Grading Bias: The Impact of Demographic Indicators on LLM Evaluations of Persuasive Essays

TARA SHUKLA* and PAIGE VEGNA*

Large Language Models (LLMs) are being integrated into the education system at a rapid pace, leading to a need for an equally rapid investigation into their potential ethical implications and their effect on equitable learning. We seek to investigate the presence or absence of demographic bias in general-purpose LLMs as persuasive essay evaluators. Utilizing the PERSUADE 2.0 dataset, which contains over 25,000 student essays with demographic information and human-generated scores, we evaluate five state-of-the-art (SOTA) LLMs for demographic bias. To do so, we use both Demographic-Blind and Demographic-Aware prompting approaches complemented by a counterfactual analysis. Our findings reveal significant biases across multiple different models and demographic variables. We found concerning patterns around racial and socioeconomic factors, but the magnitude and direction of these biases varied substantially across models, suggesting that architecture and training influence bias. Our results highlight the need for responsible integration of LLM-based grading in education and emphasize the need for robust evaluation and bias-mitigation strategies.

ACM Reference Format:
Tara Shukla and Paige Vegna. 2024. Grading Bias: The Impact of Demographic Indicators on LLM Evaluations of Persuasive Essays. 1, 1 (December 2024), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models are quickly becoming increasingly influential in many aspects of modern life due to their promise to boost productivity with–seemingly–little cost to integrity. Education is one such platform for which LLMs have begun to be employed in recent years that is particularly ethically sensitive and liable to public scrutiny. LLMs have the potential to play a wide variety of roles in the education sector for the benefit of both students and instructors. For instance, LLMs can serve as graders of student writing–providing a tool that can decrease the workload on instructors and examiners, allow students to view evaluations of their work before they turn it in for grading, and eliminate personal biases and instructor-to-instructor variations in grading standards that may exist in human graders. The usage of LLMs for augmenting or replacing subjective human judgements is called LLM-as-a-judge, and it is a quickly emerging field of research [11].

Despite their potential benefits, LLMs are not guaranteed or generally understood to be free of bias due to their dependence on their pre-training corpora and their black-box nature. Before LLMs can be responsibly deployed in educational assessment, it is important to identify any potential biases in their judgements to and explore possible mitigation strategies. Towards this goal, we seek to investigate the degree of fairness and biases of LLM graders using persuasive essays. Specifically, we seek to determine how and to what extent LLMs exhibit demographic biases in the

*Both authors contributed equally to this proposal.

Authors' Contact Information: Tara Shukla, ts6796@princeton.edu; Paige Vegna, pvegna@princeton.edu.

Manuscript submitted to ACM

task of essay scoring. We employ the PERSUADE 2.0 dataset [4], which comprises more than 25,000 persuasive essays paired with human-generated labels rating their quality over multiple categories as well as demographic information about the student who wrote each essay. We implement a multi-pronged analysis in which we ask LLMs to grade essays given varying levels of demographic information about the student. Our analysis reveals significant variations in how different models responded to demographic information, with some showing concerning patterns of statistically significant bias across multiple demographic categories. These findings have important implications for the ethical considerations for the deployment of LLMs in educational settings and highlight the need for careful consideration of fairness and bias in automated grading systems. Our project data and scripts can be found on Google Drive [1].

2 Literature Review

The evaluation of LLM fairness in the past has been explored previously; studies have investigated whether LLM judgment echoes the human biases that are inevitably present in training corpora and proposed strategies for evaluating or combating bias in LLM judgements.

LLM bias has been examined in many different fields; for instance, Ayoub et.al [1] report on racial bias in the diagnosis and treatment recommendations of LLMs used in healthcare. Similarly, Yang et.al [12] found that LLMs are inherently biased along racial lines for treatment predictions. Tamkin et.al [9] identified both positive and negative discrimination in LLM decision-making tasks across 70 scenarios, and analyzed how prompt-based interventions can reduce discrimination. Li et.al [6] explore existing research on LLM fairness, gathering a selection of papers and categorizing their evaluation in the areas of demographic representation, counterfactual fairness, performance, prompt completion, conversation, and more.

Researchers have also examined the effectiveness and fairness of the LLM-as-a-judge paradigm. Thakur et al. [11] conducted a comprehensive evaluation of the effectiveness of the LLM-as-a-judge paradigm, testing different LLM judges' alignment with human judgements. Their findings highlighted that larger models can align well with human judgment, but small models and lexical matching can achieve similar results. Ye et.al [13] propose CALM, a framework for evaluating different types of biases in LLM judges and conclude that the robustness of LLM decision making across different types of bias has room to improve. Finally, Chen et.al [2] propose a framework for evaluating different types of bias in both human and LLM judges, and found that LLMs and humans are both susceptible to bias, emphasizing the need for robust evaluation practices.

