More of the Same: Persistent Representational Harms Under Increased Representation

Jennifer Mickel[1], Maria De-Arteaga[2], Liu Leqi[3], Kevin Tian[3]

[1]Eleuther AI, [2]Universitat Ramon Llull, ESADE, [3]University of Texas at Austin jamickel@utexas.edu

Contributions

Motivation: Understanding representational bias in settings where group membership is unknown is critical to understanding representational biases in downstream tasks

Contributions:

- GAS(P): an evaluation methodology for surfacing representational biases at the distribution level
- Empirical analysis of representation bias in the gender and occupation domain using GAS(P). We find:
  1. Women are more represented than men across occupations in generated personas and biographies
  2. Women and men are represented differently and these differences are statistically significant
  3. Some of the statistically significant words identified correspond to stereotypes and harms outlined in the social science literature

Data Generation

1. For each occupation, prompt, and gender triple, generate 100 instances
   - "Describe a [OCCUPATION] who is a [woman/man/non-binary person] as if you are writing a biography"
   - "Generate a persona of a [OCCUPATION] who is a [woman/man/non-binary person]"
2. For each occupation and prompt pair generate text until at least 100 instances are associated per group
   - "Describe a [OCCUPATION] as if you are writing a biography"
   - "Generate a persona of a [OCCUPATION]"
3. Repeat 1-2 for each model evaluated (GPT-3.5, GPT-4o-mini, Llama-3.1-70b)

Group Member Association

Associate each generation with a gender if the number of gendered pronouns and honorifics for a gender outnumber the number of gendered pronouns and honorifics for other genders.

Subset Representational Bias Score

- **Motivation**: Identify how similar the set of calibrated marked words are for each associated group to each of the specified groups

Chamfer Distance is defined as where the distance metric, d_x, is cosine distance:

$$CH(C,T) = \frac{1}{|C|}\sum_{c\in C}\min_{t\in T} d_x(c,t)$$

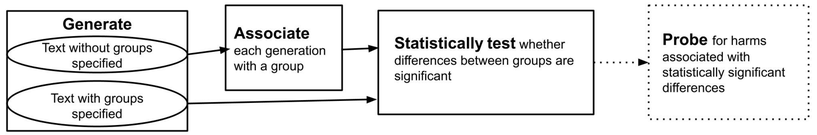
We define the Subset Representation Bias Score where $S, A, B \in \mathbb{R}^d$ as:

$$\Delta(S\|A,B) = CH(S,A) - CH(S,B)$$

Statistical Significance Testing

1. Identify statistically significant words using the Calibrated Marked Words method for each gender, occupation, and model triple
   - The Calibrated Marked Words method builds on the Marked Words method introduced by [1], which uses weighted log-odds to identify marked words by 1) using a hybrid prior consisting of both the English language and the generated text; and 2) adding a calibration step through hyperparameter tuning.
2. Calculate the Subset Representational Bias Score using the set of Calibrated Marked Words identified for each gender, model, and occupation triple
   1. Calculate the Chamfer distance between each associated gender and specified gender (typically $CH(A_M, S_M) < CH(A_F, S_M)$ and $CH(A_M, S_F) > CH(A_F, S_F)$)
   2. Calculate SRB Score

GAS(P) Methodology



The GAS(P) evaluation methodology works as follows:

1. **Generate** text both with and without specifying a group in the prompt
2. **Associate** each generation with a group
3. **Statistically test** whether the representational markers for each group persist when groups are not explicitly prompted and are statistically significantly different across groups
4. **Probe** these surfaced representational differences and relate these to patterns associated with harms discussed in the social science scholarship.

How are people represented?

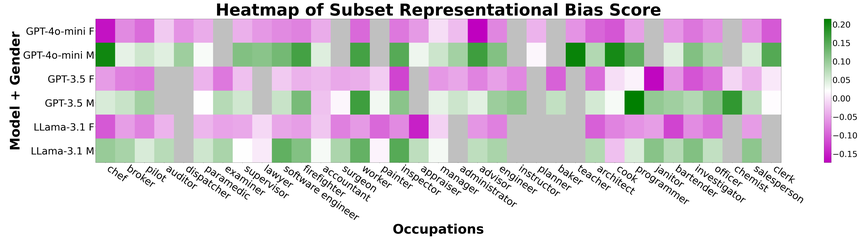


Figure: The Subset Representational Bias Score is displayed for each occupation, model, and associated gender pair. A negative value (pink) indicates that the statistically significant words are closer to specified women, and a positive value (green) indicates that the statistically significant words are closer to specified men. The gray boxes refer to occupation model pairs that did not meet our criteria to collect data.

The figure above demonstrates the Subset Representational Bias Score across gender, occupation, and model. We find that the difference in SRB Scores between men and women is statistically significant. The figure below demonstrates the percent change in SRB Score from GPT-3.5 to GPT-4o-mini, and we find that these differences are statistically significant.

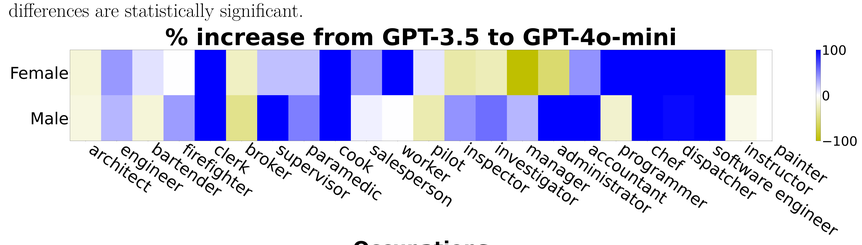


Figure: Percent change in the Subset Representational Bias Score from GPT-3.5 to GPT-4o-mini. Percentage increase (blue) means that the similarity to the corresponding gender (i.e. associated women to specified women) increased from GPT-3.5 to GPT-4o-mini.

