Declarations of interest: None.

**Title**: Can AI Provide Useful Holistic Essay Scoring?

**Abstract**: Researchers have sought for decades to automate holistic essay scoring. Over the years, these programs have improved significantly. However, accuracy requires significant amounts of training on human-scored texts—reducing the expediency and usefulness of such programs for routine uses by teachers across the nation on non-standardized prompts. This study analyzes the output of multiple versions of ChatGPT scoring of secondary student essays from three extant corpora and compares it to quality human ratings. We find that the current iteration of ChatGPT scoring is not statistically significantly different from human scoring; substantial agreement with humans is achievable and may be sufficient for low-stakes, formative assessment purposes. However, as large language models evolve additional research will be needed to continue to assess their aptitude for this task as well as determine whether their proximity to human scoring can be improved through prompting or training.

**Highlights**

- Humans and AI were substantially internally consistent (i.e., human-human, AI-AI scores)
- Mean differences between human-human and AI-human scores were not statistically significant
- Weighted Kappas (which take into account both chance and proximity of scores) showed substantial agreement (.79, Sample 1) between human scorers, and moderate to fair agreement for the AI-human comparison (.52, .23, and .52, Samples 1-3 respectively)
- AI was slightly more consistent with a human scorer (weighted Kappas of .51 and .52) compared to when papers were written by English learners (.40 and .43), but in each case much less consistent than the two human scorers (.81 for English learners and .82 for non-English learners).

**Keywords:** artificial intelligence, AI, automated scoring, writing, assessment, large language models

---

## 1. Introduction

Struggling adolescent writers are in every secondary school classroom (National Center for Education Statistics, 2012). In part, these results may be attributed to limited extended writing experiences in school (Applebee & Langer, 2011). The key to writing improvement is engaging students in recursive writing (Flower & Hayes, 1981). The more opportunities students have to write, the more their writing will improve (Tate et al., 2016). As part of this recursive process, educators frequently assign a quality score to student writing. Unfortunately, scoring papers is time consuming, and secondary school teachers face practical constraints that reduce the individualized attention to writing improvement available (Lawrence et al., 2013). Human scoring is not only resource intensive, but also results in less than optimal inter- and within-rater reliability in many cases. Human biases, including unrelated judgments due to student handwriting and bias due to English language proficiency, can lead to non-random variation in scores (e.g., Eckes, 2008; Klein & Taub, 2005). Our work is based on Graham's (2018)

writer(s)-within-community model of writing, which conceptualizes writing as a social activity situated within specific communities. Writing is shaped by the affordances and constraints of the community and the skills and characteristics of the writer, thus a combination of sociocultural and cognitive perspectives.

For decades, people have been trying to create automated tools to evaluate writing quality, and some of these have approached—or even exceeded—human reliability (for an overview, see Shermis & Burstein, 2013; Ifenthaler, 2022). This type of software is often referred to as automated essay scoring (AES). However, in order for these automated tools to work on texts responding to *any particular writing prompt*, they typically had to be trained on hundreds of samples of writing *on that exact prompt* already graded by humans (see Ifenthaler, 2022). Thus, their use cases were limited to situations where training made economic sense, and consequently, automated tools were not widely used for common writing assignments across the disciplines (e.g., history).

The AES tools' correlations and agreement with human raters have become fairly high (Ifenthaler, 2022; Link & Koltovskala, 2023; Warschauer & Ware, 2006). State-of-the-art models report quadratic weighted Kappas ranging from .57 to .80, with most in the low .70's, evidencing substantial agreement between the models and human raters (Beseiso et al., 2021; Uto et al., 2023). Many of these studies highlight the results of adjacent agreement between humans and AES systems rather than those of exact agreement (Ifenthaler & Dikli, 2015). Exact agreement is harder to achieve as it requires two or more raters to assign the same exact score on an essay while adjacent agreement requires two or more raters to assign a score within one scale point of each other (Ifenthaler, 2022). In addition, correlation studies are conducted mostly in high-stakes assessment settings rather than classroom settings; therefore, AES versus human inter-rater reliability rates may not be the same in diverse assessment settings. However, we do know that these tools can be useful even if their feedback and scores are not perfect (Grimes & Warschauer, 2010; Moore & MacArthur, 2016), with students doing additional revision though often mostly at the surface level (Warschauer & Grimes, 2008) and automated scoring being more reliable for non-struggling students (D. Chen et al., 2022).

With the emergence of ChatGPT (and the ever-increasing number of similar large language models [LLMs]), we can generate output from AI tools. Researchers have been investigating whether LLMs can provide analysis of writing without the cost-intensive training of the model. Even the earlier version of the AI tool we used showed an ability to detect low-quality content without any training (Bahri et al., 2021, using GPT-2). But, before we advise educators and students to use these new generative AI tools for instructional, assessment, or research purposes, we need to examine their output to understand how they compare to more traditional human scoring. We begin this effort by tasking AI to score a number of text sets that have already been scored by humans and then comparing human and AI scoring across a number of elements. Several other recent studies have done related work. Mizumoto and Eguchi (2023) used the text-davinci-003 GPT model to score TOEFL essays and understand the extent to which linguistic features influenced output and found that LLMs could be effectively used for automated scoring on a 10-point scale. Liang et al. (2023) evaluated the quality of feedback provided by an LLM with that of peer reviews in scientific journal evaluations of manuscripts and found an overlap of points raised by GPT-4 and human reviewers ranged from 30.85% to 39.23%, with increased overlap for the weaker papers (for other feedback studies using generative AI, see Jia, Cao, & Gehringer, 2022; Jia, Young, et al., 2022; Rashid et al., 2022; Yoon et al., 2023). Naismith et al. (2023) used GPT-4 to assess discourse coherence and found

strong potential for comparable AI and human ratings. Baffour et al. (2023) reported on the results of a Kaggle competition and found that LLM prediction of discourse effectiveness was subject to moderate bias related to English status, race/ethnicity, and economic disadvantage.

This study contributes to the current literature by comparing a one-shot, minimally prompted large language model to extant samples of holistic scoring of secondary school student essays by well-trained human raters across multiple prompts and genres. Essentially, comparing best case human scoring to the free version of a large language model as it might be used by a nonexpert teacher in the field. This provides teachers with actionable information on the extent to which they might feel comfortable using large language models for holistic essay scoring. We compare these ratings of writing quality based on discipline-specific writing rubrics and report heterogeneity based on the English language ("EL") status of the writer, an important contribution due to concerns in the field about human variation in scoring the writing of ELs.

