8 APPENDICES

8.1 Formal Definition of WEAT

We present a formal definition of Caliskan et al. [11]'s WEAT. Let X and Y be two sets of target words of equal size, and A, B be two sets of attribute words. Let $cos(\vec{a}, \vec{b})$ stand for the cosine similarity between the embeddings of words a and b. Here, the vector $\vec{a}$ is the embedding for word a. The test statistic is

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b})$$

A permutation test calculates the statistical significance of association $s(X, Y, A, B)$. The one-sided $p - value$ is

$$P = Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B))]$$

where $\{(X_i, Y_i)\}_i$ represents all the partitions of $X \cup Y$ in two sets of equal size. Random permutations of these stimuli sets represent the null hypothesis as if the biased associations did not exist so that we can perform a statistical significance test by measuring the unlikelihood of the null hypothesis, given the effect size of WEAT.

The effect size of bias is calculated as

$$ES = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std\_dev_{w \in X \bigcup Y} s(w, A, B)}$$

8.2 Formal Definition of EIBD

We first detect C_{11}'s intersectional biases W_{IB} with IBD. Then, we detect the biased attributes associated with only one constituent category of the intersectional group C_{11} (e.g., associated only with race S_{1n} - or only with gender S_{m1}). Each intersectional category C_{1n} has M constituent subcategories $S_{in}, i = 1, ...M$ and category C_{m1} has N constituent subcategories $S_{mj}, j = 1, ..., N$. S_{1n} and S_{m1} are the constituent subcategories of intersectional group C_{11}.

There are in total $M + N$ groups defined by all the single constituent subcategories. We use all $M + N$ groups to build WEFAT pairs $P_i = (S_{1n}, S_{in}), i = 1, ..., M$ and $P_j = (S_{m1}, S_{mj}), j = 1, ...N$. Then, we detect lists of words associated with each pair $W_i, i = 1, ...M$ and $W_j, j = 1, ..., N$ based on the same positive threshold t_{mn} used in IBD. We detect the attributes highly associated with the constituent subcategories S_{1n} and S_{m1} of the target intersectional group C_{11} from all $(M + N)$ WEFAT pairs. We define the words associated with emergent intersectional biases of group C_{11} as W_{EIB} and these words are identified by the formula

$$W_{EIB} = (\bigcup_{i=1}^{M}(W_{IB} - W_i)) \bigcup (\bigcup_{j=1}^{N}(W_{IB} - W_j))$$

where $W_i = \{w | s(w, S_{1n}, S_{in}) > t_{mn}, w \in W_{IB}\}$
and $W_j = \{w | s(w, S_{m1}, S_{mj}) > t_{mn}, w \in W_{IB}\}$

8.3 Random-Effects Model Details

Each effect size is calculated by

$$ES_i = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std\_dev_{w \in X \bigcup Y} s(w, A, B)}$$

The estimation of in-sample variance is V_i, which is the square of $std\_dev_{w \in X \bigcup Y} s(w, A, B)$. We use the same principle as estimation of the variance components in ANOVA to measure the between-sample variance $\sigma^2_{between}$, which is calculated as:

$$\sigma^2_{between} = \begin{cases} \frac{Q - (N - 1)}{c} & if \quad Q \geq N - 1 \\ 0 & if \quad Q < N - 1 \end{cases}$$

where

$$W_i = \frac{1}{V_i}$$

$$c = \sum W_i - \frac{\sum W_i^2}{\sum W_i} \quad \& \quad Q = \sum W_i ES_i^2 - \frac{(\sum W_i ES_i)^2}{\sum W_i}$$

The weight v_i assigned to each WEAT is the inverse of the sum of estimated in-sample variance V_i and estimated between-sample variance in the distribution of random-effects $\sigma^2_{between}$.

$$v_i = \frac{1}{V_i + \sigma^2_{between}}$$

CES, which is the sum of the weighted effect sizes divided by the sum of all weights, is then computed as

$$CES = \frac{\sum_{i=1}^{N} v_i ES_i}{\sum_{i=1}^{N} v_i}$$

To derive the hypothesis test, we calculate the standard error (SE) of CES as the square root of the inverse of the sum of the weights.

$$SE(CES) = \sqrt{\frac{1}{\sum_{i=1}^{N} v_i}}$$

Based on the central limit theorem, the limiting form of the distribution of $\frac{CES}{SE(CES)}$ is the standard normal distribution [46]. Since we notice that some CES are negative, we use a two-tailed $p - value$ which can test the significance of biased associations in two directions. The two-tailed $p - value$ of the hypothesis that there is no difference between all the contextualized variations of the two sets of target words in terms of their relative similarity to two sets of attribute words is given by the following formula, where Φ is the standard normal cumulative distribution function and SE stands for the standard error.

$$P_{combined}(X, Y, A, B) = 2 \times [1 - \Phi(|\frac{CES}{SE(CES)}|)]$$

8.4 Meta-Analysis Details for CEAT

In this section, we first construct all CEAT in the main paper (C1-C10,I1-I4) with sample size $N = 1,000$ to provide a comparison of results with different sample sizes. We report CES d and combined $p - value$ p in Table 2. We replicate these results with $N = 1,000$ instead of using the original $N = 10,000$ to show that even with $N = 1,000$, we get valid results. Accordingly, we proceed to calculate all types of biases associated with intersectional groups based on the attributes used in original WEAT. We notice that there are five tests which are significant with sample size $N = 10,000$ but insignificant with sample size $N = 1,000$. They are C10 with Bert, C4 with GPT, C7 with GPT-2, I3 with GPT-2 and I4 with GPT-2. We also notice that CES of same test can be different with different sample size but all differences are smaller than 0.1.

