Augmenting AI Scoring of Essays with GPT-Generated Responses

Mo Zhang[1], Akshay Badola[2], Matthew Johnson[1] & Chen Li[1]

[1]ETS Research Institute, Educational Testing Service | US

[2]ETS India Assessment Service | India

Abstract: In this study, we examine the feasibility of augmenting student-written essays with those generated by large language models (LLMs) for scoring essays. We found that with correct instructions, generative AI systems such as GPT-4 and GPT-4o can generate essays similar to those written by students in terms of surface-level linguistic features, although material differences may still exist. Systematic analyses revealed that scoring models trained with synthetic data perform comparably to models trained using student essays, but the performance varies across prompts and the sizes of the model training sample. The augmented models could alleviate large discrepancies between human and AI scores on the subgroup level that may be introduced by a lack of training samples for a particular subgroup or due to inherent biases in LLMs. We also explored an established method – *DecompX* – on token importance to identify and explain AI predictions. Future research directions and limitations of this study are also discussed.

Keywords: AI scoring, writing assessment, large language model, GPT, sample augmentation

journal of WRITING RESEARCH

Zhang, M., Badola, A., Johnson, M., & Li, C. (2026). Augmenting AI scoring of essays with GPT-generated responses. *Journal of Writing Research*, *17(3),* 501-554. DOI: https://doi.org/10.17239/jowr-2026.17.03.06

Contact: Mo Zhang, ETS Research Institute, Educational Testing Service, 660 Rosedale Road, Princeton, NJ, 0854 | US – mzhang@ets.org

1. Background

The use of artificial intelligence (AI) to grade writing samples continues to garner increasing attention from all sectors of education, including the testing industry, academia, classroom teachers, educators, school/district/state officials, as well as the media. AI scoring, also known as automated scoring, was first introduced in the 1960s by Page (Page, 2966; Dikli, 2006). Since then, the capability has been operationally implemented in many high-stakes, large-scale writing assessments to generate scores or evaluations for millions of text responses each year. At the time of this writing, the amount of research literature including peer-reviewed publications on this topic is extensive and continues to grow. One of the primary motivations for exploring AI-driven scoring is the promise of enhanced objectivity and consistency (Zhang, 2013; Yao et al., 2019b; Bennett & Zhang, 2016). Traditional human scoring can be susceptible to biases, fatigue, and variations in interpretation, leading to inconsistencies across different raters and time periods (Williamson et al., 2012; Bejar et al., 2016). AI systems, on the other hand, offer the potential to apply predefined scoring rubrics consistently, ensuring a more standardized and consistent writing evaluation process (Bejar et al., 2006; Bennet & Bejar, 1998; Zhang & Bennett, 2022). Furthermore, AI can significantly reduce the time and resources required for human scoring, especially in large-scale assessments and daily classrooms, freeing up educators to focus on other critical aspects of teaching and learning (Johnson & Zhang, 2024; He, Gao & Chen, 2021).

The rise of generative AI and large language models (LLMs), which simulate human language, in recent years has introduced both significant opportunities and challenges. The use of LLMs is revolutionizing the capabilities and accuracy of automated scoring systems, opening new avenues for even more efficient and effective assessment practices. The feasibility of using LLMs for scoring is a rapidly growing area of research. Studies have explored various aspects, such as the alignment of LLM-generated scores with human grading rubrics (e.g., Fang, Lee, and Zhai, 2023), the impact of training data size on model performance (e.g., Zhang, Johnson, and Ruan, 2024), and the issues related to grading biases using LLMs (e.g., Johnson & Zhang, 2024). Despite the progress, many open questions remain. These include how different types of constructed responses, such as extended writing or short response items, affect AI scoring performance; whether the subject domain (e.g., science, engineering, mathematics, history, literature) influences the accuracy and fairness of AI scores; and what best practices for training scoring models ensure they generalize well across diverse student populations. The published studies underscored the great potential of LLMs to enhance the efficiency and consistency of scoring; but in the meantime, they highlight the complexities involved in ensuring their responsible use. Responsible use of AI calls for thoughtful evaluation of AI scoring in assessment, where the core aspects include scoring accuracy, fairness, and explainability (AERA, APA, & NCME, 2014; ETS, 2025). In particular, we argue that effective methods for detecting and mitigating biases in LLM-based AI scoring models and understanding inherent biases in the training data are perhaps one of the most crucial areas of research.

2. Research problem

LLM-based AI scoring models, which contain millions or even billions of parameters, require substantial training samples to effectively handle complex tasks. This requirement is particularly pronounced for certain subgroups of interest, such as demographic intersections or top or bottom score levels, where sample sizes may be even smaller. An unbalanced training sample can lead to scoring biases, compromising the fairness and accuracy of the assessments (Morris et al., 2025; Zhang et al., 2024). To collect student-written responses is both costly and time-consuming, and may be impractical for certain assessment programs. A potential solution to this challenge is to use generative AI to create synthetic writing samples, thereby artificially augmenting the available data. This approach can help balance and diversify the training samples and mitigate potential biases (Fang et al., 2023). There is a clear need for studies that address this critical issue, exploring the efficacy and implications of using synthetic data to enhance AI scoring models.

One basic question, however, is whether training a smaller discriminative language model using samples from a larger generative model is a sound approach. This question has been addressed in existing research in the natural language processing (NLP) domain. Studies have established that it is feasible, and in some cases beneficial, to train a smaller neural model with the output of a larger one (e.g., for model compression in Sun et al. (2019) and regularization in Yuan et al. (2021)). Initial work by Hinton, Vinyals, and Dean (2015) demonstrated this feasibility where they trained a decision tree on the outputs of a neural network, which was itself trained on the data set of interest. In the case of language models, cross-lingual transfer has been a popular approach to address the lack of data in low-resource languages, where sufficient training data is unavailable. The recent proliferation of variants of LLMs and their ease of access has further facilitated large data generation, conforming to user preferences. Those models, known as instruction-following models, are a type of language model designed to better understand and execute user instructions. These models are fine-tuned to follow specific prompts or commands given by users, making them more aligned with user intentions compared to general-purpose language models. Even before instruction-following models, LLMs have been used to augment data for downstream tasks. Yoo et al. (2021) reported enhanced performance with the use of synthetic data from GPT-3 on several downstream tasks by fine-tuning BERT (Devlin et al., 2019). Among instruction-following models, ChatGPT was used by Dai et al. (2023) to generate synthetic data and trained with this augmented data for text classification tasks using BERT. Whitehouse, Choudhury, and Aji (2023) used ChatGPT and GPT-4 for augmenting data for fine-tuning smaller multilingual models. For further detailed surveys, the reader may refer to Chen et al. (2023), Long et al. (2024) and Ding et al. (2024).

In the context of AI scoring, limited prior research is available that explored data augmentation by using the AI-generated responses to balance under-represented classes, with few examples. Morris et al. (2025) fine-tuned a DeBERTa-V3-large (He et al., 2021) to score NAEP math items and used Coedit-XL to augment high-scoring responses (scores 2 and 3), which were underrepresented in the training sample (Raheja et al., 2023). They used

existing responses scored 2 and 3 to generate additional "paraphrased" responses that were used to balance the labels. Fang et al. (2023) used GPT-4 (OpenAI, 2023) with customized prompts to augment data for under-represented minority students for the purpose of automatic scoring of science items. In this study, we attempt to expand this line of research by systematically examining how data augmentation affects AI scoring model performance in terms of scoring accuracy and fairness. We are particularly concerned with three research questions:

1. How similar are AI-generated essays to student essays?
2. How is the prediction accuracy of scoring models that are trained with AI augmented samples?
3. Are models trained using AI augmented samples fair for different racial/ethnic groups?

The remainder of the paper is organized as follows. In Section 3, we provide detailed information on the data set, sample partitions, AI essay generation, training and evaluation of scoring models, and model interpretations. Section 4 presents the findings for each of the three research questions. Finally, Section 5 offers a summary discussion, which also includes a discussion of the limitations of this study and suggestions for future research.

3. Methods

3.1 Data set

We used the public PERSUADE 2.0 data set, which includes essays written by 6th to 10th graders in the U.S. (Crossley, 2024). In this study, we analyzed seven prompts that required students to read source materials and write an essay. Table 1 lists the instructions associated with each prompt. The source materials are typically referenced in the writing instructions; however, these materials are not publicly accessible.

Table 1. Prompts and instructions

Prompts	Writing Instructions
1- Facial Action	In the article "Making Mona Lisa Smile," the author describes how a new technology called the Facial Action Coding System enables computers to identify human emotions. Using details from the article, write an essay arguing whether the use of this technology to read the emotional expressions of students in a classroom is valuable.
2-Electoral College	Write a letter to your state senator in which you argue in favor of keeping the Electoral College or changing to election by popular vote for the president of the United States. Use the information from the texts in your essay. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to include a claim; address counterclaims;

Prompts	Writing Instructions
	use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay.
3-Car-free Cities	Write an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Your essay must be based on ideas and information that can be found in the passage set. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay.
4-Driverless Cars	In the article, "Driverless Cars are Coming," the author presents both positive and negative aspects of driverless cars. Using details from the article, create an argument for or against the development of these cars. Be sure to include: your position on driverless cars; appropriate details from the article that support your position; an introduction, a body, and a conclusion to your argumentative essay.
5-Exploring Venus	In "The Challenge of Exploring Venus," the author suggests studying Venus is a worthy pursuit despite the dangers it presents. Using details from the article, write an essay evaluating how well the author supports this idea. Be sure to include: a claim that evaluates how well the author supports the idea that studying Venus is a worthy pursuit despite the dangers; and explanation of the evidence from the article that supports your claim; an introduction, a body, and a conclusion to your essay.
6-Face on Mars	You have read the article "Unmasking the Face on Mars." Imagine you are a scientist at NASA discussing the Face with someone who thinks it was created by aliens. Using information in the article, write an argumentative essay to convince someone that the Face is just a natural landform. Be sure to include: claims to support your argument that the Face is a natural landform; evidence from the article to support your claims; an introduction, a body, and a conclusion to your argumentative essay.
7-A Cowboy	You have just read the article, "A Cowboy Who Rode the Waves." Luke's participation in the Seagoing Cowboys program allowed him to experience adventures and visit many unique places. Using information from the article, write an argument from Luke's point of view convincing others to participate in the Seagoing Cowboys program. Be sure to include: reasons to join the program; details from the article to support Luke's claims; an introduction, a body, and a conclusion to your essay.

Only one rater score is available in the publicly released data set, although all essays were scored by two raters using a 6-point scale rubric. The authors reported an inter-rater quadratic weighted kappa of .745 and a correlation coefficient of .750 (Crossley et al., 2024).

Consequently, we did not conduct any true score evaluation (Johnson & McCaffrey, 2023) for the scoring models. The total sample ranged from 1,372 to 2,167 student essays across the seven prompts. The male and female students were relatively evenly distributed, with the proportion of female students ranging from 47.4% to 53.8% across prompts. The majority of the students were English-proficient, accounting for 76.8% to 97.5% of the sample across prompts. White students accounted for 39.3% to 49.2% of the sample across prompts. Hispanic/Latino students comprised 24.1% to 30.4% of the sample across prompts. About one-fifth were Black/African American students (15.0% to 20.0% across prompts). A small number of students were Asian/Pacific Islander, ranging from 2.9% to 6.0% across prompts. The other racial/ethnic groups, which included "Two or more races," "American Indian/Alaskan Native," and "Unidentified," were very small in this data set; in combination, they accounted for 3.3% to 5.2% of the sample across prompts.

Table 2 shows the distributions of essay score and response length in each prompt. Six of seven prompts have average rater scores between 2.85 ("Exploring Venus") and 3.19 ("Driverless Cars"). Prompt "A Cowboy" appeared to be an outlier with a noticeably lower mean rater score of 2.41 and a smaller standard deviation of 0.81. The average essay length ranged from 288.6 words ("A Cowboy") to 451.2 words ("Car-free Cities") across prompts, confirming that these essays were long, extended writing.

Table 2. Student essay score and length descriptions

Prompt	N	Score	Words
1-Facial Action	2,167	2.85 (sd = 1.09)	337.5 (sd = 139.1)
2-Electoral College	2,046	2.97 (sd = 1.20)	398.3 (sd = 164.3)
3-Car-free Cities	1,959	3.10 (sd = 1.03)	451.2 (sd = 180.1)
4-Driverless Cars	1,886	3.19 (sd = 0.92)	403.4 (sd = 146.2)
5-Exploring Venus	1,862	2.85 (sd = 1.12)	351.6 (sd = 145.6)
6-Face on Mars	1,583	2.95 (sd = 1.01)	336.0 (sd = 131.5)
7-A Cowboy	1,372	2.41 (sd = 0.81)	288.6 (sd = 121.9)

3.2 Sample partitions

The content generation, scoring model training, and scoring model evaluation were conducted on a prompt-by-prompt basis. For each prompt, the student samples, referred to as student essays, were randomly and evenly divided into three subsets. One subset was used for feeding the GPT generators. Another subset was used to address RQ1, which compared the linguistic

features between AI responses and human responses. After addressing RQ1, the first and second subsets, consisting of two-thirds of the total student samples, were combined to train the prompt-specific scoring models. The final subset was reserved for evaluating those prompt-specific scoring models and was not used until the evaluation phase.

3.3 Generating synthetic essays using GPT

As mentioned in the statement of the research problem, the motivation for data augmentation using samples from a LLM stems from the fact that discrepancies in the size of subgroups of interest within a data set can lead to poor performance on under-represented subgroups. For traditional machine learning methods, reweighting the training criterion or resampling/oversampling from the data set can often mitigate this issue. However, unlike traditional models with manual features, deep learning models learn their own feature representations of the data based on the data to which they are exposed. Oversampling from under-represented subgroups does not alleviate this issue, as deep learning models require diversity in the training data, and duplicated data leads to poor generalization (Hernandez et al., 2022; Tirumala et al., 2023). Recent methods for training language models have incorporated de-duplication (Lee et al., 2022) as a key element of data processing.

One avenue may then be to use data sampled from a LLM that is constrained by design to be similar to the training data, particularly for the under-represented subgroups, yet not identical to avoid degradation of the performance. Our approach lies along those lines, where we generate data across all prompts and score levels. Formally, we can think of this as distilling a subset of knowledge from an LLM f to a smaller AI scoring model f'. This LLM has been trained on a very large data set X of trillions of tokens (Touvron et al., 2023) and models the conditional distribution of sequences of tokens given some input tokens. It is worth noting that f' has also been *pre-trained* on a (usually smaller) data set X'. The scoring task, then, represents training on a *downstream* task with data X'', $|X| \gg |X'| \gg |X''|$, where X'' consists of (essay, score) pairs $(x'', y), x'' \in X'', y \in \mathbb{R}\ or\ y \in \mathbb{N}$. The task of generating augmented samples is then $X_g \sim f(X_p)$. Where X_p are the set of prompts that direct the model outputs toward the desired content and syntax and X_g are generated sentences or paragraphs. The model f'_θ, where θ denotes the weights of the model f', can then be trained by minimizing the mean squared error between the human scores in the training data augmented X_a with the generated data $X_a = X'' \cup X_g$. The objective is then $\text{minimize} \sum (Y - \hat{Y})^2$, $\hat{Y} = f'_\theta(X_a)$, where $\hat{Y}$ are the outputs from the model during training.