**Research Questions**

1. How internally consistent are humans (multiple humans) versus AI (multiple conversations) when scoring texts?
2. How does AI scoring compare to that of human raters on the same data?
3. Do scores systematically vary by English language status differently for AI ratings than for human ratings?

## 2. Materials and Methods

### 2.1 Data

Our sample consists of existing student essays that were previously graded by human raters in connection with other studies and interventions. We accessed the following essays and human scores:

***Sample 1*: *WRITE Center Field Trial***

This corpus of student essays ($n$ = 493) is from the field trial of a study designed to improve source-based argument writing for students in secondary school. Students in largely Latino schools in Southern California wrote source-based arguments of causal analysis in history classes in fall and in spring. Students were given one class period to read multiple sources and a second class period to write a source-based argumentative essay on a historical question. Students were 52% female and 72% Hispanic. Grade level and language status are in the tables below (Table 1, Table 2).

**Table 1.**

*Grade level of students writing sample 1 essays.*

| Grade 6 | Grade 7 | Grade 8 | Grade 10 | Grade 11 | Grade 12 |
|---|---|---|---|---|---|
| 76 | 119 | 74 | 81 | 80 | 63 |

**Table 2.**

*English language status of students writing sample 1 essays.*

| English Only | Initially Fluent | Reclassified Fluent | English Learner |
|---|---|---|---|
| 225 | 42 | 161 | 65 |

Papers were scored by 18 trained raters on a holistic rubric from 1 (no evidence of achievement) to 6 (exceptional achievement). The rubric was developed using extant rubrics for source-based analytic writing in history and input from subject matter experts in the field (Monte-Sano, 2010, 2012; Monte-Sano & De La Paz, 2012; National Writing Project, 2010; Northwest Regional Educational Laboratory, 2011). Training occurred over the course of three hours: Raters used the holistic rubric and anchor papers to identify key features of papers scoring in the six different categories. They also engaged in rounds of scoring with the research team to calibrate their scores. Raters were secondary literacy and history teachers or graduate students majoring in education or history. All of the essays were double scored to assess reliability. Scores that disagreed by more than 1 point were scored by a third evaluator. Interrater reliability for exact agreement was reported in the initial study report as 51% (558/1,104 essays) and within 1 point agreement was 92% (1,019/1,104 essays). We calculate Cohen's Kappa for these ratings as part of Research Question 1, as it is not included in the published text, and find it is .36 unweighted (see Appendix, Table A2, Sample 1: Human-Human) and .79 weighted (see Table 8, Sample 1: Human-Human).

***Sample 2: Pathway 2.0***

This corpus of essays ($n$ = 344) is from a study (Olson et al., 2020) designed to improve argumentative writing in urban English language arts (ELA) classes (arguments of literary interpretation). Like Sample 1, the essays were written by secondary students over two days. The papers were scored by trained raters from the National Writing Project (NWP) using the NWP-Analytic Writing Continuum (AWC) and overseen by SRI. All papers received a holistic score on a 6-point scale (averaged across two graders). Over a decade, the NWP developed the AWC, which has been shown to be a valid and reliable measure of student writing (Bang, 2013). The population was primarily Latino and eligible for free and reduced lunch. Language status is indicated in the table below (Table 3). We do not have sex or grade level data on this sample.

**Table 3.**

*English language status of students writing sample 2 essays.*

| English Only | Initially Fluent | Reclassified Fluent | English Learner |
|---|---|---|---|
| 215 | 28 | 96 | 5 |

***Sample 3: Crossley Corpus***

The Crossley corpus can be found at https://github.com/scrosseye/persuade_corpus_2.0. In total, the PERSUADE 2.0 corpus comprises over 25,000 argumentative essays produced by 6th-12th grade students in the United States for 15 ELA prompts on two writing tasks: independent and source-based writing. The PERSUADE 2.0 corpus provides holistic essay scores for each persuasive essay, based on the SAT rating rubric. We used the same rubric in our AI prompt (note: we used the form for independent essays for all essays in order to facilitate batch scoring; we felt the differences were slight, e.g., "using evidence" compared to "using evidence taken from the source text"). Human inter-rater reliability before adjudication was strong, with a weighted Kappa of .745 and an $r$ of .750 (Crossley et al., 2023). The prompts can be found on GitHub, but writing conditions (e.g., time permitted for writing) are not available. We randomly selected essays ($n$ = 949) for analysis from this sample, using the RAND() function in Excel to create a random number between 0 and 1, then took the 1000 essays with the lowest random number. We ran these essays through the AI scoring analysis. After data cleaning

and removing observations without the necessary information for our analysis, we ended up with 949 in our sample.

Students were 49% female; 9% English learners; and 45% white, 25% Hispanic, and 18% Black/African American. Grade level and language status are in the tables below (Table 4).

**Table 4.**

*Grade level of students writing sample 3 essays.*

| Grade 6 | Grade 8 | Grade 9 | Grade 10 | Grade 11 | Grade 12 |
|---|---|---|---|---|---|
| 45 | 374 | 77 | 304 | 129 | 14 |

**2.2 Methods**

We randomly sampled 30 papers from Sample 1 for our first round of analyses. We ran this subsample through AI (March 23, 2023 version of GPT-4 and GPT-3.5) 3 times each (i.e., for a total of 6 scores), reporting the score for each. Our prompt for AI in all cases for Sample 1 instructed the AI to pretend it was a secondary school teacher and use a 1-6 rubric (providing the AI with the text of the top scoring criteria, e.g., "supports claim with relevant and sufficient evidence," see Appendix for exact prompt). We did not provide AI with any examples or models of papers, so this was a zero-shot exercise. We did this at two temperature settings, 0.1 (low temperature), the setting that calls for less random variation) and 1.0 (high temperature, the most common default temperature setting that allows for diverse responses).

We initially calculated Cohen's Kappa using Stata (Version 15) for the human and AI scoring to determine inter-rater reliability between human raters, AI ratings at each temperature setting (low and high temperatures), and human-AI ratings. We calculated the consistency for subgroups of corpora and EL status in the same manner. The Kappa-statistic measure of agreement is scaled to be 0 when the amount of agreement is what would be expected to be observed by chance and 1 when there is perfect agreement. For intermediate values, Landis and Koch (1977) suggest the following interpretations: below 0.0 Poor; 0.00 – 0.20 Slight; 0.21 – 0.40 Fair; 0.41 – 0.60 Moderate; 0.61 – 0.80 Substantial; 0.81 – 1.00 Almost perfect. We subsequently calculated weighted Kappa using Klein's (2016) *kappaetc* package to account for concerns that our scores were ordinal, and thus the difference between a 1 and 6 should be given more weight than that between 4 and 6. Nonweighted Kappas and agreement can be seen in Appendix.