We also construct four types of supplementary CEAT for all pairwise combinations of six intersectional groups: African American females (AF), African American males (AM), Mexican American

Table 2: CEAT from main paper (C1-C10,I1-I4) with sample size $N = 1,000$ as opposed to the $N = 10,000$ hyper-parameter in the main paper. We report the CES (d) and combined $p - values$ of all CEAT (p) in the main paper with sample size $N = 1,000$. We observe that all of the results are consistent with the CES and $p - values$ reported in the main paper on Table 1. Light, medium, and dark gray shading of combined d values (CES) indicates small, medium, and large effect size, respectively. There are five tests which are significant with sample size $N = 10,000$ but not significant with sample size $N = 1,000$. However, these have small effect sizes and as a result we don't expect statistical significance. According to our experiments, the Spearman correlation between WEAT's effect size and $p - value$ is $\rho = 0.99$. Smaller effect sizes are expected to have insignificant p-values. Accordingly, all of the results under $N = 1,000$ are consistent with the main findings. The notable yet consistent differences are C10 with Bert, C4 with GPT, C7 with GPT-2, I3 with GPT-2, and I4 with GPT-2. CES varies minimally with different sample size (N), but the differences of the results are smaller than 0.1, suggesting the degree of effect size remains consistent. In edge cases, where statistical significance or effect size is close to a significance threshold, gradually increasing N, in increments of $N = +500$ would provide more reliable results. $A\_$ stands for African Americans. $E\_$ stands for European Americans. $M\_$ stands for Mexican Americans. $\_F$ stands for females. $\_M$ stands for males.

Test	**ELMo**		**BERT**		**GPT**		**GPT-2**	
	d	p	d	p	d	p	d	p
C1: Flowers/Insects, P/U* - Attitude	1.39	$< 10^{-30}$	0.96	$< 10^{-30}$	1.05	$< 10^{-30}$	0.13	$< 10^{-30}$
C2: Instruments/Weapons, P/U* - Attitude	1.56	$< 10^{-30}$	0.93	$< 10^{-30}$	1.13	$< 10^{-30}$	-0.28	$< 10^{-30}$
C3: EA/AA names, P/U* - Attitude	0.48	$< 10^{-30}$	0.45	$< 10^{-30}$	-0.11	$< 10^{-30}$	-0.20	$< 10^{-30}$
C4: EA/AA names, P/U* - Attitude	0.16	$< 10^{-30}$	0.49	$< 10^{-30}$	0.00	0.70	-0.23	$< 10^{-30}$
C5: EA/AA names, P/U* - Attitude	0.12	$< 10^{-30}$	0.04	$< 10^{-2}$	0.05	$< 10^{-4}$	-0.17	$< 10^{-30}$
C6: Males/Female names, Career/Family	1.28	$< 10^{-30}$	0.91	$< 10^{-30}$	0.21	$< 10^{-30}$	0.34	$< 10^{-30}$
C7: Math/Arts, Male/Female terms	0.65	$< 10^{-30}$	0.42	$< 10^{-30}$	0.23	$< 10^{-30}$	0.00	0.81
C8: Science/Arts, Male/Female terms	0.32	$< 10^{-30}$	-0.07	$< 10^{-4}$	0.26	$< 10^{-30}$	-0.16	$< 10^{-30}$
C9: Mental/Physical disease, Temporary/Permanent	0.99	$< 10^{-30}$	0.55	$< 10^{-30}$	0.07	$< 10^{-2}$	0.04	0.04
C10: Young/Old people's names, P/U* - Attitude	0.11	$< 10^{-19}$	0.00	0.90	0.04	$< 10^{-2}$	-0.17	$< 10^{-30}$
I1: AF/EM, AF/EM intersectional	1.24	$< 10^{-30}$	0.76	$< 10^{-30}$	0.05	$< 10^{-3}$	0.05	0.06
I2: AF/EM, AF emergent/EM intersectional	1.24	$< 10^{-30}$	0.70	$< 10^{-30}$	-0.12	$< 10^{-30}$	0.03	0.26
I3: MF/EM, MF/EM intersectional	1.30	$< 10^{-30}$	0.69	$< 10^{-30}$	-0.08	$< 10^{-30}$	0.36	$< 10^{-30}$
I4: MF/EM, MF emergent/EM intersectional	1.52	$< 10^{-30}$	0.87	$< 10^{-30}$	0.14	$< 10^{-27}$	-0.26	$< 10^{-30}$

*Unpleasant and pleasant attributes used to measure valence and attitudes towards targets [28].

females (MF), Mexican American males (MM), European American females (EF), European American males (EM). We use two intersectional groups as two target social groups. For each pairwise combination, we build four CEAT : first, measure attitudes with words representing pleasantness and unpleasantness as two attribute groups (as in C1); second, measure career and family associations that are particularly important in gender stereotypes with the corresponding two attribute groups (as in C6); third, similar to the career-family stereotypes for gender, measure math and arts associations that are particularly important in gender stereotypes with the corresponding two attribute groups (as in C7); fourth, similar to the math-arts stereotypes for gender, measure science (STEM) and arts associations that are particularly important in gender stereotypes with the corresponding two attribute groups (as in C8). We report the CES (d) and combined $p - values$ (p) in Table 2 with sample size $N = 1,000$. All of these attributes are from the C1, C6, C7 and C8 WEAT of Caliskan et al. [11].

9 STIMULI

The stimuli used to represent targets and attributes in CEAT (C1-C10) are taken from Caliskan et al.[11]. We construct four intersection-related CEAT for African American females and Mexican American females.