3 Data and Methods

3.1 PERSUADE Dataset

We utilize the PERSUADE 2.0 dataset [4], which contains over 25,000 argumentative essays written by students in grades 8-12 in response to 15 different prompts [4]. The data set includes independent and source-based prompts [4]. In order to standardize and simplify our experiments, reduce computational and time costs, and avoid overly long and complicated input prompts to the LLM graders, we opted to only utilize non-source-based essays. Each essay is associated with a holistic score (ranging from 1-6) from a human grader and is human-annotated on various characteristics of argument effectiveness and writing quality [4]. The creators of the data set provide the exact rubric that was given to the human

[1] Our folder can be found here.

graders who scored the essays [4]. We focus only on the attainment of holistic scores. The essays are associated with demographic information about the student denoting their grade level, race/ethnicity, gender, English language learner status, economic disadvantage status, and disability status [4]. PERSUADE 2.0 is initially partitioned into train and test splits by its authors [4]. However, since we do not perform any training and instead run zero-shot inference with pretrained models, we utilize the authors' train split for our experiments. After selecting all of the independent essays from the train split, we are left with a total of 7,868 essays. For the demographic-aware portion of our analysis, we further partitioned our refined data set into six approximately equally sized random subsets, each corresponding to one of our six demographic variables.

3.2 Model Selection

We chose to evaluate five LLMs of varying sizes and architectures: Llama-3.2-1B [7], Llama-3.2-3B-Instruct [8], Qwen 2.5-1.5B-Instruct [10], Flan-T5-large [3], and Gemma 2-2B [5]. We selected these models to represent a range of model scales, model architectures, and inference types, while still approaching state-of-the-art performance. We also prioritized computational tractability, because in practice the combination of the rubric, essay prompt, and essay body yields large input token lengths and subsequently expensive computational and time costs.

3.3 Pairwise Score Comparison

In order to measure the fairness of LLM graders, we vary the amount and quality of student information that is included in the scoring prompts provided to a given LLM. Then, we compute the score shift for each individual essay under the different information settings. By computing the difference in score for every essay individually (rather than just computing the differences in the mean of all essays under one setting and the mean of all essays under another setting), we are better able to isolate the effect of changes in the provided information. We consider three such information settings: Demographic-Blind, Demographic-Aware, and Counterfactual Demographic-Aware.

3.4 Demographic-Blind Inference

Zero-shot prompting (in which an LLM is given a prompt to which it must respond without any pretraining for that specific task or any examples of desirable outputs) is a powerful and relevant tool for LLM researchers, casual experimenters, and everyday users alike. Zero-shot prompting for essay grading is a very plausible use case of LLM graders for real-life educators, whether sanctioned or not. For instance, if a teacher wished to return his or her students' essay grades more quickly, he or she might ask a chatbot to grade the essays using a zero-shot prompt. For the purpose of simulating the aforementioned use case, we employ zero-shot prompting for data collection.

Each zero-shot prompt under the Demographic-Blind setting (and likely for a real educator's automated essay scoring use case as well) includes three key elements: a grading rubric, the essay prompt, and the essay body itself. The grading rubric is standard for every experiment; we utilize the Holistic Rating Form for individual argumentative essays as provided to human graders during the formation of the PERSUADE 2.0 data set [4]. The essay prompt and essay body are copied exactly from the data source. Though we attempted to remain as consistent as possible with our model-to-model prompting language, we made slight optimizations in the structure of each model's corresponding prompts. For instance, when prompting a text-generation model like Llama-3.2-1B [7], we provide the three aforementioned elements in a single string, and end the prompt by asking for a score. When prompting an instruct model like Llama-3.2-3B-Instruct [8], on the other hand, we include the rubric in the model's system prompt and place the essay prompt and essay body in

the user prompt. We perform Demographic-Blind inference for every essay in our refined data set and for every model in our selection of LLMs.

3.5 Demographic-Aware Inference

It is possible–especially in argumentative essays where students are perhaps more likely to draw upon personal experiences in their prose–for the actual body of an essay to include identifying information. For instance, consider a hypothetical statement regarding gender-neutral bathrooms in college dormitories: "As a woman, I feel comfortable using a gender-neutral bathroom." If a student were to include such a statement in an assignment, she would clearly be identifying her gender. While essays can be anonymized to a certain extent by omitting the name of the student and refraining from using identifying language in the prompt, it may not be feasible to mask identifying language in the essay body itself without affecting the student's argument. Serious implications for the fitness and fairness of LLM graders could arise if identifying factors like gender or other sensitive demographic divisions can affect the score awarded to an essay. We attempt to simulate the effect of the presence of identifying language in the body of an essay by explicitly identifying demographic information in the prompt itself.