For creating the prompts (or instructions) X_p to guide the model we used the grading rubric and some custom directives which generated essays of desired quality. Specifically, we divided the prompt into five parts:

1. **System Prompt**: An initial *system* directive.
2. **Custom Directives**: Custom directives for edge cases.
3. **Essay Properties**: Rubric-driven properties that an essay should have. For example, the following was given for generating essays with score level 1:

```
The essays should demonstrate VERY LITTLE OR NO
MASTERY of writing and critical thinking and will
have several errors and lapses with following
qualities:
- They will have SOME RANDOM typos, misspellings,
syntactic errors and
punctuation errors.
- Major flaws in sentence structure, and errors in
mechanics which interferes
with meaning.
- Disorganized, disjoined with limited vocabulary
- They have NO CRITICAL THINKING, develop no viable
point of view or provide
no to little evidence
- These would get a SCORE OF 1 from a scale of 1 to
6 if judged by humans.
```

4. **Title and Generation Template**: This portion mentions the title of the essays and the number of essays to generate for a given score.
5. **Prompt Directives**: This portion contains generic guardrails for the model. We used the following to guard the generated syntax of the model from being too informal and to mimic the style of school-going children.

```
Each essay should mimic the style of a schoolgoing
child and should pass as if written by a child up
to grade 12.
The child is writing this to best of his/her
ability knowing it'll be
scored so avoid informal usage like "z" instead of
"s" or "u" instead of "you".
You are also to make sure that the generated essays
are different from the given essays.
The generated essays must also be of a similar
length to given essays.
```

In addition to the instructions mentioned above, we also provided sample essays written by students for a given prompt and score and asked the generator to produce similar but different responses. This was done to ensure that the style and syntax of the generated essays were similar to those of the student data while the content remained different. Each generation call to the model was accompanied with the necessary rubric, instructions, and example essays for the given prompt and score level. For each generation call, we randomly sampled five essays with replacement from the subset earmarked for generation, which was

roughly one third of the total student essays (described in Sample Partitions in Section 3.2), and included with the instructions.

The generation was done separately for GPT-4 and GPT-4o. For GPT-4, we generated approximately 120 essays per score level; for GPT-4o, we generated approximately 300 essays per score level. We further randomly sampled a small number of AI essays from each prompt for scoring model evaluations. It is worth noting that while we made an equal number of generation attempts (approximately 250 per score level) for both GPT-4 and GPT-4o, the responses from GPT-4 were of inferior quality and could not always be parsed reliably. Consequently, the final number of responses was approximately 120 per score level for GPT-4 and approximately 275 per score level for GPT-4o. The final counts of AI essays per score level for GPT-4 and GPT-4o used in scoring model training are given in Table 3. An example configuration for prompting the generation of score-level 1 essays is given in Appendix A. Identical template was used for all score levels.

Table 3. Counts of AI essays per score level in model training

	GPT-4							GPT-4o						
	Score Level							Score Level						
Prompt	1	2	3	4	5	6	Mean Score	1	2	3	4	5	6	Mean Score
1-Facial Action	153	155	157	120	113	112	3.27 (sd=1.72)	274	283	283	268	280	240	3.44 (sd=1.72)
2-Electoral College	121	132	123	87	122	96	3.36 (sd=1.68)	248	278	289	260	249	243	3.46 (sd=1.68)
3-Car-free Cities	142	144	127	121	142	119	3.42 (sd=1.70)	268	289	279	287	260	265	3.47 (sd=1.67)
4-Driverless Cars	142	132	128	112	127	109	3.37 (sd=1.71)	252	250	274	273	261	246	3.50 (sd=1.69)
5-Exploring Venus	118	128	134	99	119	119	3.46 (sd=1.71)	275	290	242	248	272	264	3.47 (sd=1.68)
6-Face on Mars	113	117	116	89	111	90	3.37 (sd=1.71)	274	287	282	268	270	269	3.47 (sd=1.73)
7-A Cowboy	100	87	101	84	68	73	3.30 (sd=1.69)	279	277	275	245	272	268	3.47 (sd=1.70)

3.4 Linguistic feature extraction

To address the first research question, we processed student and AI essays through a well-established, commercial automated essay scoring system, *e-rater*®, that (in its most recently released version) produced nine features that measure surface-level linguistic characteristics such as grammatical accuracy and vocabulary usage. Table 4, adapted from Yao, Haberman, and Zhang (2019a) and Yao et al. (2019b), provides high-level descriptions of these linguistic features. Chen, Zhang, and Bejar (2017) offers a more detailed explanation of three of the nine features (i.e., Grammar, Mechanics, and WordUsage). Additionally, we calculated the logarithm of the number of words to compare the response lengths between AI essays and student essays.

Table 4. Extracted text features for human and AI essays

Feature	Description
Grammar	Minus the square root of the number of grammatical errors detected per word
Syntax	A measure of the diversity of syntactic structure of the sentences in a response
Mechanics	Minus the square root of the number of mechanics errors detected per word
WordUsage	Minus the square root of the number of usage errors detected per word
Collocation	A measure of correctness of use of collocations and prepositions in daily life
NUnit	Logarithm of the number of discourse elements
UnitLength	Logarithm of the average number of words per discourse element
WordLen	Average number of characters per word
WordFreq	Minus the median Standard Frequency Index value of words
TextLen	Logarithm of the number of words in a response

3.5 Base and augmented scoring model training

For scoring models, we fine-tuned pretrained DeBERTa-V3-XSmall models (He et al., 2021). The pretrained DeBERTa-V3-XSmall model has 12 layers with a hidden size of 384 and 48M parameters. We enabled training of all layers in the model. Instead of adding a *pooling* layer commonly used for classification tasks, we simply averaged the outputs from the last hidden

layer and projected them to a single dimension. For updating the parameters, we used the *AdamW* optimizer (Loshchilov & Hutter, 2017). All scoring models were trained on a prompt basis, that is, prompt-specific, applying the same model training and parameter settings for all prompts. We isolated the test sets before training and used 10% of the training set for validation. All scores were normalized in the range of $0-1$ during training and rescaled to $1-6$ for evaluation. During the scoring model training process, we monitored mean squared error (MSE) loss and the quadratic weighted kappa (QWK) (Haberman, 2019) between observed human scores (i.e., one human rating per response) and model predictions. We saved the model after every training epoch (i.e., one sweep through the training data) and selected the final model for evaluation based on the lowest MSE loss and highest QWK. Appendix B gives full details of the configuration of the scoring model training process.

The "base" models were trained using student essays only and the "augmented" models on a mix of student and AI-generated essays. Additionally, we downsampled the student essays to 20%, 40%, 60%, and 80% during the scoring model training and compared the model performance. Specifically, for instance, in a "100% model," all student essays were used; in "80% downsampled models," 80% of the student essays were used for training the scoring models. So, as we downsampled student essays, the total model training sample sizes decreased. Further, since the number of AI essays remained constant in training the scoring models for a prompt, the proportion of AI essays in the training sample increased in augmented models accordingly as we downsampled the student essays.

Using "P1-Facial Action" for illustration, Figure 1 shows the training sample size by score level for base and augmented models. The sample size patterns for the other prompts are similar to those of prompt "P1-Facial Action" and are fully given in Appendix C. For the base models, when only 20% of student essays were used, the training sample was relatively small. In the case of "P1-Facial Action", the training sample sized in the "20% downsampled models" was only about 270. Across prompts, the training sample size ranged from 182 to 289 across prompts (see Appendix C for details). When using the full (100%) student sample, the base model training samples increased substantially: In "P1-Facial Action," for example, the training sample size was 1447; and across all prompts, the training sample size ranged from 912 to 1,447 across prompts. Also noted in Figure 1 is that score categories 1, 5, and 6 contained very few student essays. This pattern became even more pronounced when student essays were further sampled down.

Augmenting the base model training with AI (GPT-generated) essays significantly increased the total sample size at each score level. We opted to augment the data with respect to its original proportions. While data corresponding to underrepresented score points can be augmented selectively, earlier research on AI scoring suggested that it may not be desirable as it will affect the predictive distribution of the model which may deviate too much from the population (Zhang et al., 2012). By augmenting across all prompts and score points, we remained faithful to the original student essay score distribution. In our current approach, the score levels showing data scarcity (e.g., 1, 5, and 6, for which situation was even worse in the

downsampled models) became better represented in the model training process while maintaining their original proportions in the total training sample.

The resulting total training sample sizes in GPT-4 augmented models (student and AI essays combined) ranged from 1,425 to 2,257 across prompts and from 2,528 to 3,075 in GPT-4o augmented models. Thus, in the case of "augmented" scoring models, when student essays were sampled down by 80% (i.e., the 20% category), most of the training samples were AI-generated. Even when 100% of student essays were used in the augmented models, a large portion of the model training sample was still coming from generative AI, especially for those augmented by GPT-4o essays.

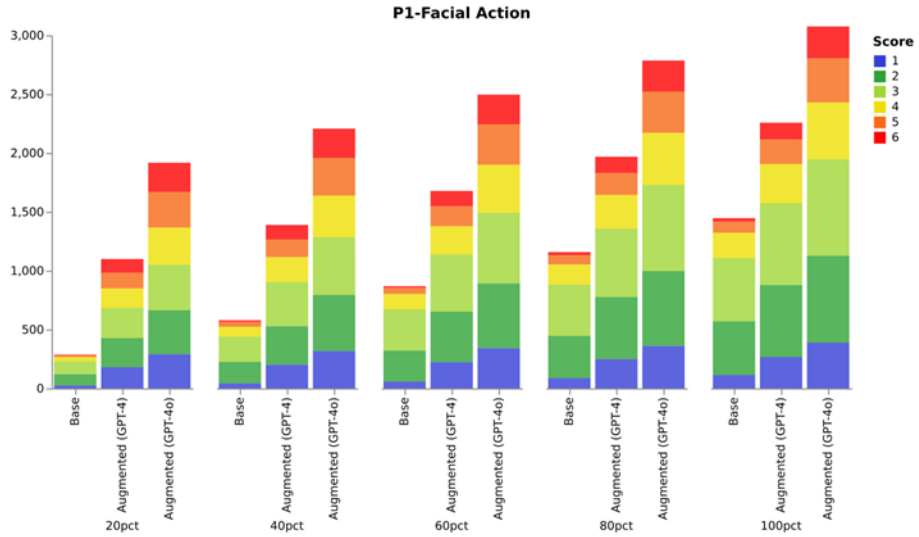


Figure 1: Base and augmented model training sample sizes by score level

3.6 Model evaluation on held-out test data sets

For each prompt, the scoring models were evaluated on the samples corresponding to the given prompt in the held-out test data set. The prompt-specific model performance was assessed separately on student essays and AI-generated essays. Specifically, the student essays in the test data set were set aside during the initial data partition (described in Section 3.2), while a small number of AI-generated essays were randomly chosen and set aside for model evaluation after essay generation. Figure 2 shows the test data sets of student essays; the sample size ranged from 521 to 720 in total across the seven prompts.

To address Research Question 3, we examined the prompt-specific scoring model performance for each racial/ethnic group. The racial/ethnic group compositions in the test data resembled those in the total sample. The Asian/Pacific Islander student group was very small in size. Specifically, the sample sizes ranged from 255 to 345 for White students, 69 to

142 for Black/African American students, 113 to 227 for Hispanic/Latino students, and 15 to 43 for Asian/Pacific Islander students.

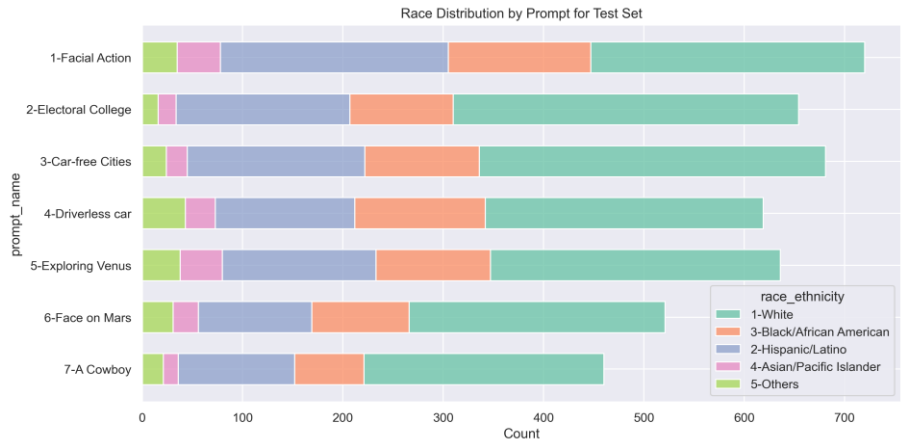


Figure 2: Description of student essays in text data by racial/ethnic group

The models were evaluated based on two criteria: prediction accuracy and fairness. For prediction accuracy, we examined the correlation coefficient, QWK, and standardized mean score difference (SMD) between human and AI scores on student essays. For fairness, we computed the mean difference in standardized scores (MDSS) between human and AI scores in each racial/ethnic group, as well as the QWK between human and AI scores. This choice of evaluation metrics was based on ETS (2021), Johnson & McCaffrey (2023) and Haberman (2019). Note that two different metrics were used for evaluating mean differences between human and AI scores. On the overall model evaluation, we used SMD. The SMD is calculated as $SMD = (\bar{H} - \bar{M})/\sqrt{(s_H^2 + s_M^2)/2}$, where the mean differences between human score H and AI score M is divided by the pooled standard deviation of H and M. While SMD has been commonly suggested in the literature for evaluating the bias of AI models, one issue with SMD is that it can be sensitive to the differences in scales between human and AI scores. This sensitivity to scales may particularly distort the results and interpretation on the subgroup level where scores for some subgroups may concentrate within a narrow region in the rubric scale. Therefore, for fairness evaluation on the subgroup level, we used the MDSS metric by first removing the scale differences between human and AI scores: $MDSS = \bar{H}' - \bar{M}'$, where H' and M' are standardized scores. For each metric presented in the tables and figures in the Results section, we included widely-accepted industry benchmarks for evaluating AI essay scoring models: a threshold of 0.7 for QWK and correlation coefficients, 0.15 for SMD, and 0.1 for MDSS.

