# Bias in Large Language Models

Poomrapee Chuthamsatid {pchuthamsatid@uvic.ca}

## ABSTRACT

Word embedding is a word representation that converts words into numeric vectors based on their meanings derived from co-occurrence statistics in text corpora. Word embeddings often contain unintended social and cultural biases within the numerical vectors, reflecting stereotypes such as associating certain professions with one gender or race over another. Understanding these biases is crucial for developing strategies to mitigate them in large language models (LLMs) and ensure fairer systems. This study analyzes gender and race biases across modern English word embedding models, including those from OpenAI, Cohere, Google, Microsoft, and Beijing Academy of Artificial Intelligence (BGE). In this work, we uncover hidden biases and demonstrate the differences in (1) the frequency of gender- and race-associated words, (2) bias by frequency range and effect size across models, (3) the semantic categories of gender- and race-associated words, and (4) bias in the tech industry and higher education. Our methodology includes calculating the Semantic Clustering Word Embedding Association Test (SC-WEAT) to determine the relative association between words and using K-means clustering and t-Distributed Stochastic Neighbor Embedding (T-SNE) visualizations to categorize biased words.

First, we filter the most frequent 100,000 words to generate embeddings from five models. Word frequency analysis reveals that 4 out of 5 models show more frequent associations with men than women. All 5 models also demonstrate significant bias favoring Caucasian and Asian over Black attributes. Next, Applying SC-WEAT to find frequency range and effect size, we identify more medium and strong male biases in 3 out of 5 models. While most models show insignificant bias between Caucasians and Asians, significant bias persists against Blacks.

In semantic categories, we use K-means clustering and T-SNE on the top 1,000 highly associated words from each gender and race group to identify related concepts. For example, female-associated words are more related to fashion, beauty, and wellness, while male-associated concepts include technology, sports, and conflict. By computing SC-WEAT on the sets of selecting Big Tech and Top University, Big Tech shows a stronger preference for Asian males. At the same time, Higher Education demonstrates a significant bias toward Caucasian males.

Ultimately, our study highlights bias in word embeddings and exemplifies the impact on real-world applications. By highlighting areas requiring focused attention, we aim to raise awareness and guide the development of mitigation strategies in natural language processing, ensuring fairness and reducing technology bias.

## CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Machine learning**; **Artificial intelligence**; **Lexical semantics**; **Cognitive computing**.

## KEYWORDS

Word Embeddings, Gender Bias, Semantic Analysis, Ethics in AI, Fairness in AI, Representation Learning

**ACM Reference Format:**

Poomrapee Chuthamsatid {pchuthamsatid@uvic.ca}. 2024. Bias in Large Language Models. In . ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The race of Large language models (LLMs) development has led to the widespread use of natural language processing (NLP) in a diverse range of applications from simple text generations to intelligent chatbots. Word embeddings are a core of these applications, which convert words into numeric vectors based on their meanings derived from co-occurrence statistics in text corpora. These embeddings are powerful but have their limitations and flaws. Ideally, word embeddings would be bias-free and contribute to fairness. However, they inevitably inherit biases from the human-generated texts that they are trained on, which reflect demographic factors such as race, gender, and other identities [17].

Understanding these biases is essential for guiding the development of LLMs, as it helps identify strategies to mitigate stereotypes in natural language processing, leading to fair and unbiased technologies. Past research has revealed that word embeddings often reflect unintended social and cultural biases [3, 4]. Caliskan A. [4] showed that pretrained GloVe embeddings reflected human biases, including age, race, and gender associations. These social biases in NLP can foster harmful stereotypes and lead to widespread societal prejudice, affecting opportunities in the form of hiring, treatment, and education. For example, systems trained on past hiring data can reproduce existing prejudices, favoring males over females. This issue manifests in Amazon's recruiting automation, which showed a preference for male candidates [2].

A series of past works aims to detect and measure these biases to mitigate them. One of the studies by Michael Smith E. [17] introduces HOLISTICBIAS, a new and more inclusive bias measurement dataset that has 13 different demographic axes such as age, gender, and race. The study demonstrates the utility of HOLISTICBIAS

*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

in uncovering undetectable biases, targeting GPT-2, RoBERTa, DialoGPT, and BlenderBot 2.0. Another study by Caliskan A. [5] measures gender bias in 100,000 frequently used words in pretrained GloVe and FastText embeddings. However, we can go beyond gender bias analysis with more modern targeted LLMs.

Our work addresses bias in four areas (1) How modern English word embedding models differ in terms of the frequency of gender- and race-associated words, (2) What the differences in bias by frequency range and effect size across embedding models, (3) What the semantic categories of gender- and race-associated attributes, (4) How word embeddings reflect biases present in the tech industry and higher education.

After filtering out non-meaningful terms, our approach involves analyzing the 100,000 most frequent words from the GloVe dataset. We examine the gender and race biases in modern word embedding models, including those from OpenAI, Cohere, Google, Microsoft, and BGE. We apply the SC-WEAT test to quantify biases from the 100, 1,000, 10,000, and 100,000 most frequent words. This test measures word associations with specific attributes, showing positive or negative scores that indicate the direction and magnitude of biases. Our analysis further identifies semantic categories of gender- and race-associated words using k-mean clustering and t-Distributed Stochastic Neighbor Embedding (T-SNE) visualizations. We employ a bottom-up approach, starting with clustering the top 1,000 word associations and then using GPT-3.5 to identify cluster concepts, to visualize how LLMs encode these biases. Finally, we compute the cosine similarity of words associated with Big Tech and Top University to identify each attribute's top 1,000 word associations. This analysis provides insight into how biases in word embeddings are reflected in real-world contexts.

Overall, our work expands on past research by broadening the scope of word embedding model analysis. While Caliskan et al. [4] focused on static embedding models like GloVe and FastText, we extend the analysis to five modern contextual word embedding models—OpenAI, Cohere, Google, Microsoft E5, and BGE. Building on Michael Smith E.'s HOLISTICBIAS framework [17], we employ the SC-WEAT approach to not only conduct an in-depth examination of gender biases but also extend the analysis of racial biases. Furthermore, we go beyond theoretical analysis to investigate how word embedding biases in word embeddings manifest in real-world contexts. Specifically, we examine how biases affect women, men, Caucasians, Asians, and Black individuals in the tech industry and higher education. Ultimately, our study paves the way for future comprehensive analyses of other demographic axes across different word embedding models. By broadening the scope, our research provides a more nuanced understanding of how word embedding biases affect different domains and demographic groups.

## 2 RELATED WORK

**Word Embedding Algorithms.** Word embedding algorithms greatly enhanced the capabilities of natural language processing (NLP) by capturing semantic relationships between words in a vector space. Our work is grounded on Global Vectors for Word Representation (GloVe). GloVe, developed by Pennington, J., Socher, R., Manning, C. D. (2014) [15], generates a static word embedding that produces a single vector representation for each word regardless of its context. It constructs word vectors such that the dot product between any two vectors corresponds to the logarithm of their co-occurrence probability [15]. This method reflects the statistical relationships between words based on their co-occurrence patterns.

Our work extends beyond static embedding, like GloVe and FastText, to explore modern contextual (dynamic) embedding algorithms including, OpenAI, Cohere, Google, Microsoft (E5), and the Beijing Academy of Artificial Intelligence (BGE-M3). Unlike GloVe, these embeddings generate different vector representations for words based on the specific context in which they appear [16]. This method allows a more nuanced understanding of language for words with multiple meanings. Despite their advanced capabilities, dynamic embeddings often reflect cultural stereotypes and social biases [4].

**Measuring Bias in Word Embeddings.** Embeddings often encode social biases presented in the data they are trained on. Michael Smith E. introduced the HOLISTICBIAS framework, which includes over 450,000 unique sentence prompts, designed to identify biases in 13 different demographic axes in generative models [17]. This framework effectively uncovers bias by analyzing the patterns in the models' responses to these prompts. The Word Embedding Association Test (WEAT), introduced by Caliskan et al. [4], is another method to quantify biases in word embeddings by measuring the differential association of two sets of target words (e.g. engineer and household) with two sets of attribute words (e.g female and male). Studies have extended this method to detect different demographic biases, such as age, gender, and race, in the field of word embeddings [6, 13].

Our work uses the Single-Category Word Embedding Association Test (SC-WEAT), an extension of WEAT, to examine gender and race biases. While WEAT measures the relative association between two sets of target words and two sets of concept words, SC-WEAT focuses on the association between a single target word (e.g. home) with two concepts (e.g. female and male) [4, 9, 18]. Each concept in the SC-WEAT must have at least eight stimuli to ensure that the group gives an accurate statistical representation of the concept [18]. SC-WEAT can be expressed with the following formula:

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

$\vec{w}$ is the single target stimulus in SC-WEAT. $\cos(\vec{a}, \vec{b})$ is the cosine of the angle between the vectors $\vec{a}$ and $\vec{b}$. $A = [\vec{a_1}, \vec{a_2}, \ldots, \vec{a_n}]$ and $B = [\vec{b_1}, \vec{b_2}, \ldots, \vec{b_n}]$ are the two equal-sized ($n \geq 8$) sets of attributes representing concepts [5].

SC-WEAT returns an effect size metric ($ES$ in Cohen's $d$) and a p-value (The p-value highly correlates with effect size.) The effect size indicates the strength of the association, and the p-value determines statistical significance [5]. According to Cohen's $d$, an effect size of 0.20 is small; 0.50 is medium; and 0.80 is large [9]. The sign of the effect size indicates the direction of the association: a positive effect size ($d$) corresponds to an association with concept $A$, whereas a negative effect size ($d$) corresponds to an association with concept $B$ [5].

