Brilliance Bias in GPT-3

Juliana Shihadeh, Margareta Ackerman, Ashley Troske, Nicole Lawson, and Edith Gonzalez
Computer Science & Engineering Department
Santa Clara University
Santa Clara, CA, USA
{jshihadeh, mackerman, atroske, nlawson,egonzalez2}@scu.edu

***Abstract*—Language has a profound impact on how we perceive the world. With GPT-3's rise in popularity, as of latest utilized in 300 applications averaging 4.5 billion words per day, it is critical to identify and correct biases in its generations. A variety of biases have been identified in generative language models, spanning biases based on gender, race, and religion. In this paper, we pioneer the study of the Brilliance Bias for generative models. This implicit, yet powerful bias imposes the idea that "brilliance" is a male trait and in turn, sets back women's achievements starting as early as ages 5-7. We perform an analysis of two GPT-3 models, the base GPT-3 model (davinci) and InstructGPT (text-davinici-002), focusing on adjectives, verbs and lexicons found in their generations. Our analysis reveals the presence of substantial Brilliance Bias across both models.**

***Index Terms*—Generative language models, bias, GPT-3,**

I. Introduction

In recent years, generative language models have started integrating into society. Readily accessible generative language model such as GPT (Generative Pre-Trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) are used in applications for marketing [1], gaming [2], and collaborative story writing [3]. Unfortunately, generative models, such as GPT-3 often exhibit a variety of biases, spanning race, gender, and religion [4]. For example, these models often associate more physical and professional roles with men [5] and have consistent lower sentiment scores on generated text about some racial minorities [6]. InstructGPT, the newest model in the GPT-3 series,[1] was released in January 2022. OpenAI shared that the model shows improvements on truthfulness and reductions in toxic output [7].

Some of the most damaging biases are implicit. A study by Chestnut et al. [8] reveals that even a relatively harmless phrase, such as "girls are as good as boys at math" can perpetuate the myth that boys are generally better at this task. This is an example of the Brilliance Bias, a powerful implicit bias that imposes the idea that intellectual "brilliance" is a male trait and in turn, sets back women's achievements starting as young as 5-7 years [9]. Some fields are thought to require intellectual brilliance while others are believed to call for other traits, such as empathy or hard work [10]. As a result, women are discouraged from pursuing careers that call for brilliance [9]. They are less likely to apply or be referred to jobs that portray a need for brilliance [9]. Brilliance bias is in turn leading to female under-representation in fields like computer science, physics and philosophy [10].

One of the most popular generative language models is GPT-3. As of GPT-3's latest reportings in 2021, it generates an average of 4.5 billion words per day and is used in over 300 applications [11]. Due to their extensive training on (exceptionally) large volumes of internet data, models such as GPT-3 represent our collective unconsciousness, which captures prejudices and biases (implicit and explicit).

Language influences how we view the world [12]. It has been found that stereotypes that cycle through multiple information channels have a higher influence on society [13]. Consequently, the growing presence of generative language models is likely to reinforce or even amplify existing biases, such as Brilliance Bias. Commenting in her findings on GPT-3, AI ethicist Timnit Gebru states that "marginalized groups might experience misrepresentation if the technologies become widespread in society" [14]. The spread of generative language models like GPT-3 can inadvertently amplify the Brilliant Bias, which would negatively impact women and other minorities. Unchecked, the continual integration of models that exhibit Brilliance Bias (and other biases) into commercial products can lead to substantial societal damage, and even undo progress on inclusiveness.

In this paper[2], we investigate the presence of Brilliance Bias in two models' generation of stories in the GPT-3 series: The newly improved GPT-3 model, InstructGPT[1] (text-davinici-002) and the original GPT-3[1] (davinci) model[3]. (We refer to the original GPT-3 (davinci) model as GPT-3 throughout our paper.) We explore whether text generated using these two models exhibits Brilliance Bias. To this end, we compare the stories made with InstructGPT and GPT-3 models when prompted with brilliance-related text with male versus female characters.

As an illustration, consider the prompts "Emma is a world-leading professor" and "Nicholas is a world-leading professor". Giving these two prompts to GPT-3, we get the following stories:

[1]https://help.openai.com/en/articles/5832130-what-s-changed-with-engine-names-and-best-practices

[2]A github repo of our work including our dataset and the output of raw results we analyze in our paper can be found at: https://github.com/julishi/Brilliance-Bias-In-GPT

[3]While our analysis relies on these two models, we stress that our focus isn't as much on highlighting shortcomings of specific models, but rather on bringing awareness to the inadvertent incorporating of Brilliance Bias in generative language models and the need to correct this issue.