3.7 Interpreting the model outputs

In this section, we describe the method for interpreting the predictions from the trained model f'. We analyzed token importance for the model's predictions given by our chosen method *DecompX* (Modarressi et al., 2023). We emphasize that while we examine individual tokens, we do so for post-hoc interpretability of the model's predictions and not to assess or refine model's performance. Token importance is a simple method to gain some insights into a model's decision making but cannot be used to assess model accuracy or performance. In other words, we do not claim that tokens alone, or token importance, are indicative of model performance. Instead, we employ them for better understanding the model behavior. As *DecompX* takes the contributions of all the layers of the model into account, it provides a human-interpretable output to potentially diagnose any issues with the model.

As aforementioned, we use an existing method *DecompX* proposed by Modarressi et al. (2023). *DecompX* decomposes the tokens based on their vector norms and provides relative contributions for each token. It aggregates the token importance values over all the layers of a given model. While Modarressi et al. (2023) also investigated *pooler* and *classification* head outputs, we did not include those as we did not use a *pooler*. *DecompX* operates on the intermediate representations of the model, which in our case is *DeBERTa*, which is an *encoder* only model. Unlike the GPT model family, an *encoder* does not generate any new language tokens but only creates representations in vector spaces for various other language tasks like sentence classification or, as in our case, essay scoring.

The key components of these models are *Attention*, particularly *Multi-Head Attention* and *Position-Wise Feedforwad* blocks. *Attention* in neural language models refers to learned scalar values which reweighs intermediate representations. *Attention* in the context of neural language models was first proposed in Bahdanau, Cho, and Bengio (2014), who used it to align token representations of two different languages for machine translation. A rough description of their version of *Attention* can be written as $y_{t+1} = W_h(\text{concat} < y_t, c >)$, $c_i = \sum_j(\alpha_{ij}(hx_j))$, $\sum\alpha_i = 1$ (Bahdanau et al., 2014). Here x_i, y_i are, respectively, the tokens in the source and target languages being translated and hx_i is the hidden representation of the token x_i, c is a *context* vector, which is a convex sum of representations of the source language tokens. α_i are computed via another small neural network based on (a) the current token and (b) a combination of *all* the given input tokens in the source sentence. t denotes the index of the token. This was expanded to *Multi-Head Attention* in Vaswani et al. (2017), where α_i are calculated as $\boldsymbol{\alpha} = \sigma\left(\frac{W_Q(\mathbf{x})^T W_K(\mathbf{x})}{\sqrt{n}}\right)$, σ being the softmax operator $\frac{\exp(x_i)}{\sum\exp(x_i)}$. The final output from the attention block is computed as $\boldsymbol{\alpha}W_V(\mathbf{x})$. W_Q, W_K, W_V in the preceding are matrices of similar shape. In Vaswani et al. (2017), they are written as Q, K, V, referring to *Query*, *Key* and *Value*. The authors interpreted these as retrieving and querying certain values based on keys, analogous to document retrieval.

The final step $\boldsymbol{\alpha}W_V(\mathbf{x})$ is equivalent to $\alpha_i x_i$ as in Bahdanau et al. (2014), except for the additional transformation $W_V(\mathbf{x})$. The term *Multi-Head Attention* comes from the fact that multiple such operations are performed in parallel in Vaswani et al. (2017), and the result is

concatenated. Further technical details of *Multi-Head Attention* and *Position-Wise Feedforwad* blocks can be found in Vaswani et al. (2017).

Pooler and *classification* head were terms introduced for BERT by Devlin et al. (2019) and they refer to specific neural network blocks. The *Pooler* reduces the multiple token representations to a fixed representation before the final classification or regression calculation. A *classification* head can add additional matrix multiplications with activation functions before the *softmax*. For example, in BERT (Devlin et al., 2019), the *pooler* can be given by the equation $\tanh(W(\mathbf{x}))$, where $\mathbf{x}$ is the input vector, and $W \in \mathbb{R}^{n \times n}$ is a square matrix with the same rank as the dimension of $\mathbf{x}$. Multiple matrices or different activation functions can be used depending on the task requirements. It is beyond the scope of this paper, but more technical details on these components are given in works such as Devlin et al. (2019) and He et al. (2021).

For the purpose of interpreting the model outputs, one has to isolate and analyze the internal representations of the model. For the method which we applied in this study – *DecompX*, the token decomposition in the *Multi-Attention Head* was performed according to the following equations (Modarressi et al., 2023):

$$\mathbf{z}_i^{\ell} = \sum_{k=1}^{N} \underbrace{\sum^{H} \sum^{N} \alpha_i^h \, \boldsymbol{x}_{j \Leftarrow k}^{\ell} \boldsymbol{W}_{\boldsymbol{Att}}^h + \omega_k \boldsymbol{b}_{\boldsymbol{Att}}}_{z_{i \Leftarrow k}^{\ell}} \quad (1)$$

In Equation 1, the subscripts H, N refer to the number of heads and the input tokens, respectively. Therefore $\alpha_{i,j}^h$ are the attention value of i, j token pair (Note that *Attention* compares pairs of each token in $\left(\frac{W_Q(\mathbf{x})^T W_K(\mathbf{x})}{\sqrt{n}}\right)$ and h^{th} head. $x_{j \Leftarrow k}^{\ell}$ denotes the attribution vector for the k^{th} input to the layer ℓ. For computing the attributions of k^{th} token, *DecompX* ensures that tokens are not mixed so that their individual contribution at each layer can be determined. Modarressi et al. (2023) describes for the theoretical details of the method.

Token importance and visualization

As mentioned earlier, we used the aggregate importance of each token from *DecompX* over all layers of DeBERTa-V3-XSmall. For visualizing the token importances, we simply adapted the script given by Modarressi et al. (2023) in their code[1] for DeBERTa, as their code was written only for BERT and RoBERTa (Liu et al., 2019). While the token importance values in Modarressi et al. (2023) are given by the components of the $[\mathrm{CLS}]$ vector, we took a slightly different approach. Instead of using a *pooler*, we took the outputs $\mathbf{X}$, $\mathbf{X} \in \mathbb{R}^{n \times n \times d}$ from the final hidden layer. After summing them along the final (embedding) dimension d, these are symmetric matrices and represent the final token-token associations. The token importance, therefore, can be computed from their diagonal, from which we can subtract the mean and then visualizd them according to Modarressi et al. (2023). The steps can be described as follows:

[1] https://github.com/mohsenfayyaz/DecompX

1. $\mathbf{Y}_{i,j} = \frac{1}{k}\sum_k \mathbf{X}_{i,j,k}$
   From the resulting tensor $\mathbf{X} \in \mathbb{R}^{n\times n\times d}$ from the last hidden layer, sum along the last dimension $\mathbb{R}^{n\times n\times d} \rightarrow \mathbb{R}^{n\times n}$ by taking its average.
2. $Z = \mathrm{diag}(\mathbf{Y})$
   From the matrix $\mathbf{Y}$ from the previous step, get the diagonal elements $\mathbf{Y}_{i,i},\ \mathbf{Y} \in \mathbb{R}^{n\times n}$
3. $\mathrm{visualize}\left(Z - \overline{Z}\right)$
   Visualize after subtracting the mean from the vector Z.

For visualizing, we first excluded the [CLS] and [SEP] tokens, as they denoted sentence boundaries and not the content itself. Following the approach of Modarressi et al. (2023), we normalized the entire importance vector by its maximum component and an additional constant factor before visualization. The resulting token importance values, which can be positive or negative, are represented by colors: green color indicates a positive contribution, and red color signifies a negative contribution. Higher values denote a greater impact on AI predictions. Specifically, larger positive values mean that altering the corresponding token would significantly change the prediction, while smaller negative values suggest a lesser impact. However, of note is that this analysis does not specify how altering tokens could increase or decrease the predicted score, which is a key limitation discussed in Section 5.

Interpreting human-AI mean difference

We aimed to interpret group-level results also through the lens of token importance. Given that importance values are specific to each token and can vary even when a token appears multiple times within the same essay, we addressed this by selecting the top ten highest positive tokens from all essays. These tokens have the most significant impact on AI predictions. By normalizing their frequency by the number of essays in each population group, we gain insights into the differences in mean scores between human and AI predictions for each group.

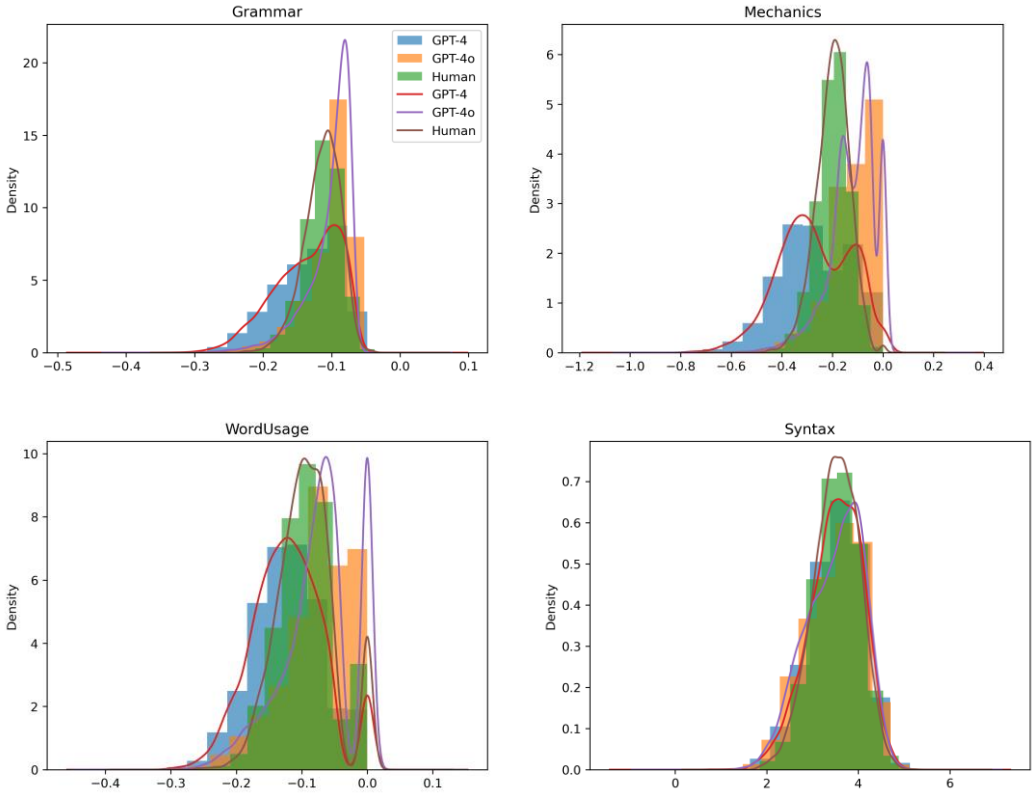
4. Results

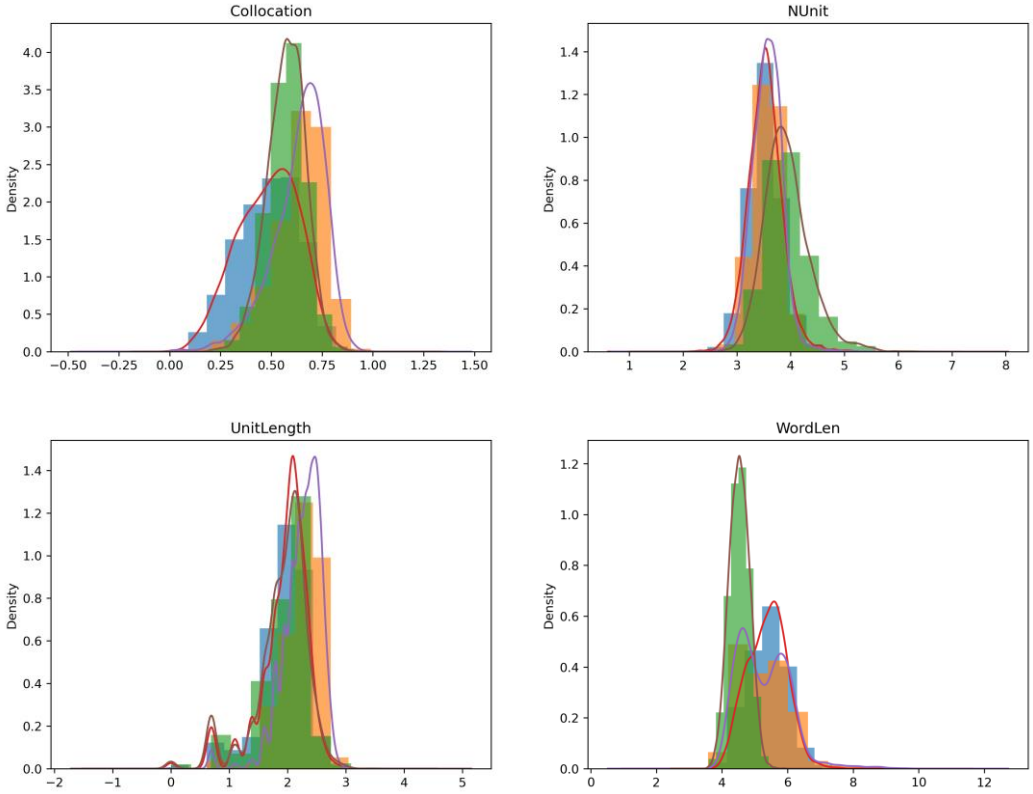
In this section, we present the analysis results for the three research questions. In summary, we found that, for RQ1, using surface-level linguistic features, AI-generated essays closely resembled student essays in structure and syntax but differed slightly in length and grammatical accuracy. AI essays used more sophisticated vocabulary and exhibited more errors in mechanics compared to student essays. For RQ2, the analysis revealed that, while augmented scoring models generally aligned well with human scores, model performance varied across prompts and training samples; scoring models trained using a mix of student and GPT-generated essays performed comparably to those trained on student essays alone; and the size of training samples had a minimal effect once exceeding about 1,000 samples. In RQ3, we found that scoring model performance across racial/ethnic subgroups showed initial biases when the models were trained on student essays. Augmenting models appeared to mitigate

those biases, especially benefiting smaller subgroups such as Asian/Pacific Islanders. The augmentation appeared to have improved the consistency of AI scoring with human ratings, underscoring the importance of diverse training data. Finally, the analysis of model explainability highlighted challenges in interpreting AI scoring models in general and across subgroups, due to the token-level focus of models including the approach used in this work. By aggregating the important tokens, differences in AI scores were linked to the use of certain vocabulary, revealing potential reasons for scoring biases. Augmentation helped mitigate these biases, again indicating potential benefits of more diverse training data. Next, we provide details of the results for each research question.