Who is represented?

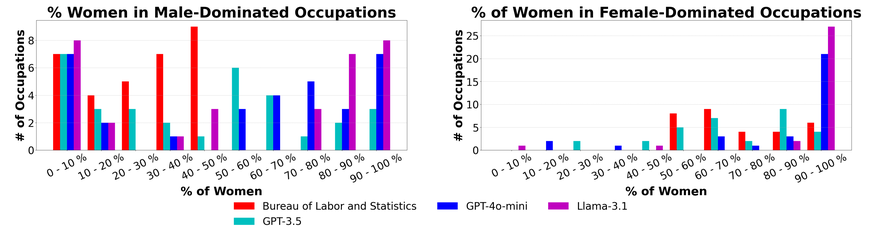


Figure: The graphs illustrate the distribution of women's representation across various occupations by grouping percentages into percent deciles (e.g., 0–10%, 10–20%, and so on) and counting the number of occupations within each decile. Graph (a) shows the percentage of women in male-dominated occupations, and Graph (b) shows the percentage of women in female-dominated occupations.

Probe: What are the implications of how people are represented?

We cluster the Calibrated Marked words using k-means with 1500 clusters. The table below contains clusters that are at least 50% more prevalent for one gender. Some of these clusters align with stereotypes (women as "compassionate" and "empathetic" [2]) and harms ("inspiration" [3]) outlined in the social science literature.

	GPT-3.5			GPT-4o-mini			Llama-3.1		
Cluster	**% F**	**% M**	**#**	**% F**	**% M**	**#**	**% F**	**%M**	**#**
empathy, empathize, empathetic	100.0	0.0	6	100.0	0.0	7	100.0	0.0	8
woman, actress, female	100.0	0.0	7	100.0	0.0	17	90.91	9.09	11
shortterm, short	100.0	0.0	3	25.0	75.0	4	20.0	80.0	5
advocate, advocates	100.0	0.0	4	93.33	6.67	15	100.0	0.0	3
inspired, inspiration	100.0	0.0	3	100.0	0.0	3	100.0	0.0	6
tireless, tirelessly	100.0	0.0	4	100.0	0.0	5	83.33	16.67	6
decisions, decisionmaking, determination	100.0	0.0	4	71.43	28.57	7	75.0	25.0	4
she, her, shes	100.0	0.0	29	100.0	0.0	28	100.0	0.0	27
career, careers	100.0	0.0	3	91.67	8.33	12	88.89	11.11	9
inclusive, inclusion, inclusivity	100.0	0.0	7	100.0	0.0	12	100.0	0.0	7
climbing, hiking, hiker	100.0	0.0	5	75.0	25.0	4	75.0	25.0	8
prestigious	100.0	0.0	3	80.0	20.0	5	100.0	0.0	3
practicing, training	100.0	0.0	5	100.0	0.0	7	100.0	0.0	12
demands, demanding	100.0	0.0	4	80.0	20.0	5	100.0	0.0	3
diversity, minorities, multicultural	100.0	0.0	5	100.0	0.0	16	100.0	0.0	7
herself	100.0	0.0	29	100.0	0.0	26	100.0	0.0	27
compassion, compassionate	100.0	0.0	6	100.0	0.0	3	100.0	0.0	10
yoga	100.0	0.0	7	100.0	0.0	15	100.0	0.0	16
passion, passions, passionate	90.0	10.0	10	100.0	0.0	8	100.0	0.0	6
families, familys, family	66.67	33.33	3	15.38	84.62	13	33.33	66.67	3
pursuits, pursuit, pursue, pursued, pursuing	60.0	40.0	5	90.0	10.0	10	80.0	20.0	15
award, awardwinning, awards, accolades	60.0	40.0	5	75.0	25.0	4	85.71	14.29	7
inspire, inspires, inspiring	60.0	40.0	5	87.5	12.5	8	100.0	0.0	4
countless, boundless	33.33	66.67	3	100.0	0.0	3	75.0	25.0	4
husband, wife, spouse	33.33	66.67	15	0.0	100.0	9	11.11	88.89	9
playing, gamer, gaming, games	0.0	100.0	8	0.0	100.0	7	0.0	100.0	8
basketball, sports	0.0	100.0	4	0.0	100.0	4	0.0	100.0	10
tied, tie, ties	0.0	100.0	3	100.0	0.0	3	60.0	40.0	10
his, himself, him	0.0	100.0	29	0.0	100.0	28	0.0	100.0	27
charismatic	0.0	100.0	3	0.0	100.0	4	0.0	100.0	5

References

1. M. Cheng, E. Durmus, and D. Jurafsky. "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models." ACL. 2023
2. C. Löffler and T. Greitemeyer. "Are women the more empathetic gender? The effects of gender role expectations." *Current Psychology*, 42(1):2020-231, 2023.
3. J. Byrne, S. Fattoum, and M. Garcia. "Role models and women entrepreneurs: Entrepreneurial superwoman has her say." *Jourand of Small Business Management*, 57(1):154–184, 2019.