**Bias in Word Embeddings.** Past research has extensively studied bias in word embeddings and uncovered how the embedding models capture and amplify societal stereotypes that are present in their training data. Caliskan et al. [4] demonstrated that word embeddings reflect gender and race biases by showing that embeddings trained on large text corpora replicate human biases. WEAT was introduced in the paper to measure bias in word embeddings. The study by Bolukbasi et al. [3] addressed the issue of gender biases in word embeddings. It underscored the associations between gender and certain occupations and proposed a methodology to reduce these biases [3]. Guo W. and Caliskan A. [13] extended the WEAT to explore how contextualized word embeddings reflect intersectional biases (e.g African American females.) SC-WEAT, a variant of WEAT, was used to identify the presence of gender biases across different dimensions in static English word embeddings [5]. The findings highlighted the widespread prevalence of gender biases in static word embeddings.

## 3 DATA

**Most Frequent Words.** Our work selects target words from the most frequently used words of the GloVe embedding, which includes 2.2 million words [15]. While GloVe exhibits existing bias [5], it is widely used by researchers, practitioners, developers, and students in LLMs. To increase the accuracy of our analysis, we apply word filterings on the GloVe words and select only the 100,000 most frequently used words.

**Word Embedding Models.** We use the best five contextual embedding models as of 2024, including those from OpenAI, Microsoft E5, Google, Cohere, and Beijing Academy of Artificial Intelligence (BGE-M3). The OpenAI (text-embedding-3-small) embedding model offers a 1,536-dimensional representation. Despite the newest and most performant models, it is the most cost-effective option available from OpenAI [14]. The Microsoft E5 (E5-large-v2) model [19], developed by Microsoft, is an English-language pre-trained model that features 24 layers and an embedding size of 1024 dimensions. The Google (text-embedding-004) model, the latest version of Google Generative AI on Vertex AI as of 2024, provides 300-dimensional embeddings for semantic similarity tasks [8]. Additionally, we use the Cohere (embed-english-v3.0) model, an English-language pre-trained embedding model that provides 1,024-dimensional representations [10]. The BGE-M3 (BAAI/bge-m3-unsupervised) model [7], the latest fine-tuned multilingual from BGE as of 2024, features 1024-dimensional embedding. We utilize the Dense Retrieval method of the BGE-M3 model to manage high-dimensional data and capture nuanced semantic similarities.

**Stimuli Words.** We use the two-word sets of gender stimuli (See Table 1), described by Caliskan et al [4], in SC-WEAT to measure the relative gender association of a target word with female and male attribute groups. In all gender bias analyses, the positive effect size indicates a female association whereas the negative indicates a male association.

We use three sets of race-related stimuli (See Table 2), consisting of common descriptors and nationalities generated by ChatGPT-3.5, in SC-WEAT to measure the relative race association of a target word with Caucasian, Asian, and Black attribute groups in a pairwise manner. In the Caucasian vs. Asian and Caucasian vs. Black bias analyses, a positive effect size indicates an association with Caucasians, while a negative effect size indicates an association with Asian or Black, respectively. In the Asian vs. Black bias analysis, a positive score indicates an Asian association, while a negative score indicates a Black association.

Each set contains at least eight stimuli to ensure that the group gives an accurate statistical representation [18].

**Big Tech Words.** The Big Tech companies are listed by Mohamed Abdalla and Moustafa Abdalla [1]. From their list, we select the Big Tech names that appear in the 100,000 most frequently used words from GloVe. The final list of Big Tech includes Google, Amazon, Facebook, Microsoft, Apple, Nvidia, Intel, IBM, Huawei, Samsung, Uber, and Alibaba.

**Top University words.** The top universities are selected from the top-ranking 50 universities as of 2024 from Times Higher Education [12]. We remove special characters from their names and convert abbreviations to full names for consistency. Specifically for the Top University Analysis, we append these 50 names to the 100,000 most frequently used words from GloVe, as they are not present in the original list.

## 4 APPROACH AND EXPERIMENTS

**Most Frequency of Words.** Obtaining the most frequent word from GloVe embedding [5], we filter out non-meaningful words with the following cleanup steps: 1. Remove stopwords, punctuation, non-English characters, and digits, 2. Exclude any word containing punctuation, non-English characters, or digits, 3. Filter out words with fewer than three characters, 4. Include the stimuli words in the frequency word list to establish a verification benchmark for accuracy (See Tables 1, 2). The experiment uses these most frequent words as a targeting word set, which inputs into five different embedding models: Namely OpenAI, Cohere, Google, E5 Microsoft, and BGE (BAAI General Embedding), to retrieve word embeddings.

**Table 1: Gender Stimuli**

| Female | Male |
|---|---|
| Female, Woman, Girl, Hers, Sister, She, Her, Daughter | Male, Man, Boy, Brother, He, Him, His, Son |

and for race:

**Table 2: Race Stimuli**

| Caucasian | Asian | Black |
|---|---|---|
| American, Australian, British, Canadian, Caucasian, European, French, German, Italian, White | Asian, Brown, Chinese, Japanese, Indonesian, Indian, Korean, Pakistani, Thai | African African-American, Black, Congolese, Egyptian, Ethiopian, Haitian, Jamaican Kenyan, Nigerian |

**Frequencies of Gender-Associated and Race-Associated Words.** To quantify the frequency association between the most frequent words and gender or race class, we apply the SC-WEAT to observe gender and race biases within the top 100, 1,000, 10,000, and 100,000 most frequent words generated from each embedding model. Specifically, we examine gender bias between female and male groups, and race bias among Caucasian, Asian, and Black groups. In the gender association analysis, the report assigns a positive effect size to words associated with females and a negative effect size to words related to males. We perform pairwise comparisons in the race association analysis to thoroughly examine associations: Caucasian vs. Black, Caucasian vs. Asian, and Asian vs. Black. Positive effect sizes indicate associations with Caucasians in the first two comparisons and Asians in the third, while negative effect sizes indicate associations with Blacks, Asians, and Blacks, respectively.

**Gender and Race Bias by Frequency Range and Effect Size Across Embedding Models.** We further classify and report the bias strength, computed from SC-WEAT, to examine the bias within the embedding models. As defined by Cohen [9], bias strength is categorized into four effect size thresholds: 0.00 − 0.19 (null), 0.20 − 0.49 (small), 0.50 − 0.79 (medium), and ≥ 0.80 (large). These thresholds are applied and reported for the top 100, 1,000, 10,000, and 100,000 most frequent words. As we discussed, an embedding model with higher bias will produce a greater discrepancy in word distribution between each pair in the groups.

**Semantic Categories of Gender- and Race-Associated Words.** Identifying strong associations between groups and specific stereotypes is crucial for determining the nature of gender and race biases. We use effect size and p-value obtained from SC-WEAT to identify two bias groups. The first group consists of the 1000 most frequent words with an effect size $\geq 0.50$ and a p-value $< 0.05$, while the second group consists of the 1000 most frequent words with an effect size $\leq -0.50$ and a p-value $< 0.05$. In gender association, the first group indicates a female attribute group, whereas the second group represents a male group. In race association, the first group indicates associations with the Caucasian attribute group (in Caucasian vs. Black and Caucasian vs. Asian comparisons) and the Asian attribute group (in Asian vs. Black comparisons). The second group indicates associations with the Blacks (in Caucasian vs. Black and Asian vs. Black comparisons) and the Asians (in Caucasian vs. Asian comparison). For each pairwise comparison of these two groups of 1000 words, we applied K-means clustering using the Elkan algorithm to reveal a cluster of biased words. We use the elbow method to determine $k = 11$ which is found to be optimal. T-SNE is used to reduce the dimensions of the clustered embeddings for 2D visualization. We use ChhatGPT 3.5 and Gemini, a Google AI model to examine and assign common theme concepts to each cluster.

**Gender and Race Bias in Big Tech Industry and Higher Education.** To understand how these word-embedding biases play a role in the real world, we examine them in concepts such as Big Tech and Higher Education.

In the Big Tech analysis, we obtain a list of Big Tech words from the 'Mohamed Abdalla and Moustafa Abdalla' paper [1]. We select the common Big Tech words that are in our 100,000 most frequency words: Google, Amazon, Facebook, Microsoft, Apple, Nvidia, Intel, IBM, Huawei, Samsung, Uber, Alibaba. We calculate the cosine similarity between each embedding and the Big Tech embeddings and select the top 10,000 most associated words. To ensure 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.

In the Higher Education concept, we obtain a list of the Top 50 University names from Times Higher Education 2024 [12]. For consistency, we convert all university abbreviations into their full name. Since all university names are absent from our 100,000 most frequent words, we obtain their embedding vectors from each model and append them to our most frequent word list. We calculate cosine similarities with the Top 50 University embeddings and select 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.

# 5 RESULT

**5.1 Frequencies of Gender-Associated and Race-Associated Words. Gender Bias.** The result from 4 out of 5 word embedding models consistently shows stronger associations with men than women across common word sets of 100, 1,000, 10,000, and 100,000 words. The Microsoft E5, Google, and Cohere embedding models exhibit a significant male preference (See Figures 10, 12, 13). The BGE model also prefers male associations but to a lesser extent (See Figure 1). In contrast, the OpenAI model demonstrates a stronger association with female attributes (See Figure 11).