We perform Demographic-Aware experiments for each of the six demographic variables available in the PERSUADE 2.0 [4] data set: grade level, race/ethnicity, gender, English language learner status, economic disadvantage status, and disability status. The Demographic-Aware prompts are identical to the Demographic-Blind prompts except for the addition of a statement that introduces exactly one demographic value. For instance, for every essay in the data split that considers a student's gender, the statement "The student is female." would be added to the prompt for all female students. Due to computational constraints, we do not perform Demographic-Aware inference with every possible demographic variable for every essay. Instead, we randomly choose one variable per essay. An example Demographic-Aware prompt format for an instruct model is shown in Appendix 4.

3.6 Counterfactual Demographic-Aware

LLMs can be sensitive to the slightest change in prompt. In order to evaluate whether score shifts that occur between Demographic-Blind and Demographic-Aware model outputs are instigated by the actual value of the demographic identifier or are instead a result of the *act of inclusion* of any identifier at all, we employ counterfactual values. In other words, for each *true* Demographic-Aware inference, we also perform a counterfactual inference for the same essay and demographic variable with every possible alternative demographic identifier for the given variable. For instance, if the variable under inspection is grade-level, and a student's true grade level is 8, we will prompt the model four additional times, asserting that the student is in 9th, 10th, 11th or 12th grade, respectively. For the counterfactual analysis, we omit Flan-T5 [3] due to computational costs.

3.7 Analysis

3.7.1 Demographic Blind vs Demographic-Aware. For each essay, we calculate score differences between Demographic-Blind and Demographic-Aware settings, as described above. For each true demographic value, we measure the average difference in the scores when essays are graded with an explicit identification of the value included in the prompt versus the scores when the essays are graded in the anonymous setting. We use a paired t-test with a null hypothesis of $p < 0.05$ between the actual and demographic average scores to test for a statistically significant difference.

3.7.2 Counterfactual Demographic-Aware. For every essay with a given true demographic identifier, we calculate the difference between the alternative score and the true score for all of the alternative identifiers of the same category. For an identical pair of identifier values, we take care to differentiate by the actual-to-counterfactual shift direction in order to account for any confusion that might occur due to the possibility of conflicting identifying language with the counterfactual values in the essay body itself. For every actual-to-counterfactual pair, we compute the mean of the pairwise differences (counterfactual less actual), the standard deviation of the pairwise differences, the effect size, and the cumulative actual and counterfactual score means. We use a paired t-test with a null hypothesis of $p < 0.05$ between the actual and counterfactual scores to test for a statistically significant difference.

4 Results

4.1 Demographic Blind vs Demographic-Aware Scoring

Model	Mean Diff	Std Diff	T-test Statistic	T-test P-value
Gemma	1.902	1.187	-142.20	**0.0**
Llama	-0.351	1.479	21.027	**1.5e-95**
Llama-Instruct	0.037	0.772	-4.18	**2.8e-4**
FlanT5	0.226	2.37	-0.799	0.423
Qwen	0.479	1.424	-29.81	**3.38e-18**

Table 1. Impact and Significance of Demographics on Scoring

Our analysis reveals significant variation in how models respond to the inclusion of demographic information. As shown in Table 1, four out of the five models demonstrated statistically significant changes in their scoring behavior when demographic information was included; only Flan-T5 showed no significant shift ($p = 0.423$). The magnitude and direction of these changes varied substantially across models. In particular, Gemma showed the most dramatic responses–on average, it increased scores by 1.9 rubric points ($p < 0.001$)–while Llama had an overall tendency to decrease scores, with a negative mean difference of -0.351 points ($p < 1.5e-95$). Llama-Instruct showed a slight positive shift (mean difference = 0.037, $p < 2.8e-4$) with the smallest standard deviation (0.772) among all models, suggesting more consistent but still statistically significant scoring changes in response to demographic information. Finally, Qwen displayed a moderate positive shift (mean difference = 0.479, $p < 3.38e-18$). Notably, with the exception of Llama-Instruct, all models show high standard deviation of above 1.1 rubric points, indicating overall substantial variability in their reactions to demographic information.

This variability is reflected in Figure 1; every model exhibited behavior changes when given true demographic information on an essay, but varied in how much the presence of any demographic information changed their behavior. As shown in the figure, models did not exhibit consistent behavior among demographic variables; for instance, Llama is seen penalizing economically disadvantaged essays, and Llama-instruct is seen slightly penalizing *White* labels while adding almost 0.5 points to *American Indian/Alaskan Native* essays.

The inclusion of race/ethnicity information produced large scoring variations; Qwen favored *Hispanic/Latino* students, while Gemma gave relatively more score boosting to *Asian/Pacific Islanders* students, and Llama penalized the presence of *American Indian/Alaskan Native* demographic information. The full results of our impact analysis for each model can be found in Appendix 2 and 3.