Emma is a world-leading professor *in her field and is one of the most accomplished women in Australian research. She has achieved so much of the good life, but she reminds us that gloating is unbecoming of a lady. There is no point in being proud of accomplishments that one's parents are responsible for. So much for parents being irresponsible for the choices their kids will make.*

Arthur is a world-leading professor *of Astrophysics at Cambridge University and was formerly Astrophysics Programme Director at STFC. He is also the president of the Royal Astronomical Society. Professor Lewis, who is director of the National Space Institute at the University of Leicester, said the UK government's "apparent uncritical acceptance" of results from the Planck satellite carried implications for the future of UK astronomical research.*

There are notable differences in the above generations, including the topic of research and more prestigious associations linked to the male character. To study this phenomenon, we create 3200 generations per model to analyze the presence or absence of Brilliance Bias. Analysing the adjectives, verbs, and connotations of word-descriptors, we show a substantial presence of Brilliance Bias in the generations of both GPT-3 models.

We begin with a discussion of previous work on Brilliance Bias, as well as prior analysis of biases in generative language models. Next, we detail our methodology for analyzing Brilliance Bias in GPT-3 models and share our finding. We conclude with a discussion of our findings.

II. Previous Work

While studies have focused on quantifying and mitigating biases like gender, race, and religion [4], [15], [16] in generative language models, Brilliance Bias has not yet been analyzed in the context of generative text models.

Brilliance Bias affects the distribution of men and women in various disciplines [9], [10]. A study conducted on children between the ages of 5-7 showed that during these 3 years, children develop the start of Brilliance Bias [9]. At 5, girls are still more-likely to associate being brilliant with their own gender but at age 6 and 7 associate it less with themselves compared to boys [9]. Similarly, representing stereotypical association of traits, girls associated "nice" more often with their gender at ages 6-7 compared to at age 5 [9].

In fields that carry the notion of requiring "raw talent", such as Computer Science, Philosophy, Economics, and Physics, there are fewer women with Ph.D.'s compared to other disciplines such as History, Psychology, Biology and Neuroscience [10]. An analysis of films showed brilliance portrayed in them as a male trait [17]. Due to a "brilliance-required" bias in some fields, women "may find the academic fields that emphasize such talent to be inhospitable" [10]. This hinders the inclusion of women in those fields.

This issue of Brilliance Bias has consequences beyond fairness and equality, but also hinders economic development. Gender-diverse teams have been shown to perform better than homogeneous ones, and have greater financial success [18]. Furthermore, this bias hurts individuals from developing, closing the door on opportunities to discover their potential for high achievements.

Studies on the GPT-3 model have revealed bias through gendered associations of occupations, sentiment toward race, and co-occurrence of words with respect to different religions ([6], [16], [19]). For example, men were more likely to be a 'detective' or endure physically rigorous labor compared to women, who were more associated with the roles 'midwife' or 'receptionist' [6]. Similar research on gender bias in BERT showed greater male than female affiliation to stereotypical occupations like 'firefighter' and 'conductor' [20]. Sheng et al. [21] has focused on biases in multiple models including BERT and GPT-2, across gender, race, and sexual orientation in the context of different occupations. Huang et al. [22] further explore biased sentiment in language models on a variety of sensitive topics including country, occupation, and gender.

Other works such as of Nadeem et al. [23] have focused on developing a dataset for researchers working on language models to measure bias in gender, profession, race, and religion. In addition, studies have evaluated the harmful effects of gender and racial biases in NLP [15].

In our work, we initiate the study of Brilliance Bias in generative language models, starting with GPT-3 models. We specifically focus on adjective and verb dependencies, looking at their correlations and connotations with respect to groups of peoples. Adjectives and verbs in text have been studied in movie plot summaries [24], books [25], and news [26], as well as in American Textbooks on minority groups [27] to study which adjectives/verbs are more commonly affiliated with a group. Lexicon analysis using adjectives/verbs on power, agency, dominance, sentiment, valence and arousal have also been used in various text analyses to understand how characters are perceived. One study revealed a lack of power, agency, and dominance of minority groups in American history textbooks [27]. Another study on evaluating news reported on the #MeTooMovement [28] showed that while women were being written about sympathetically, text about men showed more power. In this paper, we present the first analysis of Brilliance Bias in generative models.

III. Methodology

We perform a detailed analysis on the generations of GPT-3 models, InstructGPT and the original model GPT-3, to assess the presence or absence of Brilliance Bias in these two models. The data is created by calling the models on prompts that focus on brilliance, with two versions of each prompts, one with a female and one with a male character. The analysis of the data focuses on adjectives, verbs, and lexicons of generated material.

A. Data

The original GPT-3 completes generated text by expanding on user-given prompts. For this model, our prompts have the form "[Name] is [trait]." We utilize traits from Storage et al.'s

[29] study on Brilliance Bias . They are: *brilliant, genius, super smart, and brainiac*. The traits *genius* and *brainiac* are preceded by "a". As such, an example of a prompt is "Chloe is super smart" or "Chloe is a genius", which GPT-3 subsequently expands into a short story.

