Analyzing Bias in Large Language Model Solutions for Assisted Writing Feedback Tools: Lessons from the Feedback Prize Competition Series

Perpetual Baffour, Tor Saxberg and Scott Crossley

Abstract

This paper analyzes winning solutions from the Feedback Prize competition series hosted from 2021-2022. The competitions sought to improve Assisted Writing Feedback Tools (AWFTs) by crowdsourcing Large Language Model (LLM) solutions for evaluating student writing. The winning LLM-based solutions are freely available for incorporation into educational applications, but the models need to be assessed for performance and other factors. This study reports the performance accuracy of Feedback Prize-winning models based on demographic factors such as student race/ethnicity, economic disadvantage, and English Language Learner status. Two competitions are analyzed. The first, which focused on identifying discourse elements, demonstrated minimal bias based on students' demographic factors. However, the second competition, which aimed to predict discourse effectiveness, exhibited moderate bias.

1 Introduction

Assisted writing feedback tools (AWFTs) are a promising example of educational applications using Natural Language Processing (NLP) algorithms that can innovate and accelerate student learning (Nunes, Cordeiro, Limpo, & Castro, 2022). Recent advances in large language models (LLMs) have increased AWFTs' capabilities to process and provide feedback on student writing with human-like sophistication (Kasneci et al., 2023). The Feedback Prize competition series, hosted on Kaggle in 2021-2022, was an important step in advancing AWFTs potential by crowdsourcing innovative LLM solutions for assessing and evaluating student writing that were open science (The Learning Agency Lab, n.d.).

The competitions were a success with over 6,000 teams participating and over 100,000 open-source algorithms developed. (The Learning Agency Lab, n.d.) However, these algorithms have not been reported outside of the Kaggle interface, limiting knowledge of their use and minimizing potential adoption into educational applications. Additionally, the algorithms have not been assessed for bias, which may limit their effectiveness in a classroom setting, especially if that bias is aimed towards student populations that have been historically marginalized. The purpose of this study is to report initial performance for the winning Feedback Prize models and to disaggregate performance accuracy in demographic factors including race/ethnicity, economic disadvantage, and English Language Learner (ELL) status.

2 PERSUADE Corpus

The first two competitions in the Feedback Prize series were based on the PERSUADE (Persuasive Essays for Rating, Selecting, Analyzing, and Understanding Discourse Elements) corpus, a collection of ~25,000 argumentative essays written by students in the U.S. in grades 6 through 12 (Crossley et al., 2022). The essays were annotated by experts for discourse elements and the effectiveness of the discourse elements. Discourse elements refer to a span of text that performs a specific rhetorical or argumentative function, while discourse effectiveness is a rating of the quality of the discourse element in supporting the writer's overall argument. The effectiveness scale included Ineffective, Adequate, and Effective ratings. The annotation scheme for discourse elements is based on an adapted or simplified version of the Toulmin argumentative framework (Stapleton & Wu, 2015).

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 242–246

The discourse elements that were annotated for each essay were:

- **Lead**. An introduction begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis.
- **Position**. An opinion or conclusion on the main question.
- **Claim**. A claim that supports the position.
- **Counterclaim**. A claim that refutes another claim or gives an opposing reason to the position.
- **Rebuttal**. A claim that refutes a counterclaim.
- **Evidence**. Ideas or examples that support claims, counterclaims, rebuttals, or the position.
- **Concluding Statement**. A concluding statement that restates the position and claims.

The essays were annotated using a rigorous, double-blind rating process with 100 percent adjudication, such that each essay was independently reviewed by two expert raters and adjudicated by a third rater. Overall inter-rater agreement for discourse elements assessed using a weighted Cohen's Kappa was 0.73, which indicates relatively high reliability. While the experts who annotated the corpus for discourse elements also rated each element's effectiveness in supporting the writer's argument, misalignment in segmentation between the raters in the discourse elements make it difficult to calculate inter-rater reliability for the effectiveness labels.

3 Feedback Prize 1.0 Models

The first Feedback Prize competition, (Feedback Prize 1.0: Evaluating Student Writing) was hosted on Kaggle and involved the tasks of segmenting essays into smaller sections and assigning each section a discourse label such as lead, position, claim, and evidence. To evaluate performance, submissions were assessed based on the word overlap between ground truth and predicted outputs. A model prediction was considered correct (true positive) if there was at least a 50% word overlap between the machine-segmented section and the human-segmented section, as well as a match between their discourse label. False negatives were unmatched ground truths, and false positives were unmatched predictions. The final score was calculated by determining the number of true positives, false positives, and false negatives for each class (i.e., discourse label) and taking the macro F1 score across all classes.