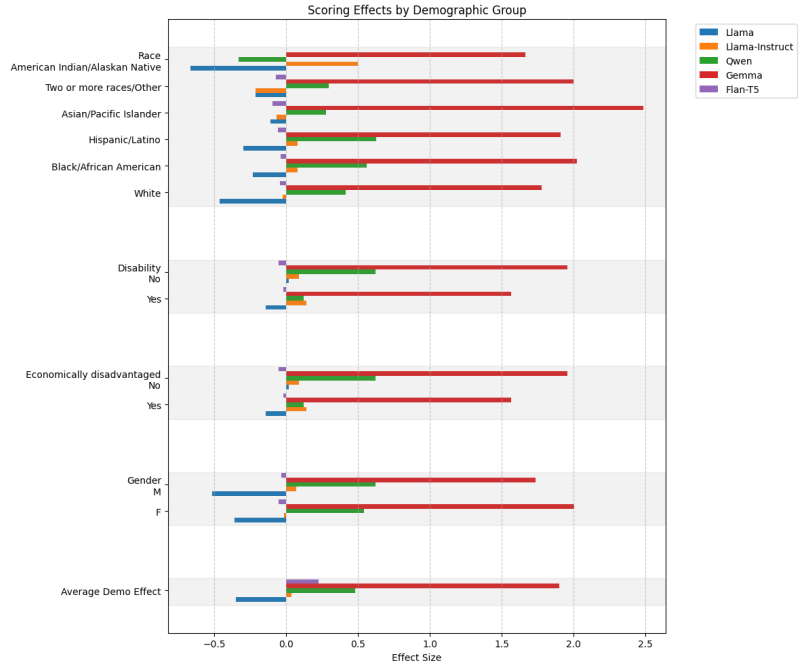

Fig. 1. Effect Sizes of True Demographic Variables

4.2 Counterfactual Demographic-Aware Scoring

We found varied results for both the statistical significance and direction of actual-to-counterfactual score differences across models and variables.

Gender was the least informative demographic variable of the six available to us; only two of eight experiments produced statistically significant results (both of which indicated a tendency to lower the score when shifting from *true* female to *false* male). The results of our statistical analysis for counterfactual gender shifts can be found in Appendix 4.

We found a variety of interesting and significant shifts for race/ethnicity identifiers. Most true race values produced consistent results across models in terms of the direction of the mean counterfactual shift. We found that when measuring true values versus all alternatives, essays written by *Hispanic/Latino*, *Black/African American*, and *Asian/Pacific Islander* students were likely to experience a decrease in score. For example, this could mean that for the same essay, if a *Hispanic/Latino* student is falsely identified as some race other than *Hispanic/Latino*, they are likely to receive a lower score than they would if they were correctly identified as *Hispanic/Latino*. Conversely, essays written by *White* students were more likely to experience an increase in score when substituting another race/ethnicity. Shifts for students who identified as *American Indian/Alaskan Native* or *Two or more races/Other* were inconsistent in direction and significance. The most consistent shifts occurred for *White* and *Black/African American* students, with all four models producing significant positive and negative shifts, respectively. The results of our statistical analysis of counterfactual race/ethnicity shifts can be found in Appendix 5. A visualization of the shifts between each true and false race variable