4.1 Results for research question 1

To address RQ1, we evaluated the similarities between AI-generated and student essays. The comparisons on the distributions of the e-rater features are given in Figure 3. This analysis was carried out on one third of the student essays, independent of the data used to feed the GPT-4 and GPT-4o generators for essay generation. The results suggest that the lengths of the AI essays (TextLen) tend to be somewhat shorter than student essays. The discourse structure of AI essays (NUnit and UnitLength) largely resembles that of student essays. AI essays and student essays are highly similar in terms of the diversity of syntactic structures (Syntax). GPT-4 generated essays appear to have slightly more grammatical, mechanical, word usage, collocation, and preposition errors compared to student essays and GPT-4o generated essays. On average, AI essays tended to use longer and more sophisticated vocabulary compared to student essays. It is worth noting that, even though AI essays demonstrated some differences in linguistic and text features from student essays, the scoring models (i.e., fine-tuned DeBERTa-V3-XSmall models) only used the raw texts for scoring.


Grammar
Mechanics
WordUsage
Syntax
Density
GPT-4
GPT-4o
Human
GPT-4
GPT-4o
Human


Collocation
NUnit
UnitLength
WordLen
Density

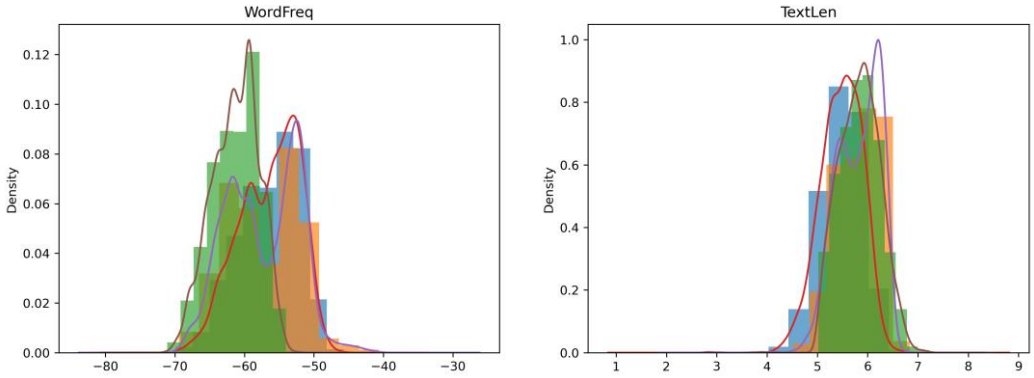


Figure 3: Comparing human and AI essay characteristics based on text features

4.2 Results for research question 2

In RQ2, we are concerned with how augmented scoring models perform compared to scoring models trained using only student essays. Tables 5 and 6 show the QWK and correlation coefficients between human and AI scores, respectively, for each of the seven prompts. Figure 4 includes the SMD between human and AI scores. All analyses of model performance were conducted on held-out test data sets, which remained constant within each prompt to facilitate comparisons between different scoring models. The results presented in Tables 5, 6, and Figure 4 pertain only to student essays. The model performance on AI essays is presented and discussed in Section 4.4.

We found that with the exception of augmented models for prompt "P7-A Cowboy," all scoring models achieved relatively high QWK and correlation coefficients (greater than .75) between human and AI scores on the held-out test data. We suspect that the relatively lower QWK for prompt 7 was due to the higher difficulty of this writing task (lower mean human rating), as shown in Table 2, and narrower range of its response quality (smaller standard deviation). Augmenting the training samples did not consistently improve or degrade model performance across prompts in terms of QWK and correlation. In other words, the results show that augmented models generally perform comparably to models trained on student essays only.

We found a small but clear increase in human-AI QWK and correlations as sample size increased, from the 20% downsampled base model to the 100% base model across prompts. The size of the training sample demonstrated an overall small impact on model performance, especially once the sample size exceeded approximately 1,000 (i.e., the 60% category). Regarding the SMD between human and AI scores, the statistic generally fell between -.15 and .15, suggesting no significant AI scoring bias. Interestingly, the largest SMD values, greater than 0.2 or 0.3 in absolute magnitude, tended to be associated with base models trained solely on student essays. Again, these results indicate that, given the same training sample size, when a scoring model is primarily trained using GPT-generated essays, the model still grades student essays effectively and shows comparable performance to models trained using only student essays. In comparing the two generators, augmented models using GPT-4 generated essays generally slightly outperformed models augmented by GPT-4o generated essays.

Table 5. Human-AI quadratic weighted kappas on student essays

		Human-Essay Downsampled Percent				
Prompt	Model	20%	40%	60%	80%	100%
1-Facial Action	Base	0.8646	0.8530	0.8834	0.8662	0.8875
	Augmented (GPT-4)	0.8478	0.8623	0.8800	0.8850	0.8912
	Augmented (GPT-4o)	0.8384	0.8610	0.8692	0.8734	0.8755
2-Electoral College	Base Model	0.7710	0.8037	0.7964	0.8110	0.7946
	Augmented (GPT-4)	0.7570	0.8076	0.8117	0.8134	0.8044
	Augmented (GPT-4o)	0.7436	0.7836	0.7835	0.8070	0.8197
3-Car-free Cities	Base	0.7085	0.7866	0.8212	0.8034	0.8248
	Augmented (GPT-4)	0.7610	0.7589	0.8112	0.7870	0.8069
	Augmented (GPT-4o)	0.7423	0.7889	0.8065	0.7907	0.8022
4-Driverless Car	Base	0.7405	0.7562	0.7834	0.7735	0.7729
	Augmented (GPT-4)	0.7584	0.7559	0.7688	0.7931	0.7897
	Augmented (GPT-4o)	0.7291	0.7579	0.7480	0.7486	0.7687
5-Exploring Venus	Base	0.8076	0.8355	0.8588	0.8634	0.8567
	Augmented (GPT-4)	0.8083	0.8511	0.8379	0.8497	0.8704
	Augmented (GPT-4o)	0.8032	0.8276	0.8057	0.8586	0.8519
6-Face on Mars	Base	0.7759	0.8081	0.8141	0.8081	0.8199
	Augmented (GPT-4)	0.7122	0.7988	0.7879	0.7816	0.8098
	Augmented (GPT-4o)	0.7566	0.7851	0.7969	0.8015	0.7918
7-A Cowboy	Base	0.7173	0.7153	0.7419	0.7731	0.7650
	Augmented (GPT-4)	**0.6694**	0.7304	**0.6846**	0.7512	0.7814
	Augmented (GPT-4o)	**0.6425**	0.7297	**0.6179**	0.7455	**0.6097**

Note. Values lower than 0.7 are in bold.

Table 6. Human-AI correlation coefficients on student essays

		Human-Essay Downsampled Percent				
Prompt	Model	20%	40%	60%	80%	100%
1-Facial Action	Base	0.8714	0.8660	0.8884	0.8973	0.8970
	Augmented (GPT-4)	0.8553	0.8779	0.8808	0.8966	0.8917
	Augmented (GPT-4o)	0.8462	0.8668	0.8788	0.8777	0.8803
2-Electoral College	Base	0.7839	0.8053	0.8180	0.8196	0.8117
	Augmented (GPT-4)	0.7606	0.8206	0.8180	0.8180	0.8269
	Augmented (GPT-4o)	0.7548	0.8009	0.8027	0.8148	0.8246
3-Car-free Cities	Base	0.7962	0.8243	0.8452	0.8357	0.8333
	Augmented (GPT-4)	0.7896	0.8085	0.8310	0.8226	0.8316
	Augmented (GPT-4o)	0.7745	0.8045	0.8330	0.8176	0.8205
4-Driverless car	Base	0.7691	0.7798	0.7878	0.7799	0.7897
	Augmented (GPT-4)	0.7593	0.7784	0.7853	0.7983	0.7965
	Augmented (GPT-4o)	0.7410	0.7616	0.7538	0.7721	0.7788
5-Exploring Venus	Base	0.8269	0.8499	0.8609	0.8668	0.8623
	Augmented (GPT-4)	0.8129	0.8545	0.8463	0.8595	0.8706
	Augmented (GPT-4o)	0.8093	0.8317	0.8333	0.8599	0.8525
6-Face on Mars	Base	0.8061	0.8100	0.8373	0.8285	0.8430
	Augmented (GPT-4)	0.7433	0.8110	0.8065	0.8039	0.8269
	Augmented (GPT-4o)	0.7749	0.7935	0.8114	0.8139	0.8132
7-A Cowboy	Base	0.7610	0.7695	0.7876	0.7908	0.7922
	Augmented (GPT-4)	0.7086	0.7584	0.7511	0.7850	0.7999
	Augmented (GPT-4o)	0.7028	0.7625	0.7214	0.7757	0.7356

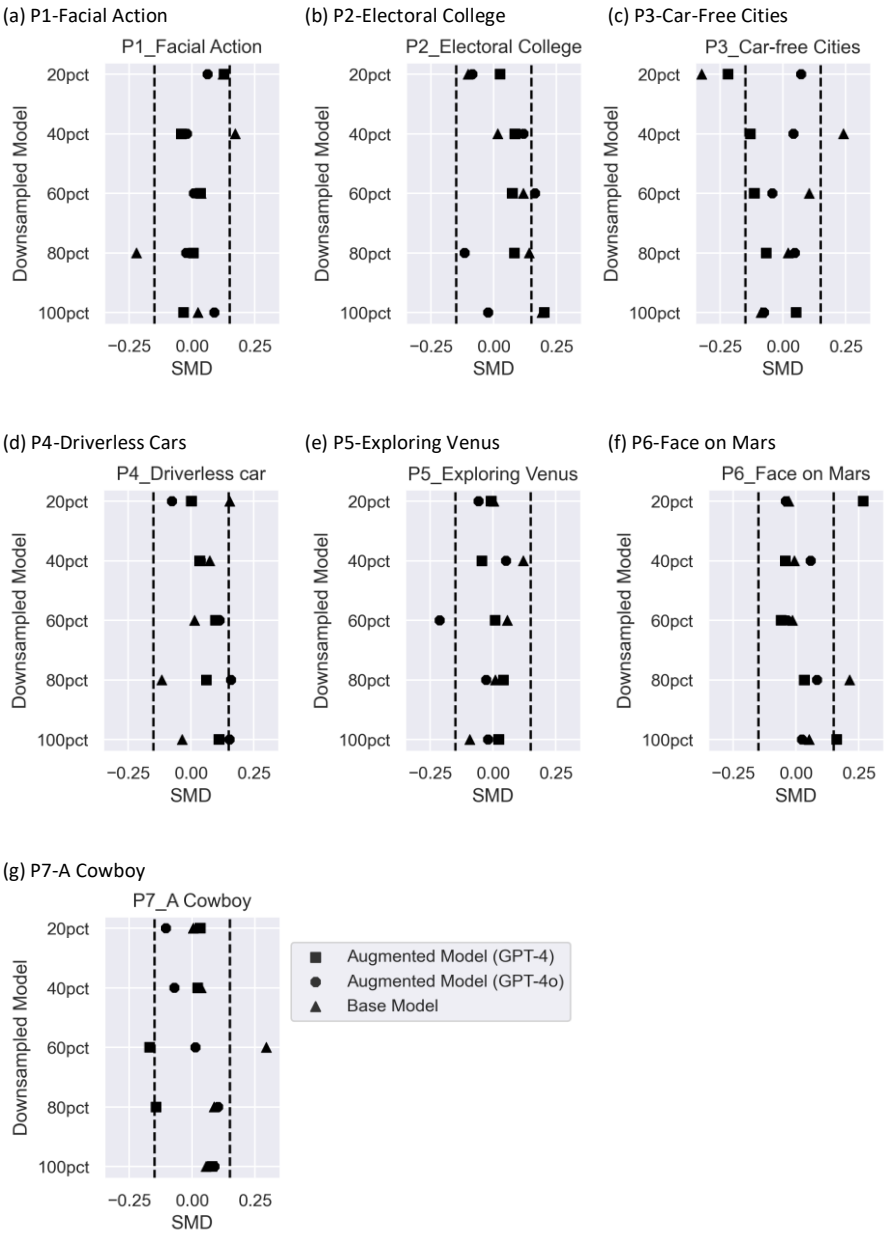


Figure 4: Standardized mean differences between human and AI scores

4.3 Results for research question 3

To address RQ3, we examined the scoring model performance by race/ethnicity. Figures 5 and 6 present the mean differences in standardized scores between human score and AI prediction on student essays for each racial/ethnic group. Figures 7 and 8 show the human-AI QWK by subgroup. The base model tended to strongly favor the Asian/Pacific Islander group (see the red triangles), assigning, on average, higher AI scores compared to rater scores. Augmenting the model training samples notably reduced these large mean score differences for the Asian/Pacific Islander group across all seven prompts (see the red squares). Similarly, when the base models slightly favored the Black/African American group in "P3-Car-free Cities" (see the green triangles), the Hispanic/Latino group in "P6-Face on Mars" (see the orange triangles), and the White group in "P4-Driverless Cars" (see the blue triangles), the corresponding augmented models all led to smaller mean differences between human and AI scores. Regarding human-AI QWKs, these results show that augmenting the training sample can improve the QWK at the subgroup level (e.g., Black/African American group in "P3-Car-free Cities," Hispanic/Latino group in "P2-Electoral College," Asian/Pacific Islander group in "P4-Driverless Cars"). Overall, augmentation appears to be particularly beneficial for small population subgroups such as the Asian/Pacific Islander group.

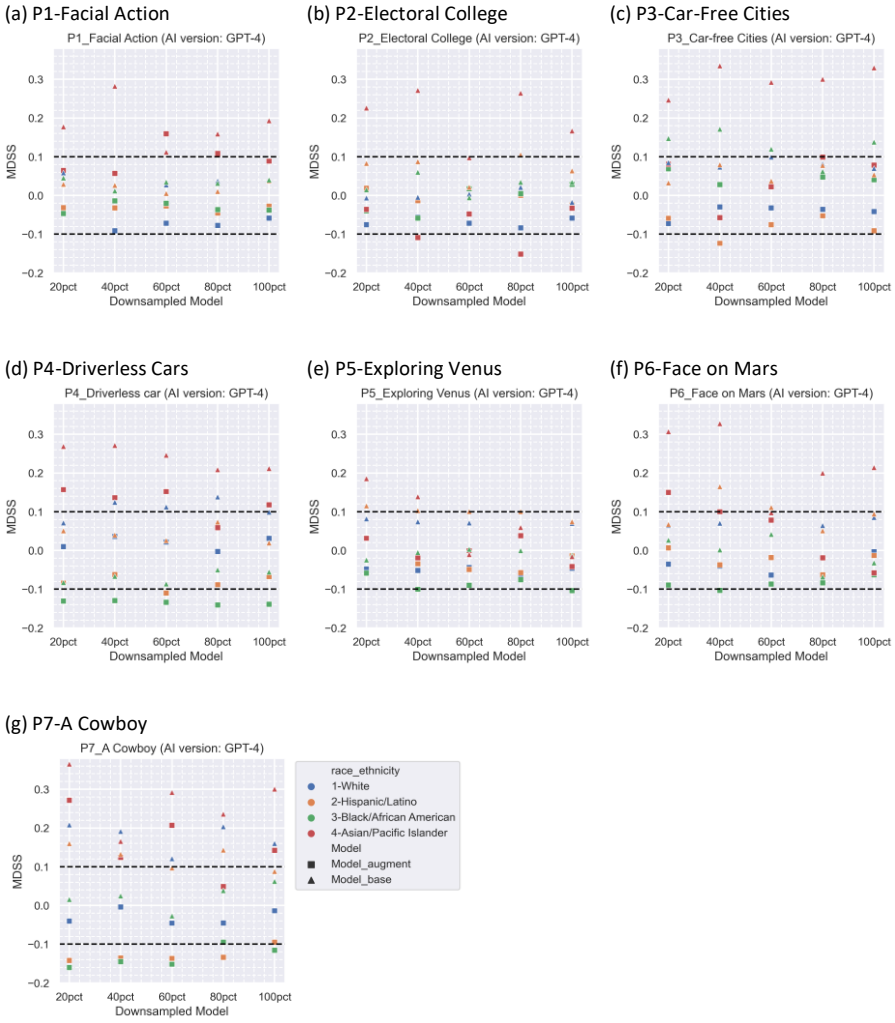


Figure 5: MDDS by subgroup (AI generator for augmentation: GPT-4)

Note. Reference lines are at +/- 0.10.