Table 1: True positive rate (TPR) by English Language Learner status of student writer, Feedback Prize 1.0 2nd place

Status	**N**	**TPR**	**SD**
ELL	7,565	0.717	0.235
Not ELL	81,207	0.726	0.220
All	*88,772*	*0.725*	*0.221*

Table 2: True positive rate (TPR) by economic status of student writer, Feedback Prize 1.0 2nd place

Status	**N**	**TPR**	**SD**
Disadvantaged	35,696	0.713	0.226
NDA	42,698	0.743	0.214
All	*78,394*	*0.729*	*0.221*

*Note: NDA refers to non-disadvantaged students.

The analysis in this paper examines the second-place, third-place, and sixth-place winning solutions from this competition. Overall, the winning solutions were broadly based on ensembles of large-scale, pre-trained Transformers, paired with custom pre-processing and post-processing techniques to improve accuracy. The first-place model was not analyzed because its complexity made it difficult to replicate and impractical in educational settings. The overall macro F1 score did not differ significantly between the second-place, third-place, and sixth-place solutions, with values of .740, .740, and .732, respectively.

To assess potential bias in the models, performance accuracy was further disaggregated by demographic factors (race/ethnicity, English Language Learner status, and economic disadvantage) and discourse effectiveness (Ineffective, Adequate, Effective). Specifically, T-tests and ANOVAs indicated that the average true positive rate (TPR) per essay of the second-place, third-place, and sixth-place models significantly varied based on demographic factors, but the effect sizes were small (see Tables 1-3). None of the t-tests or ANOVA tests reported any results with a p-value < 0.01 and a Cohen's $d > 0.2$. For instance, the t-test comparing TPR differences between ELL and non-ELL writing showed a p-value of 0.03 and Cohen's d of 0.103 for the second-place model,

Table 3: True positive rate (TPR) by race/ethnicity of student writer, Feedback Prize 1.0 2nd place

Race/Ethnicity	N	TPR	SD
White	42,197	0.723	0.217
Black	17,060	0.722	0.228
Hispanic	23,055	0.712	0.229
Asian	6,814	0.777	0.198
American Indian	574	0.728	0.226
Multiple	3,884	0.743	0.197
All	*93,584*	*0.726*	*0.221*

suggesting a negligible difference in model performance.

4 Feedback Prize 2.0 Models

The second Feedback Prize competition (Feedback Prize 2.0: Predicting Effective Arguments) also hosted on Kaggle required models to predict the effectiveness rating of discourse labels, using multi-class logarithmic loss as the evaluation metric. More specifically, for each discourse label, the model had to submit the probabilities (or the likelihood) that the label belongs to each of the three effectiveness ratings (Ineffective, Adequate, Effective). The closer the predicted probabilities were to the actual true label, the higher the model score would be. Feedback Prize 2.0 also prioritized computationally efficient algorithms, with a prize-incentivized "Efficiency Track" that evaluated submissions for both accuracy and speed.

Feedback Prize 2.0 comprised a smaller subset of the data from the first competition (around 6,900 out of the 26,000 essays), due to a need for greater balance in effectiveness scores. In the complete PERSUADE corpus, only 4% of discourse elements were labeled Ineffective while 80% were labeled Adequate and 16% were labeled Effective. The subset used in Feedback Prize 2.0 corpus had a distribution of 18% Ineffective, 24% Effective, and 58% Adequate, resulting in greater balance.

The analysis presented in this paper examines the performance of the winning models (first, second, and third place) in the Efficiency Track on the competition test set. A common trend among winning solutions from the Efficiency Track was to fine-tune a single pre-trained Transformer model on the competition dataset to minimize space and runtime requirements. The authors did not analyze the winners from the non-efficiency track because performance was similar, but computational demands were much higher. The

Figure 1: Performance accuracy by ELL status of student writer and discourse effectiveness label, Feedback Prize 2.0 Efficiency Track 1st place

