Comparison of AI and Human Scoring on A Visual Arts Assessment

Ning Jiang, Yue Huang & Jie Chen
Measurement Incorporated
Correspondence: jchen@measinc.com

Abstract

This study investigates the comparability and reliability of scores generated by Artificial Intelligence, specifically a large language model, GPT-4, against scores assigned by trained human raters on a visual arts assessment. Two types of performance tasks, art writing and drawing, were selected from the South Carolina Arts Assessment Program. Responses from 358 fourth-grade students to Task 1 and from 190 students to Task 2 were evaluated independently by both GPT-4 and trained human raters. Both exact and adjacent agreement rates, as well as the Quadratic Weighted Kappa, were examined by task between the two human raters and for GPT-4 versus the first human rater. Additionally, these statistics were compared across tasks to explore whether task characteristics (i.e., text-based vs. drawing-based) contributed to differences in rater agreement. The findings highlight that 1) GPT-4 is more lenient and consistent in grading than human raters for both tasks; 2) the agreement between the human rater and GPT-4 is slightly lower than that between two human raters; and 3) human-GPT-4 scoring agreement remains consistent across visual arts performance tasks. These findings highlight the potential and limitations of using LLMs in arts-based assessment contexts.

1 Introduction

Artificial intelligence (AI) has become increasingly prominent in educational assessment, offering scalable and efficient solutions for scoring student responses. Much of the early work in this area has centered on text-based automated essay scoring (AES), where machine learning and natural language processing techniques have been used to replicate human scoring (Lim et al., 2021). The development of large language models (LLMs), such as OpenAI's GPT-4, Meta's LLaMA, and Google's Gemini, marks a new phase in AI-driven scoring. These models demonstrate advanced language understanding, generative abilities, and even processing both text and image inputs. Consequently, recent studies have explored their application in AES tasks, highlighting their potential for providing consistent and timely scoring. However, a substantial portion of this research remains focused on traditional writing tasks. There is still limited understanding of how LLMs perform in evaluating more complex responses, such as those required in visual and performing arts assessments. Filling this gap is especially important given the subjective and interpretive nature of art-based responses.

In visual arts education, assessing students' abilities to interpret, critique, and create artwork is an essential part of the learning process (Eisner, 1999). Performance tasks in this area often involve written critiques and creative visual outputs, which demand a nuanced understanding of artistic principles, expressive intent, and contextual interpretation. These tasks require subjective judgment and domain-specific expertise, making them challenging to score consistently—even among trained human raters (Perlman, 2003).

While AI technologies have shown promise in cognitive assessments, their application in art assessments remains limited. The rapid development of LLMs introduces new possibilities for scoring complex responses that combine text and imagery. AI scoring systems offer potential benefits such as efficiency, scalability, and reduced rater fatigue (Vetrivel, et

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 147–154

al., 2025). However, concerns persist about their reliability and validity in domains that rely on interpretive, aesthetic, or culturally embedded judgments (Clauser et al., 2014). This study tackles this gap by evaluating the use of GPT-4 to score fourth-grade students' visual arts performance tasks and comparing its performance to that of trained human raters.

By comparing AI-generated scores with those assigned by trained human raters, the study seeks to evaluate the reliability of AI-assisted scores in the context of art education. This research is guided by the following questions:

1. How do scores given by GPT-4 compare to those given by trained human raters on a text-based visual arts performance task?
2. How do scores given by GPT-4 compare to those given by trained human raters on a drawing-based visual arts performance task?
3. How does scoring agreement between GPT-4 and trained human raters differ across tasks (i.e., text- vs. drawing-based tasks)?

2 Data

2.1 Instrument

The South Carolina Arts Assessment Program (SCAAP) was used to evaluate fourth-grade students' visual arts achievement during the 2015–2016 school year. Designed to align with the South Carolina Academic Standards for the Visual and Performing Arts, SCAAP delivers technically sound assessments to students enrolled in schools funded through the Distinguished Arts Program. The assessment consists of 45 multiple-choice items and two performance tasks aimed at measuring students' understanding and application of visual arts concepts.