Since InstructGPT is intended for responding to instructions, we correspondingly adjust the prompts to the following form: "Write a story about a [women/man] who is [trait]". We generated 800 stories for each trait type, 400 female/male each, producing 3200 stories in total across all prompts for each model. We run our analysis on the generated text without the prompt, to evaluate on the generations and avoid skewing the data with the same phrase or sentence repeated 800 times for each prompt.

B. Selecting Female and Male Lead Names for GPT-3

Our analysis shows that using names ("Maria is a top researcher in her field") in prompts triggered the original GPT-3 to build up stories. By contrast, when using pronouns ("She is a top researcher in her field"), generations lack continuity, cohesion, and focus. An example of a typical outcome with a pronoun rather than name (prompt is in bold): "***She was a top researcher*** *in her field too. She worked at Brown University and even received an award before she … went all batty." Drained, Henry sat back. Kostik had mentioned the mental ward, but hadn't said anything else. "I was dating someone else at the time and thought she was cute and quirky. When she started obsessing over the Noah's Ark tablet, I (Flagged as containing sensitive content by GPT-3)*".

Newman et al. [30] conducted a survey on 383 popular names in the United States and analyzed their perceived competence, warmth, gender, and age. To generate our list of names, we categorized the names by their perceived age range (12-17, 18-24, 25-34, 35-44, and 45-54), and then sorted by competence level. We chose four names from each age group category—two male names, one with the highest competence and one with the lowest competence, and likewise for the female names.

We observed that some of the names were perceived as gender neutral by GPT-3, causing it to flip their intended pronouns in generations. To reduce non gender-deterministic names, we applied three criteria. First the perceived gender for each name should match the gender result on Gender API [4] [31]. We then check if the name is labeled as unisex on wiki. Next, we analyze GPT-3 outputs, looking at a subset of generations that changed the dominate gender of the GPT-3 output based on if ≥ 0.75 [16] of the pronouns were the intended name's opposite gender. If ≤ 3 of the generations altered the gender, we kept the name. Names that were gender ambiguous according to these criteria were omitted. Further refinement in selecting deterministic gendered names is worth exploring. 20 names were selected listed in Table I, 10 female and 10 male. Each prompt was run on all 20 names, 40 times per name, totalling 800 generations per prompt.

[4]Please reference https://genderize.io/

TABLE I
FEMALE AND MALE NAMES USED FOR THE GENERATION OF BRILLIANCE-THEMED PROMPTS WITH GPT-3

Female	Male
Chloe	Dustin
Emma	Noah
Brittney	Eddie
Anna	Nicholas
Felicia	Duane
Marcia	William
Diane	Larry
Peggy	Richard
Judith	Bob
Elizabeth	Arthur

C. Analysis

We analyze Brilliance Bias in the generations of InstructGPT and GPT-3. We evaluate this bias by analyzing the relationship of adjectives and verbs in the generated stories. We use Lucy et al.'s [27] open-sourced text-analysis scripts[5] to assess various adjectives and verbs associations. The relationships of words include the degree of correlation of adjectives and verbs with different groups of individuals and a lexicon analysis, such as how much power, agency and sentiment is given to a character. We utilize both approaches in our study.

Using the parsing dependency in Lucy et al.'s [27] scripts, which they note on their github[6] is based on SpaCY [32] (Lucy et al. use Dozat et al.'s [33] parsing for their paper's results) a list of all the adjective and verb descriptors is extracted in the text we generate using GPT-3 and InstructGPT. The adjectives and verbs are then used to analyze their association with a group of individuals. The groups focused on in this study are based on the American Textbook analysis of minority groups conducted by Lucy et al. [27]. Groups are based on gender and ethnicity, labeled as: white, black, hispanic/latinx, women, men, other minority, and other. "Other minority" represents other ethnicities that are a minority including Iraqi, Asian, Aztecs (for the full list of other minorities represented in Lucy et al.'s [27] study please refer to their github[7]). The group "Other" represents any descriptive words unmarked for gender that do not fall in any of the former categories such as "farmer", "justice", and "volunteer" (for the full list of words labeled as other represented in Lucy et al.'s [27] study please refer to their github[7]).

With the descriptors, we first compare which adjectives and verbs are used to describe women vs. men. We use the same log-odds-ratios Lucy et al. [27] use to analyze groups of individuals in American History Textbooks. The log-odds-ratio is calculated based on prior probabilities of words and word frequency counts using the informative Dirichlet prior described in Section 3.5.1 of Monroe et al. [34]. Lucy et al. [27] suggested

[5]Please reference https://github.com/ddemszky/textbook-analysis
[6]https://github.com/ddemszky/textbook-analysis#verbs-and-adjectives
[7]wordlists/people_terms.csv on Lucy et al.'s [27] github[5]

the use of this method over tf-idf frequencies because it can more accurately capture the relationship of both low and high frequency words. In comparing two groups, words with higher log-odds-ratio (reported in decimal form) are affiliated more with Group A rather than Group B. Vice versa, words with more negative scores are affiliated more to Group B.