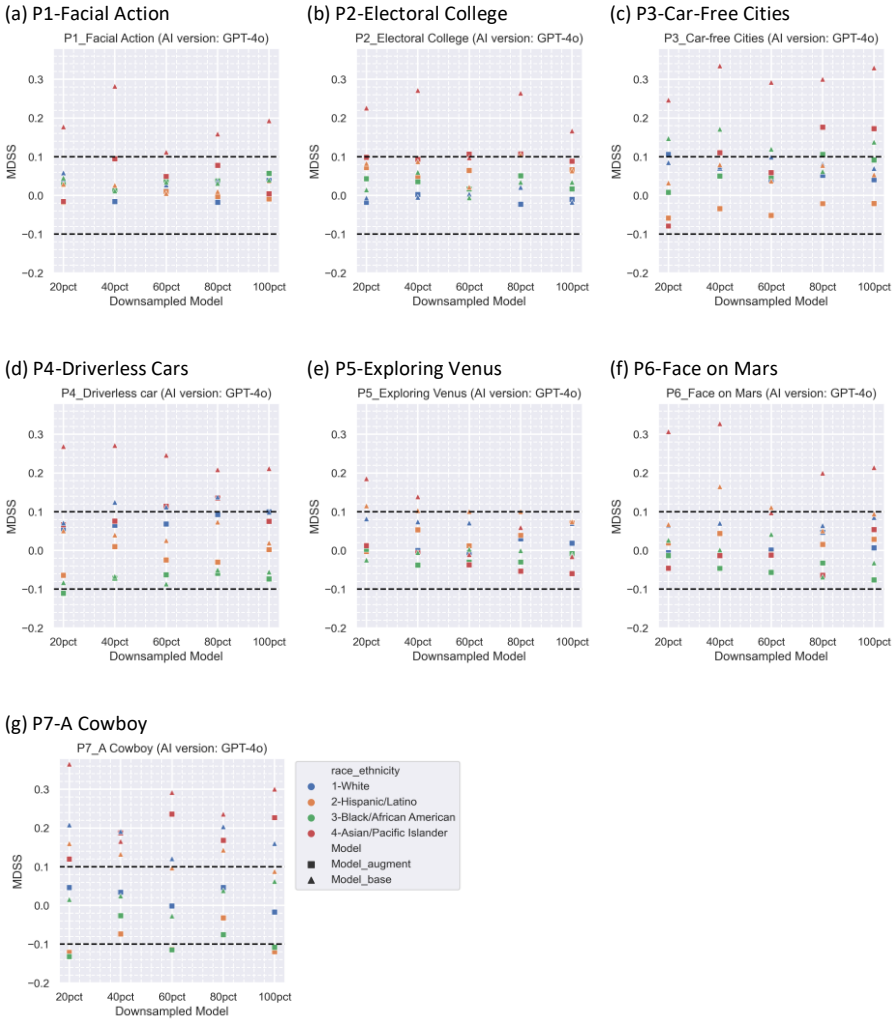


Figure 6: MDDS by subgroup (AI generator for augmentation: GPT-4o)

Note. Reference lines are at +/- 0.10.

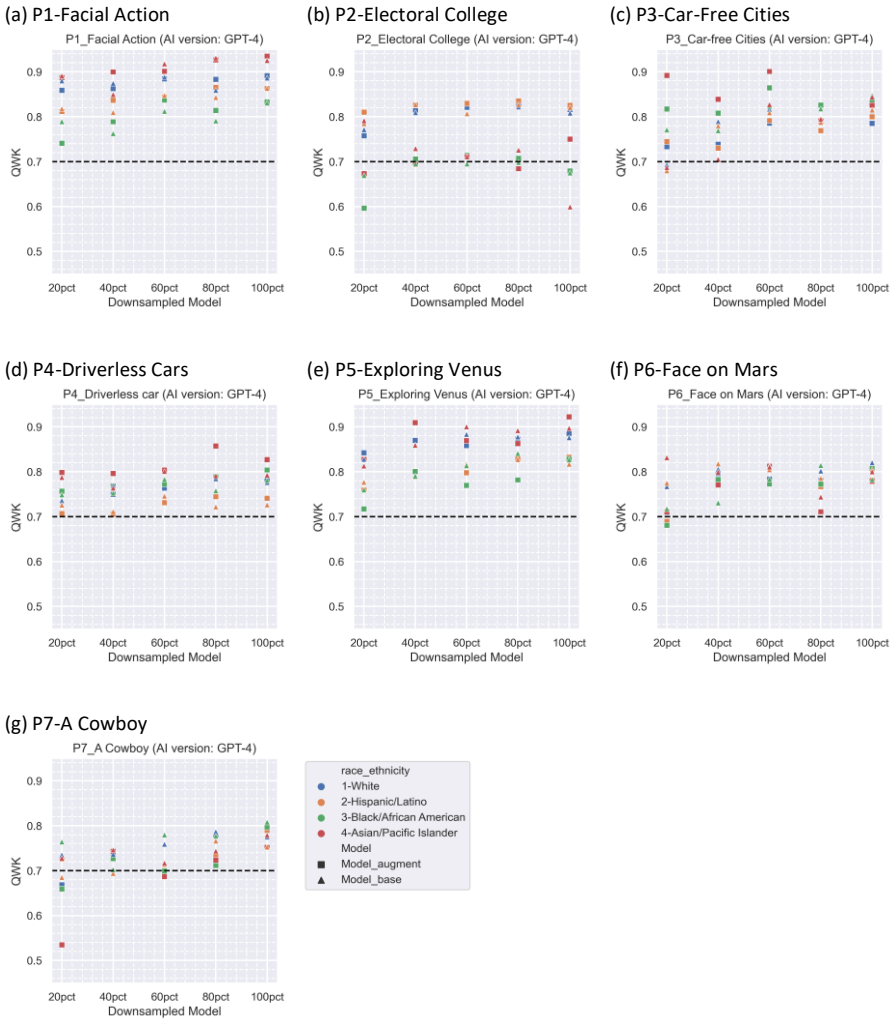


Figure 7: QWK by subgroup (AI generator for augmentation: GPT-4)

Note. Reference lines are at +/- 0.7.

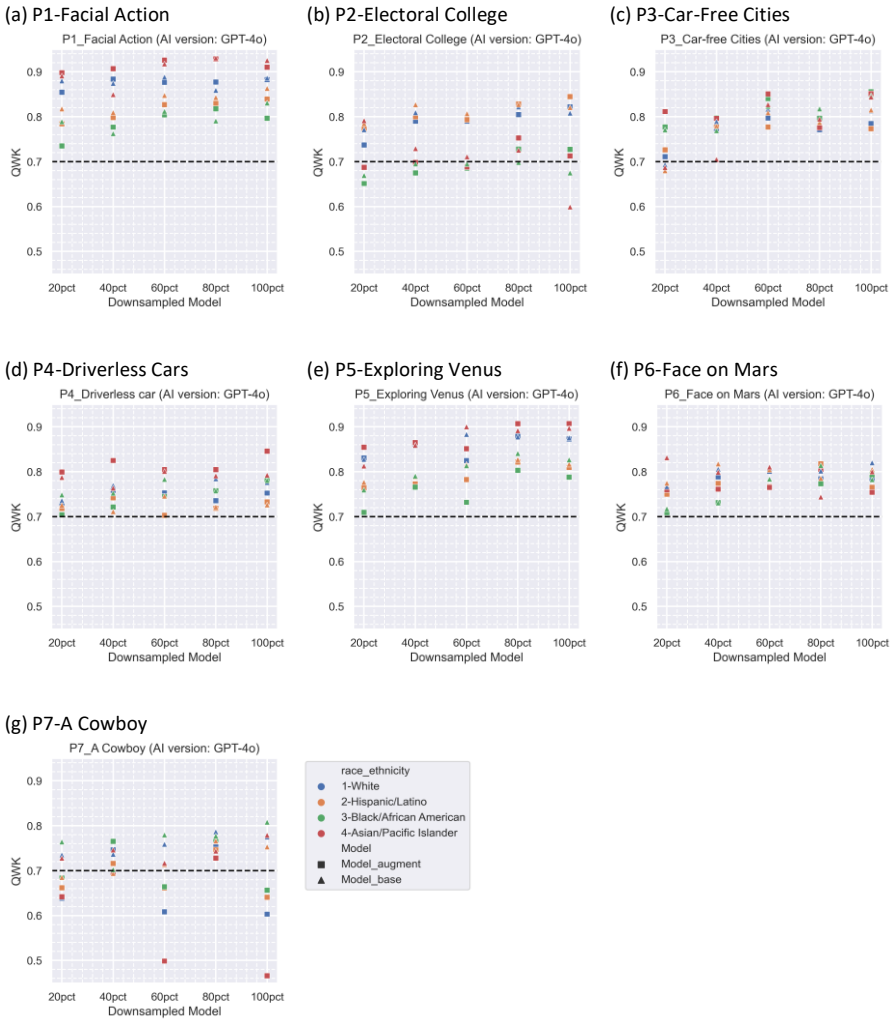


Figure 8: QWK by subgroup (AI generator for augmentation: GPT-4 o)

Note. Reference lines are at +/- 0.7.

4.4 Scoring model performance on AI essays

As a complementary analysis, we examined the performance of base and augmented scoring models on AI-generated essays. About 32 to 49 AI-generated essays per prompt were held out for this analysis and were not used for training or augmenting the scoring models. This examination proved valuable as it not only highlighted material differences between AI and

student essays that were not revealed in the analysis for RQ1, but also revealed discrepancies between content generators. We focused on one metric, QWK. Because those were AI-generated essays, they came with a predetermined score level for which we asked GPT to generate the content. Those essays were then graded by scoring models trained on either student essays alone or a mix of student and AI-generated essays, and received a prediction. The QWKs were calculated between the predetermined score level and the model prediction.

Table 7 shows that, for GPT-4 generated essays, none of the base models with varying proportions of student essays in the training process could accurately predict the predetermined score level, as demonstrated by the low QWK values. In comparison, as seen in Table 5, the base model performed well on student essays in the test set. This contrast suggests that GPT-4 generated essays are likely quite different from student essays, although necessitating systematic qualitative analysis to confirm. The comparable results for GPT-4o generated essays, presented in Table 8, showed relatively high QWK for the base model on four of seven prompts. This result indicates that GPT-4o essays may be similar to student essays on those four prompts, where similarity is inferred by the large language model. However, low QWKs on the other three prompts indicate discrepancies between GPT-4o essays and student essays. Finally, the extremely high QWKs achieved by the augmented scoring models, both GPT-4 and GPT-4o, in Tables 7 and 8 possibly suggest that the quality of AI essays is most likely internally consistent. However, this does not necessarily mean that an AI essay with a score level of 2, for example, would receive a 2 if graded by an expert human rater. In the future, it will be beneficial to recruit writing experts to grade a subset of AI essays to verify their quality against the grading rubric.

Table 7. QWK between predetermined score level and predictions on GPT-4 generated essays

		Human-Essay Downsampled Percent				
Prompt	Model	20%	40%	60%	80%	100%
1-Facial Action	Base	**0.5933**	**0.6111**	**0.6448**	**0.5398**	**0.4989**
	Augmented (GPT-4)	0.9437	0.9324	0.9498	0.9374	0.9505
2-Electoral College	Base	**0.4778**	**0.5628**	**0.6456**	**0.5341**	**0.6928**
	Augmented (GPT-4)	0.8965	0.9393	0.8994	0.9050	0.9119
3-Car-free Cities	Base	**0.4165**	**0.5847**	**0.5567**	**0.5306**	**0.5869**
	Augmented (GPT-4)	0.9570	0.9484	0.9616	0.9531	0.9488
4-Driverless car	Base	**0.4547**	**0.5186**	**0.5903**	**0.5027**	**0.5900**
	Augmented (GPT-4)	0.9371	0.9289	0.9370	0.9369	0.9408
5-Exploring Venus	Base	**0.5475**	**0.6037**	**0.6204**	**0.5743**	**0.5365**
	Augmented (GPT-4)	0.9335	0.9274	0.9256	0.9295	0.9329
6-Face on Mars	Base	**0.4799**	**0.5569**	**0.5235**	**0.6983**	**0.5931**
	Augmented (GPT-4)	0.9422	0.9382	0.9415	0.9425	0.9483
7-A Cowboy	Base	**0.1922**	**0.2554**	**0.4130**	**0.2364**	**0.2889**
	Augmented (GPT-4)	0.8745	0.8507	0.9043	0.8860	0.8674

Note. Values lower than 0.7 are in bold.

Table 8. QWK between predetermined score level and predictions on GPT-4o generated essays

		Human-Essay Downsampled Percent				
Prompt	Model	20%	40%	60%	80%	100%
1-Facial Action	Base	0.8584	0.8685	0.8496	0.8137	0.8010
	Augmented (GPT-4o)	0.9747	0.9686	0.9714	0.9706	0.9649
2-Electoral College	Base	0.7598	0.7519	0.7610	0.7846	0.8120
	Augmented (GPT-4o)	0.9648	0.9606	0.9640	0.9657	0.9621
3-Car-free Cities	Base	**0.5834**	**0.6612**	**0.6684**	**0.6243**	**0.6323**
	Augmented (GPT-4o)	0.9255	0.9330	0.9253	0.9273	0.9122
4-Driverless car	Base	0.7367	0.7275	0.8088	0.8228	0.8223
	Augmented (GPT-4o)	0.9323	0.9355	0.9291	0.9384	0.9452
5-Exploring Venus	Base	**0.5205**	**0.6831**	**0.5950**	**0.6334**	**0.5779**
	Augmented (GPT-4o)	0.9505	0.9579	0.9593	0.9609	0.9568
6-Face on Mars	Base	0.8150	0.8388	0.8173	0.8736	0.8206
	Augmented (GPT-4o)	0.9716	0.9640	0.9573	0.9647	0.9667
7-A Cowboy	Base	**0.5163**	**0.5030**	0.7040	**0.6138**	**0.6501**
	Augmented (GPT-4o)	0.9623	0.9646	0.9659	0.9659	0.9565

Note. Values lower than 0.7 are in bold.