We decided to use GPT-3.5 to analyze the full samples (Sample 1, n = 493; Sample 2, n = 344; and Sample 3, n = 949) because it was the model widely available and free of charge. It was also the only model with an available API (we used the API to run the large samples through the model in a batch process). We hypothesized that teachers in the field would have the easiest access to this version of the model, thus making it the most relevant for our application. We used a temperature of .1. Human scores are the average of multiple humans in Samples 1 and 2, while human is a single score in Sample 3, and AI is a single iteration in all. With respect to all samples, we dropped observations that did not include a human score and AI score, indication of corpus, and indicator of EL status.

We ran diagnostics on each sample and compiled descriptive information. Specifically, we calculated the mean, standard deviation, minimum, and maximum for AI and human quality scores for each corpus as a whole, as well as for English learners and non-English learners. We also indicate confidence intervals so that the statistical significance of mean differences is visible. We then calculated Cohen's Kappa (including using weights due to the ordinal nature of

our data) between the human and AI scores for each sample to understand how closely AI matched human scores. This was also done by separating English learner students in Samples 1 and 3. With respect to Sample 2, we did not have the statistical power to analyze heterogeneity. Sample 3 was run subsequently using the June 13, 2023 version of GPT 3.5. One challenge of researching AI in the current environment is that models are changed without warning or notice.

We next ran an empty model, a random effects ANOVA, to determine the ICC between corpora. We also tested the difference between human and AI scores with a random effects ANOVA. Then we used a two-way random effects model to determine the difference between human and AI ratings in all of our samples. We then used a linear regression model to examine the effect of the independent variable (scorer type, human or AI) on the dependent variable (essay quality scores), controlling for the fixed effects of the three different corpora and English learner status. While our pre-registration planned on a multilevel model, we instead chose to simply use linear regression with clustered errors to account for the non-independent nature of our error term (this study was preregistered at https://osf.io/5gkfe, with the noted difference). We regressed the writing achievement score on the dichotomous variable indicating human vs. AI scoring, a categorical variable indicating the corpus, and the student's English learner status. Finally, we regressed the scores on variables for AI versus human ratings and corpus, clustering the random errors, and tested the interaction between AI scoring and English learner status.

## 3. Results

***Research question 1: How internally consistent are humans (multiple humans) versus AI (multiple conversations) when scoring texts?***

Table 5 sets out the weighted Kappa and agreement for each variation (e.g., GPT-4, temperature 0.1). GPT-4 consistently shows better internal consistency compared to the other AI models, but the temperature is less consistently predictive of increased consistency. We would have expected the less random temperature to be more internally consistent but did not find that to be uniformly true. Overall, we see that GPT-4 is generally more consistent than humans, but humans are more consistent than GPT-3.5, though at the high temperature GPT-3.5 ties with human consistency as to Kappa and beats it in weighted agreement. These trends hold true across the non-English learners who make up the large majority of our sample, but the low temperature GPT-3.5 was more consistent than humans for English learners.

With respect to unweighted scores (Appendix, Table A1), humans score papers exactly the same as another human 43% of the time. AI scores the same as another iteration of AI between 59% and 82% of the time, depending on the model and temperature setting. All versions of the AI, GPT-3.5 and 4, at both high and low temperatures, are more internally consistent at scoring essays, whether on a percent agreement or Kappa basis, than humans are consistent with other humans. We had hypothesized that ChatGPT scoring would be highly reliable (> .80). We found that GPT-4 approached the .80 consistency level, whereas GPT-3.5 and human raters (who we hypothesized would be between .70 and .80) did not for this subsample.

**Table 5**

*Internal Consistency: Weighted Cohen's Kappa and Agreement for Subsamples of AI-AI and Human-Human Scoring (using kappaetc function with wgt(ordinal))*

| Scoring condition | Subsample 1 (all) | Subsample 1 (non-English learners) | Subsample 1 (English learners) |
|---|---|---|---|

| | Cohen's/ Conger's Kappa (wtd) | Agreement | Cohen's/ Conger's Kappa (wtd) | Agreement | Cohen's/ Conger's Kappa (wtd) | Agreement |
|---|---|---|---|---|---|---|
| Human | .73 (.09)*** | 95.24% (.01)*** | .71 (.11)*** | 94.98%*** | .73 (.11)* | 91.11% (.04)*** |
| Low temp: GPT-4 | .88 (.05)*** | 97.47% (.01)*** | .84 (.06)*** | 96.91% (.01)*** | NA | NA |
| High temp: GPT-4 | .80 (.09)*** | 97.51% (.01)*** | .76 (.12)*** | 97.10% (.02)*** | .93 (.09)*** | 97.78% (.02)*** |
| Low temp: GPT-3.5 | .63 (.11)*** | 95.06% (.01)*** | .59 (.14)*** | 94.85% (.12)*** | .74 (.12)** | 93.33% (.03)*** |
| High temp: GPT-3.5 | .73 (.07)*** | 96.37% (.01)*** | .74 (.07)*** | 96.30% (.01)*** | .46 (.35) | 88.89% (.08)*** |

*Note*. Standard errors are in parentheses. wtd = weighted.
*$p < .05$. **$p < .01$. ***$p < .001$.

***Research question 2: How does AI scoring compare to that of human raters on the same data?***

Looking first at the descriptive statistics for our subsample (we have multiple human scores for Sample 1 only), we find that scores by humans and the various AI iterations are quite similar on average (Table 6).

