Calibrating Social Bias in LLM and Human Responses in Postsecondary Online Discussion Forums

Daniel March
Columbia University
New York, NY, USA
dm3597@tc.columbia.edu

Zhen Xu
Columbia University
New York, NY, USA
zx2393@tc.columbia.edu

Chengyuan Yao
Columbia University
New York, NY, USA
cy2706@tc.columbia.edu

Renzhe Yu
Columbia University
New York, NY, USA
renzheyu@tc.columbia.edu

Abstract

Recent research has increasingly documented that large language models (LLMs) can exhibit social bias in consequential social contexts such as education. Yet this literature often evaluates LLM bias in isolation without calibrating it against human bias, even though human bias has been extensively documented and constitutes one important source of LLM bias through pretraining on human-generated text. Such calibration is important for understanding when and how LLMs may reproduce human patterns of bias, with strong practical implications for deploying general-purpose LLMs and customizing models on task-specific datasets. The current study examines these issues in postsecondary education contexts, comparing racial and gender bias in human- and LLM-generated responses to postsecondary online discussion posts. Using a small pilot dataset sampled at a public four-year institution in the United States, we find that humans exhibit small and insignificant levels of social bias and that LLM bias does not differ significantly from this human baseline. These analyses motivate a larger-scale investigation that may inform debates about the appropriateness of AI in online learning environments and the risks of fine-tuning LLMs on local educational data.

CCS Concepts

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Natural language generation**; • **Applied computing** → **Interactive learning environments**.

Keywords

Social Bias; Large Language Models; Generative AI; Postsecondary Education; Online Discussion Forums

ACM Reference Format:
Daniel March, Zhen Xu, Chengyuan Yao, and Renzhe Yu. 2026. Calibrating Social Bias in LLM and Human Responses in Postsecondary Online Discussion Forums. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale (L@S '26), June 29-July 03, 2026, Seoul, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3774398.3811591

1 Introduction

Social bias refers to disparate treatment based on social group membership. Such groups can be distinguished by race, gender, religion, age, disability status, national origin, sexual orientation, and other salient categories [4]. Although discriminatory treatment is prohibited in many institutional contexts, including employment and education, decades of research show that human judgments and interactions can still reflect systematic social bias [3, 5]. In recent years, the rapid adoption of generative AI and large language models (LLMs) has raised parallel concerns about whether these systems reproduce similar forms of bias. Indeed, a growing body of work has shown that LLMs can exhibit biased behavior across a range of social domains [2, 9, 16].

However, much of the existing literature evaluates LLM bias as an isolated model property rather than calibrating it against human bias. This comparison is theoretically important because human-generated text is a major source of LLM training data, meaning that human bias is not only a relevant benchmark but also a likely pathway through which bias enters model behavior. It is also practically important because humans and LLMs are increasingly being asked to perform similar social, evaluative, and communicative tasks. Many of these tasks, including evaluating job candidates [1] and conducting medical consultations [11], are high-stakes. It is clear that LLMs can be biased in absolute terms; what is more relevant is whether, and under which circumstances, they are more biased than humans.

In education, one context where social bias may be consequential is asynchronous online discussion forums, which are widely used for student interaction, peer support, question clarification, and collaborative learning. While research suggests that LLMs can serve as learning assistants in such contexts [8], they may also generate biased responses to student posts [6]. Furthermore, LLM bias may change when models are adapted to local institutional contexts; a debiased general-purpose LLM may later be fine-tuned on task-specific data that contains biased patterns of human interaction. If LLMs reproduce or even amplify existing levels of human bias, vulnerable groups of students may receive low-quality support, which could exacerbate educational inequities.

In this work, we investigate these challenges first by measuring social bias in existing student interactions in online discussion forums. Next, we prompt an LLM to generate responses to student posts and compare bias between LLM-generated and student-generated responses. Finally, we examine whether exposure to biased task-specific data affects bias in LLM outputs. Formally, we pose three research questions:

RQ1: Do student responses in online discussion forums exhibit bias against female or URM posters?
RQ2: Do LLM-generated responses exhibit more or less bias than student-generated responses?
RQ3: Do LLMs generate more or less biased responses when exposed to biased historical data?

L@S '26, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2293-6/2026/06
https://doi.org/10.1145/3774398.3811591

By calibrating LLM bias against human bias in the same educational task, this study contributes to a contextualized understanding of social bias in AI-mediated learning environments. The findings have implications for the use of general-purpose LLMs in online learning support and the design of fine-tuning pipelines for educational contexts.