When conducting intersection-related CEAT , we use the names from Caliskan et al. [11] and Parada et al. [50] to represent the target intersectional groups. Caliskan et al.'s WEAT provides the female and male names of African Americans and European Americans from the first Implicit Association Test in 1998 [28]. Parada et al. provide the female and male names of Mexican Americans [50]. To determine and verify the gender of names, we use three gender checkers [38]. We only use the name as a target word in our experiments, if the name is categorized to belong to the same gender by all of the three checkers. Human subjects provide the validation set of intersectional attributes with ground truth information [25]. We use this validation set for evaluating the intersection-related CEAT, IBD and EIBD experiments. To follow the order of stereotype-congruity, we use European American males as the second target

Table 3: CEAT for intersectional groups with sample size $N = 1,000$. We construct 4 types of new CEAT with all pairwise combinations of intersectional groups. We use two intersectional groups as two target social groups. We use 1) pleasant/unpleasant 2) career/family 3) math/arts 4) science/arts as two attribute groups. We report the CES d and combined $p-value$ p. Light, medium, and dark gray shading of combined d values (CES) indicates small, medium, and large effect size respectively. $A\_$ stands for African Americans. $E\_$ stands for European Americans. $M\_$ stands for Mexican Americans. $\_F$ stands for females. $\_M$ stands for males.