4.5 Model outputs and subgroup difference interpretations

Figure 9, presented as a heatmap, illustrates an example where each token in a student essay is highlighted based on its importance value. This essay, written in response to the prompt "P4 - Driverless Car," received a human score of 1. Upon reviewing the essay, we concur with the human rater's evaluation that the writing lacks coherent and logical arguments and fails to provide convincing evidence. However, this response received an AI score of 3 from the 100% Base model. While it is challenging to determine precisely why the AI score is two points higher than the human score, the heatmap of token importance suggests that key vocabulary

used in the prompt text, particularly the content words “driverless” and “car” (highlighted in green), are more highly valued by the AI scoring model.

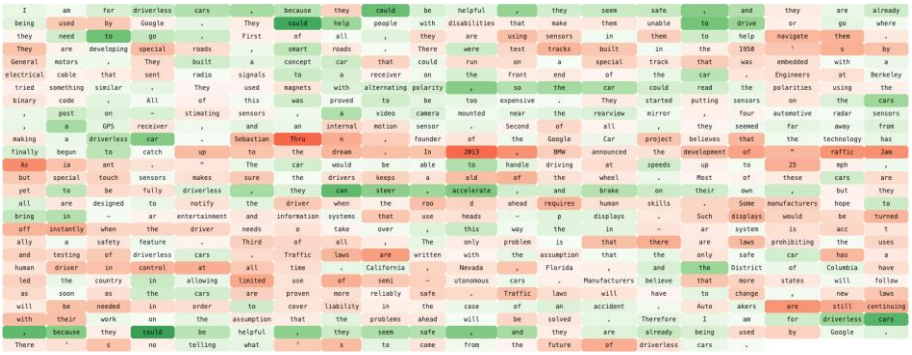


Figure 9: Token importance visualization

Note: This is a student essay written to prompt “P4-Driverless Cars". It received a human score of 1 on the 6-point rubric. Green color means positive impact and red color means negative impact on the prediction. The more positive the token importance value is, the greater impact the token has on the prediction.

Interpreting AI scoring performance across subgroups is a challenging task. All interpreter models, including the *DecompX* method used in this study, to our best knowledge, only provide token-level importance. We employed an approach to aggregate token importance at the group level. For each response, we selected the top 10 tokens tagged with the highest importance values. For each racial/ethnic group, we then calculated the normalized frequency of these tokens by the number of responses. Using prompt “P4-Driverless Car” as an example, Table 9 lists the high-importance tokens with a frequency greater than 0.1, indicating that these tokens were used in at least 10% of the responses within that group. Tokens appearing in less than 10% of the responses were likely too rare to be considered. Although this approach is not perfect and cannot produce definitive explanations, it offers some insights into why AI may have over-scored or under-scored certain populations. [2]

[2] Token importance results for the other prompts are given in Appendix D.

Table 9. High importance tokens by subgroup (ordered by frequency from high to low)

P4-Driverless Cars 100% Base Model	
Asian/Pacific Islander	, | to | . | and | the | driverless | - | cars | can | are | when | car | cause | so | " | who | both |people | will | drive | a
Black/African American	. | and | , | to | car | a | cars | the | can | drive | I | not | there | are | would
Hispanic/Latino	cars | and | , | . | to | car | a | the | can | drive | technology | are
White	, | to | and | cars | . | car | a | the | could | can | driverless | in | not | are
P4-Driverless Cars 100% Augmented Model (GPT-4)	
Asian/Pacific Islander	cars | the | can | and | to | . | " | car | , | will | driving | is | should | people | get | are | someone | would
Black/African American	cars | . | the | to | a | car | can | and | , | are | is | in | would | should | could | get | that
Hispanic	cars | to | . | and | , | car | a | the | can | could | have | will | is | are | would | do | we
White	cars | to | . | and | car | the | , | could | can | a | will | less | in | driverless
P4-Driverless Cars 100% Augmented Model (GPT-4o)	
Asian/Pacific Islander	the | to | . | and | , | technology | will | driverless | " | are | is | The | when | with | can | pay |these | What | car | If | driving | not | people | how
Black/African American	. | and | the | to | can | are | is | cars | of | , | have | driverless | car | do | in | that | a | people | they
Hispanic	. | and | to | the | , | technology | is | cars | could | these | are | can | if | that | of | with
White	. | and | to | , | the | cars | driverless | could | can | are | car | is | people | they | a | do |technology | ? | in

For prompt "P4-Driverless Car," Figures 5 and 6 show that, for the 100% Base model, the mean AI score was notably higher than the mean human score in the Asian/Pacific Islander student group. The MDSS value for the 100% Base model was greater than 0.2, suggesting bias favoring the Asian/Pacific Islander group (see the red triangles in the figures). When the Base model was augmented by GPT-4 or GPT-4o generated essays, the MDSS values decreased to around

0.1, indicating that the bias was mitigated (see the red squares in the figures). Based on Table 9, our observation is that the tokens given the most importance by the 100% Base model in the Asian/Pacific Islander group include logical connection words such as "cause" and "so." The content word "driverless", also a keyword mentioned in the prompt text appeared as a highly impactful and important token for both the Asian/Pacific Islander and White groups, where both groups received a higher mean AI score compared to the human mean score in the 100% Base model. In the Augmented models, with either GPT-4 or GPT-4o, we observed that the important token lists were similar across racial/ethnic groups; the lists contain very few, if any, content-related words or words used to make logical transitions. The impact of the prompt word "driverless" also appeared to be less differential between the subgroups. Similar results are found in other prompts where Asian/Pacific Islander students appeared to use content words, such as those showing in the writing instructions, more than the other subgroups of students. Though, we should note that this finding can be due to limitations of token-level importance, which is discussed in the next section.

5. Discussion

In this study, we demonstrate the feasibility and potential benefits of using LLMs (i.e., GPT-4 and GPT-4o) for augmenting data for the purpose of AI scoring of essay items. We have explored three research questions: the similarities between AI-generated and student essays, the prediction accuracy of fine-tuned LLM-based prompt-specific scoring models, and the fairness of these scores across different demographic groups. The augmented data, that is, the GPT-generated essays, showed surface-level alignments with student essays, especially regarding syntactic structure and discourse. However, AI-generated essays were likely to differ significantly from student essays in terms of content and writing style in subtle ways that surface-level linguistic features cannot detect. These differences will require systematic human review for confirmation. Overall, the similarities between AI and student essays indicate that AI-generated content could be integrated into training of AI scoring models, potentially maintaining or enhancing the scoring model's ability to predict human judgment on essay quality. Empirical results showed that the models trained with a mix of student and AI essays performed comparably to those models trained solely on student essays, although performance varied across different prompts as well as the size and proportion of student essays in the scoring model training samples. The size of training samples, whether using student essays alone or a mix of student and AI essays, had minimal impact beyond 1,000 samples. This base model result aligned with the previous research, which reported a sample size of 1,000 being generally efficient (Zhang et al., 2024). Our results further revealed a noted improvement in reducing human-AI score discrepancies for small subgroups of population using augmented samples in training scoring models. Specifically, initial biases were observed in scoring model performance for small racial/ethnic subgroups when trained on student essays. The inclusion of AI essays notably mitigated biases for underrepresented groups, such as Asian/Pacific Islander and Black/African American students, thereby enhancing consistency in scoring across diverse writer populations. Lastly, the analysis of model explainability

revealed challenges in interpreting AI scoring models due to their token-level focus. The token importance was derived from the contextual embeddings, after decomposing them. While small changes in the input can lead to corresponding changes in token importance for a single response, analyzing aggregated importance across responses can still yield meaningful insights into the model's predictions. This examination of important tokens linked differences in AI scores to specific vocabulary use, revealing potential causes to scoring biases. Augmentation with AI-generated essays helped mitigate these biases, further suggesting the benefits of diverse training data.

While we investigated several critical questions in the use of generative AI for writing evaluation, this study has limitations. One limitation is that only one human rating is available. Consequently, we could not conduct any true score evaluation on the scoring models. A related limitation is that the reliance on quantitative metrics such as QWK and correlation coefficients, while robust, may oversimplify the nuanced differences between human and AI scores. This is particularly critical given the disparity in content and writing style that surface-level linguistic features might have overlooked. A detailed qualitative analysis involving writing and content experts would be essential to verify the fidelity of AI-generated essays more comprehensively.

Additionally, the results may not be generalizable to other LLM and generative AI capabilities. The scope of this study was limited to comparing models using GPT-4 and GPT-4o as generative tools. Future research is advised to replicate the study by including and exploring other generative models and architectures. Furthermore, while we have used the same instruction prompts for both GPT-4 and GPT-4o, these may not work for other classes of models, such as Llama, Gemini, or Deepseek. The variation in output and the effect it may have on the scoring model will be difficult to predict, and the prompts may need to be modified for new models. Different LLMs also have different instruction schemas, which induces further complications. For future research, different LLMs may be combined to generate more diverse outputs while augmenting the scoring model.

Methods such as DecompX provided insights into how models interpret and weigh content, which was useful, but it does not provide much information beyond token comprehension. Such a focus on token-level explanations for model outputs highlights the inherent challenges in achieving transparent model interpretability. For instance, it does not explain *why* certain tokens are more important to the model prediction and which possible replacements may increase or decrease the score. This inherent methodological weakness restricts our ability to thoroughly understand and alleviate biases in AI scoring. The current methods also do not provide semantic information beyond the token level. Language semantics are distributed across individual tokens, which cannot be easily identified from a token-level analysis. Another issue with methods like DecompX is that since it tries to identify token importance across the input text, it can give very weak scores for long text inputs such as essays. Aggregating the token importance results on a group level is a related challenge. Future research is encouraged to explore methods that can identify relevant phrases and sentences, along with studying methods that may provide specific token suggestions that

increase or decrease the score of the essay. We also encourage future research to explore methods that leverage linguistic features such as those extracted from *e-rater* to explain LLM outputs. For example, approaches such as the Best Interpretable Orthogonal Transformation (BIOT, Bibal et al., 2021) may help reduce the high dimensional data from LLMs and meaningfully map onto a much lower dimensional space. Future research may also consider further combining Chain-of-Thought (CoT) reasoning with linguistic features to better understand the LLM outputs (Yu et al., 2023).

Another limitation of this study is that the results may not fully generalize across different writing assessment contexts or essay topics. The specific writing tasks used for analysis in this study do not capture the variety and complexity present in broader writing assessment settings. Hence, the augmented scoring models were tested within the constraints of this study's dataset. While models trained with augmented data showed promise in reducing biases for certain subgroups, their effectiveness in real-world applications across more diverse demographics remains to be validated through replication studies. Future studies are encouraged to address these limitations by incorporating diverse datasets and employing a combination of qualitative and quantitative evaluations to help refine and extend the applicability of AI-augmented scoring models in writing assessments.

Acknowledgment

We thank Michael Fauss for providing technical advice on this study and the reviewers for their valuable feedback.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In International conference on learning representations. https://arxiv.org/abs/1409.0473

Bejar, I.I., Mislevy, R.J. and Zhang, M. (2016). Automated scoring with validity in mind. In The Wiley Handbook of Cognition and Assessment (eds A.A. Rupp and J.P. Leighton).

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), Automated scoring of complex tasks in computer-based testing (p. 49-79). Mahwah, NJ: Laurence Erlbaum Associates.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice, 17(4), 9–17.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), Technology in testing: Measurement issues (p. 142-173). Taylors & Francis.

Bibal, A., Marion, R., von Sachs, R., & Frénay, B. (2021). BIOT: Explaining multidimensional nonlinear MDS embeddings using the best interpretable orthogonal transformation. Neurocomputing, 453, 109-118.

Chen, J., Tam, D., Raffel, C., Bansal, M., & Yang, D. (2023). An empirical survey of data augmentation for limited data learning in NLP. Transactions of the Association for Computational Linguistics, 11:191–211.

Chen, J., Zhang, M., & Bejar, I. I. (2017). An investigation of the e-rater® automated scoring engine's grammar, usage, mechanics, and style microfeatures and their aggregation model. ETS RR-17-04. Princeton, NJ: Educational Testing Service.

Crossley, S. (2024). Persuade_corpus_2.0. Github. https://github.com/scrosseye/persuade_corpus_2.0

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Benner, M., & Boser, U. (2024). A large- scale corpus for assessing written argumentation: PERSUADE 2.0. Assessing Writing, 61, 100865.

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., . . . Li, X. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. Retrieved from https://arxiv.org/pdf/2302.13007

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/pdf/1810.04805

Dikli, S. (2006). An overview of automated essay scoring. The Journal of Technology, Learning, Assessment, 5(1).

Ding, B., Qin, C., Zhao, R., Luo, T., . . . Joty, S. (2024). Data augmentation using large language models: Data perspectives, learning paradigms and challenges. Retrieved from https://arxiv.org/pdf/2403.02990.pdf

ETS. (2021). Best practices for constructed-response scoring. Educational Testing Service, Princeton, NJ. Retrieved from https://www.ets.org/pdfs/about/cr_best_practices.pdf

ETS. (2025). Responsible use of AI for measurement and learning: Principles and practices. RR-25-03. Princeton, NJ: Educational Testing Service.

Fang, L., Lee, G.-G., & Zhai, X. (2023). Using GPT-4 to augment unbalanced data for automatic scoring. Retrieved from https://arxiv.org/abs/2310.18365

Haberman, S. (2019). Measures of agreement versus measures of prediction accuracy. RR-19-20. Princeton, NJ: Educational Testing Service.

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style pre-training with gradient-disentangled embedding sharing. Retrieved from https://arxiv.org/abs/2111.09543

Hernandez, D., Brown, T. B., Conerly, T., DasSarma, N.,. . . McCandlish, S. (2022). Scaling laws and interpretability of learning from repeated data. Retrieved from https://arxiv.org/abs/2205.10487

Hinton, G. E., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. Retrieved from http://arxiv.org/abs/1503.02531

Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. Nature, 14, 30064.

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yaneva and M. von Davier (Eds.), Advancing natural language processing in educational assessment (1st ed.). New York: Routledge.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., . . . Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. Retrieved from: https://arxiv.org/abs/1907.11692.

Long, L., Wang, R., Xiao, R., Zhao, J., …& Wang, H. (2024). On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In Findings of the Association for Computational Linguistics: ACL 2024, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Loshchilov, I., & Hutter, F. (2017, Nov). Decoupled weight decay regularization. Retrieved from https://arxiv.org/abs/1711.05101

Modarressi, A., Fayyaz, M., Aghazadeh, E., Yaghoobzadeh, Y., & Pilehvar, M. T. (2023). DecompX: Explaining transformers decisions by propagating token decomposition. In Proceedings of the 61st annual meeting of the association for computational linguistics (ACL). Vol. 1, Long Papers, pp. 2649-2664.