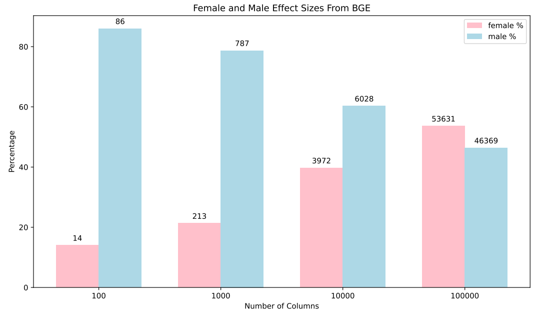


**Figure 1: Most Frequent Gender-Associated Words for the BGE Model**

**Race Bias (Caucasian vs Black).** Our result indicates that 5 out of 5 models are strongly associated with Caucasian attributes, whereas Black attributes are primarily underrepresented. Although the patterns are consistent across different models, the BGE model exhibits the strongest male association, with 95,832 out of 100,000 words (See Figure 2), while the OpenAI model shows the weakest, with 80,654 out of 100,000 male-related words (See Figure 14).

**Race Bias (Caucasian vs Asian).** Our result indicates that 3 out of 5 models show little bias between Caucasian and Asian attributes (See Figures 17, 18, 19). The Microsoft E5 model reports a stronger

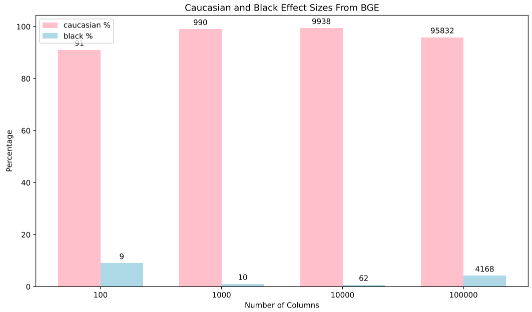


**Figure 2: Most Frequent Race-Associated (Caucasian vs Black) Words for the BGE Model**

association with Caucasians, with 74,870 out of 100,000 words (See Figure 3). In comparison, the Google model shows a stronger association with Asians, with 67,737 out of 100,000 words (See Figure 16).

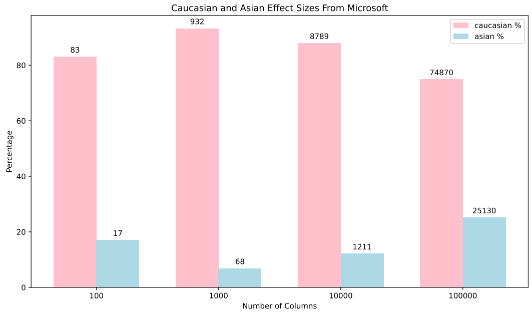


**Figure 3: Most Frequent Race-Associated (Caucasian vs Asian) Words for the Microsoft Model**

**Race Bias (Asian vs Black).** Results from all five models indicate a significant association with Asians, whereas Black attributes are primarily underrepresented. The BGE model shows the strongest association, with 92,803 out of 100,000 words (See Figure 4), while the OpenAI model reports the lowest, with 78,270 out of 100,000 words (See Figure 15).

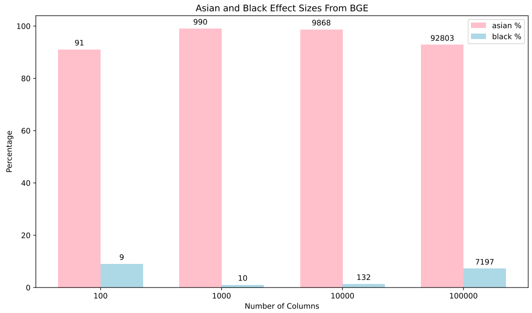


**Figure 4: Most Frequent Race-Associated (Asian vs Black) Words for the BGE Model**

### 5.2 Gender and Race Bias by Frequency Range and Effect Size Across Embedding Models.

**Gender (Female vs Male).** Tables 3, 4, 11, 12, 13 show the full details of gender association by frequency range and effect size for BGE, Cohere, Google, Microsoft E5, and OpenAI embedding models. Within the top 100,000 words, the OpenAI model has the highest number of strongly female-associated words (+0.8), with 20,174 words (See Table 4). In contrast, the Google model reports the most strongly male-associated words (-0.8), with 34,546 words (See Table 3).

**Table 3: Gender-Associated (Female vs Male) by Range and Effect Size - Google Model**

| Gender-Associated (Female vs Male) by Range and Effect Size - Google Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | |
| num_words | female_0 | female_0.2 | female_0.5 | female_0.8 | male_0 | male_0.2 | male_0.5 | male_0.8 |
| 100 | 31 | 22 | 12 | 8 | 69 | 62 | 51 | 37 |
| 1000 | 276 | 182 | 102 | 49 | 724 | 633 | 466 | 307 |
| 10000 | 2956 | 2188 | 1271 | 671 | 7044 | 6140 | 4582 | 2896 |
| 100000 | 29288 | 22778 | 14951 | 9275 | 70712 | 63081 | 49655 | 34546 |

**Table 4: Gender-Associated (Female vs Male) by Range and Effect Size - OpenAI Model**

| Gender-Associated (Female vs Male) by Range and Effect Size - OpenAI Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | |
| num_words | female_0 | female_0.2 | female_0.5 | female_0.8 | male_0 | male_0.2 | male_0.5 | male_0.8 |
| 100 | 62 | 39 | 23 | 8 | 38 | 27 | 15 | 4 |
| 1000 | 643 | 476 | 250 | 104 | 357 | 242 | 102 | 38 |
| 10000 | 6691 | 5527 | 3646 | 1980 | 3309 | 2316 | 1142 | 454 |
| 100000 | 62902 | 51678 | 34871 | 20174 | 37098 | 27198 | 14912 | 6629 |

**Race Bias (Caucasian vs Black).** Tables 5, 6, 14, 15, 16 show the full details of race association between Caucasian and Black attributes by frequency range and effect size for BGE, Cohere, Google, Microsoft E5, and OpenAI embedding models. Within the top 100,000 words, the BGE model has the highest number of strongly Caucasian-associated words (+0.8), with 52,707 words (See Table 5). In contrast, the OpenAI model reports the highest number of strongly Black-associated words (-0.8), with 3,028 words (See Table 6), however, the Black attributes are still underrepresented in all the models.

**Table 5: Race-Associated (Caucasian vs Black) by Range and Effect Size - BGE Model**

| Race-Associated (Caucasian vs Black) by Range and Effect Size - BGE Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Black | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 91 | 90 | 89 | 83 | 9 | 8 | 4 | 0 |
| 1000 | 990 | 989 | 982 | 941 | 10 | 8 | 4 | 0 |
| 10000 | 9938 | 9837 | 9455 | 7963 | 62 | 36 | 15 | 1 |
| 100000 | 95832 | 91344 | 77736 | 52707 | 4168 | 1879 | 544 | 134 |

**Table 6: Race-Associated (Caucasian vs Black) by Range and Effect Size - OpenAI Model**

| Race-Associated (Caucasian vs Black) by Range and Effect Size - OpenAI Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Black | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 85 | 82 | 70 | 49 | 15 | 12 | 9 | 9 |
| 1000 | 959 | 928 | 785 | 538 | 41 | 26 | 11 | 9 |
| 10000 | 9158 | 8587 | 7007 | 4688 | 842 | 482 | 177 | 67 |
| 100000 | 80654 | 72159 | 54904 | 33649 | 19346 | 12889 | 6604 | 3028 |

**Race Bias (Caucasian vs Asian).** Tables 7, 8, 17, 18, 19 show the full details of race association between Caucasian and Asian attributes by frequency range and effect size for BGE, Cohere, Google, Microsoft E5, and OpenAI embedding models. Within the top 100,000 words, the OpenAI model has the highest number of strongly Caucasian-associated words (+0.8), with 7,961 words (See Table 8). In contrast, the Google model reports the most strongly Asian-associated words (-0.8), with 19,465 words (See Table 7).

**Table 7: Race-Associated (Caucasian vs Asian) by Range and Effect Size - Google Model**

| Race-Associated (Caucasian vs Asian) by Range and Effect Size - Google Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Asian | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 |
| 100 | 20 | 13 | 9 | 2 | 80 | 72 | 52 | 38 |
| 1000 | 178 | 109 | 41 | 8 | 822 | 724 | 504 | 267 |
| 10000 | 2712 | 1746 | 749 | 238 | 7288 | 6091 | 4093 | 2095 |
| 100000 | 32263 | 22406 | 11439 | 4485 | 67737 | 56365 | 37510 | 19465 |

**Table 8: Race-Associated (Caucasian vs Asian) by Range and Effect Size - OpenAI Model**

| Race-Associated (Caucasian vs Asian) by Range and Effect Size - OpenAI Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Asian | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 |
| 100 | 51 | 36 | 16 | 2 | 49 | 44 | 27 | 9 |
| 1000 | 534 | 372 | 168 | 44 | 466 | 323 | 135 | 46 |
| 10000 | 4919 | 3471 | 1637 | 514 | 5081 | 3654 | 1769 | 645 |
| 100000 | 50882 | 38086 | 20833 | 7961 | 49118 | 36798 | 20833 | 9147 |

**Race Bias (Asian vs Black).** Tables 9, 10, 20, 21, 22 show the full details of race association between Asian and Black attributes by frequency range and effect size for BGE, Cohere, Google, Microsoft E5, and OpenAI embedding models. Within the top 100,000 words, the BGE model has the highest number of strongly Asian-associated words (+0.8), with 57,440 words (See Table 9). In contrast, the OpenAI model reports the highest number of strongly Black-associated words (-0.8), with 3,404 words (See Table 10), however, the Black attributes are still underrepresented in all the models.