Test	ELMo		BERT		GPT		GPT-2	
	d	p	d	p	d	p	d	p
EM/EF, P/U* - Attitude	-0.49	$< 10^{-30}$	-0.33	$< 10^{-30}$	-0.01	0.60	-0.53	$< 10^{-30}$
EM/EF, Career/Family	1.15	$< 10^{-30}$	0.73	$< 10^{-30}$	0.34	$< 10^{-30}$	0.41	$< 10^{-30}$
EM/EF, Math/Arts	0.44	$< 10^{-30}$	0.34	$< 10^{-30}$	0.13	$< 10^{-25}$	-0.41	$< 10^{-30}$
EM/EF, Science/Arts	0.37	$< 10^{-30}$	-0.11	$< 10^{-30}$	0.07	$< 10^{-6}$	-0.04	0.02
EM/AM, P/U* - Attitude	0.57	$< 10^{-30}$	0.40	$< 10^{-30}$	0.04	$< 10^{-2}$	-0.34	$< 10^{-30}$
EM/AM, Career/Family	0.32	$< 10^{-30}$	0.16	$< 10^{-30}$	-0.36	$< 10^{-30}$	0.42	$< 10^{-30}$
EM/AM, Math/Arts	-0.28	$< 10^{-30}$	-0.04	$< 10^{-2}$	-0.05	$< 10^{-30}$	-0.45	$< 10^{-30}$
EM/AM, Science/Arts	0.02	0.10	-0.18	$< 10^{-30}$	0.17	$< 10^{-30}$	-0.20	$< 10^{-30}$
EM/AF, P/U* - Attitude	-0.35	$< 10^{-30}$	0.10	$< 10^{-11}$	-0.12	$< 10^{-30}$	-0.60	$< 10^{-30}$
EM/AF, Career/Family	1.10	$< 10^{-30}$	0.90	$< 10^{-30}$	0.20	$< 10^{-30}$	0.62	$< 10^{-30}$
EM/AF, Math/Arts	0.11	$< 10^{-19}$	0.72	$< 10^{-30}$	0.14	$< 10^{-23}$	-0.62	$< 10^{-30}$
EM/AF, Science/Arts	0.56	$< 10^{-30}$	0.29	$< 10^{-30}$	0.24	$< 10^{-30}$	-0.19	$< 10^{-30}$
EM/MM, P/U* - Attitude	-0.15	$< 10^{-30}$	0.42	$< 10^{-30}$	-0.17	$< 10^{-30}$	-0.20	$< 10^{-30}$
EM/MM, Career/Family	0.01	0.46	0.28	$< 10^{-30}$	-0.32	$< 10^{-30}$	0.33	$< 10^{-30}$
EM/MM, Math/Arts	0.06	$< 10^{-5}$	-0.22	$< 10^{-30}$	0.45	$< 10^{-30}$	-0.38	$< 10^{-30}$
EM/MM, Science/Arts	0.21	$< 10^{-30}$	-0.27	$< 10^{-30}$	0.62	$< 10^{-30}$	-0.37	$< 10^{-30}$
EM/MF, P/U* - Attitude	-0.82	$< 10^{-30}$	-0.19	$< 10^{-30}$	-0.34	$< 10^{-30}$	-0.60	$< 10^{-30}$
EM/MF, Career/Family	1.14	$< 10^{-30}$	0.68	$< 10^{-30}$	0.09	$< 10^{-11}$	0.68	$< 10^{-30}$
EM/MF,Math/Arts	0.69	$< 10^{-30}$	0.27	$< 10^{-30}$	0.28	$< 10^{-30}$	-0.78	$< 10^{-30}$
EM/MF, Science/Arts	0.33	$< 10^{-30}$	0.11	$< 10^{-13}$	0.41	$< 10^{-30}$	-0.29	$< 10^{-30}$
EF/AM, P/U* - Attitude	0.95	$< 10^{-30}$	0.70	$< 10^{-30}$	0.06	$< 10^{-5}$	0.09	$< 10^{-17}$
EF/AM, Career/Family	-0.98	$< 10^{-30}$	-0.62	$< 10^{-30}$	-0.63	$< 10^{-30}$	0.11	$< 10^{-21}$
EF/AM, Math/Arts	-0.66	$< 10^{-30}$	-0.41	$< 10^{-30}$	-0.15	$< 10^{-30}$	-0.10	$< 10^{-30}$
EF/AM, Science/Arts	-0.30	$< 10^{-30}$	-0.08	$< 10^{-30}$	0.11	$< 10^{-13}$	-0.19	$< 10^{-30}$
EF/AF, P/U* - Attitude	0.09	$< 10^{-22}$	0.50	$< 10^{-30}$	-0.15	$< 10^{-30}$	-0.20	$< 10^{-30}$
EF/AF, Career/Family	0.04	$< 10^{-7}$	0.22	$< 10^{-30}$	-0.16	$< 10^{-30}$	0.33	$< 10^{-30}$
EF/AF, Math/Arts	-0.33	$< 10^{-30}$	0.39	$< 10^{-30}$	-0.01	0.44	-0.35	$< 10^{-30}$
EF/AF, Science/Arts	0.23	$< 10^{-30}$	0.43	$< 10^{-30}$	0.18	$< 10^{-30}$	-0.20	$< 10^{-30}$
EF/MM, P/U* - Attitude	0.38	$< 10^{-30}$	0.70	$< 10^{-30}$	-0.19	$< 10^{-30}$	0.32	$< 10^{-30}$
EF/MM, Career/Family	-1.10	$< 10^{-30}$	-0.45	$< 10^{-30}$	-0.65	$< 10^{-30}$	-0.02	0.14
EF/MM, Math/Arts	-0.34	$< 10^{-30}$	-0.55	$< 10^{-30}$	0.37	$< 10^{-30}$	-0.02	0.28
EF/MM, Science/Arts	-0.18	$< 10^{-30}$	-0.21	$< 10^{-30}$	0.54	$< 10^{-30}$	-0.36	$< 10^{-30}$
EF/MF, P/U* - Attitude	-0.42	$< 10^{-30}$	0.19	$< 10^{-30}$	-0.33	$< 10^{-30}$	-0.15	$< 10^{-30}$
EF/MF, Career/Family	-0.09	$< 10^{-30}$	-0.07	$< 10^{-30}$	-0.23	$< 10^{-30}$	0.43	$< 10^{-30}$
EF/MF, Math/Arts	0.30	$< 10^{-30}$	-0.05	$< 10^{-30}$	0.17	$< 10^{-30}$	-0.55	$< 10^{-30}$
EF/MF, Science/Arts	-0.01	0.40	0.25	$< 10^{-30}$	0.37	$< 10^{-30}$	-0.30	$< 10^{-30}$
AM/AF, P/U* - Attitude	-0.79	$< 10^{-30}$	-0.32	$< 10^{-30}$	-0.19	$< 10^{-30}$	-0.24	$< 10^{-30}$
AM/AF, Career/Family	0.94	$< 10^{-30}$	0.84	$< 10^{-30}$	0.50	$< 10^{-30}$	0.17	$< 10^{-30}$
AM/AF, Math/Arts	0.34	$< 10^{-30}$	0.79	$< 10^{-30}$	0.16	$< 10^{-30}$	-0.17	$< 10^{-30}$
AM/AF, Science/Arts	0.50	$< 10^{-30}$	0.47	$< 10^{-30}$	0.07	$< 10^{-7}$	-0.02	0.15
AM/MM, P/U* - Attitude	-0.72	$< 10^{-30}$	0.02	0.10	-0.20	$< 10^{-30}$	0.20	$< 10^{-30}$
AM/MM, Career/Family	-0.28	$< 10^{-30}$	0.16	$< 10^{-30}$	0.07	$< 10^{-7}$	-0.12	$< 10^{-30}$
AM/MM, Math/Arts	0.33	$< 10^{-30}$	-0.16	$< 10^{-30}$	0.51	$< 10^{-30}$	0.08	$< 10^{-9}$
AM/MM, Science/Arts	0.13	$< 10^{-30}$	-0.13	$< 10^{-30}$	0.45	$< 10^{-30}$	-0.16	$< 10^{-30}$
AM/MF, P/U* - Attitude	-1.15	$< 10^{-30}$	-0.57	$< 10^{-30}$	-0.38	$< 10^{-30}$	-0.22	$< 10^{-30}$
AM/MF, Career/Family	0.96	$< 10^{-30}$	0.56	$< 10^{-30}$	0.41	$< 10^{-30}$	0.27	$< 10^{-30}$
AM/MF, Math/Arts	0.87	$< 10^{-30}$	0.36	$< 10^{-30}$	0.31	$< 10^{-30}$	-0.38	$< 10^{-30}$
AM/MF, Science/Arts	0.30	$< 10^{-30}$	0.30	$< 10^{-30}$	0.27	$< 10^{-30}$	-0.14	$< 10^{-30}$
AF/MM, P/U* - Attitude	0.26	$< 10^{-30}$	0.33	$< 10^{-30}$	-0.04	$< 10^{-30}$	0.46	$< 10^{-30}$
AF/MM, Career/Family	-1.07	$< 10^{-30}$	-0.64	$< 10^{-30}$	-0.54	$< 10^{-30}$	-0.31	$< 10^{-30}$
AF/MM, Math/Arts	-0.03	0.03	-0.90	$< 10^{-30}$	0.37	$< 10^{-30}$	0.29	$< 10^{-30}$
AF/MM, Science/Arts	-0.38	$< 10^{-30}$	-0.56	$< 10^{-30}$	0.43	$< 10^{-30}$	-0.18	$< 10^{-30}$
AF/MF, P/U* - Attitude	-0.43	$< 10^{-30}$	-0.33	$< 10^{-30}$	-0.19	$< 10^{-30}$	-0.01	0.48
AF/MF, Career/Family	-0.15	$< 10^{-30}$	-0.31	$< 10^{-30}$	-0.06	$< 10^{-30}$	0.15	$< 10^{-30}$
AF/MF, Math/Arts	0.59	$< 10^{-30}$	-0.42	$< 10^{-30}$	0.16	$< 10^{-30}$	-0.25	$< 10^{-30}$
AF/MF, Science/Arts	-0.20	$< 10^{-30}$	-0.18	$< 10^{-30}$	0.22	$< 10^{-30}$	-0.15	$< 10^{-30}$
MM/MF, P/U* - Attitude	-0.77	$< 10^{-30}$	-0.59	$< 10^{-30}$	-0.15	$< 10^{-30}$	-0.44	$< 10^{-30}$
MM/MF, Career/Family	1.11	$< 10^{-30}$	0.40	$< 10^{-30}$	0.44	$< 10^{-30}$	0.42	$< 10^{-30}$
MM/MF, Math/Arts	0.62	$< 10^{-30}$	0.50	$< 10^{-30}$	-0.18	$< 10^{-30}$	-0.49	$< 10^{-30}$
MM/MF, Science/Arts	0.18	$< 10^{-30}$	0.41	$< 10^{-30}$	-0.19	$< 10^{-30}$	0.02	0.18