This study focuses on the two visual arts performance tasks. Each student received a test booklet containing written instructions and designated spaces for completing the tasks. The tasks were administered in group settings by trained test administrators following a standardized manual. Although untimed, each task typically took students at least 30 minutes to complete.

Performance Task 1 (hereafter referred to as Task 1) was designed to assess two key standards:

- Standard 2 – Using structures and functions in visual arts
- Standard 5 – Interpreting works of visual arts

In this task, students were asked to select four vocabulary terms from a provided word bank and write a paragraph for each term, explaining how it applied to the given artwork. Each paragraph was expected to consist of several descriptive sentences that demonstrated the student's understanding of the selected art concepts.

Performance Task 2 (hereafter referred to as Task 2) required students to complete a drawing based on a given prompt and was designed to assess two key standards:

- Standard 1 – Creating art
- Standard 2 – Using structures and functions in visual arts

2.2 Scoring Rubric

Responses to both tasks were scored holistically using a rubric with five proficiency levels, ranging from 0 to 4. Each level clearly described the degree of proficiency demonstrated in students' responses. Raters could assign augmentation scores using "+" or "–" to adjust the base score by 0.33 points. To compute quadratic weighted kappa (QWK), augmentation scores were rounded to the nearest integers, as QWK analysis requires categorical or ordinal-level data.

2.3 Participants

Three student samples were used to address the three research questions, with demographic details provided in Table 1.

- Sample 1 comprised 358 students whose Task 1 responses were double-scored and was used to address Research Question 1 (RQ 1).
- Sample 2 comprised 190 students whose Task 2 responses were double-scored and was used to address Research Question 2 (RQ 2).
- Sample 3 comprised 166 students whose responses to both tasks were double-scored and was used to address Research Question 3 (RQ 3).

3 Methods

	Gender			Race			
Sample	Male (%)	Female (%)	Missing (%)	Black (%)	White (%)	Other (%)	Missing (%)
1	150 (42)	188 (53)	20 (6)	123 (34)	182 (51)	9 (3)	44 (12)
2	73 (38)	101 (53)	16 (8)	71 (37)	85 (45)	18 (9)	16 (8)
3	67 (40)	86 (52)	13 (8)	62 (37)	74 (45)	17 (10)	13 (8)

Note. Percentages may not sum to 100% due to rounding.

Table 1: Student Demographic Summary by Sample.

3.1 Human Scoring

Student responses were scanned and saved as JPEG files and uploaded to the SCAAP web-based rating system for remote scoring. All raters were trained and required to pass a qualifying test before scoring the student work. In 2016, five trained raters were employed to score responses to Tasks 1 and 2. Considering the change of Task 1 during the 2015-2016 school year, all responses to Task 1 were double-rated (i.e., scored by two raters), and about 50% of responses to Task 2 were double-rated. In instances of non-adjacent scores, an expert rater was brought in to provide a third score. For double-rated responses, the final score was calculated as the average of the two raters' scores. If an expert rater's score was needed due to disagreement, that score would be used as the final score instead. However, in the current study, all samples consisted only of double-scored responses, without applying score averaging adjustments.

3.2 GPT-4 Scoring

GPT-4 was used to assess the same set of students' responses to performance tasks. GPT-4 is OpenAI's latest multimodal model that can process and generate text, images, and audio. It offers fast performance, improved reasoning, and seamless handling of multiple input types. The version adopted for scoring in this study is gpt-4o-2024-11-20. All de-identified student responses were scanned and input to LLMs through the model API interaction in Python.

Specifically, we designed a prompt engineering framework to simulate the human scoring process. Each prompt includes the following components: a description of the context, a description of the performance task description and scoring rubric identical to those used in training human raters, the chain-of-thought prompt for scoring the student's response. Every individual response was scored using the same prompt framework to control possible drifts in LLM output.