**Table 6**
*Descriptive Statistics for Subsample Populations*

| Scoring condition | Subsample 1 (all) | | | | Subsample 1 (non-English learners) | | | | Subsample 1 (English learners) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SE* | 95% CI | | *M* | *SE* | 95% CI | | *M* | *SE* | 95% CI | |
| | | | *LL* | *UL* | | | *LL* | *UL* | | | *LL* | *UL* |
| Human | 2.83 | .18 | 2.47 | 3.19 | 2.98 | .20 | 2.57 | 3.39 | 2.17 | .29 | 1.50 | 2.83 |
| Low temp: GPT-4 | 2.89 | .21 | 2.46 | 3.32 | 3.12 | .22 | 2.68 | 3.57 | 1.80 | .37 | .76 | 2.84 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High temp: GPT-4 | 3.02 | .22 | 2.57 | 3.47 | 3.22 | .22 | 2.77 | 3.68 | 2.07 | .55 | .53 | 2.60 |
| Low temp: GPT-3.5 | 2.64 | .16 | 2.30 | 2.97 | 2.76 | .17 | 2.40 | 3.12 | 2.08 | .41 | .94 | 3.23 |
| High temp: GPT-3.5 | 2.46 | .17 | 2.10 | 2.81 | 2.59 | .20 | 2.18 | 2.99 | 1.85 | .22 | 1.24 | 2.46 |

*Note*. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

GPT-4 mean scores are higher than human scores with respect to the full sample, and GPT-3.5 scores are lower than human scores, but these differences are not statistically significant (see confidence intervals). The AI models tended to give fewer absolute high and low scores compared to the human raters. We find variation in the averages by language status: For English learners, all of the AI scores are lower than human scores. For non-English learners, the higher temperature scores are higher than human scores, while the lower temperature scores are lowers.

Turning to our full samples (Table 7), we see that the descriptive statistics for human and AI scores are similar; only Sample 1 (both the full sample and the non-English learner sample) had a significant difference.

**Table 7**

*Descriptive Statistics for Sample 1, 2, and 3 as a Whole and Sample 1 and 3 by English Learner Status*

| | Sample 1 (all) | | Sample 2 (all) | | Sample 3 (all) | | Sample 1 (non-English learners) | | Sample 1 (English learners) | | Sample 3 (non-English learners) | | Sample 3 (English learners) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Human | AI | Human | AI | Human | AI | Human | AI | Human | AI | Human | AI | Human | AI |
| Mean | 3.16 | 3.44 | 2.69 | 2.71 | 3.38 | 3.36 | 3.27 | 3.56 | 2.47 | 2.62 | 3.43 | 3.40 | 2.82 | 2.90 |
| Conf. Int. | 3.05-3.28 | 3.32-3.55 | 2.60-2.78 | 2.63-2.79 | 3.30-3.46 | 3.30-3.42 | 3.15-3.39 | 3.44-3.68 | 2.22-2.72 | 2.35-2.88 | 3.35-3.51 | 3.34-3.47 | 2.60-3.04 | 2.74-3.06 |
| Std. Err. | .06 | .06 | .05 | .04 | .04 | .03 | .06 | .06 | .13 | .13 | .04 | .03 | .11 | .08 |
| Std. Dev. | 1.25 | 1.27 | .85 | .77 | 1.20 | .94 | 1.25 | 1.26 | 1.02 | 1.09 | 1.20 | .95 | 1.01 | .74 |
| Minimum | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| Maximum | 6 | 6 | 5 | 5 | 6 | 5.5 | 6 | 6 | 4.5 | 5 | 6 | 5.5 | 6 | 4 |
| Skewness | .03 | -.10 | .10 | .70 | .14 | .06 | -.05 | -.22 | .34 | .74 | .13 | .05 | .22 | .15 |
| Kurtosis | 2.29 | 1.78 | 2.77 | 2.75 | 2.44 | 2.21 | 2.33 | 1.85 | 2.26 | 2.75 | 2.40 | 2.19 | 2.98 | 1.86 |
| *n* | 493 | 493 | 344 | 344 | 949 | 949 | 428 | 428 | 65 | 65 | 865 | 865 | 84 | 84 |

The distributions of scores for Sample 1, the history sample, by AI and humans were dissimilar, with a peak at a score of 2 for AI and at 3 for humans. Neither distribution was normal (AI skewness was -.10 and kurtosis was 1.78, while human was .03 and 2.29, respectively). For Sample 2, distributions of AI and human scores were also quite different, with a peak of 2 and virtually no lower score for the AI scores, compared to a more normal-looking peak of around 3 for humans (AI skewness was .70 and kurtosis was 2.75, while human was .10 and 2.77, respectively). Sample 3 is also not normally distributed, but it shows the most similar distribution across human and AI scores, with AI having a skewness of .90 and kurtosis of 2.20 and human of .16 and 2.44, respectively. We note that non normality can impact the ability of an automated model to score the less represented scores well, which may have impacted the quality of scoring and in future work we may seek to augment the low-frequency scores or balance the representation of scores across the levels in order to train AI to more effectively score papers (see, e.g., Fang, Lee, & Zhai, 2023).

Beyond mean scores, it is important that at the *student* level the scores are valid. Educators want to know whether the score an essay receives from AI is consistent with the score given on the same paper by a well-trained human. The agreement and Cohen's Kappa for human and AI scores (Table 8, for weighted scores, Appendix Table A2, for unweighted) for all samples provide information on the consistency of AI and human scores.

**Table 8**

*Human-Human (for Purposes of Comparison, Sample 1) and Human-AI Consistency for All Samples*

| | Cohen's/ Conger's Kappa (*SE*), wtd | Weighted agreement | Agreement within 1 point |
|---|---|---|---|
| Sample 1: Human-Human | .79 (.01)*** | 96.61%*** | 74% |
| Sample 1: Human-AI | .52 (.03)*** | 93.03%*** | 76% |
| Sample 2: Human-AI | .23 (.05)*** | 92.47%*** | 83% |
| Sample 3: Human-A1 | .52 (.02)*** | 95.65%*** | 89% |

*Note*. wtd = weighted.

***$p < .001$.

Weighted Kappas showed substantial agreement (.79) between human scorers, and moderate to fair agreement for the AI-human comparison. The results show that (unweighted) exact agreement is hard for humans (49%), but much worse for the AI. Similarly, the human consistency with other humans is better than the AI-human consistency in any of our samples, whether looking at the weighted agreement or Kappa. Interestingly, Sample 3, which is one of the ELA samples and was scored by a later version of ChatGPT-3.5, approaches the human level of exact agreement (42%), and the weighted agreement of humans (96.61%) is only slightly better than the AI (Sample 3, 95.65%, Sample 1, 93.03%).

We had hypothesized that ChatGPT would be within 1 point of human scoring 90% or more of the time for non-English learners on a 1-6 scale. When looking at agreement within one point, we see that the AI performs better than the humans, presumably due to the lack of the scores of 1 and 6 by AI noted above.

Combining the three samples, our ANOVA found no statistically significant difference in the scores of humans and AI. However, there were statistically significant differences in scores by sample. Given the differences in genre (history versus ELA), prompts, and conditions, this is not surprising. Similarly, the differences in scores between non-English learners and English learners were statistically significant. Our two-way random effects model found a significant difference between human and AI ratings in all of our samples. The consistency of our human and AI ratings is moderate: .57 when we analyze the ICC. Finally, we regressed the scores on variables for AI versus human ratings and corpus, clustering the random errors, and found no significant effect of AI, when controlling for corpus and English learner status (Table 9).

