and associations with areas such as big-tech, [5] discovered deeper gender biases not present in their previous work [4]. Additionally, they introduced the SC-WEAT method as a central tool for quantifying biases.

Our work significantly extends the scope of prior research on bias in word embeddings by addressing four key areas: (1) the frequency of gender- and race-associated words in modern large language models (LLMs); (2) bias variation across different frequency ranges and effect sizes in these models; (3) the identification and clustering of gender- and race-associated thematic concepts; and (4) the manifestation of biases in the tech industry and higher education. Unlike Caliskan et al. in [5], which focused on older embedding models like GloVe and FastText and examined only gender bias, we expand our analysis to include modern LLM embeddings and race-associated biases. Moreover, while Caliskan et al. provided a limited conceptual analysis, we offer a more detailed examination of the thematic clusters related to both gender and race.

More specifically, we analyzed gender and race biases in the 100,000 most frequent words from the GloVe dataset, focusing on five modern contextual word embedding models–OpenAI, Cohere, Google, Microsoft E5, and BGE. Using the SC-WEAT test, we quantified biases across various frequency ranges (top 100, 1,000, 10,000, and 100,000 words), measuring word associations with specific attributes to determine the direction and magnitude of bias. Additionally, we used k-means clustering and t-Distributed Stochastic Neighbor Embedding (T-SNE) visualizations to identify the semantic categories of gender- and race-associated words. We used a bottom-up approach to cluster 10,000 word associations and utilized GPT-3.5 to visualize key bias concepts in LLMs. We further analyzed the cosine similarity of words associated with Big Tech and top universities, identifying the top 1,000 word associations for each attribute. Our analysis provides critical insights into how biases in word embeddings are reflected in real-world contexts, particularly in the tech industry and higher education. Our study takes an important step in revealing hidden biases within popular large language models, moving toward understanding their impact on model behavior.

2 Related Work

2.1 Word Embedding Algorithms

Word embeddings, which map words to numeric vectors, have advanced natural language processing (NLP) by capturing semantic relationships within text. While GloVe [16] and similar models like Word2Vec [14] laid the groundwork for word embedding, modern LLM-based embeddings have since emerged, such as those from OpenAI, Cohere, Google, Microsoft (E5), and the Beijing Academy of Artificial Intelligence (BGE-M3) (among others). These models adapt word representations based on their surrounding context, providing a more nuanced understanding of language [11]. Despite their sophistication, these embeddings

still exhibit cultural stereotypes and biases, which remain challenging to mitigate [2]. Our work extends beyond [5] to explore these modern LLM embedding frameworks.

2.2 Measuring Bias in Word Embeddings

The Word Embedding Association Test (WEAT), introduced by Caliskan et al. [4], quantifies biases in word embeddings by measuring the differential association of two sets of target words with two sets of attribute words. For instance, it has shown that terms like "engineer" and "scientist" are often associated with male attributes, while "nurse" and "teacher" are associated with female attributes. Extensions of WEAT have been applied to detect various demographic biases, including gender, race, and age [6, 13].

Our study leverages the Single-Category Word Embedding Association Test (SC-WEAT) [5] to investigate biases related to gender and race. SC-WEAT, an extension of the Word Embedding Association Test (WEAT) [4, 18], quantifies the bias of a single target word relative to two sets of attribute words. The test computes an effect size (ES), using the formula:

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

Here, $\boldsymbol{w}$ is the target word, and A and B are two sets of words, each containing at least eight terms [5]. The formula calculates the mean cosine similarity between the target word and the terms in each set, normalized by the overall standard deviation. A higher positive effect size indicates a stronger association with set A (e.g., female-association), while a negative effect size indicates a stronger association with set B (e.g., male-association).

The output is an effect size, expressed as Cohen's d [9], along with a p-value. The effect size indicates the strength of the association, and the p-value tests its statistical significance [5]. Cohen's benchmarks define effect sizes of 0.2 as small, 0.5 as medium, and 0.8 as large [9]. For example, if a target word like "nurse" shows a large positive effect size (e.g., 0.8) with the female-association set, it suggests a strong gender bias associating nursing with women. Conversely, a negative effect size (e.g., -0.8) would imply a stronger male-association.