**Table 9: Race-Associated (Asian vs Black) by Range and Effect Size - BGE Model**

| Race-Associated (Asian vs Black) by Range and Effect Size - BGE Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Asian | | | | Black | | | |
| num_words | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 91 | 90 | 88 | 86 | 9 | 7 | 3 | 0 |
| 1000 | 990 | 985 | 973 | 947 | 10 | 8 | 3 | 0 |
| 10000 | 9868 | 9715 | 9299 | 8310 | 132 | 59 | 16 | 5 |
| 100000 | 92803 | 87553 | 75710 | 57440 | 7197 | 3862 | 1171 | 270 |

**Table 10: Race-Associated (Asian vs Black) by Range and Effect Size - OpenAI Model**

| Race-Associated (Asian vs Black) by Range and Effect Size - OpenAI Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Asian | | | | Black | | | |
| num_words | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 89 | 86 | 68 | 54 | 11 | 11 | 9 | 5 |
| 1000 | 950 | 897 | 730 | 540 | 50 | 32 | 17 | 7 |
| 10000 | 9062 | 8430 | 6858 | 4743 | 938 | 563 | 236 | 100 |
| 100000 | 78270 | 69567 | 53272 | 34445 | 21730 | 15070 | 7779 | 3404 |

## 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-, Caucasian-, Asian-, and Black-biased words, each with an effect size greater than 0.50 and a p-value less than 0.05. We then input these cluster groups into ChatGPT 3.5 to classify and title each cluster, obtaining the following results:

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

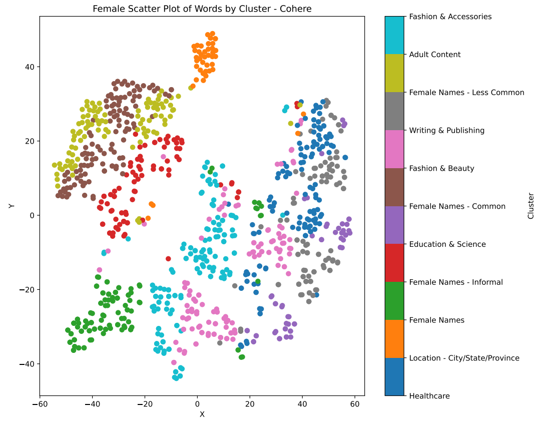


**Figure 5: Semantic Categories of Gender-Associated (Female) - Cohere Model**

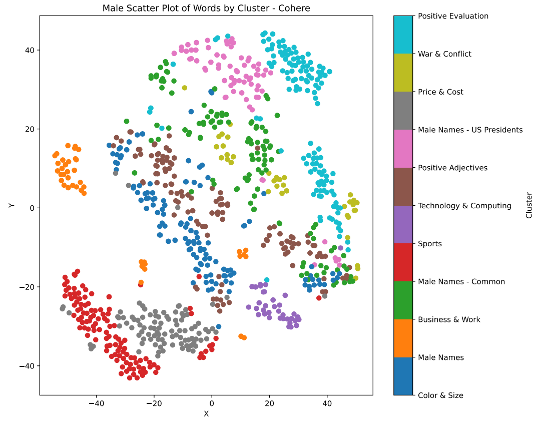


**Figure 6: Semantic Categories of Gender-Associated (Male) - Cohere Model**

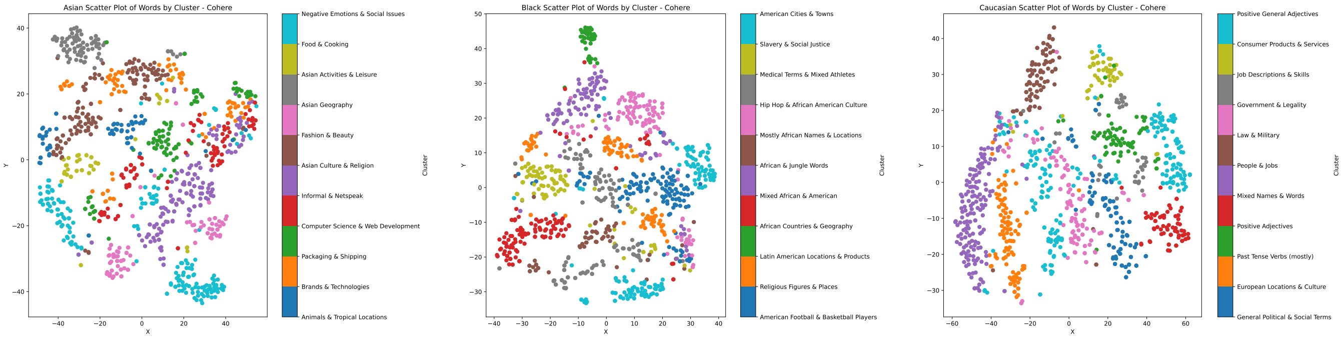


**Figure 7: Semantic Categories of Race-Associated Asian, Black, and Caucasian Clusters - Cohere Model**

**Race Bias.** We identified ten Caucasian-associated, ten Asian-associated, and ten Black-associated cluster sets from pairwise race analyses across five different models. Common Caucasian-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.

**5.4 Gender and Race Bias in Big Tech Industry. Gender - Big Tech.** 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 E5 models, over 50% of the total 622 big tech words are moderately (0.5) associated with men, while fewer than 10% are associated with women (See Figure 20). In contrast, the OpenAI model reports a significant association between big tech words and women (See Figure 23), while the BGE model indicates minimal gender bias in the tech field (See Figure 8).

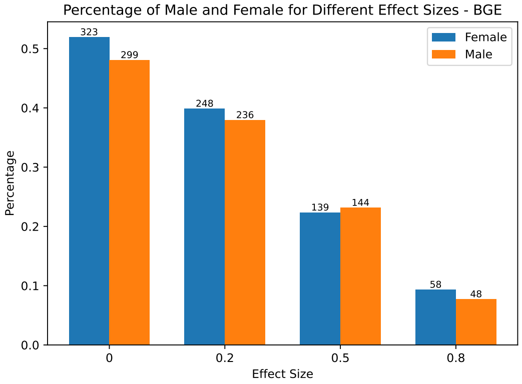


**Figure 8: Big Tech Gender-Associated (Female vs Male) words for BGE**

**Race - Big Tech.** In pairwise comparisons, 4 out of 5 models show that big tech words are primarily associated with Asians rather than Caucasians (See Figure 9). The exception is the OpenAI model, which slightly favors Caucasians over Asians (See Figure 24). All five models indicate a significant association of big tech words with Asians and Caucasians compared to Blacks. Notably, Black attributes have a minimal association with big tech (See Figures 21, 22).

**5.5 Gender and Race Bias in Higher Education. Gender - Higher Education.** Our results indicate that 3 out of 5 models show a stronger association between top university words and males. In the Cohere, Google, and Microsoft E5 models, over 30%, 50%, and 40%, respectively, of the 1,120 top university words are moderately (0.5) associated with men, while only 5% or less are associated with women (See Figure 25). In contrast, the BGE and OpenAI models report a higher association between top university words and women rather than men (See Figures 26, 27).

**Race - Higher Education.** In pairwise comparisons, 3 out of 5 models reveal a stronger association between top university words and Caucasians rather than Asians (See Figure 29). The remaining two models prefer Asians over Caucasians. Across all five models, top university words are significantly more associated with Asians and Caucasians rather than Black attributes, which show minimal association with these words (See Figures 30, 31).

## 6 DISCUSSION

**Gender Bias.** Our results indicate gender bias across five models. The 4 out of 5 word embedding models show a stronger association with males than females. The Cohere, Google, and Microsoft models exhibit a similar pattern of significant male bias. The BGE model shows a similar pattern but with a lower male bias while the OpenAI model demonstrates a higher female association. This finding suggests that different models embed gender biases differently, potentially influenced by their training data. Past research on static word embeddings, including GloVe and FastText in 2022 [5], indicates a significant male bias. This suggests that some word embedding models as of 2024 have attempted to improve the representation of women in their training data.

**Race Bias.** The examination reveals that all word embeddings exhibit racial biases. Our findings indicate that the BGE and OpenAI