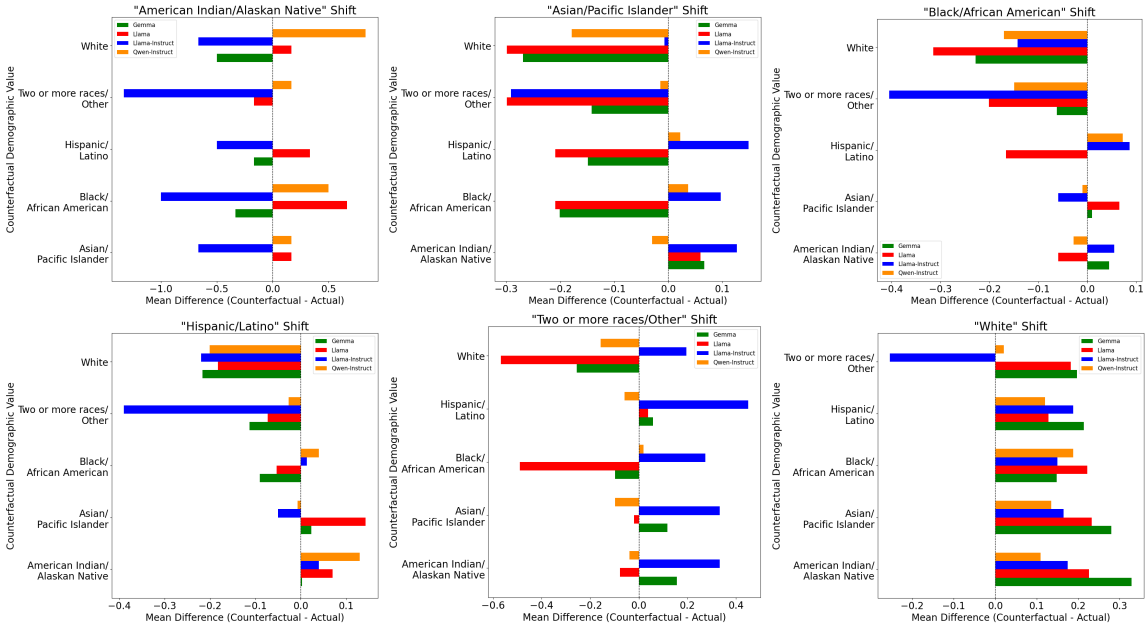

Fig. 2. Mean Counterfactual Race/Ethnicity Shifts. The actual value is identified in the title of each graph, and the counterfactual values are enumerated along the y-axis.

can be found in Figure 2.

ELL status, economic disadvantage status, and disability status also yielded significant and interesting counterfactual shifts. We found that when students were not actually disadvantaged by the aforementioned metrics, but they were falsely identified as being disadvantaged, they were likely to experience a decrease in score. Inversely, when students were actually disadvantaged by the aforementioned metrics, but they were falsely identified as not being disadvantaged, they were likely to experience an increase in score. As seen in Figure 3, the counterfactual effects for the two directions of shift (not-disadvantaged-to-disadvantaged and disadvantaged-to-not-disadvantaged) mirror each other fairly consistently. The results of our statistical analysis of the above shifts can be found in Appendix 6 7 8.

5 Discussion

Our research revealed significant evidence that the inclusion of demographic information can introduce bias in LLM-based essay grading across multiple models and demographic variables. Our findings address our initial research question about the presence and extent of bias in LLMs as evaluators of student writing. We discover large but inconsistent biases across sensitive demographic categories such as race, disability status, and economic disadvantage. Specifically, we expand on the work by Tamkin, et.al [9], in which the authors find consistent patterns of discrimination across prompts and demographics for Claude models. In our work, we see that across model architectures, there are patterns of bias within a model (for instance, Gemma tended to bias positively with any demographic information), but the nature of that bias is variable based on model architecture.

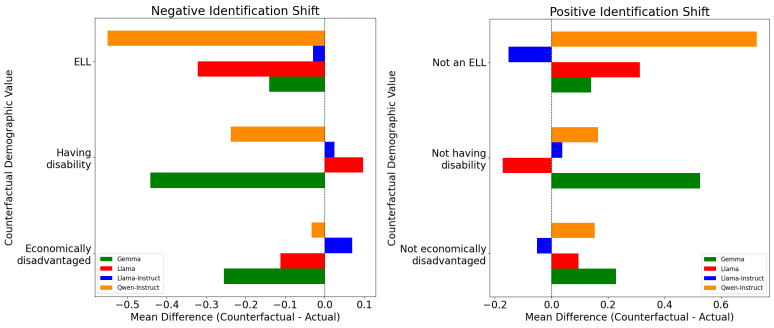

Fig. 3. Mean Counterfactual ELL, Economic Disadvantage, and Disability Shifts. The actual value is identified in the title of each graph, and the counterfactual values are enumerated along the y-axis. A *Positive Identification* of the actual value means that students are identified as being disadvantaged by that value.

Our counterfactual demographic analysis also reveals large implications for bias evaluation. White students' essays consistently received higher scores after their racial identity was altered, while essays written by Hispanic/Latino, Black/African American, and Asian/Pacific Islander students were significantly likely to experience score decreases when assigned to counterfactual racial variables. The consistency of these effects across multiple models could indicate that these biases are rooted in the pretraining data; in addition, the directionality of these score changes provides additional insight into how LLMs process and utilize demographic information. Many of these biases relate to ongoing questions about how models adapt to implicit and explicit patterns in their training data. We found above that counterfactual assignment of ELL status, economic disadvantage, and disability status demographics benefited students who were not disadvantaged. Our findings raise important questions about the ethical implications of using LLMs in high-stakes educational assessment contexts, where these biases could contribute to perpetuating existing educational disparities.