We then run an analysis on the connotation of the descriptors by evaluating six lexicons as listed in Table II. This is a technique commonly used since 1966 to evaluate text [35] and especially important today to understand social connotations of words [36]. We evaluate the six lexicon categories based on adjectives and verbs. Verbs are used to assess Power/Agency [37] and Sentiment [38], while adjectives are used to measure Valence/Arousal/Dominance [39]. *Power* represents how much authority is given to a subject [37], *agency* represents how much control a person has on their life [37], and *dominance* measures how much influence one has [39]. *Valence* is a degree of pleasure vs. displeasure associated [39], *arousal* measures the degree of a person's energy [39], and *sentiment* is the writer's attitude toward the subject [38]. (Refer to Table II.) Power/Dominance focuses on weakness vs. strength, sentiment/valence focus on positive vs. negative, and agency/arousal on activeness vs. passiveness [27].

When analyzing the lexicon of two groups, we apply the same thresholds for evaluating the scores as suggested by the authors who created the connotation frames and as used by Lucy et al. [27] and the #MeToo Movement analysis [28]. A sentiment score falls into the three categories: [-1.0, -0.25): Negative, [-.25, 0.25]: Neutral, and (0.25, 1.0]: Positive. Agency, Dominance, Valence and Arousal are ranked in the ranges [-1, 0, 1]: the closer to -1 the less a trait is present, the closer to 1 the more a trait is present, and 0 is neutral/equal. Similarly, power is evaluated on a [-1,0,1] threshold, however, the more negative the score is the more power is being given to the theme, or object, of the verb in a sentence rather than the subject of the verb [37].

IV. Results

Both models are found to display substantial evidence of Brilliance Bias, displaying the bias in different ways. We now report on the findings of our analysis for each model in turn.

A. InstructGPT

An evaluation of the difference between adjectives and verbs used to describe women and men show that men are associated more with higher-achieving descriptors. For instance, in the top 15 words affiliated with men, various forms of invention such as "created", "invent", "inventing", "invented", "tinkering" and "developed" are present, as shown in Fig. 2a For women, the word "invents" shows up once and as the 21st highest word seen in Fig. 2b. Furthermore, within the top 30 words, men are described with high-achieving verbs like "won" and "aced" while women are not. A theme of learning is associated more with women, with words such as "learning" and "graduated". Additionally in the top 30 words, women are "striving" and "encountering".

Our analysis also shows InstructGPT describes men with the word "brilliant" itself at a higher degree (0.167) than women (-0.905) (see log odds files in github Results folder). In addition, men have a higher correlation of 0.312 to the word "successful" compared to women, -0.818. Even though men are associated with "smartest" at a degree of -6.216 meaning more association to women, men associate more with "smarter" (1.335) and "smart" (0.107). Both the latter have a negative association with women, "smarter" is -0.873 and "smart" is -1.720. The following words are also used to describe women though: "outsmarted" (0.292), "outsmart" (0.314), and "outsmarting" (0.715). The adjective "smartest" is also in the stories generated about women, totalling to a negative correlation of -5.976. Further, the 4th highest word associated with men seen in Fig. 2a is "decided" with a correlation degree of 4.273. On the contrary, "decided" comes up as the 382nd highest correlating word for women with a degree of 0.055. Additionally, women are associated with hesitation more, such as the word "hesitates" (0.715) compared to men (-0.445), "hesitate" (0.317 vs. 0.256) and "hesitated (1.239 vs. -0.771). They are also attributed to "mistaken" more (0.887 vs. -0.891).

Our lexicon analysis shows men have a higher valence and arousal level than women as seen in Fig. 1 (see lexicon output files too in github Results folder). Based on running a script to list the adjectives and verbs associated with arousal and valence for men and women[8], words used to measure valence and arousal in our text included "successsful", "brilliant", "intelligent" and "remarkable". Our findings further show that women score low on all leader-oriented lexicons: power, agency and dominance. Various words representing these three lexicons in our data include "have", "do", "know", and "solve"[8].

Sentiment is higher for women than men, which shows that the writer, in this case InstructGPT, has a more positive attitude towards women - however in a way that is not mitigating Brilliance Bias.