**Table 9**

*Regression, With Clustered Errors, of AI Scoring and Corpus 1 and 2 Compared to Corpus 3 (Model 1), Adding English Learner Control (Model 2), and an Interaction Between AI Scoring and English Learner Status (Model 3)*

| Variable | *B* | *SE* | *t* | *p* |
|---|---|---|---|---|
| Model 1 | | | | |
| Constant | 3.34 | .05 | 73.68 | 0.000 |
| AI scoring | .07 | .09 | .73 | 0.541 |
| Corpus 1 | -.07 | .00 | .00 | 0.000 |
| Corpus 2 | -.67 | .00 | .00 | 0.000 |
| Model 2 | | | | |
| Constant | 3.36 | .06 | 60.78 | 0.000 |
| AI scoring | .07 | .09 | .73 | 0.541 |
| Corpus 1 | -.09 | .02 | -4.97 | 0.038 |
| Corpus 2 | -.51 | .12 | -4.37 | 0.049 |

| | | | | |
|---|---|---|---|---|
| English learner | -.30 | .22 | -1.38 | 0.302 |
| Model 3 | | | | |
| Constant | 3.36 | .07 | 51.18 | 0.000 |
| AI scoring | .07 | .11 | 0.65 | 0.584 |
| Corpus 1 | -.09 | .02 | -4.97 | 0.038 |
| Corpus 2 | -.51 | .12 | -4.36 | 0.049 |
| English learner | -.28 | .26 | -1.11 | 0.382 |
| AI x EL | -.03 | .12 | -.25 | 0.829 |

*Note*. EL = English learner.

***Research question 3: Do scores systematically vary by English language status differently for AI ratings than for human ratings?***

We hypothesized that ChatGPT scoring would be significantly different from human scoring and would exhibit different patterns from human scoring by English learner status and text corpus. Indeed, we found some interesting differences. Tables 5, 6, and 7 above show the descriptive statistics for both English learners and non-English learners for Samples 1 and 3. With only 5 English learners in Sample 2, we did not calculate English learner data. Table 10 shows the weighted Kappa and agreement by population for Sample 1 and Sample 3.

**Table 10**

*Human-AI Consistency: Weighted Cohen's Kappa and Exact Agreement for All Samples, by Population (using kappaetc function with wgt(ordinal))*

| Sample | Non-English learner | | | English learner | | |
|---|---|---|---|---|---|---|
| | Cohen's/ Conger's Kappa (*SE*), wtd | Weighted agreement | Agreement within 1 point | Cohen's/ Conger's Kappa (*SE*), wtd | Weighted agreement | Agreement within 1 point |
| Sample 1: Human-Human | .82 (.01)*** | 97.01%*** | 73% | .81 (.01)*** | 97.08%*** | 80% |
| Sample 1: Human-AI | .51 (.04)*** | 92.90%*** | 77% | .40 (.10)*** | 90.73%*** | 75% |
| Sample 3: Human-AI | .52 (.02)*** | 95.58%*** | 89% | .43 (.07)*** | 95.24%*** | 92% |

***$p < .001$.

Human consistency did not differ much between our English learners and their peers (.82 and .81 weighted Kappas). AI, however, was better (i.e., was closer to human scores) with respect to the non-English learners (.51 and .52) compared to the English learners (.40 and .43). Turning to agreement within one point, 6 shows in Sample 3 we actually see the English learner population showing more consistency between human and AI scoring than the non-English learner population. Our regression (using clustered standard errors) found no significant effect of the interaction between AI scoring and English learner status. Combining the three samples, our ANOVA found a statistically significant difference in the scores of English learners and their peers, which was not surprising, but our regression (Table 9) did not find an interaction between English learner status and the use of AI for scoring.

## 4. Discussion

Particularly for summative assessments, humans should score student essays, at least compared to extant untrained OpenAI models. Humans' weighted Kappa of .79 (Sample 1, see Table 8) shows substantial agreement with other human scorers, compared to AI's weighted Kappas of .52 (Sample 1), .23 (Sample 2), and .52 (Sample 3) which showed fair to moderate agreement with human scores. Our AI model for Research Questions 2 and 3 was ChatGPT-3.5, the current free version. We suspected, given the results of Research Question 1, that as the AI improves, it may be able to match—or even exceed—human reliability. In connection with the review process, we were able to take advantage of the availability of the API for GPT-4 and run a robustness check on our initial results (March 14, 2024). Using the combined corpus of all 3 samples, we tested the weighted Kappa of both GPT-3.5 and GPT-4. For non-English learners GPT-3.5 had a weighted Kappa of .52*** and GPT-4 of .58***, indicating additional improvement in scoring this group of students over human-human scores. Results for English learners also remained lower than that of humans with GPT-3.5 having a weighted Kappa of .36*** and with GPT-4 barely improving to .37***. For now, the AI is not consistent enough to substitute for human ratings for higher stakes assessments, especially when English learners are part of the sample.

Our subsample analysis (Research Question 1) showed that GPT-4 was much more reliable based on *internal* consistency (i.e., AI consistency with AI) than GPT-3.5 and humans, with GPT-4 in exact agreement with itself over 80% of the time, compared to GPT-3.5 (approximately 60%) and humans (43%). Internal consistency was AI's strongest positive trait, regardless of the model or temperature used. This is consistent with prior research showing the fallibility of human scoring (see, e.g., Brown, 2009; Cohen et al., 2018; Saal et al., 1980; Weigle, 1999). Higher temperatures in AI (the amount of randomness permitted) made the AI scoring slightly more consistent on the full sample, but there were differences when we compared English learners to their peers. Ultimately, the temperature setting had much less impact on reliability than the model.

Comparing the AI scores to human scores, we saw that the mean scores for each corpus were similar across models and temperatures, but the differences were largely small and not statistically significant. This finding is consistent with a similar study in which researchers found that multiple AI scores had a variation of one to two points on a 10-point scale and that AI scores were consistent with the 3-point scale of human scoring on the TOEFL11 (Mizumoto & Eguchi, 2023). We did notice, though, that the AI was less likely to score papers on the edges, fewer ones and sixes, than humans, which may have contributed to its better internal consistency. This may be related to the training dataset, e.g., if the AI had more average writing in its training data than

extremely good and bad writing. We think this tendency is worth further investigation to understand whether it remains true in later models and what its impact is on scores.