6 Conclusion and Future Work

Overall, our work contributes to an ongoing field of research into bias in LLMs, particularly in LLM-as-a-judge tasks. Our project has provided evidence that current LLMs exhibit significant patterns of demographic biases when evaluating student persuasive essays, with particularly pronounced effects related to race, economic status, and disability status. Our findings have significant implications for the educational technology industry's responsible deployment of LLM-based assessment tools. While LLMs show promise for reducing teacher workload or providing tools for scalable and standardized assessment, it is important to note the potential for these tools to perpetuate or exacerbate existing educational inequities. Future research should focus on developing robust LLM bias detection and mitigation strategies through techniques like chain-of-thought prompting, training data de-biasing, and counterfactual testing. Chain-of-thought reasoning in particular could be utilized to decode an LLM essay grader's decision making process and to investigate the model's justifications for these score differences, Further investigation is needed into how different architectures and training approaches influence bias patterns.

References

[1] Noel F. Ayoub, Karthik Balakrishnan, Marc S. Ayoub, Thomas F. Barrett, Abel P. David, and Stacey T. Gray. 2024. Inherent Bias in Large Language Models: A Random Sampling Analysis. *Mayo Clinic Proceedings: Digital Health* 2, 2 (June 2024), 186–191. https://doi.org/10.1016/j.mcpdig.2024.03.003

[2] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the Judge? A Study on Judgement Biases. (2024). arXiv:arXiv:2402.10669

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. https://doi.org/10.48550/ARXIV.2210.11416

[4] S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing* 61 (July 2024), 100865. https://doi.org/10.1016/j.asw.2024.100865

[5] Google. 2024. Gemma-2-2b. https://huggingface.co/google/gemma-2-2b

[6] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A Survey on Fairness in Large Language Models. (2023). arXiv:arXiv:2308.10149

[7] Meta. 2024. Llama-3.2-1B. https://huggingface.co/meta-llama/Llama-3.2-1B

[8] Meta. 2024. Llama-3.2-3B-Instruct. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[9] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and Mitigating Discrimination in Language Model Decisions. (2023). arXiv:arXiv:2312.03689

[10] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[11] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. (2024). arXiv:arXiv:2406.12624

[12] Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine* 4, 1 (Sept. 2024). https://doi.org/10.1038/s43856-024-00601-z

[13] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. (2024). arXiv:arXiv:2410.02736

7 Appendix

```
[{"role": "system", "content":  """You are an essay grader who assigns a holistic score for
the essay based on the rubric below. For the following evaluations you will need to use a
grading scale between 1 (minimum) and 6 (maximum). {HOLISTIC_RATING_FORM}"""},
{"role": "user", "content": """Consider a student's essay response. The student is
{DEMOGRAPHIC_IDENTIFIER}. Based on the following prompt: '{ESSAY_PROMPT}', Assign a holistic
grade to the essay: '{ESSAY_BODY}'. Do not explain your reasoning. Only respond with the
score."""}]
```

Fig. 4. Example Instruct Prompt Format (Demographic-Aware)

Model	Avg	Female	Male	Econ Dis	Not Econ Dis	Disability	No Disability
Llama	-0.35	-0.36	-0.52	-0.44	-0.24	-0.29	-0.47
Llama-Instruct	0.04	-0.02	0.07	0.04	0.00	-0.01	0.04
Qwen	0.48	0.54	0.62	0.40	0.23	0.18	0.52
Gemma	1.9	2.0	1.7	1.4	2.1	1.6	2.2
Flan-T5	0.23	-0.05	-0.04	-0.09	2.8	-0.07	-0.08

Table 2. Score effect across general demographic groups

Model	White	Hispanic/Latino	Black	Asian/PI	Multiple	Native
Llama	-0.46	-0.30	-0.23	-0.11	-0.22	-0.67
Llama-Instruct	-0.02	0.08	0.08	-0.07	-0.22	0.50
Qwen	0.42	0.63	0.56	0.28	0.29	-0.33
Gemma	1.8	1.9	2.0	2.5	2.0	1.7
Flan-T5	-0.04	-0.06	-0.04	-0.09	-0.07	0.00

Table 3. Score effect across racial/ethnic demographic groups

Actual Gender vs. Counterfactual Value							
Actual Value	False Value	Model	Mean Diff	Actual Mean	False Mean	Effect Size	P-Value
M	F	Llama-3.2-1B	0.018	4.459	4.477	0.014	0.778
		Llama-3.2-3B-Instruct	-0.036	3.929	3.892	-0.053	0.174
		Qwen2.5-1.5B-Instruct	0.002	5.17	5.172	0.001	0.98
		Gemma-2-2b	0.029	2.736	2.764	0.027	0.074
F	M	Llama-3.2-1B	-0.07	4.616	4.546	-0.053	0.27
		Llama-3.2-3B-Instruct	-0.047	3.986	3.939	-0.074	0.059
		Qwen2.5-1.5B-Instruct	-0.087	5.162	5.075	-0.06	**0.025**
		Gemma-2-2b	-0.08	3.003	2.924	-0.073	**0.000**

Table 4. Counterfactual Analysis for Gender.