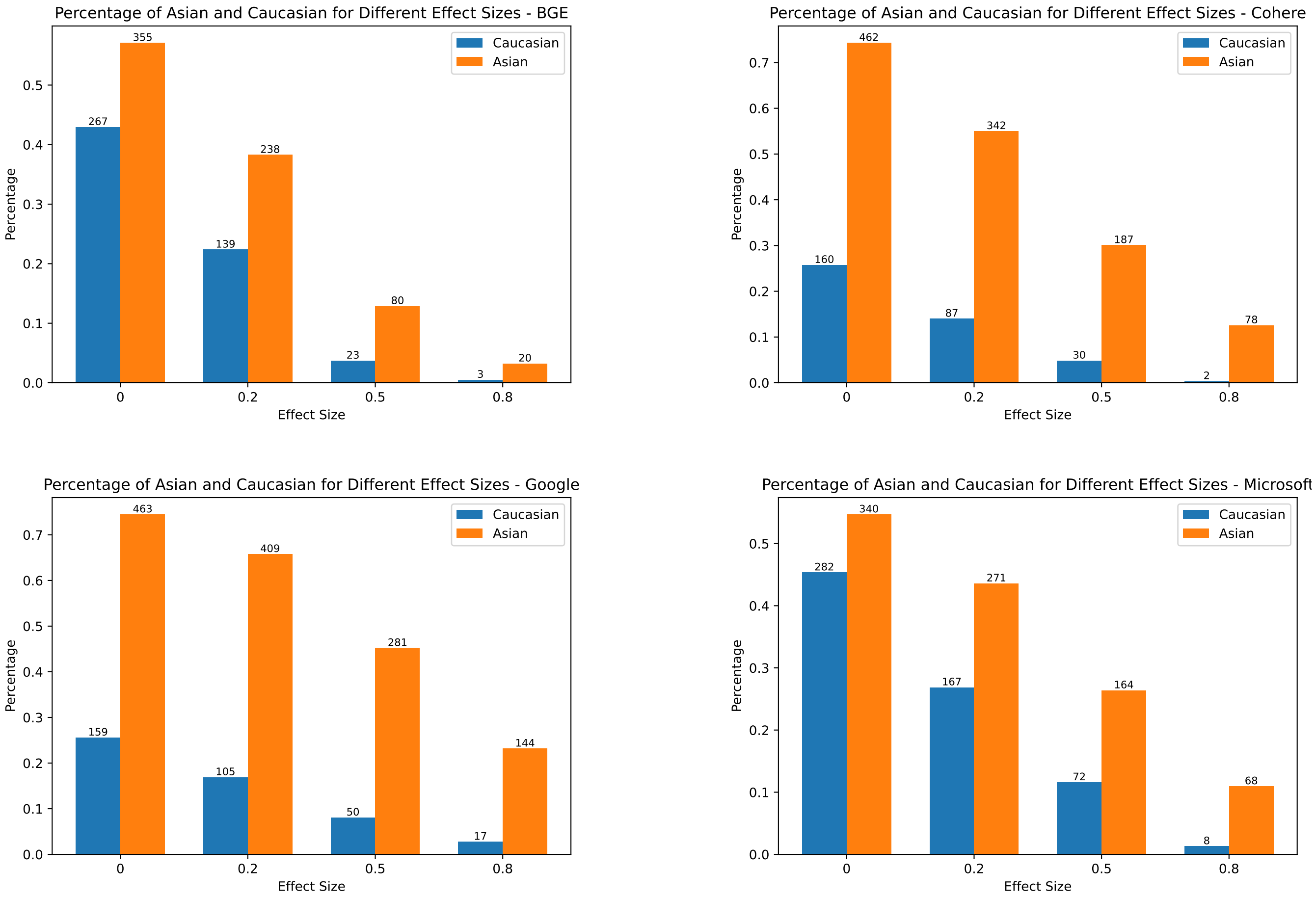


**Figure 9: Big Tech Race-Associated (Caucasian vs Asian) words for BGE, Cohere, Google, Microsoft**

models show less bias in associating target words with Caucasian and Asian attributes (See Tables 8, 17). Conversely, the Cohere and Google models show a strong (0.8) association of target words with Asian attributes (See Tables 7, 18), while the Microsoft model demonstrates a strong (0.8) association with Caucasian attributes. The Black attributes are underrepresented across all different models. This indicates that BGE and OpenAI have improved the inclusivity and fairness of their training data for Caucasian and Asian groups. However, other models still need further development to mitigate persistent biases among different racial groups. Unfortunately, all models exhibit some degree of misrepresentation of Black individuals in their training data.

**Semantic Catagory.** The clusters identified for the 1,000 female-associated words include topics like healthcare, beauty, and fashion, while the 1,000 male-associated words are linked to technology, business, and sports. These patterns highlight persistent stereotypes within the embedding models. Our findings across five contextual embedding models mirror past research on static embeddings (GloVe and FastText) [5]. It suggests ongoing challenges in eliminating these persistent social stereotypes from embedding models.

Similarly, the pairwise comparison of race attributes in word embeddings shows clustering patterns that reflect underlying stereotypes. The clusters identified for the 1,000 Caucasian-associated words include those topics from business, technology, and education. Likewise, the cluster identified 1,000 Asian-associated words are related to business, technology, and software engineering. This result suggests a stereotype of Caucasians and Asians being linked with knowledge, professional careers, and academic fields. In contrast, the clustering for the 1,000 words associated with Black attributes differs with a focus on religion, music, and ethnicity. This perception indicates a tendency to associate Blacks with cultural and religious contexts rather than professional or academic fields. Ultimately, these patterns may reinforce stereotypes that limit the representation of Black individuals in various domains such as employment and education.

**Big Tech and Higher Education.** Our findings, obtained from big tech and higher education analyses, align with the result from semantic category analysis. Mirroring the male bias pattern from Caliskan et al's work [5], the 3 out of 5 models indicate that men are more associated with the big tech field. Cohere, Google, and Microsoft models show significant male associations with both big tech and higher education contexts. While BGE is less gender-biased, the OpenAI model reveals a stronger association with women in these contexts. This result reinforces the stereotype of men being more linked with technology and education. Furthermore, our analysis indicates significant bias against Blacks compared to Caucasians and Asians in the big tech and higher education fields. All models indicate that Black attributes are underrepresented. Ultimately, social biases may limit the opportunities for individuals from different demographic groups. Our findings underscore the need for continuous efforts to improve embedding models to ensure fairness across all domains and demographic dimensions.

The resources are available to the public on GitHub.[1]

## 7 CONCLUSION

This work analyzes gender and race biases in five popular word embedding models in various contexts. Our findings not only show that gender and race biases manifest across different models but quantify the level of bias in different contexts. We find that the 100,000 most frequently used words in English are strongly associated with males more than females, though the degree of bias varies among models. Similarly, all models show an immense degree of bias favoring Caucasian and Asian attributes over Black attributes. The semantic clustering further exposes gender stereotypes associated with each attribute group. Female-associated words cluster around healthcare, beauty, and fashion, while male-associated words are related to business, technology, and sports. The clustering of Caucasian-associated words with business, technology, and education, Asian-associated words with business, technology, and software engineering, and Black-associated words with religion, music, and ethnicity can demonstrate racial stereotypes. Our tech and higher education analysis examines the impact in real-world applications. The findings reveal that male attributes are more frequently associated with both big tech and higher education terms. Big tech words are closely linked to Asian attributes, while higher education terms are more related to Caucasians. In contrast, Black attributes are significantly underrepresented in both contexts. Ultimately, our findings address the need to refine bias mitigation strategies and provide guidelines to improve fairness in natural language processing.

## 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*. 2021, pp. 287–297.

[2] J Stewart Black and Patrick van Esch. "AI-enabled recruiting: What is it and how should a manager use it?" In: *Business Horizons* 63.2 (2020), pp. 215–226.

[3] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 4349–4357.

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[5] Aylin Caliskan et al. "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*. 2022, pp. 156–170.

[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*. 2019, pp. 25–32.

[7] Jianlv Chen et al. "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation". In: *University of Science and Technology of China and BAAI* (2024). URL: https://arxiv.org/pdf/2402.03216.

[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] Emily Dinan et al. "Multidimensional gender bias classification". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 314–331.

[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*. 2021, pp. 122–133.

[14] OpenAI. *Embeddings Guide*. https://platform.openai.com/docs/guides/embeddings. 2023.

[15] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[16] Karahan Sarıtaş, Cahid Arda Öz, and Tunga Güngör. "A Comprehensive Analysis of Static Word Embeddings for Turkish". In: *arXiv preprint* arXiv:2405.07778 (2024).

[17] 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.

[18] Autumn Toney-Wails and Aylin Caliskan. "ValNorm quantifies semantics to reveal consistent valence biases across languages and over centuries". In: *Empirical Methods in Natural Language Processing (EMNLP)* (2021).

[19] Liang Wang et al. "Text Embeddings by Weakly-Supervised Contrastive Pretraining". In: *arXiv* 2212.03533 (2022). URL: https://arxiv.org/abs/2212.03533.

## A ADDITIONAL DATA

**A.1 Female Stimuli.** daughter, female, girl, her, hers, she, sister, woman

**A.1 Male Stimuli.** boy, brother, he, him, his, male, man, son

**A.3 Caucasian Stimuli.** american, australian, british, canadian, caucasian, european, french, german, italian, white

**A.4 Asian Stimuli.** asian, brown, chinese, filipino, indian, indonesian, japanese, korean, pakistani, thai

**A.5 Black Stimuli.** african, african-american, black, congolese, egyptian, ethiopian, haitian, jamaican, kenyan, nigerian

**A.6 Big Tech Words.** Alibaba, Amazon, Apple, Facebook, Google, Huawei, IBM, Intel, Microsoft, Nvidia, Samsung, Uber

**A.7 Top University Words.** "University of Oxford", "Stanford University", "Massachusetts Institute of Technology", "Harvard University", "University of Cambridge", "Princeton University", "California Institute of Technology", "Imperial College London", "University of California, Berkeley", "Yale University", "ETH Zurich", "Tsinghua University", "The University of Chicago", "Peking University", "Johns Hopkins University", "University of Pennsylvania", "Columbia University", "University of California, Los Angeles", "National University of Singapore", "Cornell University", "University

[1] Source code available at: https://github.com/Poomon001/Bias-in-Word-Embeddings.

of Toronto", "University College London", "University of Michigan-Ann Arbor", "Carnegie Mellon University", "University of Washington", "Duke University", "New York University", "Northwestern University", "The University of Tokyo", "University of Edinburgh", "Technical University of Munich", "Nanyang Technological University, Singapore", "École Polytechnique Fédérale de Lausanne", "University of California, San Diego", "University of Hong Kong", "Georgia Institute of Technology", "University of Melbourne", "King's College London", "LMU Munich", "Paris Sciences et Lettres – PSL Research University Paris", "University of British Columbia", "University of Illinois at Urbana-Champaign", "Shanghai Jiao Tong University", "Fudan University", "KU Leuven", "London School of Economics and Political Science", "Universität Heidelberg", "Delft University of Technology", "McGill University", "Karolinska Institute"