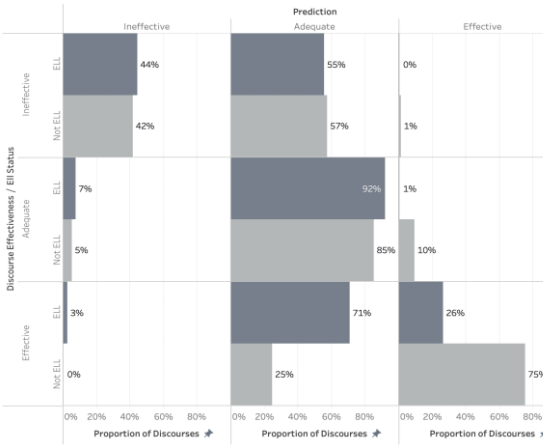
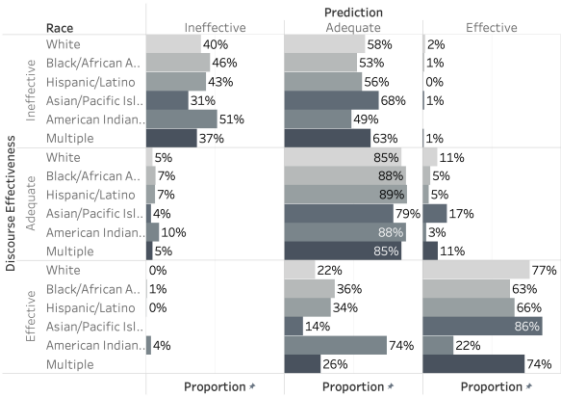


Figure 2: Performance accuracy by race/ethnicity of student writer and discourse effectiveness label, Feedback Prize 2.0 Efficiency Track 1st place



analysis consists of two parts. The first part examines the accuracy of the models in predicting the three original effectiveness ratings (Ineffective, Adequate, Effective). In the second part, the winning models' predictions were evaluated by grouping Ineffective and Adequate labels into a Non-Effective label, creating a binary outcome variable (Effective, Non-Effective). This analysis recoded the labels 'post hoc,' after the model submitted probabilities for all three original ratings. In both analyses, the model's predicted label was determined as the label with the highest predicted likelihood among the outputted probabilities.

4.1 Analysis of accuracy using original effectiveness ratings

The first part of the Feedback Prize 2.0 bias analysis found that the selected winning models

showed higher levels of bias for certain students compared to the winning models from Feedback Prize 1.0. This disparity can be attributed to patterns in the label distribution of the data. The data sample for the Feedback Prize 2.0 competition had a more balanced representation of minority and historically disadvantaged students in the overall sample, but there were roughly twice as many discourse elements labeled Ineffective from economically disadvantaged students and almost three times as many Effective discourses from non-disadvantaged students.

As a result, effective writing discourses from white, non-ELL, and economically advantaged students were more likely to receive higher ratings and the models amplified the existing disproportionate representation of effective writing found in the human-rated dataset. As shown in Figure 1, the first-place model was more accurate in identifying effective discourses in non-ELL writing (76% vs 27% accurate) with a statistically significant difference in likelihood scores (p-value ~0.000) and a larger effect size (Cohen's d ~0.671), as shown in Table 4. As shown in Table 5, the first-place model was also less accurate in predicting effective writing for economically disadvantaged students, and a t-test revealed that the difference in likelihood scores for effective discourses was statistically significant (p-value ~0.000) and the effect size was moderate (Cohen's d ~0.263). Similarly, accuracy disaggregated by the race/ethnicity of each student writer also showed statistically significant differences (p-values ~ 0.000), but with small effect sizes (Cohen's d ~ 0.15), as shown in Table 6 and Figure 2.

Table 4: Likelihood scores for effective discourses by English Language Learner status of student writer, Feedback Prize 2.0 Efficiency Track 1st place

Status	N	Likelihood	SD
ELL	2,623	0.028	0.083
Not ELL	19,853	0.246	0.321
All	*22,476*	*0.221*	*0.311*

Table 5: Likelihood scores for effective discourses by economic status of student writer, Feedback Prize 2.0 Efficiency Track 1st place

Status	N	Likelihood	SD
Disadvantaged	10,268	0.113	0.224
NDA	9,805	0.338	0.353
All	*20,073*	*0.223*	*0.315*

*Note: NDA refers to non-disadvantaged students.

