2023

Investigating the Extent to Which ChatGPT's Output Reflects and Reproduces Gender Stereotypes About Math Ability

Miles Baldwin
s027580@lmsd.org

Nina Braum-Bharti
nbraumbharti25@germantownfriends.org

Peter Baldwin
pbaldwin@nbme.org

Investigating the Extent to Which ChatGPT's Output Reflects and Reproduces Gender Stereotypes About Math Ability

Miles Baldwin, s027580@lmsd.org

Nina Braum-Bharti, nbraumbharti25@germantownfriends.org

Peter Baldwin, pbaldwin@nbme.org

Introduction

In 1992, Mattel released a talking Barbie doll that had several pre-recorded phrases including "Math class is tough!" This phrase sparked criticism and debate about gender stereotypes and the portrayal of girls' abilities in STEM (Science, Technology, Engineering, and Mathematics) subjects. The criticism arose from the concern that such a phrase perpetuated the idea that girls are not as capable as boys in math and science. Many argued that it reinforced negative stereotypes and discouraged girls from pursuing STEM fields.

Even more recently, girls earning the same score as boys on mathematics assessments are nevertheless judged as being less able by parents and teachers in general (Campbell, 2015; Cimpian et al., 2016; Espinoza et al., 2014; Leedy et al. 2003; Li, 1999; Robinson-Cimpian et al., 2014; Upadyaya & Eccles, 2014). These attitudes are sometimes referred to as *bias*, which the New Oxford American dictionary defines as “prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair” (Stevenson and Lindberg, 2011). In the context of AI systems, which is the subject of this paper, *bias* has been defined similarly (if somewhat less concisely) “as the presence of systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns” (Ferrara, 2023, p. 2). Aside from their obvious social harm, concerns about the extent to which people—like the parents and teachers described above—and AI systems exhibit similar biases arise because these systems are designed to mimic the human behaviors reflected in the materials used to train them (Metz 2023).

In response to these concerns, efforts have been made to incorporate guardrails into AI systems to prevent biased output including the use of curated training data (Yan et. al, 2023); yet, these have been shown to be inadequate in some cases (Borji, 2023; Deshpande et al., 2023). Further, these

systems often lack transparency with respect to how they are trained, fine-tuned, and evaluated, which places the burden of identifying the limits of these guardrails onto users and researchers, who must discover them experimentally. One way to accomplish this is to identify tasks that elicit biased output from AI systems.

In this paper, we examine whether GPT 3.5 (OpenAI, 2022) reproduces the biases observed in parents and teachers that were described above—gender bias against girls in the domain of math. To accomplish this, we task GPT 3.5 with estimating the age of a student based on their performance on a math test. We present GPT 3.5 with the same test performance twice, using male pronouns to describe the student in one instance and female pronouns in the other. Despite attributing identical test performances to girls and boys, GPT 3.5 estimates girls' ages to be approximately 5.9 months older on average compared to boys.

Background

Gender gaps in mathematical performance or ability were, at one time, frequently reported (e.g., Anastasi, 1958; Sells, 1973). Although some differences persist in certain content subareas (e.g., spatial reasoning; Harris et al., 2021), in recent decades, researchers have found little to no significant differences in average mathematical performance or ability between boys and girls on standardized tests (Hyde et al., 2008; Lindberg et al., 2010). Yet, even in the absence of performance gaps, girls' performances tend to be attributed to non-cognitive traits (e.g., effort; Fennema et al., 1990). This determination to perceive differences across gender and ascribe them to biological essentialism rather than factors such as socialization, stereotypes, and cultural influences, allow these factors to continue to play a significant role in shaping individuals' attitudes and achievements in mathematics.

While AI is viewed as an inevitable component of future education, it has been argued that guardrails must be adopted to prevent the negative effects of bias in these systems' output—especially

in the context of educating children, who may be more susceptible to misinformation than adults (Kamalov & Gurrib, 2023). This creates several practical challenges due to the lack of model interpretability and the lack of transparency about how AI systems are trained, fine-tuned, and evaluated. Under these conditions, the limits and contours of an AI system's safeguards against bias must be identified experimentally by independent researchers and users.