Table 4: CEAT for intersectional groups with sample size $N = 1,000$. We construct 4 types of new CEAT with all pairwise combinations of intersectional groups. We use two intersectional groups as two target social groups. We use 1) pleasant/unpleasant 2) career/family 3) math/arts 4) science/arts as two attribute groups. Each one of the four experiments with the neural language models is conducted using the same sample of sentences. We report the CES d and combined $p - value$ p. Light, medium, and dark gray shading of combined d values (CES) indicates small, medium, and large effect size respectively. $A\_$ stands for African Americans. $E\_$ stands for European Americans. $M\_$ stands for Mexican Americans. $\_F$ stands for females. $\_M$ stands for males.

Test	ELMo		BERT		GPT		GPT-2	
	d	p	d	p	d	p	d	p
EM/EF, P/U* - Attitude	-0.62	$< 10^{-30}$	-0.17	$< 10^{-30}$	-0.11	$< 10^{-30}$	-0.28	$< 10^{-30}$
EM/EF, Career/Family	1.07	$< 10^{-30}$	0.40	$< 10^{-30}$	0.27	$< 10^{-30}$	0.33	$< 10^{-30}$
EM/EF, Math/Arts	0.07	$< 10^{-30}$	0.12	$< 10^{-30}$	0.23	$< 10^{-30}$	-0.22	$< 10^{-30}$
EM/EF, Science/Arts	0.21	$< 10^{-30}$	-0.04	$< 10^{-30}$	0.12	$< 10^{-30}$	0.01	$< 10^{-2}$
EM/AM, P/U - Attitude	0.50	$< 10^{-30}$	0.37	$< 10^{-30}$	0.13	$< 10^{-30}$	-0.15	$< 10^{-30}$
EM/AM, Career/Family	0.19	$< 10^{-30}$	0.10	$< 10^{-30}$	-0.35	$< 10^{-30}$	0.30	$< 10^{-30}$
EM/AM, Math/Arts	-0.47	$< 10^{-30}$	-0.18	$< 10^{-30}$	0.13	$< 10^{-30}$	-0.30	$< 10^{-30}$
EM/AM, Science/Arts	-0.11	$< 10^{-30}$	-0.03	$< 10^{-14}$	0.14	$< 10^{-30}$	-0.20	$< 10^{-30}$
EM/AF, P/U - Attitude	-0.24	$< 10^{-30}$	0.32	$< 10^{-30}$	-0.26	$< 10^{-30}$	-0.26	$< 10^{-30}$
EM/AF, Career/Family	1.12	$< 10^{-30}$	0.40	$< 10^{-30}$	0.23	$< 10^{-30}$	0.66	$< 10^{-30}$
EM/AF, Math/Arts	0.07	$< 10^{-30}$	0.30	$< 10^{-30}$	0.28	$< 10^{-30}$	-0.52	$< 10^{-30}$
EM/AF, Science/Arts	0.55	$< 10^{-30}$	0.47	$< 10^{-30}$	0.21	$< 10^{-30}$	-0.35	$< 10^{-30}$
EM/MM, P/U - Attitude	-0.18	$< 10^{-30}$	0.37	$< 10^{-30}$	-0.36	$< 10^{-30}$	-0.12	$< 10^{-30}$
EM/MM, Career/Family	-0.08	$< 10^{-30}$	0.04	$< 10^{-21}$	-0.26	$< 10^{-30}$	0.23	$< 10^{-30}$
EM/MM, Math/Arts	-0.08	$< 10^{-30}$	-0.22	$< 10^{-30}$	0.54	$< 10^{-30}$	-0.47	$< 10^{-30}$
EM/MM, Science/Arts	0.16	$< 10^{-30}$	-0.09	$< 10^{-30}$	0.56	$< 10^{-30}$	-0.45	$< 10^{-30}$
EM/MF, P/U - Attitude	-0.73	$< 10^{-30}$	0.09	$< 10^{-30}$	-0.24	$< 10^{-30}$	-0.24	$< 10^{-30}$
EM/MF, Career/Family	1.06	$< 10^{-30}$	0.35	$< 10^{-30}$	0.06	$< 10^{-30}$	0.66	$< 10^{-30}$
EM/MF,Math/Arts	0.58	$< 10^{-30}$	0.09	$< 10^{-30}$	0.49	$< 10^{-30}$	-0.61	$< 10^{-30}$
EM/MF, Science/Arts	0.24	$< 10^{-30}$	0.26	$< 10^{-30}$	0.48	$< 10^{-30}$	-0.43	$< 10^{-30}$
EF/AM, P/U - Attitude	0.96	$< 10^{-30}$	0.51	$< 10^{-30}$	0.24	$< 10^{-30}$	0.12	$< 10^{-30}$
EF/AM, Career/Family	-1.00	$< 10^{-30}$	-0.31	$< 10^{-30}$	-0.57	$< 10^{-30}$	0.00	0.86
EF/AM, Math/Arts	-0.53	$< 10^{-30}$	-0.30	$< 10^{-30}$	-0.10	$< 10^{-30}$	-0.13	$< 10^{-30}$
EF/AM, Science/Arts	-0.28	$< 10^{-30}$	0.00	0.42	0.03	$< 10^{-13}$	-0.24	$< 10^{-30}$
EF/AF, P/U - Attitude	0.27	$< 10^{-30}$	0.47	$< 10^{-30}$	-0.17	$< 10^{-30}$	0.01	0.27
EF/AF, Career/Family	0.13	$< 10^{-30}$	0.01	0.037	-0.05	$< 10^{-30}$	0.45	$< 10^{-30}$