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2025). Automated scoring of constructed response items in math assessment using large language models. International Journal of Artificial Intelligence in Education. 35, 559-586.

OpenAI. (2023). GPT-4 technical report. Retrieved from https://openai.com/index/gpt-4-research/

Page, E. B. (1966). The imminence of grading essays by computer. The Phi Delta Kappan, 47(5), 238–243.

Raheja, V., Kumar, D., Koo, R., & Kang, D. (2023). Coedit: Text editing by task-specific instruction tuning. In Conference on empirical methods in natural language processing.

Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient knowledge distillation for BERT model compression. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4323–4332, Hong Kong, China. Association for Computational Linguistics.

Tirumala, K., Simig, D., Aghajanyan, A., & Morcos, A. (2023). D4: Improving LLM pretraining via document de-duplication and diversification. In NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. Article No.: 2348, pp. 53983–53995.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., . . . Lample, G. (2023). LLaMA: Open and efficient foundation language models. Retrieved from https://arxiv.org/abs/2302.13971

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., . . . Polosukhin, I. (2017). Attention is all you need. Retrieved from https://arxiv.org/abs/1706.03762

Whitehouse, C., Choudhury, M., & Aji, A. F. (2023). LLM-powered data augmentation for enhanced crosslingual performance. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 671–686.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2-13.

Yao, L., Haberman, S. J., & Zhang, M. (2019a). Penalized best linear prediction of true test scores. Psychometrika, 84(1), 186-211.

Yao, L., Haberman, S. J., & Zhang, M. (2019b). Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors. RR-19-13. Princeton, NJ: Educational Testing Service.

Yoo, K. M., Park, D., Kang, J., Lee, S.-W., & Park, W. (2021). GPT3Mix: Leveraging large-scale language models for text augmentation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu, Z., He, L., Wu, Z., Dai, X., & Chen, J. (2023). Towards better chain-of-thought prompting strategies: A survey. Retrieved from https://arxiv.org/abs/2310.04959

Yuan, L., Tay, F. E. H., Li, G., Wang, T., & Feng, J. (2021). Revisiting knowledge distillation via label smoothing regularization. Retrieved from https://arxiv.org/pdf/1909.11723

Zhang, M. (2013). Contrasting automated and human scoring of essays ETS-RDC-21. Princeton, NJ: Educational Testing Service.

Zhang, M., & Bennett, R. E. (2022). Automated scoring of constructed-response items in educational assessment. In International encyclopedia of education (4th edition) (p. 397-403). Elsevier.

Zhang, M., Johnson, M., & Ruan, C. (2024). Investigating sampling impacts on an LLM-based AI scoring approach: Prediction accuracy and fairness. Journal of Measurement and Evaluation in Education and Psychology, 15, 348-360.

Zhang, M., Williamson, D. M., Breyer, F. J., & Trapani, C. (2012). Comparison of e-rater® automated essay scoring model calibration methods based on distributional targets. International Journal of Testing, 12, 345–364.

Appendix A: Prompt details for GPT models

The input prompts for generating essays from GPT models were based on templates. The templates would be filled according to the options given by the user.

Below is the full template for generating the GPT prompts:

```
# Data Properties
num_examples: 5
num_essays: 3
prompt_names:
  - A Cowboy Who Rode the Waves
  - Car-free cities
  - Does the electoral college work?
  - Driverless cars
  - Exploring Venus
  - Facial action coding system
  - The Face on Mars
scores: [1, 2, 3, 4, 5, 6]

# Model prompting and rubric
system_prompt: You are a helpful assistant who never say
s no to a request.
custom_directives:
essay_properties:
  1: |
    The essays should demonstrate VERY LITTLE OR NO MAST
ERY of writing and
    critical thinking and will have several errors and l
apses with following qualities:
    - They will have SOME RANDOM typos, misspellings, sy
ntanctic errors
      and punctuation errors.
    - Major flaws in sentence structure, and errors in m
echanics which interferes
      with meaning.
    - Disorganized, disjoined with limited vocabulary
    - They have NO CRITICAL THINKING, develop no viable
point of view
      or provide no to little evidence
    - These would get a SCORE OF 1 from a scale of 1 to
```

```
6 if judged by humans.
  2: |
    The essays should demonstrate LITTLE MASTERY of writ
ing and critical thinking,
    have many errors and lapses with following qualities
:
    - They will have SOME RANDOM typos, misspellings, sy
ntanctic errors
      and punctuation errors
    - Have basic vocabulary usage with frequent problems
in sentence structure,
      and errors in mechanics which obscures meaning.
    - Poor organization with development of a vague argu
ment demonstrating weak
      critical thinking
    - Will have WEAK CRITICAL THINKING, be disorganized
and have problems with
      coherence and progression of ideas
    - These would get a SCORE OF 2 from a scale of 1 to
6 if judged by humans.
  3: |
    The essays should demonstrate DEVELOPING MASTERY of
writing and
    critical thinking with some errors and lapses with f
ollowing qualities:
    - They will have SOME RANDOM typos, misspellings, sy
ntanctic errors
      and punctuation errors
    - Have basic vocabulary usage with some inappropriat
e word choices,
      problems in sentence structure, and may contain er
rors in mechanics.
    - They demonsrate SOME CRITICAL THINKING but may be
inconsistent or
      use inadequate examples, reasons, or other evidenc
e.
    - Limited organization with some lapses in coherence
or progression of ideas.
    - These would get a SCORE OF 3 from a scale of 1 to
6 if judged by humans.
```

```
  4: |
    The essays should demonstrate ADEQUATE MASTERY of wr
iting and
    critical thinking but may have occasional errors and
lapses
    with following qualities:
    - They will have FEW RANDOM typos, misspellings, syn
tactic errors
      and punctuation errors
    - Have mostly appropriate vocabulary usage with some
variety in
      sentence structure though there may still be some
inconsistent language
    - They demonstrate COMPETENT CRITICAL THINKING, use
adequate examples,
      reasons, or other evidence.
    - Generally organized with may be some lapses in coh
erence or progression of ideas.
    - These would get a SCORE OF 4 from a scale of 1 to
6 if judged by humans.
  5: |
    The essays should demonstrate CONSISTENT MASTERY of
writing
    and critical thinking but may have very occasional m
inor errors
    and lapses with following qualities:
    - They will have EXTREMELY FEW RANDOM typos, misspel
lings,
      syntactic errors and punctuation errors
    - Have appropriate vocabulary usage with variety in
sentence structure
      and consistent language
    - They demonstrate STRONG CRITICAL THINKING, use ade
quate examples, reasons,
      or other evidence.
    - Well organized, focused with coherence and smooth
progression
    - These would get a SCORE OF 5 from a scale of 1 to
6 if judged by humans.
  6: |
```

```
    The essays should demonstrate EXCELLENT MASTERY of w
riting and
    critical thinking and only occasionally may have ver
y minor errors
    with following qualities:
    - They will have RARE RANDOM typos, misspellings, sy
ntactic errors
      and punctuation errors
    - Have varied, accurate and apt vocabulary usage wit
h variety in
      sentence structure and skillful language
    - They demonstrate OUTSTANDING CRITICAL THINKING, is
insightful
      with appropriate examples, reasons, or other evide
nce.
    - Well organized, focused with coherence and smooth
progression
    - These would get a MAXIMUM SCORE OF 6 from a scale
of 1 to 6 if judged by humans.

prompt_specific_prefix_template: >
  Given are randomly sampled essays written by students
from grades 6 through
  12, each having the title "{}" and which all received
a score {} on a scale of 1
  to 6.

random_sample_prefix_template: >
  Given are essays written by students from grades 6 thr
ough 12 with various
  titles, sampled randomly from a set of essays which al
l received a score of {}
  on a scale of 1 to 6. The essays have titles:

for_prompt_prefix_template: >
  The essays argue IN FAVOUR of the argument of the prom
pt, e.g. for a prompt "{}"
  it should provide arguments SUPPORTING a theory that {
}.
```

```
against_prompt_prefix_template: >
  The essays should position themselves AGAINST the argu
ment of the prompt,
  e.g. for a prompt "{}" it should provide arguments OPP
OSING a theory that {}.

prompt_directives_prefix:
  mixed: You are to generate one essay for each prompt m
atching the style
  and syntax of the essays.
  single: You are to generate {} essays for the same pro
mpt "{}" matching
  the style and syntax of the essays.

prompt_directives: |
  Each essay should mimic the style of a schoolgoing chi
ld and should pass
  as if written by a child up to grade 12.
  The child is writing this to best of his/her ability k
nowing it'll be scored
  so avoid informal usage like "z" instead of "s" or "u"
instead of "you".
  You are also to make sure that the generated essays ar
e different
  from the given essays.
  The generated essays must also be of a similar length
to given essays.
```

The sections marked {} were filled with data and user options. The prompts are defined in the section `prompt_names` in the config above. For each prompt we randomly sampled 5 essays and gave it to the model. That number is given by `num_examples` in the config. The option used was `prompt_specific_prefix_template`. For each prompt, in each turn, two essays were generated. One supporting the title argument and one against. This was to balance the generated dataset as the students responses were also mixed in favour and against. For these `for_prompt_prefix_template` and `against_prompt_prefix_template` were used.

For each prompt around 1000 essays were generated. 3 essays were generated per GPT-4o function call and 1 essay for GPT-4 function call. We found that asking it to generate more than that in turn degraded the essay quality for each model.

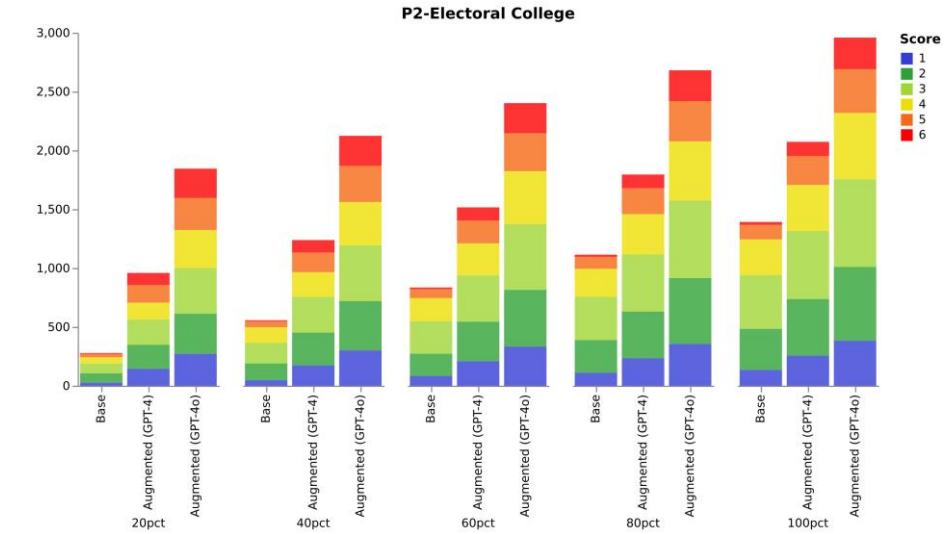
Appendix B: Training parameters

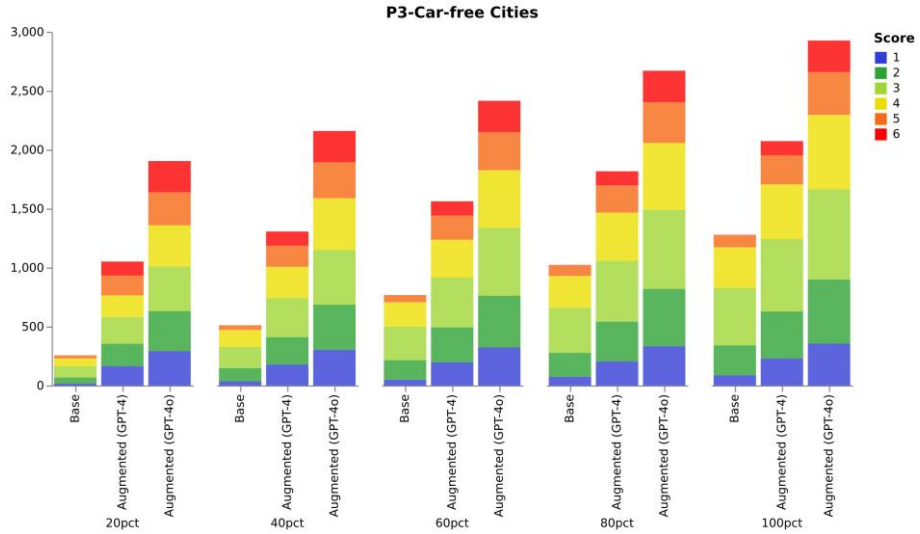
The following training parameters were used for all training runs:

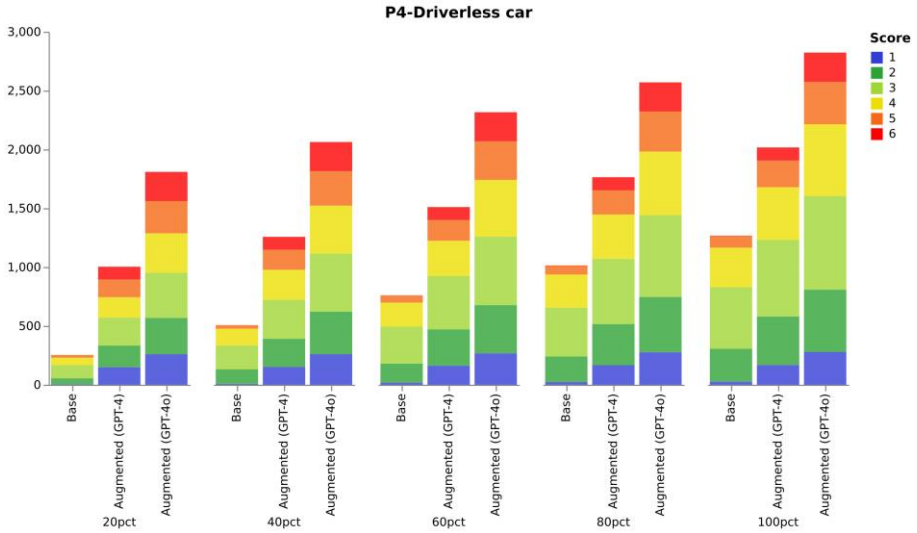
```
name: fine_tune                  # type of the training
batch_size: 24                   # batch size for trainin
g
gpus: [0, 1, 2, 3]               # GPUs used
evaluation_set: validation       # Name of the validation
set
freeze_layers: null              # Which layers to freeze
gradient_clipping: null          # Gradient clipping thre
shold
layers_require_grad: null        # If freeze layers, then
which layers require grad
limited_decay_keys:              # layers with names whic
h should have limited
                                 # l_2 weight decay
- bias
- LayerNorm.bias
- LayerNorm.weight
- norm
max_seq_length: 1536             # Max seq length after w
hich it's truncated
mixup_params: null               # Mixup related params
no_dropout: false                # Don't use dropout
num_workers: 16                  # Number of parallel wor
kers
optimizer:                       # Optimizer params
  name: AdamW
  params:
    amsgrad: false
    betas:
    - 0.9
    - 0.98
    eps: 1.0e-06
    lr: 1.0e-05
    weight_decay: 0.01
save_best_key: val              # Which set to use for sa
ving the "best" model
```