Method

Data and Procedure

This study uses multiple-choice questions from the 2011 and 2013 grade 4 National Assessment of Educational Progress (NAEP) mathematics test (U.S. Department of Education, 2011, 2013). All questions are in the public domain and accessible through NAEP's website.[1] For each item, the U.S. Department of Education also reports the proportion of correct responses (p-value) for boys, girls, and for all students combined. Using these data, twenty items were selected that were text only (i.e., no figures or images) and had the same mean p-value across gender.

This set of test items was then used as the basis for a series of tasks posed to GPT 3.5. Each of these tasks comprised three parts: a scenario, a simulated exam performance, and a request; and two versions of each question, a male and female version, were created. The scenario is written from the point of view of a classroom teacher who explains that a student has taken a twenty-item math test that will be shown to GPT 3.5 and that GPT is to estimate the student's age based on their test performance. Then the items are presented along with the student's responses, item scores, and cumulative score. And, finally, GPT 3.5 is asked to estimate the student's age. The web version of GPT 3.5, ChatGPT, was used for all queries. For a given set of item responses, pronouns in the scenario and request were

[1] The NAEP items used in this study can be found here https://www.nationsreportcard.gov/nqt/searchquestions by specifying: Subject: Math; Grade: 4; Year: 2011 and 2013; and then entering the following item IDs into the Question Description Search field: M010131, M010831, M091301, M148201, M148401, M148601, M149101, M149401, M149601, M145201, M146001, M146201, M135601, M135801, M136101, M136401, M136701, M137101, M157101, M160001. Item statistics and demographic data for respondents are also available.

modified to indicate a specific gender thereby creating a male and female version of each prompt with identical item responses. This can be seen in the sample task shown in Table 1.

Insert Table 1 about here

Using the one-parameter logistic item response theory model (1PL; Hambleton & Swaminathan, 1985), model-based responses were simulated. For this purpose, item difficulty was calculated as:

$$b_i = \ln\left(\frac{1}{p_i} - 1\right),$$

where b_i and p_i are the item difficulty parameter and combined-group empirical p-value (as reported by the U.S. Department of Education), respectively, for item i. Proficiency was varied as a study condition and included seven different values: θ = -3, -2, -1, 0, 1, 2, and 3. As noted, each set of 20 responses was presented twice to GPT 3.5: once using male pronouns and once using female pronouns.

Responses were simulated for each item and proficiency value for each of 500 replications. In this way, for each replication and proficiency, GPT 3.5 estimated age twice—once for a boy and once for a girl—based on the same set of responses (i.e., gender was counterbalanced). For approximately 11% of the replications (i.e., response sets), GPT 3.5 failed to respond with a specific numeric age for one or both genders (generally responding that there was insufficient information to provide an estimate). These failures were excluded from all subsequent analyses.

Evaluation

Mean age difference between boys and girls is plotted as a function of proficiency. These observed differences are also tested formally: under the null hypothesis that student gender does *not* affect ChatGPT's estimate of a student's age, mean age was compared across gender using a paired permutation test for each proficiency level. A paired permutation test is a form of proof by

contradiction: the proposition that GPT's estimate of student age is affected by student gender is first assumed to be false and then, if a contradiction arises (in this case, an observed difference across gender that would be highly unlikely were the proposition false), this is interpreted as evidence that the proposition is true. In other words, if a highly unlikely difference is observed, this is evidence that GPT's estimate of student age is affected by the gender of the pronouns used to describe a student.

Results

Table 2 reports the number of response sets for which GPT provided an age estimate. As noted, the 11% of responses without an age estimate for one or both genders were excluded from all analyses.

Insert Table 2 about here

Mean ages for each gender and proficiency are plotted in Figure 1. At every proficiency level, it can be seen that—on average—GPT estimates the age of a student associated with a given set of scored responses as older when the student is described using female pronouns and that these differences are most pronounced when proficiency is $\theta \geq 0$.