EF/AF, Math/Arts	0.00	0.85	0.19	$< 10^{-30}$	0.06	$< 10^{-30}$	-0.40	$< 10^{-30}$
EF/AF, Science/Arts	0.34	$< 10^{-30}$	0.50	$< 10^{-30}$	0.11	$< 10^{-30}$	-0.44	$< 10^{-30}$
EF/MM, P/U - Attitude	0.47	$< 10^{-30}$	0.52	$< 10^{-30}$	-0.25	$< 10^{-30}$	0.17	$< 10^{-30}$
EF/MM, Career/Family	-1.10	$< 10^{-30}$	-0.35	$< 10^{-30}$	-0.50	$< 10^{-30}$	-0.09	$< 10^{-30}$
EF/MM, Math/Arts	-0.15	$< 10^{-30}$	-0.34	$< 10^{-30}$	0.37	$< 10^{-30}$	-0.26	$< 10^{-30}$
EF/MM, Science/Arts	-0.05	$< 10^{-30}$	-0.06	$< 10^{-30}$	0.47	$< 10^{-30}$	-0.51	$< 10^{-30}$
EF/MF, P/U - Attitude	-0.18	$< 10^{-30}$	0.26	$< 10^{-30}$	-0.14	$< 10^{-30}$	0.02	$< 10^{-2}$
EF/MF, Career/Family	-0.13	$< 10^{-30}$	-0.05	$< 10^{-30}$	-0.19	$< 10^{-30}$	0.46	$< 10^{-30}$
EF/MF, Math/Arts	0.52	$< 10^{-30}$	-0.03	$< 10^{-12}$	0.32	$< 10^{-30}$	-0.52	$< 10^{-30}$
EF/MF, Science/Arts	0.04	$< 10^{-30}$	0.30	$< 10^{-30}$	0.38	$< 10^{-30}$	-0.52	$< 10^{-30}$
AM/AF, P/U - Attitude	-0.63	$< 10^{-30}$	-0.06	$< 10^{-30}$	-0.39	$< 10^{-30}$	-0.11	$< 10^{-30}$
AM/AF, Career/Family	1.05	$< 10^{-30}$	0.32	$< 10^{-30}$	0.53	$< 10^{-30}$	0.41	$< 10^{-30}$
AM/AF, Math/Arts	0.49	$< 10^{-30}$	0.48	$< 10^{-30}$	0.16	$< 10^{-30}$	-0.24	$< 10^{-30}$
AM/AF, Science/Arts	0.57	$< 10^{-30}$	0.52	$< 10^{-30}$	0.08	$< 10^{-30}$	-0.18	$< 10^{-30}$
AM/MM, P/U - Attitude	-0.67	$< 10^{-30}$	-0.02	$< 10^{-3}$	-0.48	$< 10^{-30}$	0.04	$< 10^{-21}$
AM/MM, Career/Family	-0.27	$< 10^{-30}$	-0.06	$< 10^{-30}$	0.13	$< 10^{-30}$	-0.08	$< 10^{-30}$
AM/MM, Math/Arts	0.38	$< 10^{-30}$	-0.04	$< 10^{-30}$	0.45	$< 10^{-30}$	-0.15	$< 10^{-30}$
AM/MM, Science/Arts	0.24	$< 10^{-30}$	-0.06	$< 10^{-30}$	0.44	$< 10^{-30}$	-0.27	$< 10^{-30}$
AM/MF, P/U - Attitude	-1.03	$< 10^{-30}$	-0.28	$< 10^{-30}$	-0.37	$< 10^{-30}$	-0.09	$< 10^{-30}$
AM/MF, Career/Family	0.98	$< 10^{-30}$	0.25	$< 10^{-30}$	0.38	$< 10^{-30}$	0.42	$< 10^{-30}$
AM/MF, Math/Arts	0.91	$< 10^{-30}$	0.26	$< 10^{-30}$	0.39	$< 10^{-30}$	-0.37	$< 10^{-30}$
AM/MF, Science/Arts	0.31	$< 10^{-30}$	0.30	$< 10^{-30}$	0.34	$< 10^{-30}$	-0.28	$< 10^{-30}$
AF/MM, P/U - Attitude	0.09	$< 10^{-30}$	0.04	$< 10^{-11}$	-0.09	$< 10^{-30}$	0.16	$< 10^{-30}$
AF/MM, Career/Family	-1.15	$< 10^{-30}$	-0.36	$< 10^{-30}$	-0.47	$< 10^{-30}$	-0.47	$< 10^{-30}$
AF/MM, Math/Arts	-0.14	$< 10^{-30}$	-0.52	$< 10^{-30}$	0.31	$< 10^{-30}$	0.11	$< 10^{-30}$
AF/MM, Science/Arts	-0.39	$< 10^{-30}$	-0.56	$< 10^{-30}$	0.35	$< 10^{-30}$	-0.10	$< 10^{-30}$
AF/MF, P/U - Attitude	-0.41	$< 10^{-30}$	-0.23	$< 10^{-30}$	0.02	$< 10^{-8}$	0.02	$< 10^{-3}$
AF/MF, Career/Family	-0.27	$< 10^{-30}$	-0.05	$< 10^{-30}$	-0.15	$< 10^{-30}$	0.06	$< 10^{-30}$
AF/MF, Math/Arts	0.48	$< 10^{-30}$	-0.22	$< 10^{-30}$	0.26	$< 10^{-30}$	-0.17	$< 10^{-30}$
AF/MF, Science/Arts	-0.29	$< 10^{-30}$	-0.21	$< 10^{-30}$	0.26	$< 10^{-30}$	-0.13	$< 10^{-30}$
MM/MF, P/U - Attitude	-0.62	$< 10^{-30}$	-0.27	$< 10^{-30}$	0.11	$< 10^{-30}$	-0.15	$< 10^{-30}$
MM/MF, Career/Family	1.11	$< 10^{-30}$	0.31	$< 10^{-30}$	0.30	$< 10^{-30}$	0.49	$< 10^{-30}$
MM/MF, Math/Arts	0.63	$< 10^{-30}$	0.30	$< 10^{-30}$	-0.04	$< 10^{-30}$	-0.25	$< 10^{-30}$
MM/MF, Science/Arts	0.09	$< 10^{-30}$	0.35	$< 10^{-30}$	-0.08	$< 10^{-30}$	-0.04	$< 10^{-14}$