To evaluate the significance of the observed effect size, a permutation test is applied. In this process, the associations between sets A and B are shuffled by randomly reassigning words to each set, effectively mixing their labels. The mean difference in associations is recalculated for each shuffle, and this procedure is repeated 10,000 times. The resulting distribution of mean differences represents the effect size expected under random conditions. The p-value is obtained by comparing the observed effect size to the distribution of randomly generated effect sizes.

Caliskan et al. [5] used SC-WEAT to reveal entrenched gender stereotypes in widely used word embeddings such as GloVe and FastText.

2.3 How is the present paper different?

In this work, we extend the bias analysis of [5] to five contemporary large language model-based word embedding models: OpenAI, Cohere, Google, Microsoft, and BGE. We go beyond gender analysis to include race bias, addressing four key dimensions: the frequency of gender- and race-associated words, variations in biases across different frequency ranges and effect sizes, the clustering of related thematic concepts, and how these biases manifest in the tech industry and higher education.

3 Data

Most Frequent Words. We focus on the most frequent words from the GloVe embedding dataset, which contains 2.2 million words [16]. After filtering, we select the top 100,000 words for our analysis. We emphasize that, unlike [5], we use GloVe solely to identify the most frequent words, without using the GloVe embeddings themselves, as our focus is on the aforementioned LLM embeddings.

Word Embedding Models. We use five contextual embedding models: OpenAI's text-embedding-3-small (1,536 dimensions) [15], Microsoft's E5-large-v2 (1,024 dimensions) [19], Google's text-embedding-004 (300 dimensions) [8], Cohere's embed-english-v3.0 (1,024 dimensions) [10], and BGE's BGE-M3 (1,024 dimensions) [7]. Each model provides high-dimensional embeddings for semantic analysis.

Stimuli Words (Attribute Sets). Gender bias is measured using gender stimuli from Caliskan et al. [4], where positive effect sizes indicate a female association and negative sizes indicate male. Race bias uses ChatGPT-3.5-generated stimuli to measure biases among White, Asian, and Black groups. Positive effect sizes indicate a White or Asian association, depending on the comparison [18]. We show our stimuli words in Table 1.

Table 1: Gender and Race Stimuli

Category	**Stimuli Group**	**Stimuli Words**
Gender	Female	Female, Woman, Girl, Hers, Sister, She, Her, Daughter
	Male	Male, Man, Boy, Brother, He, Him, His, Son
Race	White	American, Australian, British, Canadian, White, Caucasian, European, French, German, Italian
	Asian	Asian, Chinese, Japanese, Indonesian, Indian, Korean, Pakistani, Thai, Filipino, Brown
	Black	African, African-American, Black, Congolese, Egyptian, Ethiopian, Haitian, Jamaican, Kenyan, Nigerian

Big Tech Words. We select Big Tech companies based on [1] that are present in the top 100,000 GloVe words, including companies such as Google, Amazon, Facebook, and Microsoft.
Top University Words. The top 50 universities from the 2024 Times Higher Education rankings [12] are added to the word list after normalizing their names.

4 Approach

Most Frequent Words Extraction. After we obtain the most frequent words, we apply the following preprocessing steps to clean up the data:

1. Remove stopwords, punctuation words with non-English characters, digits, and exclude words containing any of these.
2. Filter out words with fewer than three characters long.
3. Include the stimuli words used in SC-WEAT analysis to ensure accuracy benchmarks (See Table 1).

This cleaned set of frequent words is then input into five embedding models: OpenAI, Cohere, Google, E5 Microsoft, and BGE (BAAI General Embedding), generating the embeddings used for further analysis.
Frequency of Gender- and Race-Associated Words. We apply SC-WEAT to observe gender and race biases in the top 100, 1,000, 10,000, and 100,000 most frequent words generated from each embedding model. For gender bias, we analyze associations between female and male groups, assigning positive effect sizes to female-associated words and negative effect sizes to male-associated words. For race bias, we analyze three groups: White, Asian, and Black. Pairwise comparisons (White vs. Black, White vs. Asian, and Asian vs. Black) are performed, with positive effect sizes indicating associations with Whites (or Asians in the third comparison), and negative values indicating associations with Blacks or Asians, as relevant.
Bias Analysis by Frequency Range and Effect Size. We quantify gender and race biases across different frequency ranges by computing bias strength using SC-WEAT. Bias strength is evaluated based on effect size thresholds, following Cohen's classification [9]: Null bias: 0.00 – 0.19 Small bias: 0.20 – 0.49 Medium bias: 0.50 – 0.79 Large bias: ≥ 0.80. We apply these thresholds to the top 100, 1,000, 10,000, and 100,000 most frequent words extracted from each embedding model. A higher effect size indicates stronger bias, with greater disparities in word associations between the groups (e.g., gender or race). Models exhibiting larger effect sizes are considered to have more pronounced biases, as they display significant differences in the word distributions between demographic groups. This classification allows for a clear comparison of bias levels across embedding models and frequency ranges.