Our regression found that there were significant differences between the scores of the 3 corpuses (with Sample 3 having the highest mean score), AI scores were slightly higher than human scores (.07, but this was not statistically significant), and English learners score lower than their non-English learner peers (-.28). Thus, on a population level there was no difference between human and AI scores. Our interaction of AI and English learners was small (-.03) and insignificant, indicating no particularized harm for English learners when AI scoring is used.

Similar mean scores were insufficient for our purposes: We wanted to know how consistent AI and human scores were on an essay, not a sample, level. Thus, although the results of both our ANOVA and our regression show that there was no statistically significant difference between the mean AI and human scores, the differences on the essay level are reflected in our Kappas.

The weighted Kappa of humans was .79 (Sample 1), showing substantial agreement between two human scores. AI showed moderate (Sample 1, .52, and Sample 3, .52) and fair (Sample 2, .23) agreement with human scores on the same essays, based on the weighted Kappas that adjust for both proximity of scoring and chance agreement. Note that weighted agreement was 96.61% (Sample 1, human-human), 93.03% (Sample 1, Human-AI), 92.47% (Sample 2, Human-AI), and 95.65% (Sample 3, Human-AI), suggesting that most students would receive actual scores quite close to the human scores from AI.

Writing researchers and practitioners often consider it sufficient for two graders to be within one point of one another—the students receive approximately correct scores that provide a sense of their performance. Such approximation is most appropriate for low-stakes, formative assessments. Using this metric, we saw that the AI actually performed *better* than the humans: Sample 3 was within one point of the human scores 89% of the time, while humans only were within one point 74% of the time.

We were also concerned about any particular impact on English language learners. While we hoped that AI would be less inclined to be biased against English learners, given the known bias in AI (see, e.g., Hofmann et al., 2024), we could not be sure. Unfortunately, unweighted and weighted agreement for both human pairs and AI-human scores was better for non-English learners (fair to moderate) compared to English learners (slight); weighted Kappas showed humans were only slightly better for non-English learners (.82 versus .81), while AI was better for non-English learners (.51 and .52, Sample 1 and 3, respectively) than English learners (.40 and .43). Exact agreement was less clear cut, however, with a more than 10% decline in exact agreement for humans (51% non-English learners compared to 40%) compared to a 2% decline in AI (20% compared to 18% for English learners) in Sample 1. Sample 3 was even more interesting, with the AI scores reaching 49% exact agreement with human scores for English learners compared to 41% for non-English learners. Our regression analysis found no statistically significant interaction between English learners and AI versus human scoring in our combined sample. We conclude that there is no clear harm to English learners in using AI scoring and, in fact, it is worth additional research to consider whether AI scoring could actually be more reliable for English learners in other corpora and with improved AI models and prompts.

## 5. Limitations and conclusion

The first limitation of note is that we used only one form of generative AI, ChatGPT, and even that model itself is changing rapidly (L. Chen et al., 2023). Thus, using a future version for

scoring could change the results. This is part of the reason for our robustness check with the now-available GPT-4; however the basic findings are unchanged at this point.

We also ran our experiment as a zero-shot exercise, with no training of the model on anchor papers exhibiting what a "6" should look like compared to what a "3" should look like on a 1-6 scale. We did this because many teachers will not want to take the time to train the model for every essay prompt. However, future research should consider how much training is needed for significant improvement. Teachers might find it worthwhile to do a limited amount of training, especially if they use similar assignments every year. Similarly, we did not do a significant amount of prompt engineering, and we think that there are areas that could be improved in future iterations (e.g., to reduce the clumping of scores in the mid ranges). It would be particularly beneficial to see if researchers can improve the rating quality of text written by English learners.

We note that this research is limited to two types of essays (ELA and history) and a single population—secondary school students. Future research should consider other age groups and genres. Research should also continue to identify other types of bias that may be latent in the AI models; a recent paper found a bias in favor of the AI's own generated content and longer texts (Liu et al., 2023).

Finally, we note that the three samples we used were somewhat opportunistic and differed in size and composition. Additional samples and populations are needed to confirm that these results are comparable to other situations.

While we would not recommend that AI be used for high-stakes summative purposes nor as a replacement for teachers' evaluation, we do see it as a useful tool in the early stages of multiple draft writing. It is sufficiently reliable and valid in providing a sense of the student's achievement level and could provide an impetus for revision (Grimes & Warschauer, 2010). Researchers should test whether the provision of instant scores, even if imperfect, motivates additional writing or revision by students. Revision is very limited in secondary school writing, and the ability to motivate students to spend some time on revision is likely to positively impact writing achievement. Research on students' perception of AES systems and the effect on motivation as well as on learning processes and learning outcomes is needed, particularly given the powerful and ubiquitous nature of generative AI (Stephen et al., 2021). Despite all the doomsday headlines about AI causing the end of writing, we believe that generative AI could, instead, be harnessed as a powerful tool to improve students' writing skills and habits.

**Statements on open data and ethics**

The study was approved by an ethical committee with ID: 20195085. Informed consent was obtained from all participants in the underlying intervention, and their privacy rights were strictly observed. The participants were protected by hiding their personal information during the research process. They knew that the participation was voluntary, and they could withdraw from the study at any time. The data can be obtained by sending a request via e-mail to the corresponding author.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Applebee, A., & Langer, J. (2011). A snapshot of writing instruction in secondary and high schools. *English Journal, 100*, 14–27. https://doi.org/10.58680/ej201116413

Baffour, P., Saxberg, T., & Crossley, S. (2023). Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023),* pages 242–246. https://doi.org/10.18653/v1/2023.bea-1.21

Bahri, D., Tay, Y., Zheng, C., Brunk, C., Metzler, D., & Tomkins, A. (2021, March). Generative models are unsupervised predictors of page quality: A colossal-scale study. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 301-309). https://doi.org/10.1145/3437963.3441809

Bang, H. J. (2013). Reliability of national writing project's analytic writing continuum assessment system. *Journal of Writing Assessment, 6*(1).

Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education, 33*, 727-746. https://doi.org/10.1007/s12528-021-09283-1

Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. *Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice and Research*, 40-48.