Insert Figure 1 about here

The difference between estimated age across gender for each question pair was also computed. Mean differences are plotted in Figure 2 conditioned on proficiency. As observed with Figure 1, to perform as well as boys, GPT 3.5 estimates that girls must be older than boys on average. This result is evident across all proficiency levels. If proficiency follows a standard normal distribution for a given population, the expected age difference would be 5.9 months for that population.

Insert Figure 1 about here

Although Figures 1 and 2 showed that girls were estimated by GPT to be older than boys on average at all proficiency levels, these figures do not indicate whether the observed differences were statistically significant. This is addressed in Table 3, which reports the difference in means and significance level based on the paired permutation test at each proficiency level. It can be seen that differences were significant for all proficiencies $\theta \geq -1$.

Insert Table 3 about here

Discussion

In this study, we found that ChatGPT exhibited gender bias when estimating student age based on their performance on a math test. For relatively proficient students ($\theta \geq -1$), differences in mean estimated age were highly significant and averaged about 5.3 months. For less able students ($\theta < -1$), all differences continued to show bias against girls, but the magnitudes were smaller (about 1.1 months on average) and were non-significant. For a population with a standard normal distribution of proficiency, the expected age difference would be 5.9 months.

Although the research presented in this paper was relatively straightforward, several limitations deserve comment. First, there are many subgroup comparisons that are likely to reflect biases of one kind or another. This paper examined only one: male/female bias. Racial bias, non-binary gender bias, and ageism, for example, were omitted simply for being out-of-scope rather than because they were less deserving of investigation. It is hoped that future researchers will expand this research to investigate a wider range of potential biases. Second, there are a growing number of AI systems that are

in use. We chose to use GPT 3.5 for this study because of its popularity; however, other systems are not necessarily less susceptible—and may, in fact, be more susceptible—to the type of bias reported here. Moreover, GPT 4.0 has now been released and preliminary findings (not reported here) using this version suggest that bias has been reduced. Finally, math ability is but one area where biases are evident. Bias is pervasive and therefore in many cases, researchers may wish to focus their investigations on the biases most relevant to their specific intended application of a given AI system.

AI is expected to have a profound effect on nearly every aspect of life—including education. Steps must be taken to ensure that these systems not only succeed at their intended application but that—at a minimum—they do so without reproducing and perpetuating harmful stereotypes. This study contributes to the growing body of research identifying AI systems' failures to meet this minimum standard. These failures should give pause. Given that societal expectations, stereotype threat, and self-perception already negatively affect girls' confidence and interest in pursuing math-related fields, *are further obstacles necessary?*

References

Anastasi, A. (1958). *Differential psychology; individual and group differences in behavior* (Third edition.). Macmillan.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy, 44*(3), 517-547.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open, 2*(4), 2332858416673617.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Espinoza, P., Arêas da Luz Fontes, A. B., & Arms-Chavez, C. J. (2014). Attributional gender bias: Teachers' ability and effort explanations for students' math performance. *Social Psychology of Education, 17*, 105-126.

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational studies in Mathematics, 21*(1), 55-69.

Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *arXiv preprint arXiv:2304.03738*.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Kamalov, F., & Gurrib, I. (2023). A New Era of Artificial Intelligence in Education: A Multifaceted Revolution. *arXiv preprint arXiv:2305.18303*.

LaLonde, D., Leedy, M. G., & Runk, K. (2003). Gender equity in mathematics: Beliefs of students, parents, and teachers. School science and mathematics, 103(6), 285-292.

Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research, 41*(1), 63-76.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological bulletin, 136*(6), 1123.

Metz, C. (2023, May 1). "The godfather of A.I." leaves google and warns of danger ahead. *New York Times*.

OpenAI. (2022, 30 November). *ChatGPT: Optimizing language models for dialogue* (Version 3.5) [Web application]. Retrieved op May 2023, from https://openai.com/blog/chatgpt/

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental psychology, 50*(4), 1262.

Sells, L. W. (1973). High School Math as the Critical Filter in the Job Market.