4.1 Semantic Categories of Gender- and Race-Associated Words

To better understand the nature of gender and race biases, we identify strong associations between demographic groups and specific stereotypes. Using SC-WEAT, we categorize words into two groups based on effect size and p-value:

- **Group 1**: The 1,000 most frequent words with an effect size ≥ 0.50 and a p-value < 0.05, indicating strong positive associations.
- **Group 2**: The 1,000 most frequent words with an effect size ≤ -0.50 and a p-value < 0.05, indicating strong negative associations.

In the gender bias analysis, Group 1 represents words associated with female attributes, while Group 2 represents words associated with male attributes. For race bias, Group 1 includes words associated with the White attribute group (in the White vs. Black and White vs. Asian comparisons) and the Asian attribute group (in the Asian vs. Black comparison). Group 2 represents words associated with Black individuals (in the White vs. Black and Asian vs. Black comparisons) and with Asians (in the White vs. Asian comparison).

To further explore patterns, we applied K-means clustering (using the Elkan algorithm) to these two groups of 1,000 words each. The optimal number of clusters, determined using the elbow method, is $k = 11$. We then reduced the dimensionality of the clustered embeddings using t-SNE for 2D visualization. Finally, we employed ChatGPT 3.5 to analyze and assign thematic concepts to each cluster, revealing common patterns and stereotypes linked to the identified biases.

4.2 Bias in Big Tech and Higher Education Contexts

Big Tech Bias Analysis We examine the representation of bias within Big Tech by focusing on the common Big Tech words identified by Abdalla and Abdalla [1]. These include Google, Amazon, Facebook, Microsoft, Apple, Nvidia, Intel, IBM, Huawei, Samsung, Uber, and Alibaba. From the 100,000 most frequent words, we calculate the cosine similarity of the embeddings for these Big Tech terms and identify the top 10,000 most associated words. For consistency, we identify the most associated words by intersecting the top 10,000 most associated words from all five models, resulting in a consistent set of 622 Big Tech-associated words. We then apply SC-WEAT to this 622-word set, using effect size ranges of 0.00 – 0.19 (null), 0.20 – 0.49 (small), 0.50 – 0.79 (medium), and ≥ 0.80 (large) to observe the bias strength for each class in pairwise comparisons for each of the five embedding models we consider.

Higher Education Bias Analysis For the higher education context, we compile a list of the Top 50 universities from the Times Higher Education 2024 ranking [12]. Since these universities are absent from our frequent word list, we extract their embeddings and append them to the word set. We then compute the cosine similarity with these university embeddings, identifying the top 10,000 associated words. Intersecting these word sets from all five models produces a consistent set of 1,120 Higher Education-associated words. We then apply SC-WEAT to this set, using effect size ranges (0.00 – 0.19, 0.20 – 0.49, 0.50 – 0.79, ≥ 0.80) to observe bias strength in pairwise comparisons for each of the five embedding models we consider.

5 Results

5.1 Top-Word Association Bias

Gender Association of Top-k Words. Figure 1 presents the distribution of gender-associated words across five different word embedding models (BGE, OpenAI, Cohere, Google, and Microsoft) for various sizes of word sets (top 100, 1,000, 10,000, and 100,000 words). The chart uses blue to indicate the percentage of male-associated words and pink for female-associated words.