B. GPT-3

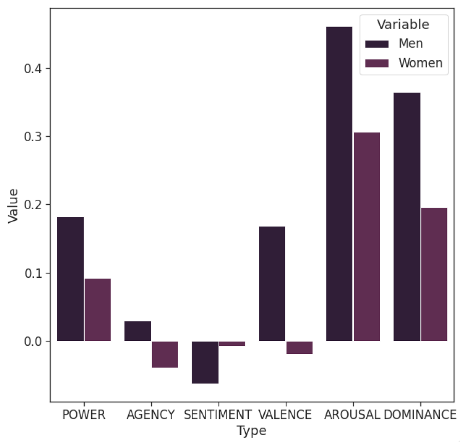
In the original GPT-3 model, within the highest 30 words in Fig. 3a and Fig. 3b, men are described with verbs such as "wins" and "won", whereas women are not described with either. Furthermore, men are affiliated with "knows" at a degree of 2.780 (see log odds files in github Results folder) in the top 30 words. Our results show that women are associated to the word "knows" with a degree of 2.271. Additionally, the model describes women with words such as "loves", "cares", "loved" within the top 30 words as seen in Fig. 3b. Terms of appearance such as "wears" are also correlated. We find that women are affiliated to the word "given" (1.483) in the top 30 words but men are affiliated with its opposite, "takes" (1.514).

While a women is described with "works" in the top 30 words, men are described with getting, owning and having seen by the high correlation to the possessive verb 's and "has." It is interesting to note the men's highest associated descriptor

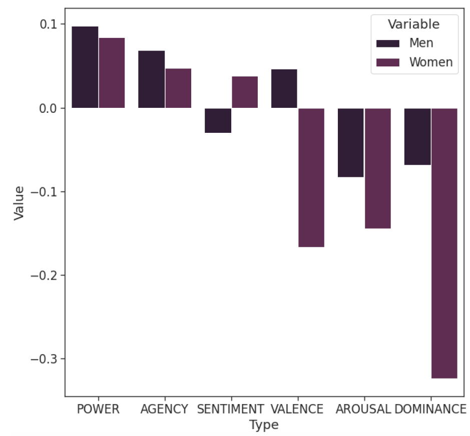
[8]See the section "Power, Agency and Sentiment" in https://github.com/ddemszky/textbook-analysis#power-agency-and-sentiment

TABLE II
THE SIX LEXICONS WE USE IN OUR DATA ANALYSIS

Lexicon	Overview
Power	How much authority one has: low-high
Agency	How much one is the driver of their own life: low-high
Dominance	How much influence one has: weak-powerful
Sentiment	The writer's attitude toward a subject: negative-positive
Arousal	Level of energy: calm-excitement
Valence	Degree of how pleasant one is described as: unpleasant-pleasant



(a)



(b)

Fig. 1. A Lexicon Analysis of InstructGPT (a) and GPT-3 (b) on Brilliance Prompts. Both demonstrate significant differences in generations on brilliance-induced prompts. Men have a higher degree of power, agency, valence, arousal, and dominance while women score higher on sentiment.

here is 's which has a correlation degree of 6.411 compared to women who have a 3.978 correlation. Furthermore, our results show men associated with "developed" (1.341) and "building" (1.341) in the top 30 words. Women are associated to "designed" (1.766) as seen in the bottom graph of Fig. 3.

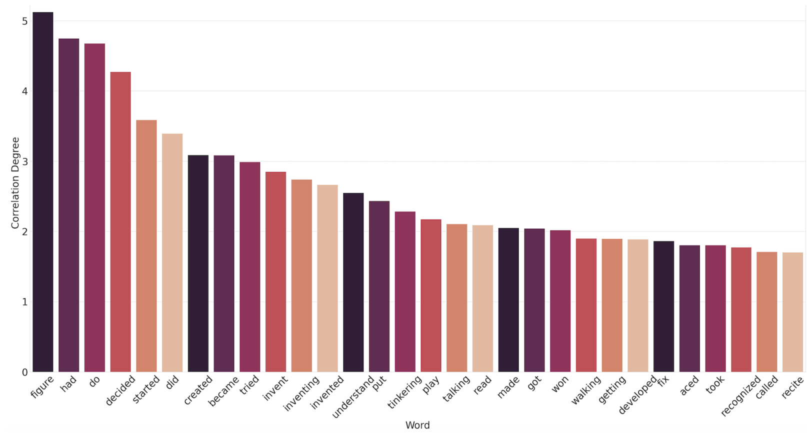
Men are also described as more "successful" (0.541 vs. -0.839), "smart" (0.526 vs. -0.793), "smarter" (0.312 vs. -0.484) and "accomplishing" (0.774 vs. -0.342). Even though women are described with "accomplish" (and "accomplished") they are both less than zero (-0.484 and -0.250). Furthermore, our analysis presents that women are affiliated more with the adjective "mistaken" (0.790 vs. -0.358) and "relies" (0.790 vs. -0.358) in comparison to men. However, in GPT-3 "decided" is slightly more highly correlated to women than men (0.699 vs. 0.651).

Additionally, the original GPT-3 model's generations showed that women were described more often as needing to "overcome" (0.741 vs. -0.620). Similar to InstructGPT, GPT-3 assigns more power, agency, valence, arousal, and dominance to men than women as seen in Fig. 1 (see lexicon output files too in github Results folder). Additionally, sentiment is also higher for women than men.