Actual Race/Ethnicity vs. All Counterfactual Values						
Actual Value	Model	Mean Diff	Actual Mean	False Mean	Effect Size	P-Value
Hispanic/Latino	Llama-3.2-1B	-0.019	4.640	4.621	-0.015	0.615
	Llama-3.2-3B-Instruct	-0.121	3.797	3.675	-0.168	**0.000**
	Qwen2.5-1.5B-Instruct	-0.013	4.883	4.870	-0.008	0.687
	Gemma-2-2b	-0.079	2.910	2.831	-0.071	**0.000**
White	Llama-3.2-1B	0.198	4.536	4.734	0.159	**0.000**
	Llama-3.2-3B-Instruct	0.085	3.874	3.959	0.095	**0.000**
	Qwen2.5-1.5B-Instruct	0.115	5.079	5.194	0.077	**0.000**
	Gemma-2-2b	0.234	2.776	3.010	0.184	**0.000**
Black/African American	Llama-3.2-1B	-0.135	4.630	4.495	-0.103	**0.002**
	Llama-3.2-3B-Instruct	-0.093	3.945	3.852	-0.138	**0.000**
	Qwen2.5-1.5B-Instruct	-0.057	5.093	5.037	-0.038	**0.048**
	Gemma-2-2b	-0.047	3.024	2.977	-0.040	**0.006**
Two or more races/Other	Llama-3.2-1B	-0.224	4.549	4.325	-0.163	**0.042**
	Llama-3.2-3B-Instruct	0.318	3.706	4.024	0.478	**0.000**
	Qwen2.5-1.5B-Instruct	-0.067	5.098	5.031	-0.043	0.371
	Gemma-2-2b	-0.004	3.000	2.996	-0.004	0.926
American Indian/ Alaskan Native	Llama-3.2-1B	0.233	4.167	4.400	0.143	0.575
	Llama-3.2-3B-Instruct	-0.833	4.500	3.667	-1.348	**0.000**
	Qwen2.5-1.5B-Instruct	0.333	4.667	5.000	0.182	0.106
	Gemma-2-2b	-0.200	2.667	2.467	-0.157	0.184
Asian/Pacific Islander	Llama-3.2-1B	-0.191	4.978	4.787	-0.183	**0.000**
	Llama-3.2-3B-Instruct	0.015	4.015	4.030	0.022	0.581
	Qwen2.5-1.5B-Instruct	-0.033	5.201	5.169	-0.023	0.389
	Gemma-2-2b	-0.139	3.485	3.346	-0.112	**0.000**

Table 5. Counterfactual Analysis for Race/Ethnicity.

Actual English Language Learner Status vs. Counterfactual Value							
Actual Value	False Value	Model	Mean Diff	Actual Mean	False Mean	Effect Size	P-Value
Yes	No	Llama-3.2-1B	0.313	4.596	4.909	0.284	**0.027**
		Llama-3.2-3B-Instruct	-0.152	3.687	3.535	-0.243	**0.031**
		Qwen2.5-1.5B-Instruct	0.727	4.131	4.859	0.453	**0.000**
		Gemma-2-2b	0.141	2.566	2.707	0.151	0.075
No	Yes	Llama-3.2-1B	-0.321	4.981	4.66	-0.285	**0.000**
		Llama-3.2-3B-Instruct	-0.029	3.937	3.908	-0.048	0.148
		Qwen2.5-1.5B-Instruct	-0.55	5.221	4.671	-0.356	**0.000**
		Gemma-2-2b	-0.141	2.958	2.818	-0.127	**0.000**

Table 6. Counterfactual Analysis for ELL Status.

Actual Economic Status vs. Counterfactual Value							
Actual Value	False Value	Model	Mean Diff	Actual Mean	False Mean	Effect Size	P-Value
Economically disadvantaged	Not economically disadvantaged	Llama-3.2-1B	0.096	4.428	4.525	0.074	0.175
		Llama-3.2-3B-Instruct	-0.051	3.741	3.690	-0.071	0.141
		Qwen2.5-1.5B-Instruct	0.154	4.452	4.606	0.089	**0.009**
		Gemma-2-2b	0.229	2.398	2.627	0.211	**0.000**
Not economically disadvantaged	Economically disadvantaged	Llama-3.2-1B	-0.112	4.820	4.708	-0.100	**0.033**
		Llama-3.2-3B-Instruct	0.070	3.970	4.040	0.102	**0.007**
		Qwen2.5-1.5B-Instruct	-0.033	5.095	5.062	-0.022	0.361
		Gemma-2-2b	-0.255	3.058	2.803	-0.217	**0.000**

Table 7. Counterfactual Analysis for Economic Disadvantage Status.