Across most models, there is a consistent trend of greater male association, which tends to decrease as the size of the word sets increases. For instance, in the BGE Model, male-associated words account for 86% at the top 100 words but decrease to 54% at the top 100,000 words. Similarly, the Cohere Model shows a decrease from 79% male association at the top 100 words to 67% at the top 100,000 words. The observed decrease in male-associated word bias as the size of the top word sets increases indicates that models become less biased with a broader selection of frequently used words.

Interestingly, the OpenAI Model exhibits a different trend. Here, female-associated words dominate, ranging from 62% at the top 100 words to 63% at the top 100,000 words. This pattern indicates a reversal of the common trend seen in the other models, suggesting a unique alignment in the OpenAI model toward female associations across all analyzed word set sizes.

The Google and Microsoft models also show a consistent male bias, though with less variation across different word sets. For example, the Microsoft Model maintains a high percentage of male-associated words, from 83% at the top 100 words to 86% at the top 100,000 words, indicating strong and persistent male bias regardless of the word set size.

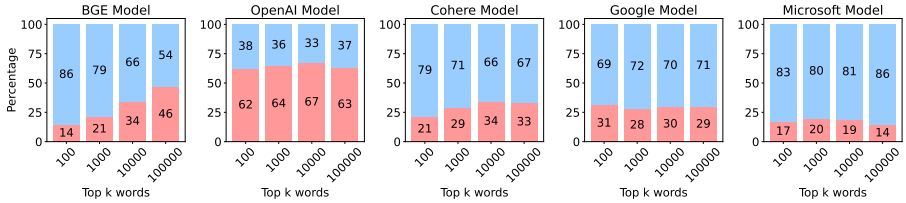

Fig. 1: Gender Association of Top Words. Male is light blue, female is pink.

Race (White vs Black) Association of Top-k Words. Figure 2 (1st row) presents the association distribution of the most frequent words between White and Black attribute sets. As before, the analysis spans different word set sizes, focusing on the top 100, 1,000, 10,000, and 100,000 words. For instance, in the BGE model, 95% of the top 100,000 words are associated with White attributes, while only 5%

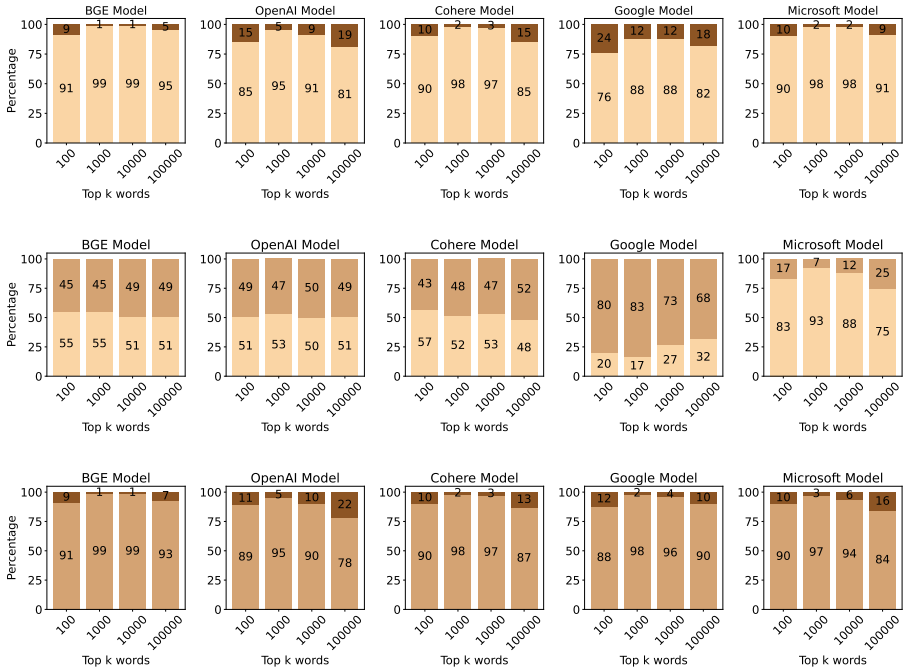

Fig. 2: Race Association of Top Words. 1st row: White (lighter color) vs Black (darker color); 2nd row: White (lighter color) vs Asian (darker color); 3rd row: Asian (lighter color) vs Black (darker color).

are associated with Black attributes. This strong skew toward White-associated words remains consistent across different word set sizes. Notably, all the models exhibit a strong association bias toward White.