## B ADDITIONAL RESULTS

### B.1 Frequencies of Gender-Associated Words.

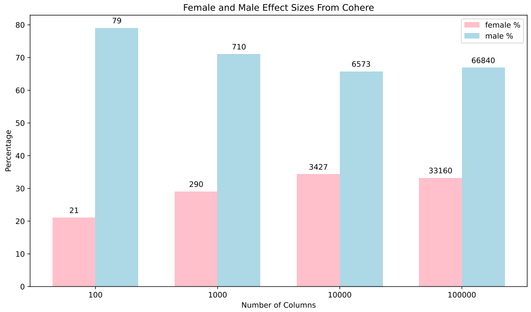


Figure 10: Most Frequent Gender-Associated Words for the Cohere Model

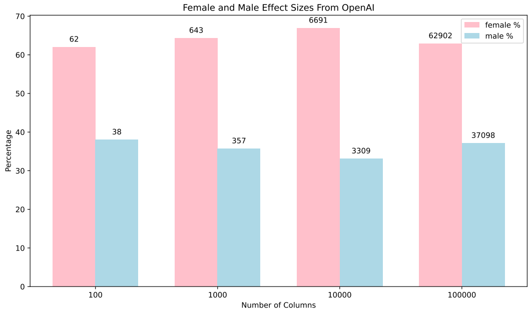


Figure 11: Most Frequent Gender-Associated Words for the OpenAI Model

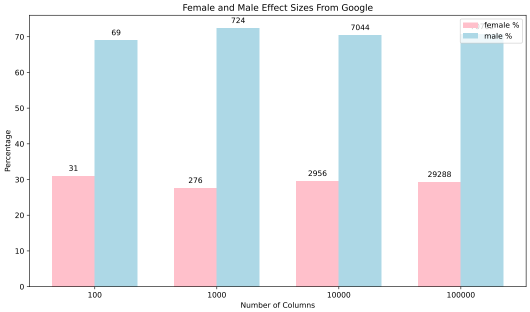


Figure 12: Most Frequent Gender-Associated Words for the Google Model

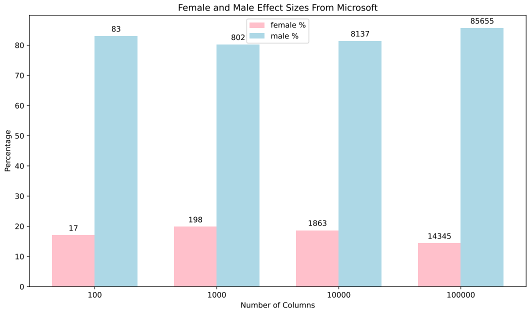


Figure 13: Most Frequent Gender-Associated Words for the Microsoft Model

### B.2 Frequencies of Race-Associated Words.

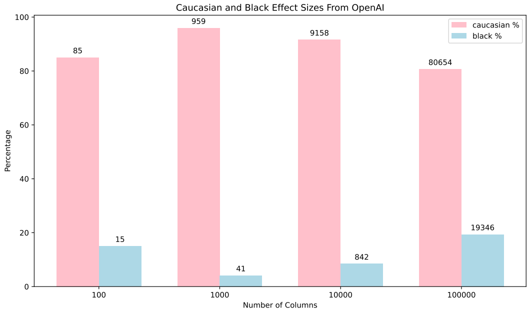


Figure 14: Most Frequent Race-Associated (Caucasian vs Black) Words for the OpenAI Model

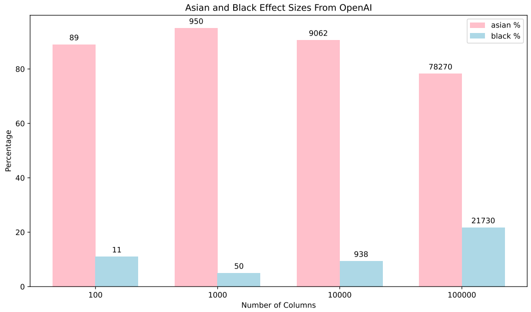


**Figure 15: Most Frequent Race-Associated (Asian vs Black) Words for the OpenAI Model**

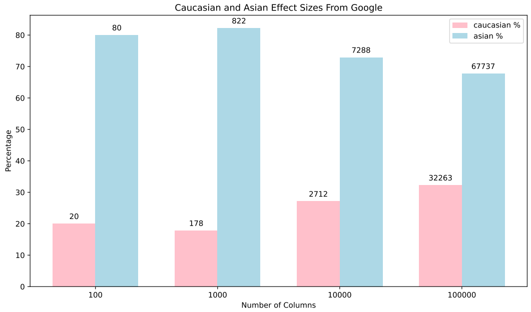


**Figure 16: Most Frequent Race-Associated (Caucasian vs Asian) Words for the Google Model**

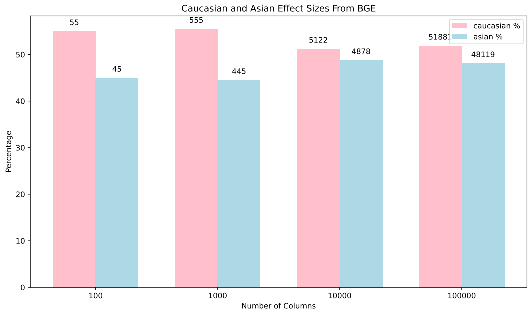


**Figure 17: Most Frequent Race-Associated (Caucasian vs Asian) Words for the BGE Model**

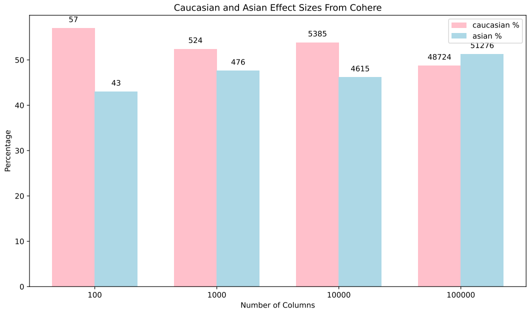


**Figure 18: Most Frequent Race-Associated (Caucasian vs Asian) Words for the Cohere Model**

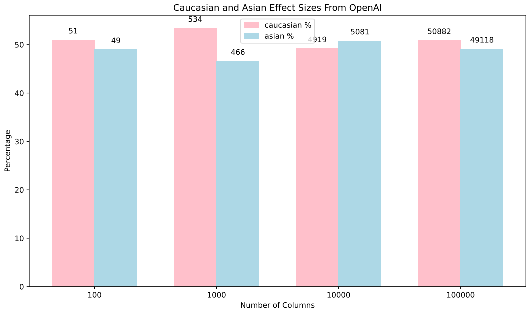


**Figure 19: Most Frequent Race-Associated (Caucasian vs Asian) Words for the OpenAI Model**

## B.3 Gender Bias by Frequency Range and Effect Size Across Embedding Models.

**Table 11: Gender-Associated (Female vs Male) by Range and Effect Size - BGE Model**

| Gender-Associated (Female vs Male) by Range and Effect Size - BGE Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | |
| num_words | female_0 | female_0.2 | female_0.5 | female_0.8 | male_0 | male_0.2 | male_0.5 | male_0.8 |
| 100 | 14 | 11 | 8 | 8 | 86 | 80 | 60 | 30 |
| 1000 | 213 | 145 | 66 | 30 | 787 | 667 | 444 | 159 |
| 10000 | 3972 | 2904 | 1614 | 815 | 6028 | 4811 | 2813 | 1151 |
| 100000 | 53631 | 42980 | 28055 | 15825 | 46369 | 35333 | 20293 | 8998 |

**Table 12: Gender-Associated (Female vs Male) by Range and Effect Size - Cohere Model**

| Gender-Associated (Female vs Male) by Range and Effect Size - Cohere Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | |
| num_words | female_0 | female_0.2 | female_0.5 | female_0.8 | male_0 | male_0.2 | male_0.5 | male_0.8 |
| 100 | 21 | 16 | 10 | 10 | 79 | 66 | 28 | 16 |
| 1000 | 290 | 177 | 86 | 36 | 710 | 569 | 289 | 112 |
| 10000 | 3427 | 2325 | 1148 | 539 | 6573 | 5240 | 3144 | 1545 |
| 100000 | 33160 | 24315 | 14378 | 8604 | 66840 | 56283 | 39141 | 23695 |

## Table 13: Gender-Associated (Female vs Male) by Range and Effect Size - Microsoft Model

| Gender-Associated (Female vs Male) by Range and Effect Size - Microsoft Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | |
| num_words | female_0 | female_0.2 | female_0.5 | female_0.8 | male_0 | male_0.2 | male_0.5 | male_0.8 |
| 100 | 17 | 10 | 8 | 4 | 83 | 66 | 28 | 3 |
| 1000 | 198 | 87 | 36 | 11 | 802 | 599 | 276 | 86 |
| 10000 | 1863 | 990 | 396 | 175 | 8137 | 6536 | 3471 | 1332 |
| 100000 | 14345 | 8553 | 4705 | 2895 | 85655 | 75282 | 50949 | 26358 |

## B.4 Race Bias by Frequency Range and Effect Size Across Embedding Models.

## Table 14: Race-Associated (Caucasian vs Black) by Range and Effect Size - Cohere Model

| Race-Associated (Caucasian vs Black) by Range and Effect Size - Cohere Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Black | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 90 | 86 | 83 | 64 | 10 | 10 | 8 | 3 |
| 1000 | 985 | 976 | 928 | 709 | 15 | 14 | 9 | 3 |
| 10000 | 9746 | 9431 | 8383 | 5824 | 254 | 119 | 36 | 12 |
| 100000 | 85446 | 77539 | 60416 | 36263 | 14554 | 9132 | 4048 | 1493 |

## Table 15: Race-Associated (Caucasian vs Black) by Range and Effect Size - Google Model

| Race-Associated (Caucasian vs Black) by Range and Effect Size - Google Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Black | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 76 | 66 | 57 | 36 | 24 | 17 | 13 | 8 |
| 1000 | 886 | 815 | 631 | 395 | 114 | 64 | 25 | 11 |
| 10000 | 8814 | 8039 | 6236 | 3846 | 1186 | 666 | 224 | 72 |
| 100000 | 82104 | 73213 | 54736 | 32319 | 17896 | 11228 | 5133 | 2010 |