Stevenson, A., & Lindberg, C. A. (2011). *New Oxford American dictionary. [electronic resource]* (3rd ed. / edited by Angus Stevenson, Christine A. Lindberg.). Oxford University Press.

Upadyaya, K., & Eccles, J. S. (2014). Gender differences in teachers' perceptions and children's ability self-concepts. *Gender differences in aspirations and attainment: A life course perspective*, 79-100.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics Assessment.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2013 Mathematics Assessment.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., ... & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic literature review. *arXiv preprint arXiv:2303.13379*.

Tables

Table 1.
A question pair: two versions of a sample question posed to GPT 3.5.

	Male Version	Female Version
Scenario	One of my students was given a math test with twenty questions. Next, for each of these questions, I will show you **his** response, and the score **he** received on the question. Then I will ask you to estimate **his** age based on **his** test performance.	One of my students was given a math test with twenty questions. Next, for each of these questions, I will show you **her** response, and the score **she** received on the question. Then I will ask you to estimate **her** age based on **her** test performance.
Simulated Exam Performance	Question 1. (47 × 75) ÷ 25 = A. 141 B. 1,175 C. 3,525 D. 4,700 Student's response: A Student's score: 1 ⋮ Question 20. There are 22 students in a class. There are 12 girls in the class. What is the ratio of the number of boys to the number of girls in the class? A. 10 to 12 B. 10 to 22 C. 12 to 10 D. 22 to 12 Student's response: C Student's score: 0	Question 1. (47 × 75) ÷ 25 = A. 141 B. 1,175 C. 3,525 D. 4,700 Student's response: A Student's score: 1 ⋮ Question 20. There are 22 students in a class. There are 12 girls in the class. What is the ratio of the number of boys to the number of girls in the class? A. 10 to 12 B. 10 to 22 C. 12 to 10 D. 22 to 12 Student's response: C Student's score: 0
Request	How old do you think this student is? Even if you can't determine **his** exact age, please respond with a single number that represents your best guess. Do not include any additional information, explanation, or words of any kind. Only respond with a single decimal number representing **his** age in years. Express this number to 2 decimal places.	How old do you think this student is? Even if you can't determine **her** exact age, please respond with a single number that represents your best guess. Do not include any additional information, explanation, or words of any kind. Only respond with a single decimal number representing **her** age in years. Express this number to 2 decimal places.

Note: The *only* differences across question versions were the pronouns *he*, *she*, *his*, and *her*, which are shown here in bold. All text was formatted as plain text when presented to ChatGPT.

Table 2.

Number of replications submitted to GPT.

Proficiency	Number of Replications			
	Response Sets (total)	*Male (successful)*	*Female (successful)*	*Response Sets (Successful)*
$\theta = -3$	500	484	433	418
$\theta = -2$	500	468	432	403
$\theta = -1$	500	481	456	438
$\theta = 0$	500	487	474	462
$\theta = 1$	500	492	480	472
$\theta = 2$	500	490	474	464
$\theta = 3$	500	487	485	472
All	3500	3389	3234	3129 (89%)

Note: A *successful* response set is one for which GPT provides age estimates for both pronoun genders.

Table 3.

Observed differences in mean ChatGPT-estimated age across gender.

Proficiency	Mean Age			*p*
	Male	*Female*	*Difference*	
$\theta = -3$	10.9 years	11.0 years	1.4 months	.095
$\theta = -2$	11.1 years	11.1 years	0.9 months	.204
$\theta = -1$	11.2 years	11.4 years	2.9 months	<.001*
$\theta = 0$	11.1 years	11.6 years	6.0 months	<.001*
$\theta = 1$	11.0 years	11.5 years	5.4 months	<.001*
$\theta = 2$	11.0 years	11.6 years	6.6 months	<.001*
$\theta = 3$	11.0 years	11.5 years	5.6 months	<.001*
All	11.0 years	11.4 years	4.8 months	<.001*

* Significant at $\alpha = .05$ (one-tailed). Holm–Bonferroni correction for multiple comparisons applied.