Interestingly, the Google model is the most balanced among all the models, with 24%, 12%, 12%, and 18% Black association for the top 100, 1,000, 10,000, and 100,000 words, respectively. This distribution shows a noticeable reduction in bias compared to the other models. Still, even for the Google model, the association bias towards White remains. The OpenAI model is next in terms of balance, exhibiting a higher Black association of top words compared to the other models (except Google, which performs the best) but still showing a significantly skewed distribution towards White-associated words in the larger sets. Such an imbalance of association toward White across all the models highlights the persistent issue of racial bias, despite varying degrees of mitigation efforts.

Race (White vs. Asian) Association of Top-k Words. Figure 2 (2nd row) shows the percentage distributions of race association of top words (White vs. Asian) across the five embedding models. The BGE and OpenAI models maintain a relatively balanced distribution between White and Asian associations. The Cohere

model is also balanced, albeit less so than the previous two models. Surprisingly, the Google and Microsoft models show the most bias. The Google model is heavily skewed toward Asian associations, while the Microsoft model favours White associations.

Race (Asian vs. Black) Association of Top-k Words. Figure 2 (3rd row) shows the percentage distributions of race associations for top words (Asian vs. Black) across the five models. All models show a strong skew toward Asian associations, with Black associations being consistently underrepresented. Among them, the BGE model exhibits the strongest bias toward Asian associations, whereas the OpenAI model displays the least. These findings highlight a noticeable disparity in how the models represent Asian and Black groups.

5.2 Top-Word Association by Effect Size.

Table 2 presents the effect size analysis of gender associations for the top 100, 1,000, 10,000, and 100,000 words using SC-WEAT across the five models, with effect sizes ranging from 0 to 0.8, indicating the strength of gender association. Similarly, Table 3 provides the effect size analysis for race associations.

Across most models, there is a clear trend where male-association of top words consistently outnumbers female-association across all effect sizes. Notably, OpenAI deviates from this pattern, showing a higher number of female-associations across all top-k sets and effect size levels. For instance, within the top 100,000 words, the OpenAI model reports the highest number of strong female-associations (+0.8), with 20,174 words. In contrast, the Google and Microsoft models show the highest number of strong male-associations for the top words.

For White vs. Black associations, the BGE model has the highest number of strong White-associations (+0.8), with 52,707 words, while OpenAI shows the highest number of strongly Black-association words (-0.8), with 3,028 words, although Black associations remain underrepresented in all models. For White vs. Asian associations, OpenAI has the highest number of strong White-association words (+0.8), with 7,961 words, while Google shows the most strong Asian-association words (-0.8), with 19,465 words. Lastly, for Asian vs. Black associations, BGE has the most strong Asian-association words (+0.8), with 57,440 words, while OpenAI has the highest number of strong Black-associations (-0.8), with 3,404 words, yet Black associations remain consistently underrepresented across all models (See also Appendix for more details[1]).

5.3 Semantic Categories of Gender and Race Associated Words.

For each attribute in the gender and race analysis, we identified eleven clusters from each set of the 1,000 most frequently used female-, male-, White-, Asian-, and Black-biased words, each with an effect size greater or equal to 0.50 and a

[1] The Appendix can be found in the full version of this paper: `https://github.com/Poomon001/Bias-in-Word-Embeddings/blob/main/BiasEmbeddingLLM_full.pdf`

Table 2: Gender-Associations by Effect Size: number of top-100,000 words associated with female and male attributes. The 0, 0.2, 0.5, 0.8 columns denote the number of words with an effect size between 0 and 0.2, 0.2 and 0.5, and so on.