Table 6: Likelihood scores for effective discourses by race/ethnicity of student writer, Feedback Prize 2.0 Efficiency Track 1st place

Race/ethnicity	N	Likelihood	SD
White	9,816	0.270	0.328
Black	4,157	0.133	0.246
Hispanic	6,218	0.149	0.261
Asian	1,721	0.398	0.370
Am. Ind.	179	0.096	0.176
Multiple	888	0.250	0.321
All	*22,979*	*0.220*	*0.310*

4.2 Analysis of accuracy using binary label of effectiveness

The second part of the analysis aimed to address the low sample size of Ineffective discourses in the dataset by recoding the effectiveness label as a binary variable. This involved combining Ineffective and Adequate discourses into a Non-Effective label. The goal was to examine whether similar levels of bias persisted in the recoded label. Combining Adequate and Ineffective discourse labels into a Non-Effective category did achieve greater balance in performance accuracy for the Non-Effective label, but there remained bias in the prediction of Effective discourses because white, non-ELL, and advantaged students remain overrepresented in this category, as shown in Figure 3.

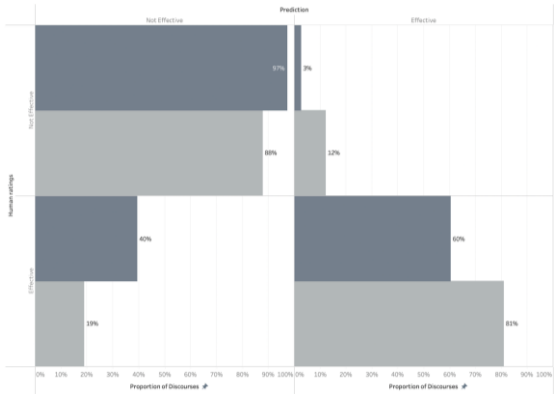
5. Discussion

The winning solutions across the first two Feedback Prize competitions reported a degree of accuracy comparable to that of humans, which is an important indicator of the models' strength. Additionally, since the models are open-source, they can quickly be adapted into educational applications to not only assess student writing at a summative level but to also provide fine-grained feedback to students at the formative level.

However, as noted in the analyses above, the winning solutions from the second competition that focused on predicting effective arguments showed a moderate degree of bias among factors related to race/ethnicity, economic status, and English Language Learner (ELL) status while the winning solutions from the first competition, which focused on annotating discourse elements, showed minimal bias.

It appears the models from Feedback Prize 2.0 amplified the biases inherent in the data despite not being explicitly trained with demographic

Figure 3: Performance accuracy for Non-Effective and Effective discourses, by student economic status, Feedback Prize 2.0 Efficiency Track 1st place



information. Data bias in label distribution, label agreement, and demographic representation in the PERSUADE corpus may have contributed to the model bias, but it is unclear how well these factors could be addressed given current writing achievement disparities in the U.S. educational system (National Center for Education Statistics, 2012). Using a binary classification for effectiveness (i.e., recoding the data as Effective or Ineffective) helped to mitigate the bias in the models to some degree. However, the use of models from Feedback Prize 2.0 for educational applications should be handled with care, especially when dealing with students from diverse populations.

These analyses demonstrate the importance of assessing algorithms for bias prior to wide-scale adoption. The results point to future work in building educational NLP applications like AWFTs to identify potential data biases in label distribution, agreement, or demographic representation before adoption to reduce bias in algorithmic outputs and help ensure fairness in systems. As can be seen with the PERSUADE corpus, bias will likely be present in any dataset that accurately represents populations in the United States because of achievement disparities in the educational systems.

References

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements corpus 1.0. *Assessing Writing*, 54. https://doi.org/10.1016/j.asw.2022.100667

Kevin A. Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology,* 8(1):23–34. https://doi.org/10.20982%2Ftqmp.08.1.p023

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, and Tina Seidel. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103. https://doi.org/10.1016/j.lindif.2023.102274

The Learning Agency Lab. (n.d.). The Feedback Prize: A case study in assisted writing feedback tools working paper. https://www.the-learning-agency-lab.com/the-feedback-prize-case-study/

National Center for Education Statistics. (2012). The Nation's Report Card: Writing 2011 (NCES 2012-470). National Center for Education Statistics.

Andreia Nunes, Carolina Cordeiro, Teresa Limpo, and São Luís Castro. 2021. Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2):599–620. https://doi.org/10.1111/jcal.12635

E. Michael Nussbaum, CarolAnne M. Kardash, and Steve Graham. 2005. The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, 97(2):157–169. https://psycnet.apa.org/doi/10.1037/0022-0663.97.2.157

Paul Stapleton and Yanming (Amy) Wu. 2015. Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, 17:12–23. https://doi.org/10.1016/j.jeap.2014.11.006