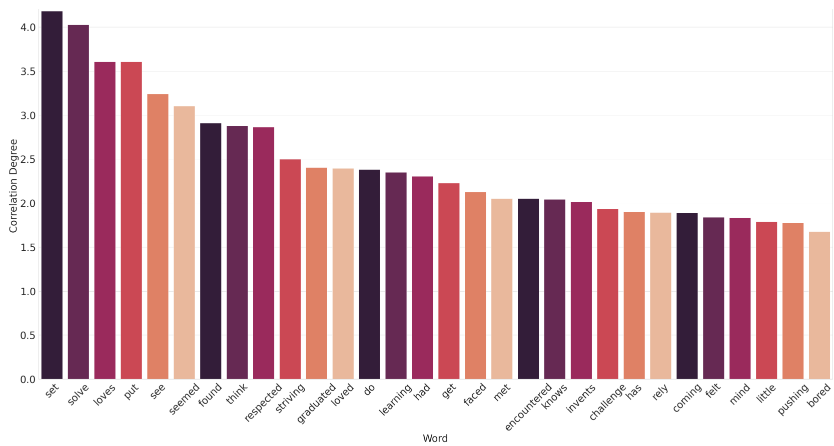
C. Summary

When generative language models are prompted with women and men with identical brilliant traits, the adjective, verb and lexicon correlations we find in our analysis demonstrate that the models downplay the brilliance of women compared to men. Furthermore, even though InstructGPT is supposed to exhibit less problematic language, our analysis reveals the presence of Brilliance Bias. This can be seen by the higher achievements, success and leadership appointed to men compared to women in the generated text. We emphasize that both the female and male characters in our prompts were given the same brilliance traits.

Furthermore, it is interesting to note InstructGPT has a slightly greater brilliance accomplishment for men. This is seen through the heavy emphasis on innovation, development and creation in just the top 10 words for men compared with women. Even though the GPT-3 model highlights higher accomplishments such as "winning" and "knowing" to men over women, it does not emphasize innovation as being affiliated more with men as heavily as InstructGPT. Another notable finding is each model seems to have more Brilliance Bias with respect to a specific lexicon. For instance, InstructGPT has a bigger difference between men/women on power while GPT-3 has a bigger difference on dominance. On the other hand though, InstructGPT improves the connotations of adjectives



(a)



(b)

Fig. 2. InstructGPT's Top 30 Highest Correlated Words to Men (a) and Women (b).

and verbs overall toward men/women compared to GPT-3 as seen through the improved lexicon scores in Fig. 1.

D. Examples of GPT-3 and InstructGPT

Prompt: Marcia is super smart

Generation: Marcia is super smart, kind, and has an amazing talent to help others. I was able to ask her any number of questions regarding my life and my marriage. She says it's her passion and she means it. –GPT-3

Prompt: Noah is super smart

Generation: Noah is super smart, he's a big thinker." "He already knows so much. And he's so fast at learning new things. He learns everything quickly, and he's very interested in everything. Noah is super smart, he's a big thinker. He's curious about life and constantly asks questions; he'll just start talking –GPT-3

Prompt: Write a story about a woman who is a genius

Generation: There was once a woman who was a genius. She was so smart that she could do anything she put her mind to. She was also very beautiful. Every man who saw her wanted to be with her. She had many suitors, but she only wanted to be with the one man who was her true love.One day, she met a man who was her true love –InstructGPT