3.3 Data Analysis

For all samples, descriptive statistics, exact and adjacent agreement rates, and the QWK were calculated by task for Rater 1 (R1) vs. Rater 2 (R2) and R1 vs. GPT-4. R1's ratings served as the reference. As a chance-corrected agreement measure that weighs disagreements based on their severity, QWK offers further insight into the extent of agreement and disagreement between the raters (Cohen, 1968). To address RQ 3, the statistics of rater agreement and the QWK regarding the two different tasks were compared to investigate the impact of task type on the agreement between human and GPT-4 scoring.

In addition, confusion matrices were constructed for each rater pair (R1 vs. R2 and R1 vs. GPT-4) by task and research question to analyze the frequency with which rating categories from one rater corresponded to those of the other. The QWK provides a single summary value that adjusts for chance agreement and the severity of disagreements, while the confusion matrix helps identify where and how raters disagree.

4 Results

Table 2 presents summary statistics of scores assigned by both human raters and GPT-4 for each task, using Sample 1 for Task 1 and Sample 2 for Task 2. Table 4 provides the same statistics based on the common sample (Sample 3). Table 3 reports agreement rates between R1 and R2 and between R1 and GPT-4 by task, while Table 5 presents the corresponding agreement rates based on the common sample. Finally, Table 6 summarizes the differences in agreement levels between Task 1 and Task 2. Figures 1–4 display confusion matrices for Task 1 and Task 2, comparing R1 with R2 and GPT-4 on the 0–4 scale. Figures 5–8 repeat these comparisons using the RQ 3 subsample.

4.1 RQ 1: Task 1 – Art Writing

Results in Table 2 show that GPT-4 produced slightly higher average scores than both human raters and with less variability on Task 1 (N=358). In terms of assigning scores, GPT-4 is the most lenient and consistent (M = 2.37, SD = 0.76), while R1 is the harshest (M = 2.08, SD = 1.31).

For Task 1, the exact agreement between human raters was 34%, with an adjacent

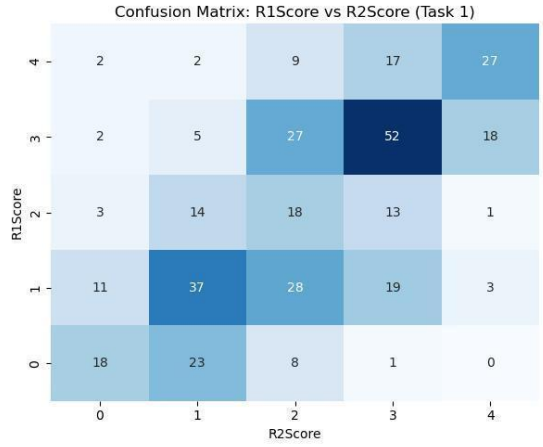

Figure 1: RQ1: Confusion Matrix for Task 1 (Human Rater 1 vs Human Rater 2).

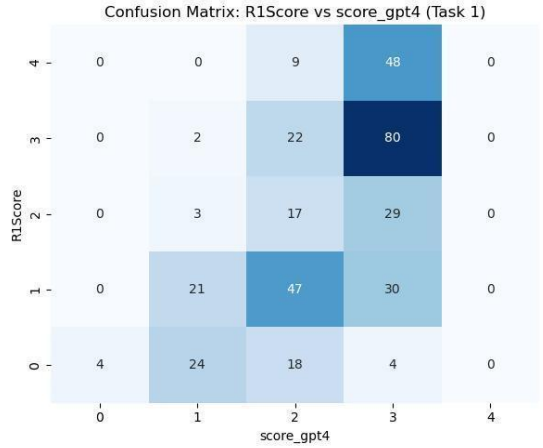

Figure 2: RQ1: Confusion Matrix for Task 1 (Human Rater 1 vs GPT-4).

agreement of 48%, totaling 82% for within-one-point agreement. Comparably, the agreement between GPT-4 and R1 yielded to 28% exact and 54% adjacent, also totaling 82% (see Table 3).