## Table 16: Race-Associated (Caucasian vs Black) by Range and Effect Size - Microsoft Model

| Race-Associated (Caucasian vs Black) by Range and Effect Size - Microsoft Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Black | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 90 | 89 | 84 | 67 | 10 | 10 | 7 | 1 |
| 1000 | 988 | 984 | 937 | 703 | 12 | 11 | 7 | 1 |
| 10000 | 9858 | 9655 | 8668 | 5887 | 142 | 73 | 23 | 2 |
| 100000 | 91012 | 83785 | 63970 | 34996 | 8988 | 4834 | 1787 | 556 |

## Table 17: Race-Associated (Caucasian vs Asian) by Range and Effect Size - BGE Model

| Race-Associated (Caucasian vs Asian) by Range and Effect Size - BGE Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Asian | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 |
| 100 | 55 | 26 | 2 | 0 | 45 | 20 | 11 | 0 |
| 1000 | 555 | 210 | 24 | 1 | 445 | 139 | 22 | 0 |
| 10000 | 5122 | 2519 | 587 | 94 | 4878 | 2293 | 487 | 74 |
| 100000 | 51881 | 31694 | 11091 | 2600 | 48119 | 28248 | 9534 | 2241 |

## Table 18: Race-Associated (Caucasian vs Asian) by Range and Effect Size - Cohere Model

| Race-Associated (Caucasian vs Asian) by Range and Effect Size - Cohere Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Asian | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 |
| 100 | 57 | 31 | 8 | 1 | 43 | 25 | 13 | 6 |
| 1000 | 524 | 258 | 56 | 6 | 476 | 211 | 46 | 13 |
| 10000 | 5385 | 3274 | 1060 | 202 | 4615 | 2643 | 870 | 212 |
| 100000 | 48724 | 32613 | 13457 | 3624 | 51276 | 35780 | 17645 | 7109 |

## Table 19: Race-Associated (Caucasian vs Asian) by Range and Effect Size - Microsoft Model

| Race-Associated (Caucasian vs Asian) by Range and Effect Size - Microsoft Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Caucasian | | | | Asian | | | |
| num_words | caucasian_0 | caucasian_0.2 | caucasian_0.5 | caucasian_0.8 | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 |
| 100 | 83 | 75 | 48 | 9 | 17 | 16 | 11 | 4 |
| 1000 | 932 | 832 | 461 | 113 | 68 | 37 | 16 | 6 |
| 10000 | 8789 | 7532 | 4098 | 981 | 1211 | 633 | 236 | 78 |
| 100000 | 74870 | 59300 | 29510 | 7497 | 25130 | 15330 | 7205 | 3178 |

## Table 20: Race-Associated (Asian vs Black) by Range and Effect Size - Cohere Model

| Race-Associated (Asian vs Black) by Range and Effect Size - Cohere Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Asian | | | | Black | | | |
| num_words | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 90 | 89 | 85 | 67 | 10 | 10 | 7 | 2 |
| 1000 | 984 | 980 | 951 | 752 | 16 | 13 | 8 | 2 |
| 10000 | 9745 | 9440 | 8498 | 5941 | 255 | 114 | 40 | 9 |
| 100000 | 86816 | 79657 | 64012 | 40730 | 13184 | 8338 | 3755 | 1300 |

## Table 21: Race-Associated (Asian vs Black) by Range and Effect Size - Google Model

| Race-Associated (Asian vs Black) by Range and Effect Size - Google Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Asian | | | | Black | | | |
| num_words | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 88 | 86 | 80 | 65 | 12 | 11 | 10 | 7 |
| 1000 | 980 | 958 | 890 | 670 | 20 | 15 | 12 | 8 |
| 10000 | 9598 | 9188 | 8043 | 5859 | 402 | 210 | 71 | 31 |
| 100000 | 90015 | 83964 | 69988 | 48838 | 9985 | 6088 | 2550 | 928 |

## Table 22: Race-Associated (Asian vs Black) by Range and Effect Size - Microsoft Model

| Race-Associated (Asian vs Black) by Range and Effect Size - Microsoft Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Asian | | | | Black | | | |
| num_words | asian_0 | asian_0.2 | asian_0.5 | asian_0.8 | black_0 | black_0.2 | black_0.5 | black_0.8 |
| 100 | 90 | 86 | 68 | 23 | 10 | 10 | 4 | 1 |
| 1000 | 967 | 923 | 666 | 195 | 33 | 17 | 6 | 1 |
| 10000 | 9476 | 8646 | 5875 | 2214 | 524 | 172 | 36 | 4 |
| 100000 | 83823 | 71859 | 46359 | 19632 | 16177 | 8562 | 2685 | 650 |

## B.3 Bias in Big Tech Industry.

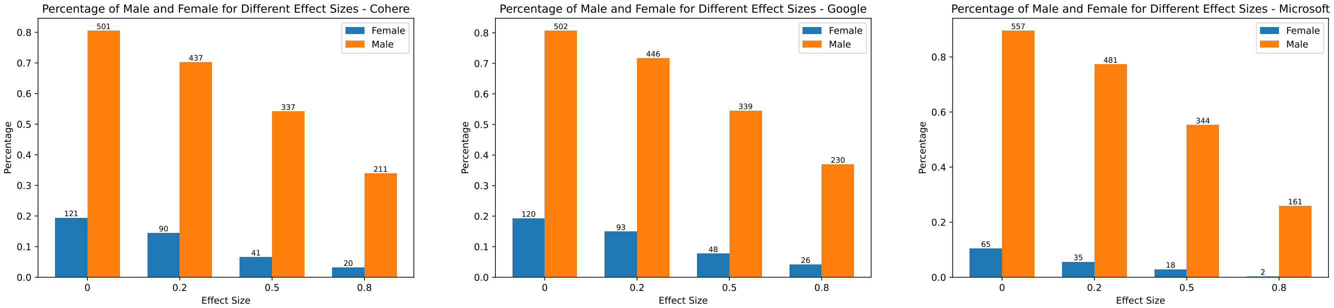


**Figure 20: Big Tech Gender-Associated (Female vs Male) words that show a more association with men rather than women: Generated from Cohere, Google, Microsoft**

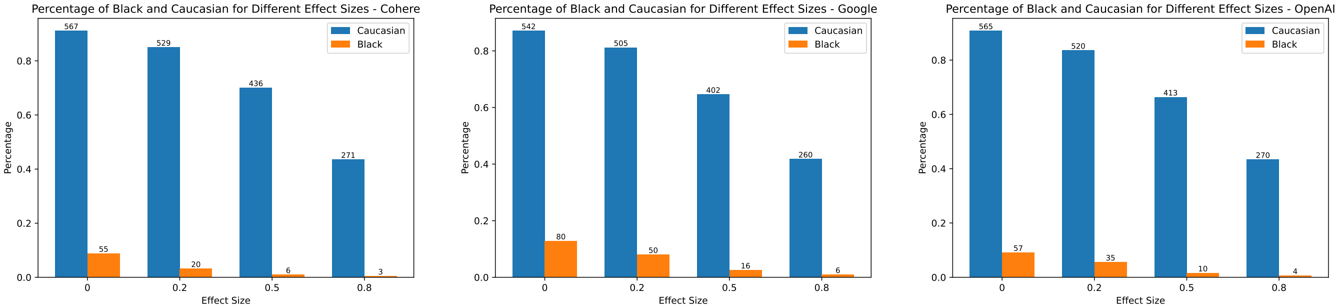


**Figure 21: Pairwise (Caucasian vs Black) comparisons of Big Tech words with Race-Associated words, showing Black individuals are underrepresented: Generated by Cohere, Google, OpenAI**

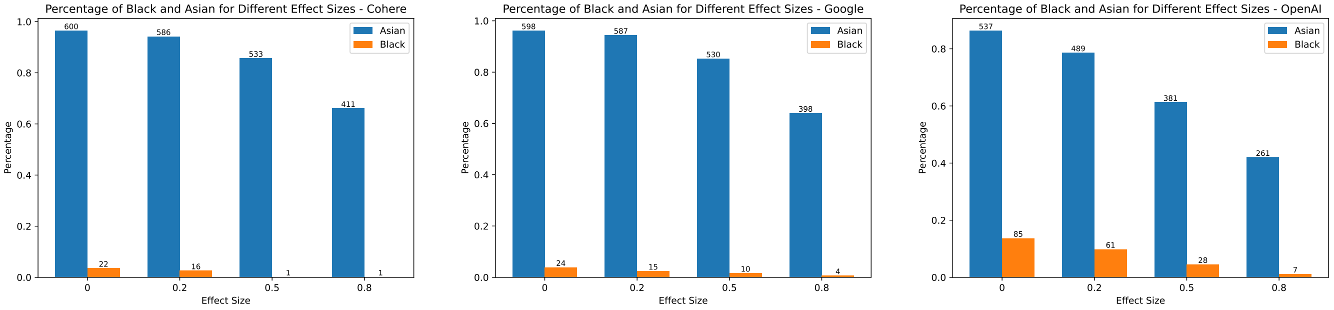
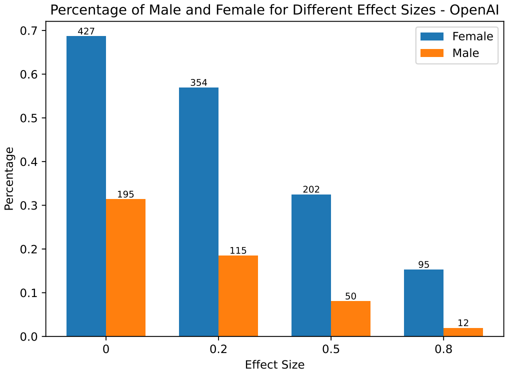