Actual Disability Status vs. Counterfactual Value							
Actual Value	False Value	Model	Mean Diff	Actual Mean	False Mean	Effect Size	P-Value
Identified as having disability	Not identified as having disability	Llama-3.2-1B	-0.173	4.414	4.241	-0.119	0.240
		Llama-3.2-3B-Instruct	0.038	3.744	3.782	0.052	0.531
		Qwen2.5-1.5B-Instruct	0.165	4.632	4.797	0.099	**0.034**
		Gemma-2-2b	0.526	2.586	3.113	0.391	**0.000**
Not identified as having disability	Identified as having disability	Llama-3.2-1B	0.097	4.538	4.634	0.075	0.052
		Llama-3.2-3B-Instruct	0.025	3.939	3.964	0.036	0.231
		Qwen2.5-1.5B-Instruct	-0.238	5.151	4.913	-0.154	**0.000**
		Gemma-2-2b	-0.442	3.189	2.747	-0.321	**0.000**

Table 8. Counterfactual Analysis for Disability Status.

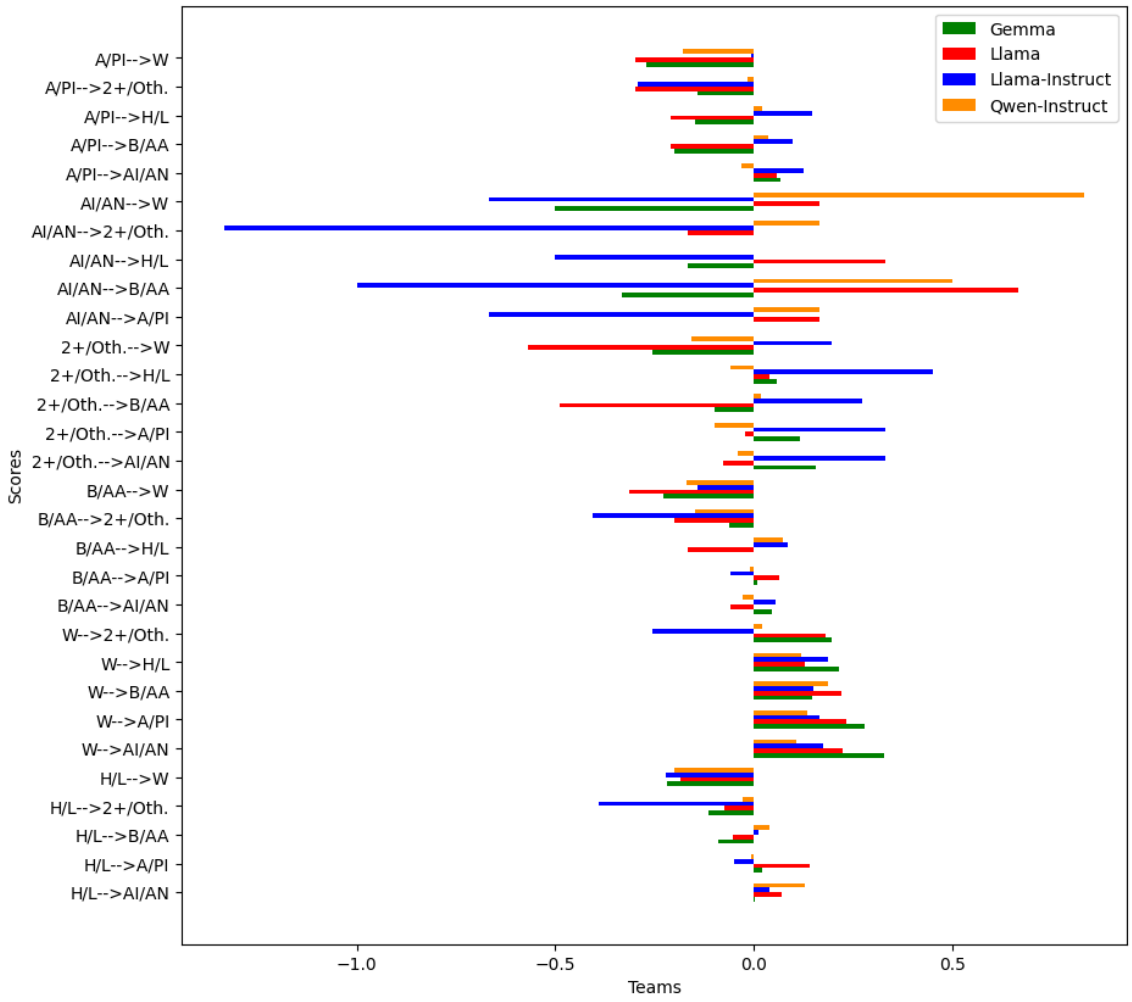

Fig. 5. Mean Counterfactual Shifts for All Race/Ethnicity Pairs.