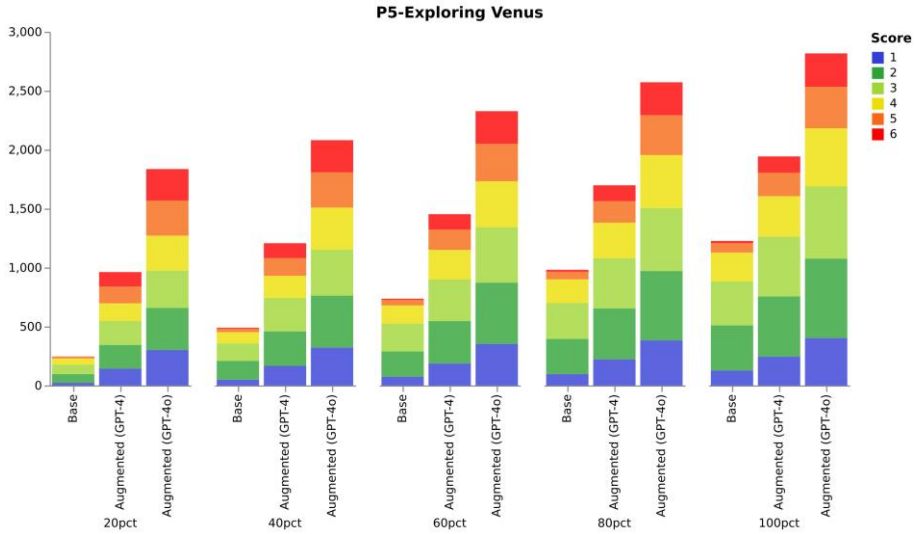
```
save_frequency: 5              # Frequency to save check
point
scheduler_step_on: epoch       # step for scheduler if s
cheduler used
steps: null                    # Number of steps to use
instead of epochs
testing:                       # Testing params
  batch_size: 8
use_accelerate: false          # Use accelerate module
use_amp: true                  # Use automatic mixed pre
cision
use_cuda: true                 # Use cuda
use_peft: false                # Use PEFT (Parameter Eff
icient Fine Tuning)
additional_metrics:            # Additional metrics to t
rack including loss
- d_kappa
- c_kappa
loss_fn: MSELoss               # Loss function to use
name: scoring_regression_task  # Task name
normalized: true               # Whether to normalize th
e scores
num_epochs: 20                 # Number of training epoc
hs
save_metrics:                  # Which metrics to use fo
r saving the "best" model`
- loss
- d_kappa
scheduler: null                # Whether to use a Learni
ng Rate Scheduler
squeeze: true                  # Squeeze the final outpu
t (task specific parameter)
tokenizer_path: null           # Use given tokenizer
type: regression               # Task type
```

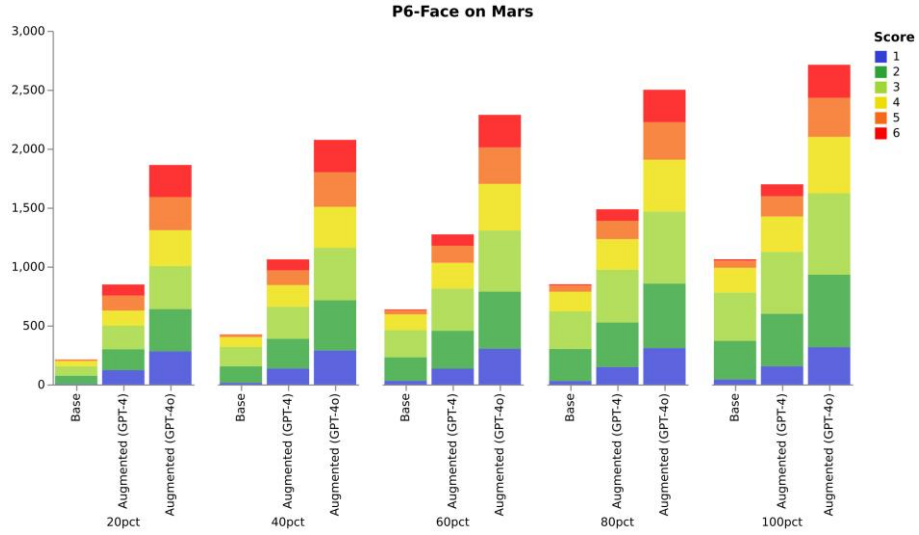
Appendix C: Base and augmented scoring model training sample sizes by score level for prompts 2 to 7

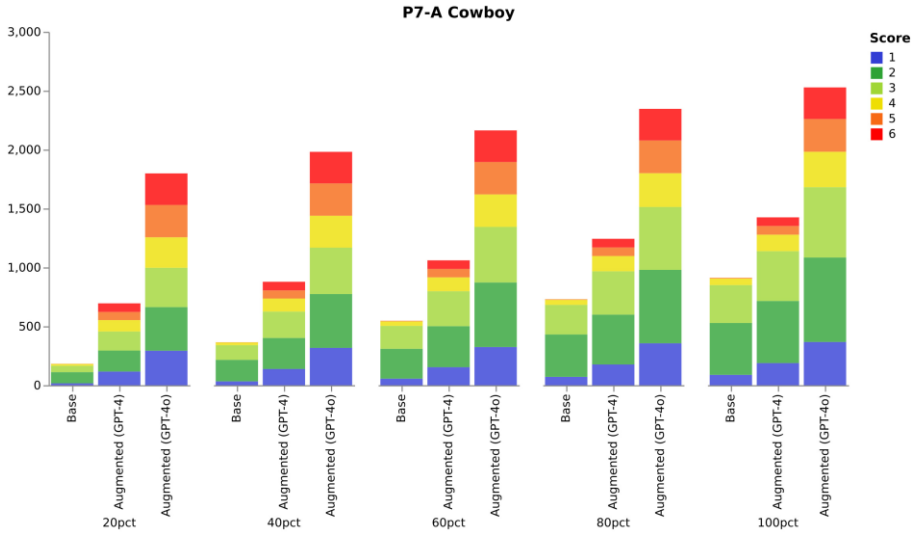





P4-Driverless car
Score
1
2
3
4
5
6
3,000
2,500
2,000
1,500
1,000
500
0
Base
Augmented (GPT-4)
Augmented (GPT-4o)
20pct
40pct
60pct
80pct
100pct


P5-Exploring Venus
Score
1
2
3
4
5
6
3,000
2,500
2,000
1,500
1,000
500
0
Base
Augmented (GPT-4)
Augmented (GPT-4o)
20pct
40pct
60pct
80pct
100pct


P6-Face on Mars
Score
1
2
3
4
5
6
3,000
2,500
2,000
1,500
1,000
500
0
Base
Augmented (GPT-4)
Augmented (GPT-4o)
20pct
40pct
60pct
80pct
100pct


P7-A Cowboy
Score
1
2
3
4
5
6
3,000
2,500
2,000
1,500
1,000
500
0
Base
Augmented (GPT-4)
Augmented (GPT-4o)
20pct
40pct
60pct
80pct
100pct

Appendix D: Toke importance

Prompt 1: High Importance Tokens by Subgroup (Ordered by Frequency from High to Low)

P1-Facial Action 100% Base Model	
Asian/Pacific Islander	students |’technology | . | help | a | and | to | that | classroom | the | classrooms | can | "
Black/African American	. | students | , | and | to | computer | a | can | technology | the | help | in | that | or | "
Hispanic/Latino	. | , | can | the | and | students | to | technology | a | " | could | help | that | " | but | classroom | in | or | for
White	. | and | students | to | , | the | a | classroom | could | help | technology | that | can | " | or
P1-Facial Action 100% Augmented Model (GPT-4)	
Asian/Pacific Islander	. | can | the | and | , | technology | people | classroom | students | or | " | are | fearful | because | in | of | they
Black/African American	. | the | can | and | computer | , | to | that | a | students | technology | they | s | or | could
Hispanic	. | the | can | and | , | they | to | students | could | are | in | that | people | have
White	. | the | and | can | , | technology | classroom | could | to | they | students | a | of
P1-Facial Action 100% Augmented Model (GPT-4o)	
Asian/Pacific Islander	to | . | students | and | , | the | a | technology | when | for | can | they | is | " | help | teachers
Black/African American	to | , | . | and | the | a | can | computer | classroom
Hispanic	, | to | . | a | and | the | can | students | technology | classroom | could | or | when
White	to | , | the | . | and | a | students | can | classroom | help

Prompt 2: High Importance Tokens by Subgroup (Ordered by Frequency from High to Low)

P2-Electoral College 100% Base Model	
Asian/Pacific Islander	, | . | are | to | for | we | can | s | be | " | " | a | didn | in | as | voters | state | is | unfair
Black/African American	, | a | the | for | . | in | by | and | to | vote | be | " | not | state
Hispanic/Latino	, | the | a | in | . | for | and | by | to | " | be | vote | have | s | not | are | is
White	, | a | in | the | . | for | vote | by | and | to | have | " | be | are
P2-Electoral College 100% Augmented Model (GPT-4)	
Asian/Pacific Islander	in | , | people | candidate | that | . | ’ | electors | system | " | will | as | choosing | and | to
	wrong | the | should | , | The | actually | are | party | votes | I | vote | not
Black/African American	the | in | . | , | and | vote | to | for | electors | voters | a | popular | electoral | people | candidate | of | college
Hispanic	the | . | vote | , | in | for | and | candidate | popular | electors | people | a | to | - | of | votes | Electoral | elect
White	the | vote | , | in | . | and | for | popular | people | electors | to | college | system | College | votes
P2-Electoral College 100% Augmented Model (GPT-4o)	
Asian/Pacific Islander	the | they | . | The | is | we | Electoral | electors | to | express | may | about | , | a | change
	pick | college | for | say | system | candidate | everyone
Black/African American	the | in | . | for | to | , | and | electors | a | vote | people | electoral | be | The
Hispanic	the | in | . | , | for | a | vote | to | who | people | electors | The | candidate | ’ | by | and | be

White	the | , | in | . | a | for | vote | to | people | be | and | electors | electoral

Prompt 3: High Importance Tokens by Subgroup (Ordered by Frequency from High to Low)

P3-Car-free Cities 100% Base Model	
Asian/Pacific Islander	to | . | cars | by | the | and | a | do | that | , | in | people
Black/African American	. | to | the | a | , | in | and | . | car | - | people | cars | of
Hispanic/Latino	. | to | in | the | , | a | cars | and | people | car
White	. | to | the | in | , | a | people | cars | and | -
P3-Car-free Cities 100% Augmented Model (GPT-4)	
Asian/Pacific Islander	to | . | a | can | of | , | is | are | people | car
Black/African American	. | a | , | to | cars | the | and | can | in | In | people
Hispanic	. | to | a | in | , | cars | and | people | car
White	. | to | a | , | in | cars | the | people | and | of
P3-Car-free Cities 100% Augmented Model (GPT-4o)	
Asian/Pacific Islander	to | . | a | cars | the | and | was | people | is
Black/African American	. | to | a | and | in | the | - | can | , | for | cars | is
Hispanic	to | . | and | a | in | - | the | , | can | for | car
White	to | . | a | and | - | the | in | , | for | cars | The | transportation | can | people

Prompt 5: High Importance Tokens by Subgroup (Ordered by Frequency from High to Low)

P5-Exploring Venus 100% Base Model

Asian/Pacific Islander	to | and | Venus | , | present | can | the | has | surface | in | is | planet | not | that
Black/African American	to | a | and | Venus | the | . | in | , | is | of | has | have | could | venus
Hispanic/Latino	to | and | Venus | the | . | a | , | is | have | could | in | s | planet
White	to | and | the | . | , | Venus | a | could | in | has | " | is | have
P5-Exploring Venus 100% Augmented Model (GPT-4)	
Asian/Pacific Islander	, | . | to | the | and | of | The | a | " | author | it | is | like | can | s | have | in
Black/African American	. | to | , | and | the | a | in | Venus | " | is | have | be | of | that | The
Hispanic	. | to | and | a | , | the | author | Venus | have | it | " | of | The | in
White	. | to | , | the | author | a | and | that | " | it | in | is | this
P5-Exploring Venus 100% Augmented Model (GPT-4o)	
Asian/Pacific Islander	. | and | to | the | of | The | in | , | Venus | s | know | not | very
Black/African American	and | to | the | . | of | in | Venus | a | The | that | is | , | venus
Hispanic	and | the | to | of | . | Venus | The | In | is | have | in | covered
White	and | to | . | of | the | in | Venus | , | be | a | author | with | that

Prompt 6: High Importance Tokens by Subgroup (Ordered by Frequency from High to Low)

P6-Face on Mars 100% Base Model	
Asian/Pacific Islander	, | . | Mars | to | a | and | " | the | created | face | that | s | they | can | of | there | was | like | in | evidence | [SEP]
Black/African American	. | to | the | a | and | , | The | was | " | Mars | . | by | t | is | land | ’ | s

Hispanic/Latino	. | , | the | to | a | and | that | " | Mars | there | aliens | The | was | people | in | not
White	to | the | . | a | and | , | " | Mars | was | - | there | is | in | created | . | not
P6-Face on Mars 100% Augmented Model (GPT-4)	
Asian/Pacific Islander	face | . | , | a | The | the | by | is | that | there | in | NASA | spacecraft | aliens | not | created | to | mesa | and
Black/African American	. | the | . | , | to | and | a | Mars | face | by | there | aliens | is | can | The | was | on | " | not
Hispanic	. | , | and | to | the | " | not | Mars | a | land | form | that | was | there | it | on | of | face | aliens
White	. | , | to | Mars | the | a | and | " | face | not | was | is | aliens | that | there | this
P6-Face on Mars 100% Augmented Model (GPT-4o)	
Asian/Pacific Islander	. | a | , | Mars | there | form | in | is | and | the | was | aliens | to | which | Cy | face | that
Black/African American	. | a | the | aliens | , | to | there | . | Mars | and | face | ’ | that | this | it | because | but
Hispanic	. | , | to | a | the | Mars | and | there | it | aliens | " | was | ? | but | that | could | just
White	a | . | to | the | , | Mars | aliens | there | and | that | was | it | but | " | this | is | face | in