**Figure 22: Pairwise (Asian vs Black) comparisons of Big Tech words with Race-Associated words, showing Black individuals are underrepresented: Generated by Cohere, Google, OpenAI**



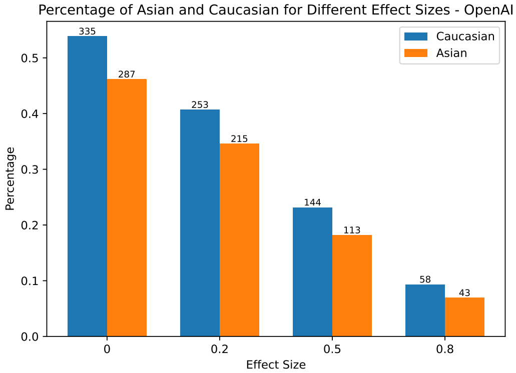


**Figure 23: Big Tech Gender-Associated (Female vs Male) words for OpenAI**

**Figure 24: Big Tech Race-Associated (Caucasian vs Asian) words for OpenAI**

### B.4 Bias in Higher Education.

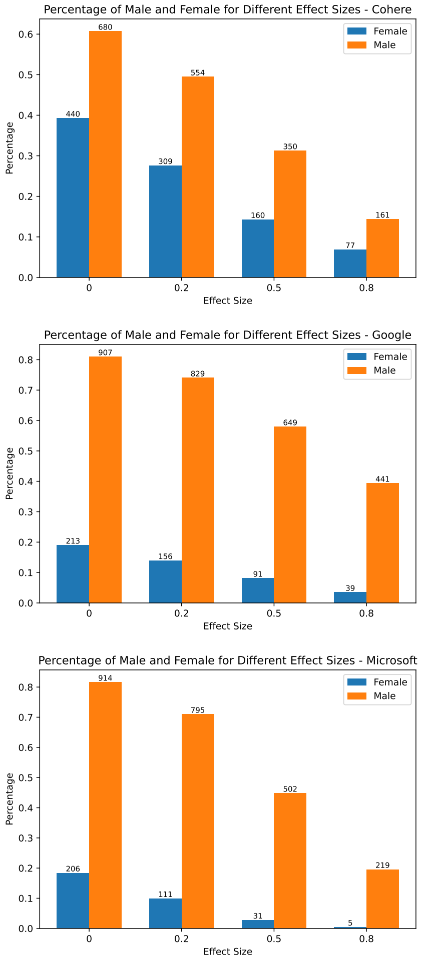


**Figure 25: Higher Education Gender-Associated (Female vs Male) words that show a more association with men rather than women: Generated from Cohere, Google, Microsoft**

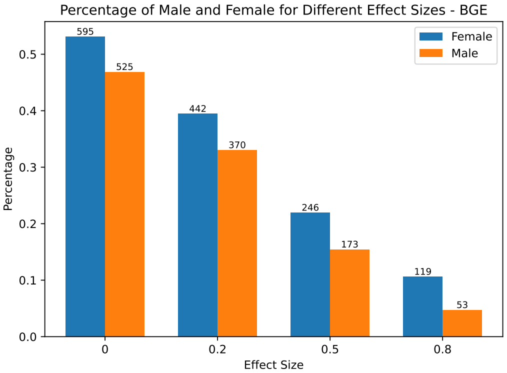


**Figure 26: Higher Education Gender-Associated (Female vs Male) words for BGE**

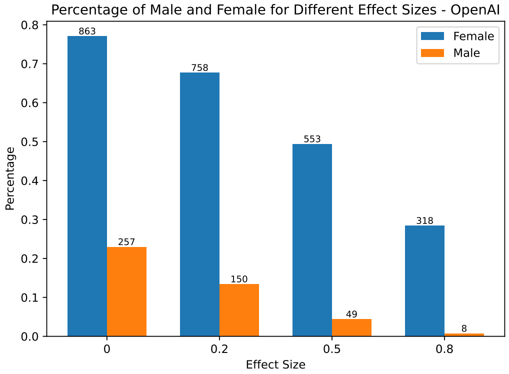


**Figure 27: Higher Education Gender-Associated (Female vs Male) words for OpenAI**

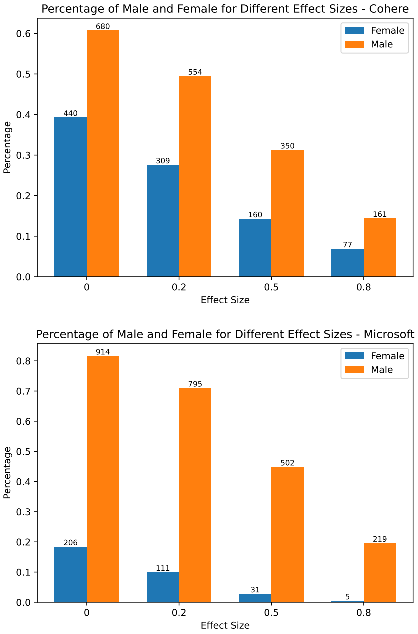


**Figure 28: Higher Education Race-Associated (Caucasian vs Black) words that show a more association with Caucasians rather than Blacks: Generated from Cohere, Google**

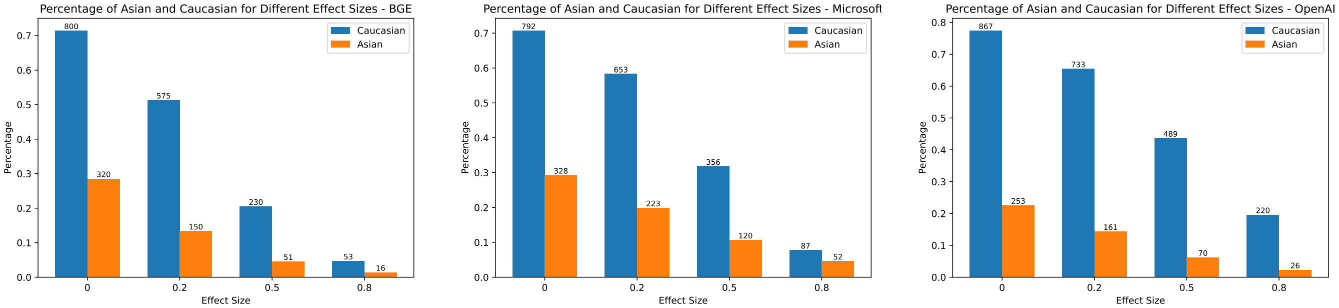


**Figure 29: Pairwise (Caucasian vs Asian) comparisons of Top University words with Race-Associated words, showing more Caucasian association rather than Asian: Generated by BGE, Microsoft, OpenAI**

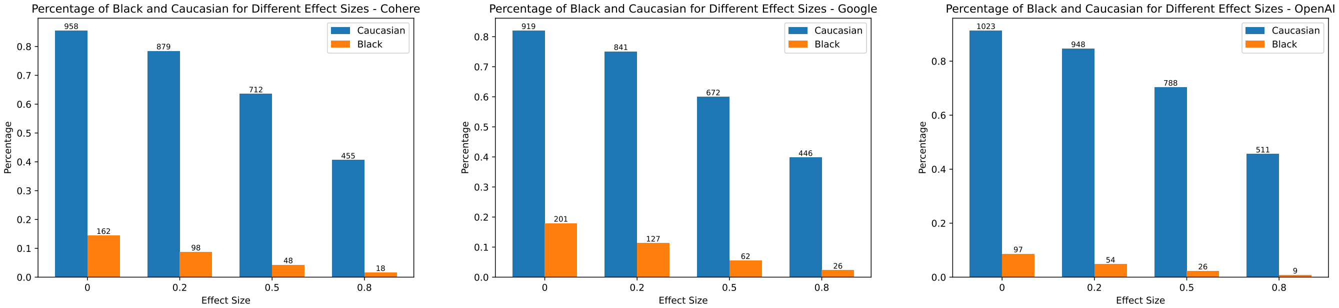


**Figure 30: Pairwise (Caucasian vs Black) comparisons of Top University words with Race-Associated words, showing Black individuals are underrepresented: Generated by Cohere, Google, OpenAI**

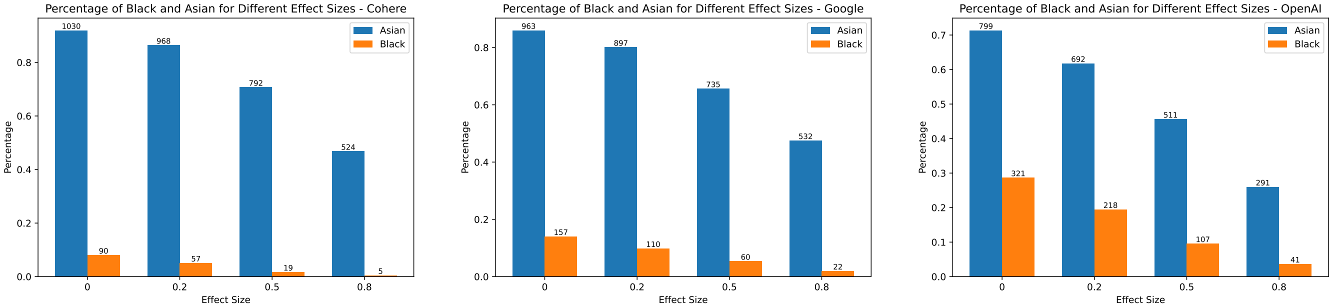


**Figure 31: Pairwise (Asian vs Black) comparisons of Big Tech words with Race-Associated words, showing Black individuals are underrepresented: Generated by Cohere, Google, OpenAI**