2 Methodology

2.1 Experimental Design

Our study includes three conditions for generating responses to student posts in online discussion forums:

- A baseline human generation condition.
- A standard LLM generation condition under which the LLM sees the original post and the poster name, similar to common discussion forum configurations.
- A bias-exposed LLM generation condition under which the LLM sees the original post, the poster name, and is exposed to a set of historical discussion threads with prominent social bias as examples to learn from.

The baseline condition draws on real-world data containing anonymized student posts and responses. Both LLM generation conditions draw on the same student posts but required us to prompt LLMs to generate responses to each post. Because student names are anonymized, we use demographic information of student posters and established name pools to simulate names that reflect real-world demographic distributions [10, 12]. This enabled name exposure for LLMs and elicited implicit social bias (if present) in the response generation process. Details about the data source and measurements are provided in the following sections. The bias-exposed LLM generation condition probes the scenario where general-purpose LLMs may be locally customized (e.g., fine-tuned) with biased educational data. For each generation, we exposed the LLMs to a different historical dataset that showed strong bias in the same dimension along which we measure bias in LLM outputs, because a biased response in one aspect may not be biased in another.

2.2 Data

We obtained access to detailed online discussion records from the Canvas Learning Management System (LMS) deployed at a public four-year university in the United States. These records include the content and response structures of individual posts and are linked to demographic information in the student information system. Individual posts are clustered under discussion threads set up by instructors. All the data were fully anonymized prior to research access in line with privacy regulations.

For our primary experiments, we obtained a small sample drawn from the discussion forum data from Fall 2021 through Fall 2022. Specifically, we restricted the sample to threads with 10-15 unique student posters, of whom 2-8 were male and 2-8 were white. This left us with ten unique threads and 129 unique post-response pairs. In addition, we curated a biased thread candidate pool for the third condition by sampling from discussion threads from Winter 2020 through Spring 2021. We used slightly more lenient constraints, sampling 51 threads with 1,277 post-response pairs.

2.3 LLM Setup

Our study evaluated two LLMs: OpenAI GPT-4o and Anthropic Claude 4.5 Sonnet. For comparability, we ran both models with a temperature setting of 0.7 and instructed each LLM to act as a reasonably prepared student in the course. This role framing was intended to simulate realistic peer discussion responses given the discussion thread, the target post, the (simulated) poster name, and historical biased threads (for the bias-exposed generation condition only).

2.4 Bias Measurement

We operationalize social bias as the difference in response quality based on poster race and gender. We measure response quality using the following linguistic features:

- **Sentiment**: Sentiment measures the emotional tone expressed in peer responses. In this study, we assessed sentiment using the Linguistic Inquiry and Word Count [7]. Scores range from -1 to 1, with higher values indicating more positive sentiment.
- **Politeness**: We assessed politeness using the framework developed by [15], which quantifies politeness through a set of linguistic markers. Scores range from 0 to 1, with higher values indicating higher politeness.
- **Analytical Thinking**: Analytical thinking reflects the degree of logical reasoning and structured cognitive processing in text and was measured using [7]. Scores range from 0 to 1, with higher values indicating higher analytical thinking.

To select biased threads for the bias-exposed generation condition, we calculated racial and gender bias for each thread in our candidate pool. Specifically, we computed the average differences in each of the linguistic features between responses to posters from a majority group (white, male) and those from a minority group (URM, female), weighted by the number of participants in the thread. To estimate social bias in human- and LLM-generated responses across all of our experiments, we treated the three linguistic features introduced above as outcome variables in regression analyses and regressed them on indicators of poster race and gender (see Section 2.5).

2.5 Analytic Approach

To address our research questions, we ran a series of linear regression models with the linguistic features as outcome variables. To understand baseline human bias (RQ1), we compare linguistic features of different human-generated responses and estimate the model below:

$$Y_{ipj} = \alpha + \beta URM_p + \gamma Female_p + \boldsymbol{\theta}'\mathbf{X}_p + \lambda_j + \varepsilon_{ipj} \quad (1)$$

where Y_{ipj} is a linguistic feature of a student-generated response i to original post p within thread j; URM_p and $Female_p$ are demographic indicators of the *original poster* (URM includes all non-white races); $\mathbf{X}_p$ is a vector of characteristics of the *original post*, including word count and the same linguistic features described in Section 2.4; λ_j represents thread-level fixed effects; and ε_{ipj} is the error term. β and γ are the coefficients of interest and capture average human bias by poster race and gender, respectively.