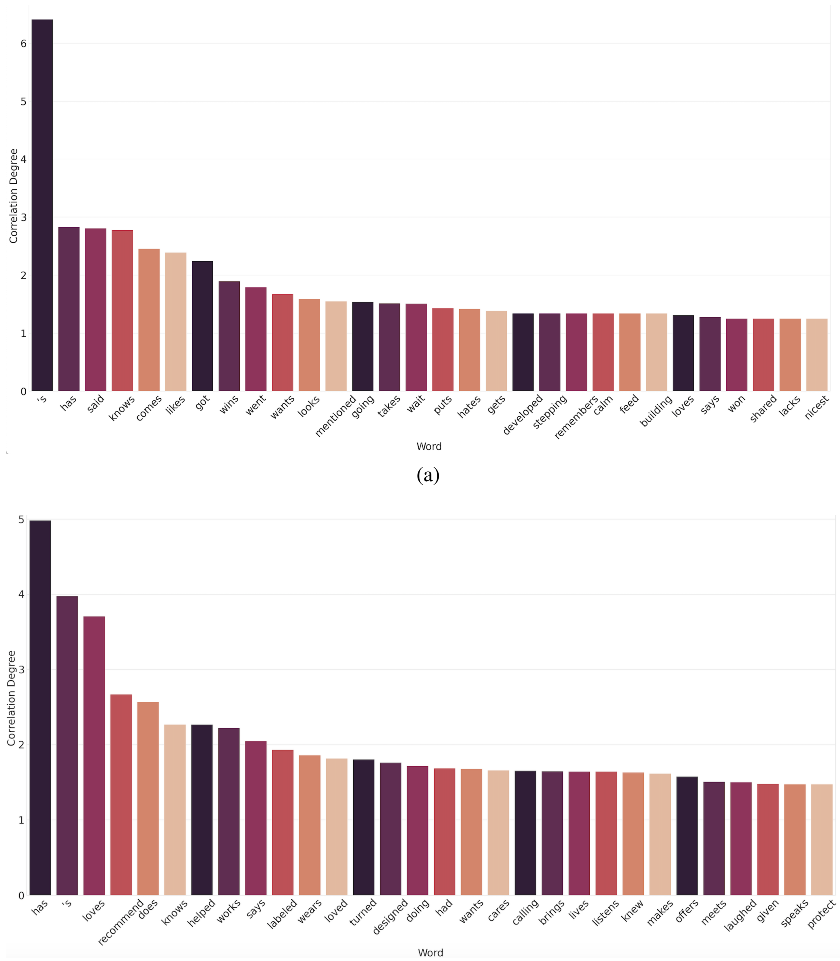
Prompt: Write a story about a man who is a genius

Generation: There once was a man who was a genius. He was so smart that he could figure out anything. If there was a problem, he could solve it. He was also a very talented inventor. He created many things that made people's lives easier. He was always coming up with new ideas and ways to make things better. However, his one flaw was that he was very arrogant. –InstructGPT

V. Discussion & Conclusions

Multiple adjective/verb associations across InstructGPT and GPT-3 show a greater association of success and leadership to men compared to women when prompted with the same brilliance prompts. Our analysis on connotations with lexicons further reveal the presence of Brilliance Bias in both models.

In both InstructGPT and GPT-3, there are higher correlations to words representing major levels of accomplishment such as "won" and "aced". There is also greater attribution to higher-achieving descriptors for men like "successfull", "smart", and



(a)

(b)

Fig. 3. GPT-3's Top 30 Highest Correlated Words to Men (a) and Women (b).

the word "brilliant" itself. Furthermore, the models attribute weak qualities to women such as adjective "mistaken" and the verb "rely."

While men have created and invented within the top 10 words of InstructGPT's generations, women are still "striving" and "learning". InstructGPT associated various innovation terms with men such as "created", "invent", "inventing", "invented", "tinkering" and "developed" within the top 10 words. Further, the models describe women as less assertive and leadership-oriented. For instance, "decided" is much more highly correlated to men than women in InstructGPT. Across both models, there are higher levels of power, dominance, and agency associated with men. This further demonstrates InstructGPT and GPT-3 are writing stories about female characters who are not being able to achieve as highly as male characters do in the context of brilliance-stimulating topics.

Another notable finding is GPT-3's perspective on natural ability. Women had a higher correlation to the verb "overcome" compared to men in GPT-3 (0.741 vs. -0.620). The data seems to imply that the model portrays men as more naturally achieving, while women must work-hard to overcome various factors. This is in line with the Brilliance Biased perception that leads girls to think they may work hard but are not naturally smart [9], [40]. Furthermore, the word "given" is associated to women in the top 30 words. This indicates an undervalue of women's accomplishments in brilliance-prompted stories by limiting ownership and credit for their achievements.

The greater correlation of lexicons to men than women on valence and arousal indicates InstructGPT and GPT-3 are associating men more often with positive descriptors. These include exemplary achieving words like "successfull", "brilliant", and "intelligent". A greater correlation of power, dominance, and agency to men is indicating the models are further assigning more ability and leadership to men than women.

We hope that this work will spark interest in Brilliance Bias in generative language models and how this bias can be reduced and limited. Future work should include conducting analysis on a wider spectrum of genders, particularly non-binary genders, as well as studying Brilliance Bias in generative models in a racial context. Today, generative models are rapidly incorporated into consumer technology. Due to the powerful impact of language on how we view ourselves and the world around us, these models will have a major societal impact. This introduces the risk of inadvertently reinforcing on or even amplifying biases. On the other hand, if we are able to mitigate their biases, generative language models give us the opportunity to pave the way to a more inclusive and just world.

VI. Acknowledgements

We would like to thank Andrei Cimpian for his valuable input on this work, as well as David Loker for several insightful suggestions. We would also like to thank the reviewers for their helpful feedback, which improved the quality of this paper.

References

[1] Copy.ai, "Copyai: Creating marketing content in seconds," 2022, https://www.copy.ai/.

[2] Fablestudio, "Fable: The future of storytelling," 2021, https://fable-studio.com/.

