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

TARA SHUKLA* and PAIGE VEGNA*

Large Language Models 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 bias in general-purpose LLMs as persuasive essay evaluators. We plan to employ the PERSUADE dataset [3] of persuasive essays that are associated with human-generated labels for quality as well and student demographic information. We will utilize counterfactual fairness and stereotypical association paradigms [4] in our prompt-engineering schema in order to search for bias.

ACM Reference Format:
Tara Shukla and Paige Vegna. 2024. Grading Bias: The Impact of Demographic Indicators on LLM Evaluations of Persuasive Essays. 1, 1 (November 2024), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models (LLMs) 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 in 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 arise from human grading. The use of LLMs to augment or replace subjective human judgements is called LLM-as-a-judge, and it is an emerging field of research in the field [6].

However, 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. Therefore, we seek to investigate the degree of fairness and biases of LLM graders using persuasive essays. To this end, we employ the PERSUADE 2.0 dataset [3], 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 propose a multi-pronged analysis in which we prompt LLMs to grade the essays for the same characteristics as the human-graders while incorporating demographic information about the student with varying degrees of explicitness and truth, and analyze the variance in the LLM-generated grades for bias. Our usage of the PERSUADE dataset differs from its intended purpose in that we seek to take advantage of its demographic information to evaluate the grading bias in existing LLMs, while the original

*Both authors contributed equally to this proposal.

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

Manuscript submitted to ACM

purpose of the dataset was for a competition in which researchers could use the human-generated quality labels to train the best LLM-grader in terms of its alignment to the human-grades without necessarily taking bias into account [3].

2 Literature Review

The evaluation of LLM fairness in the past has been previously explored through existing literature in the field; many studies have been done finding that LLM judgment echoes human biases inevitably present in training datasets. Also, many scholars have created strategies to evaluate or combat bias in LLM judgements, such as benchmarking or prompt interventions.

Firstly, the problem of biased LLMs has been examined in many different fields; for instance, in [1], Ayoub et.al report on racial bias in the diagnosis and treatment recommendations of LLMs used in healthcare; similarly, in [7], Yang et.al found that LLMs are inherently biased along racial lines for treatment predictions. Also, in [5], Tamkin et.al found both positive and negative discrimination in language model decision-making tasks across 70 scenarios, and analyzed how prompt-based interventions can reduce discrimination. In [4]„ Li et.al 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. During our project we will be reviewing all of these works to gain a deeper understanding of past work in LLM evaluation.

In addition, there exists a field of research devoted to the study of LLM-as-a-judge; in [6], Thakur et al. 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. In [8], Ye et.al propose CALM, a framework for evaluating different types of biases in LLM judges, and conclude that there is room to improve the robustness of LLM decision making across different types of bias; during our project we will understand and incorporate this paper's findings into our methodology and research. Finally, in [2], Chen et.al 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. Our research will contribute to the field's continued exploration of bias in LLM judgements by focusing on demographic bias in the task of writing evaluation.

3 Data and Methods

3.1 PERSUADE Dataset

For our project, we will be using the PERSUADE 2.0 dataset as referenced in [3]. This dataset contains over 25,000 argumentative essays, written by students in grades 6-12 about 15 prompts; these datasets are human annotated with different characteristics on argument effectiveness and writing quality. This dataset also includes student demographic information such as race, ethnicity, gender, and financial status.

3.2 Approach

In order to encourage LLMs to grade essays on the same characteristics and standards as the human-graders, we will employ the annotation scheme provided by the creators of the PERSUADE corpus, which provides descriptions and examples of what is considered effective, adequate, and ineffective for each discourse element [3]. We will incorporate

this information as a system prompt (or equivalent) for each LLM that we evaluate. In the user prompt, we will include the essay to be evaluated and, optionally, some form of demographic indicator. Each essay in our testing corpus will be independently graded four times, and we will use counterfactual fairness an stereotypical association mechanisms [4] to evaluate the LLM's bias. First, we will prompt the LLM without any additional demographic indicator. Secondly, we will prompt with the explicit addition of a true demographic profile of the student who wrote the essay. Thirdly, we will prompt with the explicit addition of a false demographic profile in which one or more of the student's covariates is altered. Fourthly, we will prompt with an implicit non-canonical indication of demographic information. For instance, we could include proxies rooted in stereotypes like fake names to indicate race or gender, or fake school type (public, private, charter, etc.) to indicate socioeconomic status. We intend to evaluate a small set of general-purpose LLMs from sources like OpenAI, Anthropic, and Microsoft.

To evaluate the presence or absence of bias in each LLM, we will run a statistical analysis of the heterogeneity of discourse element scores and overall quality for the provided demographic covariates for all the grades generated with no additional demographic information. We will then repeat this for the human-generated grades, and compare the results. Note that we do not propose that the human-generated grades are a "control" group in terms of bias. Rather, we acknowledge that there may be bias in those scores as well, and we merely wish to compare the heterogeneity of the human and AI graders. In addition, we will compare the grades generated with no demographic indicators to those generated with demographic indicators. Finally, we will evaluate the variance in LLM-generated grades between factual versus counterfactual and explicit versus implicit demographic indicators.

4 Timeline

Give a week by week break down of milestones for the project, starting in week 7, ending week 13. Describe responsibilities for each group member.

- Week 7: We refine problem formulation and methodology; Paige conducts exploratory data analysis while Tara refines the strategies through which we will define and find different types of bias. We will start on the prompt-engineering schema.
- Week 8: We will choose different open-sourced LLMs to work with, and set up and perform the experiements. We will explore several different LLMs, and Paige and Tara will each work on several models.
- Week 9: We will begin analyzing the results of the LLM judgements, including debugging and drawing out initial results. We will continue working on our separate models.
- Week 10: We will extend the project if we have time; we could examine adjacent research questions with other data or add another model. Otherwise this week is flexible– we will finalize the previous steps or begin our final analysis.
- Week 11: We will perform our final statistical Analysis. Paige will work on calculating the variance between the LLM-generated grades according to the factual/counterfactual and explicit/implicit prompting schema. Tara will work on analyzing the heterogeneity of both the LLM- and human-generated grades without any demographic indicators, and comparing the LLM-generated responses with demographic indicators to those without.
- Week 12: We will write our portions of the paper, incorporating all the methodologies we have used and the results of our statistical analysis. We will work together on the discussion and future work sections.
- Week 13: We will both finalize our written portions and work together to edit and revise our final draft.

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] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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