To address RQ2 and RQ3, we compare bias between human-generated and LLM-generated responses to the same set of posts. The following model is estimated:

$$Y_{ipk} = \alpha + \sum_{k=1}^{2} \beta_k (C_k \times URM_p) + \sum_{k=1}^{2} \gamma_k (C_k \times Female_p) + \sum_{k=1}^{2} \delta_k C_k + \lambda_p + \varepsilon_{ipk} \quad (2)$$

where Y_{ipk} is a linguistic feature of a response i to original post p under condition k, and C_k is an indicator for each condition (the two LLM generation conditions are indexed 1 and 2, respectively, in reference to the human condition indexed as 0). URM_p and $Female_p$ are the same as in Equation 1, λ_p represents post-level fixed effects, and ε_{ipk} is the error term. In this model, δ_k captures baseline between-condition differences, and β_k and γ_k estimate differential racial and gender bias levels in each condition, respectively.

3 Results

We first present the average linguistic feature values of student-generated responses overall and disaggregated by poster race and gender in Table 1. These values contextualize the interpretations of our regression coefficients.

	S	P	A
White	0.190	0.478	0.175
URM	0.132	0.509	0.167
Male	0.125	0.532	0.180
Female	0.162	0.482	0.164
Total	0.149	0.500	0.169

Table 1: Average Student Response Quality Measures by Poster Attribute

Note: S = Sentiment, P = Politeness, A = Analytical Thinking.

Table 2 reports regression results for RQ1. Estimated gender and racial biases in existing human responses are generally small in magnitude and not statistically significant in our analytical sample. For example, the coefficient for URM in the sentiment score model is -0.04, meaning that URM posters are predicted to receive responses that are 0.04 sentiment units lower than white posters, given thread-level fixed effects and controlling for post-level characteristics.

In Table 3, we present regression results for RQ2 and RQ3, including coefficients for the main effects of Condition (C) and for the interactions between Condition and Race and Condition and Gender. For example, the coefficient of C3 × URM for the Claude politeness score model is -0.05. This means that under C3, Claude's responses to URM posters are predicted to be 0.08 politeness units lower than student-generated responses to URM posters. However, this coefficient, like nearly all others in Table 3, is small and insignificant, indicating that the difference in the LLM–human gap between URM and white male posters is not significantly different from zero.

As presented, these findings do not lend strong support for racial and gender bias across the three conditions. However, we interpret them with caution due to our small sample size. Although almost all coefficients are insignificant, their magnitudes are still informative. The aforementioned coefficient for C3 × URM, for example, may be a sign of racial bias in Claude. Furthermore, for both Claude and GPT, we observe coefficients with greater magnitudes in the sentiment and politeness models than in the analytical thinking model. Such patterns may be statistically significant in a larger-scale study, potentially allowing for more robust interpretations.

	S	P	A
URM	−0.04	0.04	−0.01
	(0.03)	(0.03)	(0.01)
Female	0.00	−0.04	−0.01
	(0.03)	(0.03)	(0.01)
31-75 Words	−0.05	0.03	0.01
	(0.05)	(0.05)	(0.01)
> 75 Words	−0.04	0.05	0.02
	(0.06)	(0.06)	(0.02)
Parent Post S	−0.02	−0.18*	−0.01
	(0.07)	(0.08)	(0.02)
Parent Post P	−0.28**	0.13	0.04+
	(0.08)	(0.09)	(0.02)
Parent Post A	−0.25	0.00	0.28**
	(0.48)	(0.51)	(0.11)

Table 2: Linear Regression Results for Human Bias

Note: S = Sentiment, P = Politeness, A = Analytical Thinking. White, Male, and 10–30 Words are the reference groups. Standard errors in parentheses. Statistical significance: $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Claude			GPT		
	S	P	A	S	P	A
C2	−0.01	−0.02	0.00	0.05	−0.04	0.02+
	(0.03)	(0.04)	(0.01)	(0.04)	(0.04)	(0.01)
C3	0.01	−0.03	0.01	0.05	−0.03	0.01
	(0.03)	(0.04)	(0.01)	(0.04)	(0.04)	(0.01)
C2 x URM	0.04	−0.04	0.01	0.04	−0.03	0.01
	(0.03)	(0.04)	(0.01)	(0.03)	(0.04)	(0.01)
C3 x URM	0.04	−0.05	0.00	0.04	−0.04	0.01
	(0.03)	(0.04)	(0.01)	(0.04)	(0.04)	(0.01)
C2 x Female	−0.02	0.03	0.01	−0.02	0.04	0.01
	(0.03)	(0.03)	(0.01)	(0.03)	(0.03)	(0.01)
C3 x Female	−0.04	0.02	0.00	−0.03	0.05	0.01
	(0.03)	(0.03)	(0.01)	(0.03)	(0.03)	(0.01)

Table 3: Linear Regression Results for Human-LLM Comparison

Note: S = Sentiment, P = Politeness, A = Analytical Thinking. C1 (human generation condition), White, and Male are the reference groups. C2 = standard LLM generation condition, C3= bias-exposed LLM generation condition. Standard errors in parentheses. Statistical significance: $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

4 Discussion

LLMs have the potential to effectively support student interaction and collaborative learning in online discussion forums. However, they may also exhibit social bias in this environment, negatively impacting student experiences and outcomes. In this study, we compared bias in LLM-generated responses under different conditions to bias in student-generated responses. The results from a pilot

dataset, while largely insignificant across conditions, encourage us to conduct follow-up analyses on a larger sample.