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal, 59*(6), pp. 1122-1156. https://doi.org/10.3102/00028312221106773

Chen, L., Zaharia, M., & Zou, J. (Jul. 18, 2023). How is ChatGPT's behavior changing over time? https://arxiv.org/pdf/2307.09009.pdf

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, *31*(3), 241-250. https://doi.org/10.1080/08957347.2018.1464450

Crossley, S., Baffour, P., Yu, T., Franklin, A., Benner, M., & Boser, U. (2023). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. https://doi.org/10.1016/j.asw.2023.100667

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Fang, L., Lee, G.-G., & Zhai, X. (2023). Using GPT-4 to augment unbalanced data for automatic scoring. Preprint. arXiv: 2310.18368v2

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365-387. https://doi.org/10.58680/ccc198115885

Graham, S. (2018). A revised writer (s)-within-community model of writing. *Educational Psychologist, 53*(4), 258-279. https://doi.org/10.1080/00461520.2018.1481406

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 1-44.

Hofman, V., Kalluri, P., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. Preprint https://arxiv.org/pdf/2403.00742.pdf

Ifenthaler, D. (2022). Automated essay scoring systems. In *Handbook of open, distance and digital education* (pp. 1-15). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0351-9_59-1

Ifenthaler, D., & Dikli, S. (2015). Automated scoring of essays. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 1, pp. 64–68). Sage.

Jia, Q., Cao, Y., & Gehringer, E. (2022). Starting from "zero": An incremental zero-shot learning approach for assessing peer feedback comments. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022),* pages 46 – 50. https://doi.org/10.18653/v1/2022.bea-1.8

Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Ghringer, E. (2022). Automated feedback generation for student project reports: A data-driven approach. *Journal of Educational Data Mining, 14*, 3.

Klein, D. (2016). "KAPPAETC: Stata module to evaluate interrater agreement," Statistical Software Components S458283, Boston College Department of Economics, revised 11 Aug 2022.

Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing, 10*(2), 134-148. https://doi.org/10.1016/j.asw.2005.05.002

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374. https://doi.org/10.2307/2529786

Lawrence, J. F., Galloway, E. P., Yim, S., & Lin, A. (2013). Learning to write in secondary school? *Journal of Adolescent & Adult Literacy*, *57*(2), 151-161. https://doi.org/10.1002/JAAL.219

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., ... & Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *arXiv preprint arXiv*:2310.01783.

Link, S., & Koltovskala, S. (2023), Automated scoring of writing in (Kruse et al., editors). *Digital writing technologies in higher education: theory, research, and practice,* p. 333. https://doi.org/10.1007/978-3-031-36033-6_21

Liu, Y., Moosavi, N. S., & Lin, C. (2023). LLMs as narcissistic evaluators: When ego inflates evaluation scores. arXiv preprint arXiv:2311.09766.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *The Journal of the Learning Sciences, 19*(4), 539-568. https://doi.org/10.1080/10508406.2010.481014

Monte-Sano, C. (2012). What makes a good history essay? Assessing historical aspects of argumentative writing. *Social Education, 76*(6), 294-298.

Monte-Sano, C., & De La Paz, S. (2012). Using writing tasks to elicit adolescents' historical reasoning. *Journal of Literacy Research, 44*(3), 273-299. https://doi.org/10.1177/1086296X12450445

Moore, N. S., & MacArthur, C. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research, 8*(1), 149-175. Doi: 10.17239/jowr-2016.08.01.05 https://doi.org/10.17239/jowr-2016.08.01.05

Naismith, B., Mulcaire, P., & Burstein, J. (2023, July). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394-403). https://doi.org/10.18653/v1/2023.bea-1.32

National Center for Education Statistics. (2012). The Nation's Report Card: Writing 2011 (NCES 2012–470). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Writing Project. (2010). *The Analytic Writing Continuum: A comprehensive writing assessment system.* University of California, Berkeley; Berkeley, CA: National Writing Project.

Northwest Regional Educational Laboratory. (2011). 6+1 Trait® Writing. Retrieved from http://educationnorthwest.org/traits.

Olson, C. B., Woodworth, K., Arshan, N., Black, R., Chung, H., D'Aoust, C., & Dewar, T. (2020). The pathway to academic success: Scaling up a text-based analytical writing intervention for Latinos and English learners in secondary school. *Journal of Educational Psychology, 112*(4), 701–717. https://doi.org/10.1037/edu0000387

Rashid, M. P, Xiao, Y., & Gehringer, E. F. (2022). Going beyond "good job": Analyzing helpful feedback from the student's perspective. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 515–521, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge. https://doi.org/10.4324/9780203122761

Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice, 54*, 103085. https://doi.org/10.1016/j.nepr.2021.103085

Tate, T. P., Warschauer, M., & Abedi, J. (2016). The effects of prior computer use on computer-based writing: The 2011 NAEP writing assessment. *Computers & Education*, *101*, 115-131. https://doi.org/10.1016/j.compedu.2016.06.001

Uto, M., Aomi, I., Tsutsumi, E., & Ueno, M. (2023). Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies.* https://doi.org/10.1109/TLT.2023.3253215

Warschauer, M., & Grimes, D. (2008). Automated Writing Assessment in the Classroom, *Pedagogies: An International Journal, 3*:1, 22-36. https://doi.org/10.1080/15544800701771580

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10,* 2; pp. 157-180. https://doi.org/10.1191/1362168806lr190oa

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. *arXiv preprint arXiv:2310.06505*.

Appendix
Prompts

**Sample 1**
Pretend you are a secondary school teacher scoring class essays based on this holistic rubric from 1 (minimum) to 6 (maximum), with the distance between each number (e.g., 1-2, 2-3) considered equal. A score of 6 means that the essay presents a clear, compelling, and accurate argument that addresses all requirements of the prompt; supports claim with relevant and sufficient evidence and compelling reasoning that connects evidence to claims; integrates sufficient, appropriate evidence from multiple sources and attributes evidence to sources, citing the title, author, and/or genre; evaluates reliability of sources and discusses how they confirm each other; effectively addresses and refutes a counterclaim with strong evidence and reasoning; writing is well organized with a strong introduction, body, and conclusion and transitions to create coherence; demonstrates effective and varied sentence fluency with little to no errors in writing conventions; uses sophisticated language and academic tone.
Each essay is indicated by "" and the text immediately prior to each essay is the id of the essay.

For each following essay, provide only the overall score of 1-6 based on the above rubric. Do not provide feedback other than the score. Set temperature to .1. Format the output as JSON in the following format for each essay:

```
{
        "ID": "insert id here",
        "Grade": "insert grade here"
}
```

Repeat for each essay.