*Unpleasant and pleasant attributes used to measure valence and attitudes towards targets from Greenwald et al. [28].

group and use the attributes associated with their intersectional biases as the second attribute set in intersection-related CEAT. There are only three emergent intersectional biases associated with European American males in the validation set, which doesn't provide a sufficient number of stimuli. A small set of stimuli does not satisfy the requirements for generating statistically significant concept representation and WEATs. Related stimuli details are discussed in the dataset and stimuli sections of the main paper. In addition, if the size of the first attribute set is smaller than that of the attributes of European American males, we randomly select an equal number of attributes associated with the intersectional biases of European American males. WEAT requires equal-sized sets of attributes.

9.1 CEAT I1

We use the frequent given names of African American females and European American males as two target social groups and use the attributes associated with the intersectional biases of African American females and attributes associated with the intersectional biases of European American males as the two attribute groups.

Since 'assertive' is associated with both African American females and European American males, we do not include it in this test.

- **African American females**: Aisha, Keisha, Lakisha, Latisha, Latoya, Malika, Nichelle, Shereen, Tamika, Tanisha, Yolanda, Yvette
- **European American males**: Andrew, Brad, Frank, Geoffrey, Jack, Jonathan, Josh, Matthew, Neil, Peter, Roger, Stephen
- **Intersectional biases of African American females**: aggressive, athletic, bigbutt, confident, darkskinned, fried-chicken, ghetto, loud, overweight, promiscuous, unfeminine, unintelligent, unrefined
- **Intersectional biases of European American males**: all-American, arrogant, attractive, blond, high-status, intelligent, leader, privileged, racist, rich, sexist, successful, tall

9.2 CEAT I2

We use the frequent given names of African American females and European American males as two target groups. We use attributes associated with emergent intersectional biases of African American females and attributes associated with intersectional biases of European American males as two attribute groups.

Since 'assertive' is associated with emergent intersectional bias of African American females and intersectional bias of European American males, we do not include it in this test.

- **African American females**: Aisha, Keisha, Lakisha, Latisha, Latoya, Malika, Nichelle, Shereen, Tamika, Tanisha, Yolanda, Yvette
- **European American males**: Andrew, Brad, Frank, Geoffrey, Jack, Jonathan, Josh, Matthew, Neil, Peter, Roger, Stephen
- **Emergent intersectional biases of African American females**: aggressive, bigbutt, confident, darkskinned, fried-chicken, overweight, promiscuous, unfeminine
- **Intersectional biases of European American males**: arrogant, blond, high-status, intelligent, racist, rich, successful, tall

9.3 CEAT I3

We use the frequent given names of Mexican American females and European American males as the target groups and the words associated with their intersectional biases as the attribute groups.

Since 'attractive' is associated with intersectional biases of both Mexican American females and European American males, we do not include it in this test.

- **Mexican American females**: Adriana, Alejandra, Alma, Brenda, Carolina, Iliana, Karina, Liset, Maria, Mayra, Sonia, Yesenia
- **European American males**: Andrew, Brad, Frank, Geoffrey, Jack, Jonathan, Josh, Matthew, Neil, Peter, Roger, Stephen
- **Intersectional biases of Mexican American females**: cook, curvy, darkskinned, feisty, hardworker, loud, maids, promiscuous, sexy, short, uneducated, unintelligent
- **Intersectional biases of European American males**: all-American, arrogant, blond, high-status, intelligent, leader, privileged, racist, rich, sexist, successful, tall

9.4 CEAT I4

We use the frequent given names of Mexican American females and European American males as target groups. We use words associated with the emergent intersectional biases of Mexican American females and words associated with the intersectional biases of European American males as the two attribute groups.

- **Mexican American females**: Adriana, Alejandra, Alma, Brenda, Carolina, Iliana, Karina, Liset, Maria, Mayra, Sonia, Yesenia
- **European American males**: Andrew, Brad, Frank, Geoffrey, Jack, Jonathan, Josh, Matthew, Neil, Peter, Roger, Stephen
- **Emergent intersectional biases of Mexican American females**: cook, curvy, feisty, maids, promiscuous, sexy
- **Intersectional biases of European American males**: arrogant, assertive, intelligent, rich, successful, tall

9.5 IBD and EIBD

We detect the attributes associated with the intersectional biases and emergent intersectional biases of African American females and Mexican American females in GloVe SWE. We assume that there are three subcategories under the race category (African American, Mexican American, European American) and two subcategories under the gender category (female, male). We use the frequent given names to represent each intersectional group. Again, we note that, in future work we'd generalize this work to *n* subcategories under each category. Further, in future work, instead of categorizing people into social groups, we'd like to explore representing individuals in social data with continuous real-valued variables as opposed to associating them with category labels.