Our study is an exploratory effort to calibrate LLM bias against human bias in educational contexts, building upon existing research that measures LLM bias in writing assignment grading scenarios [13], LLM-as-teacher settings [14], and discussion forum environments similar to our own [6]. Despite the preliminary nature of our study, we highlight a few potential contributions to the literature. First, we were able to calibrate LLM bias against existing human bias, informing debates about the appropriate role of LLMs in online learning platforms. Second, we operationalized social bias along multiple dimensions, contributing more granular measurement approaches and informing more targeted mitigation techniques. Third, we included an exposure condition under which the LLMs were asked to learn from biased data before responding to student posts. This condition underscores the potential risks of customizing debiased general-purpose LLMs for task-specific contexts.

This study has several limitations. A major challenge was establishing experimental equivalence between the human generation and the LLM generation conditions. Our data contain student respondents whose demographics, academic abilities, and personalities vary widely. When we rerun the experiment on a larger sample, we will attempt to introduce similar variability within LLM generation conditions, perhaps by changing the temperature of the LLM between responses. In addition, we should acknowledge that student and LLM respondents may exhibit bias against student posters based on attributes that are not included in our data (e.g., religion) and that this bias was not captured by our models. Finally, our data came from a single institution, so we should be cautious when generalizing results to broader postsecondary contexts.

References

[1] Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2025. Measuring Gender and Racial Biases in Large Language Models: Intersectional Evidence from Automated Resume Evaluation. *PNAS Nexus* 4, 3 (2025), pgaf089. doi:10.1093/pnasnexus/pgaf089

[2] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2025. Explicitly Unbiased Large Language Models Still Form Biased Associations. *Proceedings of the National Academy of Sciences* 122, 8 (2025), e2416228122. doi:10.1073/pnas.2416228122

[3] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004), 991–1013.

[4] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (2024), 1097–1179. doi:10.1162/coli_a_00524

[5] Catherine Hill, Christianne Corbett, and Andresse St. Rose. 2010. *Why So Few? Women in Science, Technology, Engineering, and Mathematics.* American Association of University Women, Washington, DC.

[6] Zifeng Liu, Wanli Xing, Xinyue Jiao, and Chenglu Li. 2025. Exploring Fairness and Explainability in LLM-Generated Support for Online Learning Discussion Forums. *Journal of Learning Analytics* (2025), 1–26. doi:10.18608/jla.2025.8885

[7] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015.* Technical Report. University of Texas at Austin, Austin, TX.

[8] Shuying Qiao, Paul Denny, and Nasser Giacaman. 2025. Oversight in Action: Experiences with Instructor-Moderated LLM Responses in an Online Discussion Forum. In *Proceedings of the 27th Australasian Computing Education Conference.* 95–104. doi:10.1145/3716640.3716651

[9] Daiki Shirafuji, Makoto Takenaka, and Shinya Taguchi. 2025. Bias Vector: Mitigating Biases in Language Models with Task Arithmetic Approach. In *Proceedings of the 31st International Conference on Computational Linguistics.*

[10] Social Security Administration. 2025. National Baby Names Data (U.S.). Social Security Administration Actuarial Office baby names data. https://www.ssa.gov/oact/babynames/limits.html National dataset of baby names from Social Security card applications.

[11] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards Conversational Diagnostic AI. *Nature* (2024). doi:10.1038/s41586-024-07494-0

[12] Konstantinos Tzioumis. 2018. Demographic Aspects of First Names. Harvard Dataverse. doi:10.7910/DVN/TYJKEZ Dataset version V1.

[13] Melissa Warr, Nicole Jakubczyk Oster, and Roger Isaac. 2024. Implicit Bias in Large Language Models: Experimental Proof and Implications for Education. *Journal of Research on Technology in Education* (2024), 1–24. doi:10.1080/15391523.2024.2395295

[14] Iain Xie Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. 2025. LLMs Are Biased Teachers: Evaluating LLM Bias in Personalized Education. In *Findings of the Association for Computational Linguistics: NAACL 2025.*

[15] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness Package: Detecting Politeness in Natural Language. *R Journal* 10, 2 (2018).

[16] Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A Comparative Study of Explicit and Implicit Gender Biases in Large Language Models via Self-Evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).*