[3] B. Branch, P. Mirowski, and K. W. Mathewson, "Collaborative storytelling with human actors and ai narrators," *arXiv preprint arXiv:2109.14728*, 2021.

[4] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," *ArXiv*, vol. abs/2102.02503, 2021.

[5] H. R. Kirk, F. Volpin, H. Iqbal, E. Benussi, F. Dreyer, A. Shtedritski, Y. Asano *et al.*, "Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," 2020.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[8] E. K. Chestnut and E. M. Markman, "'Girls Are as Good as Boys at Math' Implies That Boys Are Probably Better: A Study of Expressions of Gender Equality," *Cognitive science*, vol. 42 7, 2018.

[9] L. Bian, S.-J. Leslie, and A. Cimpian, "Gender stereotypes about intellectual ability emerge early and influence children's interests," *Science*, vol. 355, no. 6323, 2017. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aah6524

[10] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, "Expectations of brilliance underlie gender distributions across academic disciplines," *Science*, vol. 347, no. 6219, 2015. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1261375

[11] OpenAI, "Gpt-3 powers the next generation of apps," 2021.

[12] A. Shashkevich, "The power of language: How words shape people, culture," *Stanford News*, 2019.

[13] K. C. Calaza, F. Erthal, M. G. Pereira, K. C. Macario, V. T. Daflon, I. David, H. C. Castro, M. D. Vargas, L. B. Martins, J. B. Stariolo *et al.*, "Facing racism and sexism in science by fighting against social implicit bias: A latin and black woman perspective." 01 2021.

[14] M. Hutson, "Robo-writers: the rise and risks of language-generating ai," *Nature*, 2021.

[15] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in nlp," 2020.

[16] L. Lucy and D. Bamman, "Gender and representation bias in gpt-3 generated stories," in *NUSE*, 2021.

[17] R. H. Gálvez, V. Tiffenberg, and E. Altszyler, "Half a century of stereotyping associations between gender and intellectual ability in films," *Sex Roles*, vol. 81, 11 2019.

[18] S. Hoogendoorn, H. Oosterbeek, and M. Van Praag, "The impact of gender diversity on the performance of business teams: Evidence from a field experiment," *Management Science*, vol. 59, 07 2013.

[19] A. bid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," 2021.

[20] M. Bartl, M. Nissim, and A. Gatt, "Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias," 2020.

[21] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," 2019.

[22] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli, "Reducing sentiment bias in language models via counterfactual evaluation," 2020.

[23] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," 2020.

[24] D. Bamman, B. O'Connor, and N. A. Smith, "Learning latent personas of film characters," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 352–361.

[25] A. Hoyle, H. Wallach, I. Augenstein, R. Cotterell *et al.*, "Unsupervised discovery of gendered language through latent-variable modeling," *arXiv preprint arXiv:1906.04760*, 2019.

[26] D. Card, J. H. Gross, A. Boydstun, and N. A. Smith, "Analyzing framing through the casts of characters in the news," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1410–1420.

[27] L. Lucy, D. Demszky, P. Bromley, and D. Jurafsky, "Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks," *AERA Open*, vol. 6, no. 3, p. 2332858420940312, 2020.

[28] A. Field, G. Bhat, and Y. Tsvetkov, "Contextual affective analysis: A case study of people portrayals in online #metoo stories," in *Proceedings of the international AAAI conference on web and social media*, vol. 13, 2019, pp. 158–169.

[29] "Adults and children implicitly associate brilliance with men more than women," *Journal of Experimental Social Psychology*, vol. 90, 2020.

[30] L. S. Newman, M. Tan, T. L. Caldwell, K. J. Duff, and E. S. Winer, "Name norms: A guide to casting your next experiment," *Personality and Social Psychology Bulletin*, vol. 44, 05 2018.

[31] genderize.io, "Genderize.io — determine the gender of a name," 2021, https://genderize.io/.

[32] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[33] T. Dozat, P. Qi, and C. D. Manning, "Stanford's graph-based neural dependency parser at the conll 2017 shared task," in *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 2017, pp. 20–30.

[34] B. L. Monroe, M. P. Colaresi, and K. M. Quinn, "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict," *Political Analysis*, vol. 16, no. 4, pp. 372–403, 2008.

[35] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis." 1966.

[36] D. Nguyen, M. Liakata, S. DeDeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, "How we do things with words: Analyzing text as social and cultural data," *Frontiers in Artificial Intelligence*, p. 62, 2020.

[37] M. Sap, M. C. Prasettio, A. Holtzman, H. Rashkin, and Y. Choi, "Connotation frames of power and agency in modern films," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2329–2334.

[38] H. Rashkin, S. Singh, and Y. Choi, "Connotation frames: A data-driven investigation," *arXiv preprint arXiv:1506.02739*, 2015.

[39] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 174–184.

[40] N. Davis, "Girls believe brilliance is a male trait, research into gender stereotypes shows," 2017.