LLM	**Female**				**Male**			
	0	0.2	0.5	0.8	0	0.2	0.5	0.8
BGE	53,631	42,980	28,055	15,825	46,369	35,333	20,293	8,998
OpenAI	62,902	51,678	34,871	20,174	37,098	27,198	14,912	6,629
Cohere	33,160	24,315	14,378	8,604	66,840	56,283	39,141	23,695
Google	29,288	22,778	14,951	9,275	70,712	63,081	49,655	34,546
Microsoft	14,345	8,553	4,705	2,895	85,655	75,282	50,949	26,358

Table 3: Race Associations by Effect Size (Top 100,000 words)

LLM	**White vs. Black**							
	White				**Black**			
	0	**0.2**	**0.5**	**0.8**	**0**	**0.2**	**0.5**	**0.8**
BGE	95,832	91,344	77,736	52,707	4,168	1,879	544	134
OpenAI	80,654	72,159	54,904	33,649	19,346	12,889	6,604	3,028
Cohere	85,446	77,539	60,416	36,263	14,554	9,132	4,048	1,493
Google	82,104	73,213	54,736	32,319	17,896	11,228	5,133	2,010
Microsoft	91,012	83,785	63,970	34,996	8,988	4,834	1,787	556

LLM	**White vs. Asian**							
	White				**Asian**			
	0	**0.2**	**0.5**	**0.8**	**0**	**0.2**	**0.5**	**0.8**
BGE	51,881	31,694	11,091	2,600	48,119	28,248	9,534	2,241
OpenAI	50,882	38,086	20,833	7,961	49,118	36,798	20,833	9,147
Cohere	48,724	32,613	13,457	3,624	51,276	35,780	17,645	7,109
Google	32,263	22,406	11,439	4,485	67,737	56,365	37,510	19,465
Microsoft	74,870	59,300	29,510	7,497	25,130	15,330	7,205	3,178

LLM	**Asian vs. Black**							
	Asian				**Black**			
	0	**0.2**	**0.5**	**0.8**	**0**	**0.2**	**0.5**	**0.8**
BGE	92,803	87,553	75,710	57,440	7,197	3,862	1,171	270
OpenAI	78,270	69,567	53,272	34,445	21,730	15,070	7,779	3,404
Cohere	86,816	79,657	64,012	40,730	13,184	8,338	3,755	1,300
Google	90,015	83,964	69,988	48,838	9,985	6,088	2,550	928
Microsoft	83,823	71,859	46,359	19,632	16,177	8,562	2,685	650

p-value less than 0.05. We then input these cluster groups into GPT-3.5 to label each cluster, obtaining the following results.

Gender. We identified five female-associated cluster sets from each gender analysis across five different models (see Figure 3). Common female-associated clusters are related to healthcare, home decor, beauty, fashion, and sexual content.

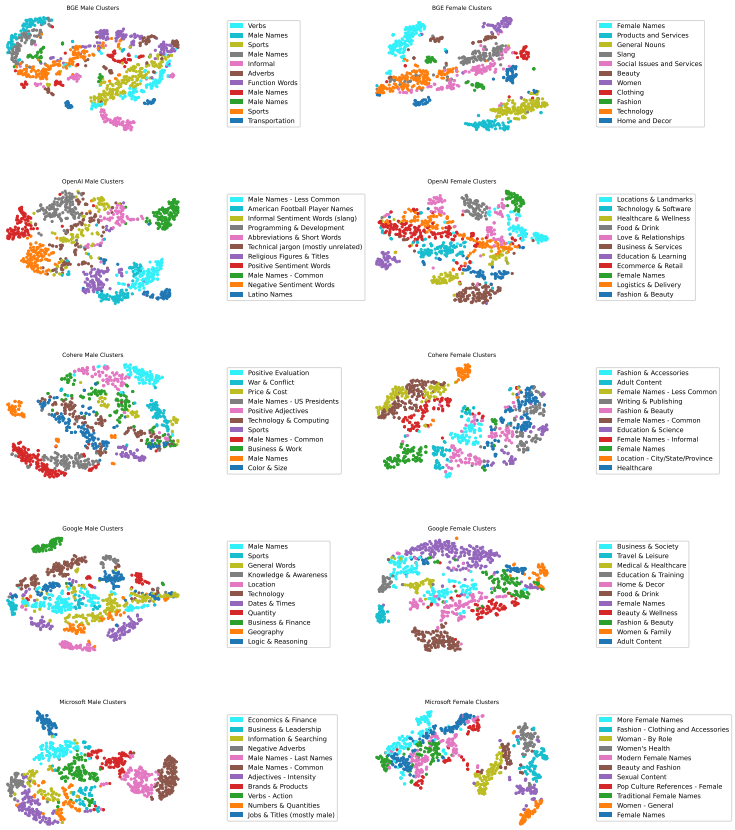

Fig. 3: Clusters for gender

Additionally, female names are consistently classified as a common theme. Conversely, male-associated clusters commonly focus on technology, sports, business, and sentiment words, with male names classified as a theme across all models. Each gender cluster set includes some noise in the form of generic titles, such as "name", "miscellaneous," "verb," or "adjective," which does not provide clear and meaningful categorization.