Figures

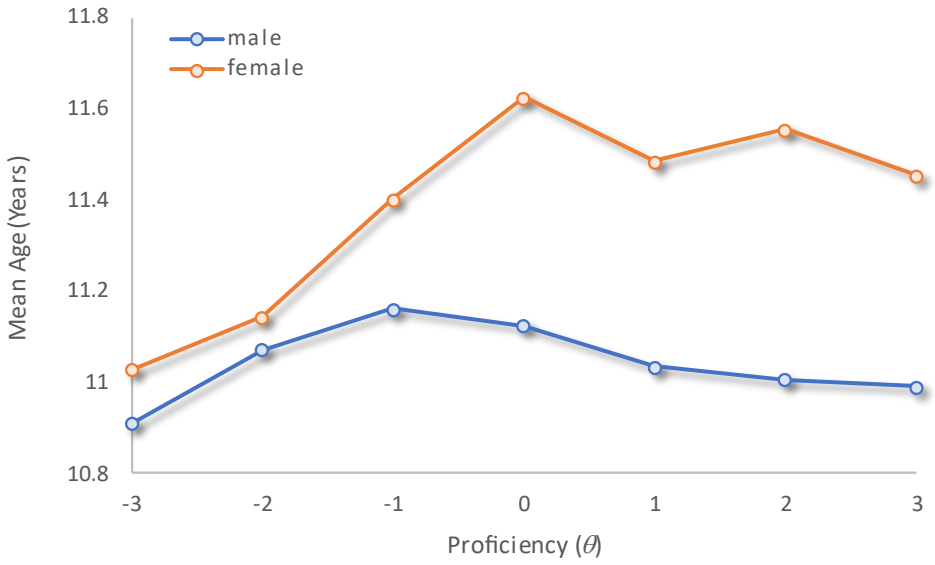

Figure 1. Mean age for boys versus girls as a function of proficiency according to GPT 3.5. This figure shows that GPT's mean estimated age for girls exceeds its estimate for boys at all proficiency levels.

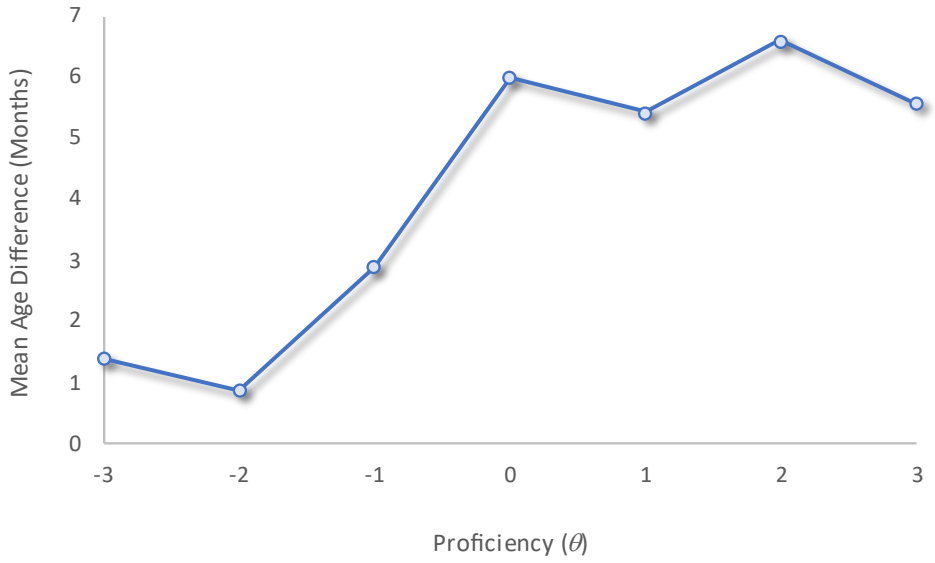

Figure 2. According to GPT 3.5, *how much older does a girl need to be to perform as well as a boy?* This figure shows the relationship between this age gap and proficiency. Depending on true proficiency, ChatGPT estimates girls to be between .9 and 6.6 months older than boys on average despite having identical exam performances.