**Sample 2**
Pretend you are a secondary school teacher scoring class essays based on this holistic rubric from 1 (minimum) to 6 (maximum), with the distance between each number (e.g., 1-2, 2-3) considered equal. A score of 6 means that the essay is clear and consistently focused; exceptionally well shaped and connected; reflects outstanding control and development of ideas and content; contains ideas that consistently and fully support and/or enhance the central theme or topic (e.g., well-developed details, reasons, examples, evidence, anecdotes, events, and/or descriptions, etc.); includes ideas that are consistently purposeful, specific, and often creative; presents an organization that enhances the central idea or theme; presents a compelling order and structure; writing flows smoothly so that organizational patterns are seamless; includes a compelling opening and an effective closure that reinforces unity and provides an outstanding sense of resolution; includes transitions that are smooth and cohesive; demonstrates a purposeful, coherent and effective arrangement of events, ideas, and/or details; consistently and powerfully demonstrates a clear perspective through tone and style; consistently demonstrates distinctive and sophisticated tone or style that adds interest and is appropriate for purpose and audience; exhibits level(s) of formality or informality very well suited for purpose and audience; demonstrates a sophisticated rhythm and cadence with very effective phrasing so that each sentence flows easily into the next; includes sentences that vary in structure and length creating an extremely effective text; fragments, if present, appear deliberately and effectively chosen for

stylistic purposes; includes sentence structures that are consistently logical and clear so that the relationships among ideas are firmly and smoothly established; contains words and expressions that are consistently powerful, vivid, varied, and precise; contains words that are usually creative and/or sophisticated, but natural and not overdone; contains lively verbs and precise nouns and modifiers that add depth and specificity; may include imagery; when present, it is consistently powerful; may include figurative language; when present, it is effective; is almost error-free and demonstrates an outstanding control of age appropriate standard writing conventions; includes spelling, usage, punctuation, capitalization, and paragraph breaks that are correct to the extent that almost no editing is needed; includes a wide range of age appropriate conventions intentionally used for stylistic effect.

For each following essay, provide only the overall score of 1-6 based on the above rubric. Do not provide feedback other than the score. Set temperature to .1. Format the output as JSON in the following format for each essay:

```
{
        "ID": "insert id here",
        "Grade": "insert grade here"
}
```

Repeat for each essay.

**Sample 3**

Pretend you are a secondary school teacher scoring class essays based on this holistic rubric from 1 (minimum) to 6 (maximum), with the distance between each number (e.g., 1-2, 2-3) considered equal. A score of 6 means that the essay is clear and consistently focused; exceptionally well shaped and connected; reflects outstanding control and development of ideas and content; contains ideas that consistently and fully support and/or enhance the central theme or topic (e.g., well-developed details, reasons, examples, evidence, anecdotes, events, and/or descriptions, etc.); includes ideas that are consistently purposeful, specific, and often creative; presents an organization that enhances the central idea or theme; presents a compelling order and structure; writing flows smoothly so that organizational patterns are seamless; includes a compelling opening and an effective closure that reinforces unity and provides an outstanding sense of resolution; includes transitions that are smooth and cohesive; demonstrates a purposeful, coherent and effective arrangement of events, ideas, and/or details; consistently and powerfully demonstrates a clear perspective through tone and style; consistently demonstrates distinctive and sophisticated tone or style that adds interest and is appropriate for purpose and audience; exhibits level(s) of formality or informality very well suited for purpose and audience; demonstrates a sophisticated rhythm and cadence with very effective phrasing so that each sentence flows easily into the next; includes sentences that vary in structure and length creating an extremely effective text; fragments, if present, appear deliberately and effectively chosen for stylistic purposes; includes sentence structures that are consistently logical and clear so that the relationships among ideas are firmly and smoothly established; contains words and expressions that are consistently powerful, vivid, varied, and precise; contains words that are usually creative and/or sophisticated, but natural and not overdone; contains lively verbs and precise nouns and modifiers that add depth and specificity; may include imagery; when present, it is consistently powerful; may include figurative language; when present, it is effective; is almost error-free and demonstrates an outstanding control of age appropriate standard writing conventions; includes

spelling, usage, punctuation, capitalization, and paragraph breaks that are correct to the extent that almost no editing is needed; includes a wide range of age appropriate conventions intentionally used for stylistic effect.

For each following essay, provide only the overall score of 1-6 based on the above rubric. Do not provide feedback other than the score. Set temperature to .1. Format the output as JSON in the following format for each essay:

```
{
        “ID”: “insert id here”,
        “Grade”: “insert grade here”
}
```

Repeat for each essay.

Appendix

**Table A1**

*Internal Consistency: (Unweighted) Cohen's Kappa and Agreement for Subsamples of AI-AI and Human-Human Scoring (using kappaetc function, Klein 2016)*

| | Subsample 1 (all) | | Subsample 1 (non-English learners) | | Subsample 1 (English learners) | |
|---|---|---|---|---|---|---|
| Scoring condition | Cohen's/ Conger's Kappa | Agreement | Cohen's/ Conger's Kappa | Agreement | Cohen's/ Conger's Kappa | Agreement |
| Human | .28 (.11)* | 42.86%*** | .25 (.13)* | 42.03%*** | .31 (.26) | 46.67% |
| Low temp: GPT-4 | .75 (.07)*** | 80.95%*** | .69 (.09)*** | 76.81%*** | NA | NA |
| High temp: GPT-4 | .77 (.09)*** | 82.14%*** | .75 (.10)*** | 81.16%*** | .81 (.19)* | 86.67%** |
| Low temp: GPT-3.5 | .40 (.12)** | 59.26%*** | .37 (.15)* | 59.09%*** | .44 (.20) | 60.00%* |
| High temp: GPT-3.5 | .48 (.10)** | 66.07%*** | .47 (.10)*** | 64.49%*** | .49 (.31) | 73.31%* |

*$p < .05$. **$p < .01$. ***$p < .001$.

**Table A2**

*Human-Human (for Purposes of Comparison, Sample 1) and Human-AI Consistency for All Samples, Unweighted (using kappaetc function, Klein 2016)*

| | Cohen's/ Conger's Kappa (*SE*) | Exact agreement | Agreement within 1 point |
|---|---|---|---|
| Sample 1: Human-Human | .36 (.03)*** | 49.18% | 74% |
| Sample 1: Human-AI | .11 (.02)*** | 20.08% | 76% |
| Sample 2: Human-AI | .01 (.02) | 19.48% | 83% |
| Sample 3: Human-A1 | .22 (.02)*** | 41.52% | 89% |

***$p < .001$.