Race. We identified ten White-associated, ten Asian-associated, and ten Black-associated cluster sets from pairwise race analyses across five different models. Common White-associated clusters include business, people & society, education, media, and technology. In contrast, Asian-associated clusters frequently relate to business, software engineering, technology, entertainment, and food and culture. Black-associated clusters typically focus on religion, music, athletes and public figures, wild animals, and ethnicity. Each race cluster set includes some noise in the form of generic titles, such as "name", "location", "noun", "adverb", or "adjective", which does not provide clear and meaningful categorization. Due to

space constraints, we do not include figures for these clusters. They can be found in our github repository `github.com/Poomon001/Bias-in-Word-Embeddings`.

5.4 Gender and Race Bias in Big Tech Industry.

Gender. Our results indicate that 3 out of 5 models show a stronger association between big tech words and males. In the Cohere, Google, and Microsoft models, over 50% of the total 622 big tech words are associated with men at an effect size of 0.5, while fewer than 10% are associated with women (see Figure 4, top). In contrast, the OpenAI model reports a significant association between big tech words and women, while the BGE model indicates minimal gender bias in the tech field.

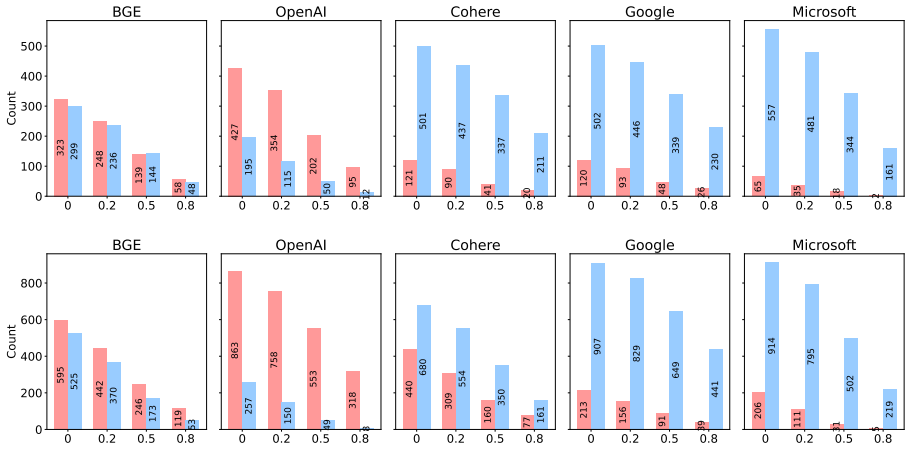

Fig. 4: Big-Tech and Top-University Gender-Association (Female vs Male). Effect sizes on the x-axis. 1st row: Big-Tech, 2nd row: Top-University.

Race. In pairwise comparisons, 4 out of 5 models show that big tech words are primarily associated with Asians rather than Whites (See Figure 5). The exception is the OpenAI model, which slightly favours Whites over Asians. All five models indicate a significant association of big tech words with Asians and Whites compared to Blacks. Notably, Black attributes have a minimal association with big tech.

5.5 Gender and Race Bias in Higher Education.

Gender. Our results indicate that 3 out of 5 models show a stronger association between top university words and males. This is similar to the results we observed