- **African American females**: Aisha, Keisha, Lakisha, Latisha, Latoya, Malika, Nichelle, Shereen, Tamika, Tanisha, Yolanda, Yvette
- **African American males**: Alonzo, Alphonse, Hakim, Jamal, Jamel, Jerome, Leroy, Lionel, Marcellus, Terrence, Tyrone, Wardell

- **European American females**: Carrie, Colleen, Ellen, Emily, Heather, Katie, Megan, Melanie, Nancy, Rachel, Sarah, Stephanie
- **European American males**: Andrew, Brad, Frank, Geoffrey, Jack, Jonathan, Josh, Matthew, Neil, Peter, Roger, Stephen
- **Mexican American females**: Adriana, Alejandra, Alma, Brenda, Carolina, Iliana, Karina, Liset, Maria, Mayra, Sonia, Yesenia
- **Mexican American males**: Alberto, Alejandro, Alfredo, Antonio, César, Jesús, José, Juan, Miguel, Pedro, Rigoberto, Rogelio

We list all the attributes used in validation experiments. These are stimuli associated with different social groups and a set of random words that are not expected to be associated with social groups. These random attributes are borrowed from the insect target category of the 'universally accepted stereotypes' IAT, which is a baseline WEAT. We use 98 words in total in the validation experiments.

We also list the probability of random chance of correct classification in parentheses next to each social group. The goal here is to present the success rate of the machine learning detection method in a meaningful way. To calculate the probability of random chance of correct classification, we treat the detection task as if this were a one-vs-all verification task, which is one type of a classification task, that aims to verify the membership of a data point in the main class. All the other data points from the open world, that don't belong to the main class we are testing for, can belong to any other class. However, in reality, our detection method is a semi-supervised non-parametric prediction method which relies on the threshold *t*. Some of the attributes of some social groups overlap. For example, 'assertive' is associated with emergent intersectional biases of African American females and also associated with emergent intersectional biases of European American males. As a result, the sum of all random chances of different groups is not 100%.

- **Biases of Females**: attractive, caring, dependent, emotional, feminine, jealous, manipulative, materialistic, motherly, petite, soft, submissive, talkative (random chance of correct classification: 13.3%)
- **Biases of Males**: aggressive, ambitious, arrogant, fixer-upper, high-status, intelligent, leader, messy, provider, respected, sexist, tall, unfaithful (random chance of correct classification: 13.3%)
- **Biases of African Americans**: athletic, criminals, dangerous, gangsters, ghetto, lazy, loud, poor, tall, uneducated, unrefined, violent (random chance of correct classification: 12.2%)
- **Emergent Intersectional Biases of African American Females**: aggressive, assertive, bigbutt, confident, darkskinned, fried-chicken, overweight, promiscuous, unfeminine (random chance of correct classification: 9.2%)
- **Intersectional Biases of African American Females**: aggressive, assertive, athletic, bigbutt, confident, darkskinned, fried-chicken, ghetto, loud, overweight, promiscuous, unfeminine, unintelligent, unrefined (random chance of correct classification: 14.3%)
- **Emergent Intersectional Biases of African American Males**: darkskinned, hypersexual, rapper (random chance of correct classification: 3.1%)
- **Intersectional Biases of African American Males**: athletic, criminals, dangerous, darkskinned, gangsters, hypersexual, lazy, loud, poor, rapper, tall, unintelligent, violent (random chance of correct classification: 13.3%)
- **Biases of European Americans**: all-American, arrogant, attractive, blond, blue-eyes, high-status, ignorant, intelligent, overweight, patronizing, privileged, racist, red-neck, rich, tall (random chance of correct classification: 15.3%)
- **Emergent Intersectional Biases of European American Females**: ditsy (random chance of correct classification: 1.0%)
- **Intersectional Biases of European American Females**: arrogant, attractive, blond, ditsy, emotional, feminine, high-status, intelligent, materialistic, petite, racist, rich, submissive, tall (random chance of correct classification: 14.3%)
- **Emergent Intersectional Biases of European American Males**: assertive, educated, successful (random chance of correct classification: 3.1%)
- **Intersectional Biases of European American Males**: all-American, arrogant, assertive, attractive, blond, educated, high-status, intelligent, leader, privileged, racist, rich, sexist, successful, tall (random chance of correct classification: 15.3%)
- **Biases of Mexican Americans**: darkskinned, day-laborer, family-oriented, gangster, hardworker, illegal-immigrant, lazy, loud, macho, overweight, poor, short, uneducated, unintelligent (random chance of correct classification: 14.3%)
- **Emergent Intersectional Biases of Mexican American Females**: cook, curvy, feisty, maids, promiscuous, sexy (random chance of correct classification: 6.1%)
- **Intersectional Biases of Mexican American Females**: attractive, cook, curvy, darkskinned, feisty, hardworker, loud, maids, promiscuous, sexy, short, uneducated, unintelligent (random chance of correct classification: 13.3%)
- **Emergent Intersectional Biases of Mexican American Males**: drunks, jealous, promiscuous, violent (random chance of correct classification: 4.1%)
- **Intersectional Biases of Mexican American Males**: aggressive, arrogant, darkskinned, day-laborer, drunks, hardworker, illegal-immigrant, jealous, macho, poor, promiscuous, short, uneducated, unintelligent, violent (random chance of correct classification: 15.3%)
- **Random (Insects)**: ant, bedbug, bee, beetle, blackfly, caterpillar, centipede, cockroach, cricket, dragonfly, flea, fly, gnat, hornet, horsefly, locust, maggot, mosquito, moth, roach, spider, tarantula, termite, wasp, weevil (random chance of correct classification: 25.5%)

10 OPEN SOURCE CODE, DATA, AND DOCUMENTATION

https://github.com/weiguowilliam/CEAT is the link to our open source git repository. Code and links to datasets are available in the project repository. In addition, answers to frequently asked questions about the details of extracting the contextualized word embeddings are documented. The extracted embeddings for the stimuli take up approximately $\sim 50GB$ memory.