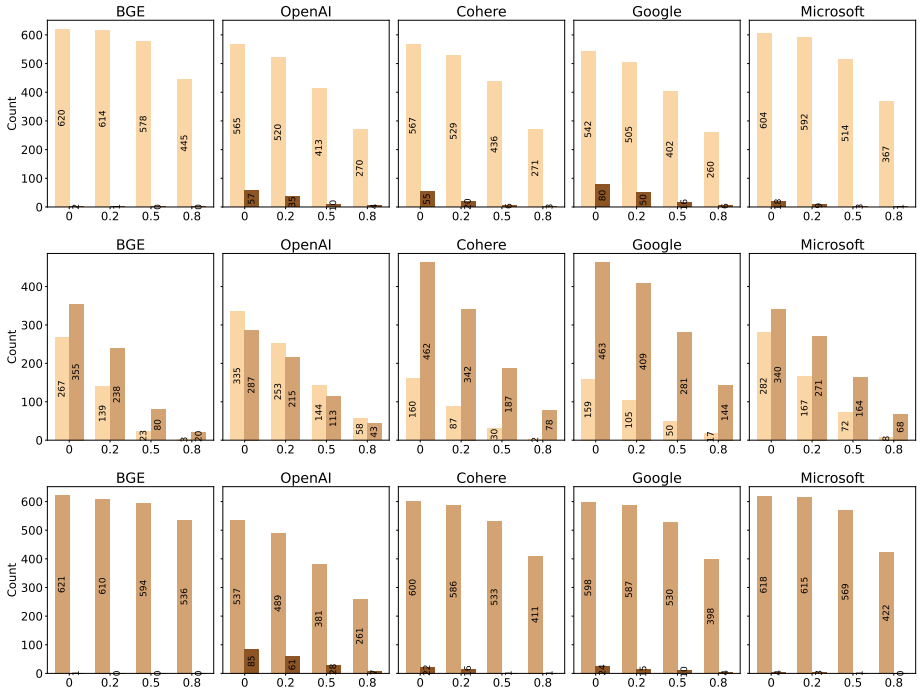

Fig. 5: Big-Tech Race Association. Effect sizes on the x-axis. 1st row: White (lighter color) vs Black (darker color); 2nd row: White (lighter color) vs Asian (darker color); 3rd row: Asian (lighter color) vs Black (darker color).

for Big Tech words. In the Cohere, Google, and Microsoft models, over 30%, 50%, and 40%, respectively, of the 1,120 top university words are associated with men at an effect size of 0.5, while only 5% or less are associated with women. In contrast, the BGE and OpenAI models report a higher association between top university words and women rather than men.

Race. In pairwise comparisons, 3 out of 5 models reveal a stronger association between top university words and Whites rather than Asians. The remaining two models prefer Asians over Whites. Across all five models, top university words are significantly more associated with Asians and Whites rather than Black attributes, which show minimal association with these words. We show the charts in Appendix.

6 Conclusions

Our study reveals several surprising patterns of gender and race bias across modern large language models, exposing clear disparities that extend beyond what previous research has shown. For instance, the analysis of gender association

highlights that, unlike most models, the OpenAI model exhibits a reversal of the common male bias trend, showing a higher proportion of female associations across all word set sizes. This unique pattern contrasts sharply with the consistent male association found in other models.

Similarly, when examining race associations, we discovered that Black attributes are strikingly underrepresented across all models, and the few strong Black associations typically involve specific domains like public figures and athletes. In contrast, Asian associations dominate in many models, particularly in the BGE model, which shows an overwhelming skew toward Asian associations across multiple word set sizes.

These findings emphasize the need for ongoing research and more nuanced debiasing methods to tackle such pervasive biases. As large language models are increasingly integrated into high-impact applications, it becomes crucial to address these imbalances to support the development of fairer and more inclusive AI systems.

References

1. Mohamed Abdalla and Moustafa Abdalla. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–297, 2021.
2. Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
3. Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29:4349–4357, 2016.
4. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
5. Aylin Caliskan, Parth Ajay Pimparkar, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170, 2022.
6. Kaytlin Chaloner and Alfredo Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, 2019.
7. Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *University of Science and Technology of China and BAAI*, 2024.
8. Google Cloud. Text embeddings api | generative ai on vertex ai. `https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api`, 2023.

9. Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 2013.
10. Cohere. Embed api reference. `https://docs.cohere.com/reference/embed`, 2023.
11. Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
12. Times Higher Education. World university rankings 2024. `https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking`, 2024.
13. Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133, 2021.
14. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
15. OpenAI. Embeddings guide. `https://platform.openai.com/docs/guides/embeddings`, 2023.
16. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
17. Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "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*, pages 9180–9211, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
18. Autumn Toney-Wails and Aylin Caliskan. Valnorm quantifies semantics to reveal consistent valence biases across languages and over centuries. *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
19. Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv*, 2212.03533, 2022.