However, the QWK coefficient for R1 vs. GPT-4 (0.49, indicating moderate agreement) was smaller than that for human-human comparison (0.62, indicating substantial agreement), suggesting slightly lower consistency between human- and machine-scoring.

Results of the confusion matrix can visually identify exact/adjacent agreement, observe patterns of disagreement, and detect systematic bias. Figure 1 displays confusion matrix comparing R1 with R2 and Figure 2 displays R1 and GPT-4 on the 0–4 scoring scale. The R1–R2 matrix shows relatively consistent scoring across the full score range, with disagreements mostly concentrated in adjacent categories, suggesting moderate to strong alignment between human raters. In contrast, the R1–GPT-4 matrix shows high agreement primarily around score 3, but GPT-4 avoids assigning extreme scores (0 and 4), resulting in compressed scoring toward the middle. This indicates that while GPT-4 is more consistent with R1 at mid-range scores than R1 and R2 agreement, it demonstrates a conservative bias, especially at the scoring extremes.

Task (N)	Rater	Mean	SD	Min	Max
1 (358)	Rater 1	2.08	1.31	0	4
	Rater 2	2.13	1.20	0	4
	GPT-4	2.37	0.76	0	3
2 (190)	Rater 1	2.19	0.77	0	4
	Rater 2	2.35	0.79	0	4
	GPT-4	2.59	0.55	1	4

Table 2: Descriptive Statistics of Scores by Task and Individual Rater.

Task	Rater Pair	Exact	Adjacent	QWK
1	Rater 1 vs Rater 2	0.34	0.48	0.62
	Rater 1 vs GPT-4	0.28	0.54	0.49
2	Rater 1 vs Rater 2	0.51	0.46	0.61
	Rater 1 vs GPT-4	0.43	0.51	0.44

Table 3: Rater Agreement and Interrater Reliability by Task and Rater Pair.

4.2 RQ 2: Task 2 – Drawing

For the drawing-based performance task (N=190), GPT-4 again yielded slightly higher average scores and a narrower distribution (M = 2.59, SD = 0.55) (see Table 2). As shown in Table 3, GPT-4 exhibited a slightly lower rate of exact agreement with the human rater compared to the human-human agreement (43% vs. 51%). The adjacent + exact agreement remained high for both pairs (97% for the R1-R2 pair vs. 94% for the R1-GPT-4 pair), but the QWK coefficient (0.44) for the R1-GPT-4 pair again fell short of the human-human benchmark (0.61).

Figure 3 and Figure 4 display confusion matrices comparing R1's scores with those from R2 and GPT-4, respectively, on Task 2. The R1–R2 matrix shows strong agreement around score 2 (n=77) but with notable dispersion at higher score levels; for example, many R1 scores of 3 were rated as 2 (n = 15) or 4 (n = 8) by R2. This indicates moderate agreement with some variability, especially at the upper end. In contrast, the R1–GPT-4 matrix shows high

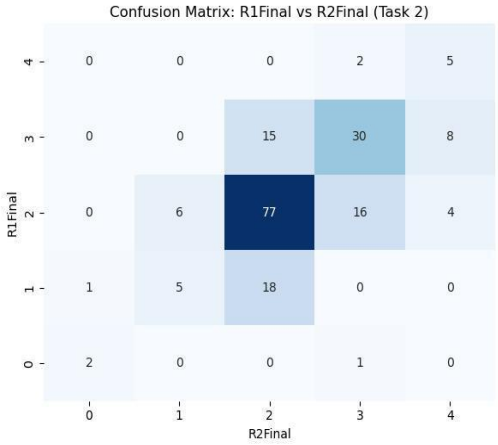

Figure 3: RQ2: Confusion Matrix for Task 2 (Human Rater 1 vs Human Rater 2).

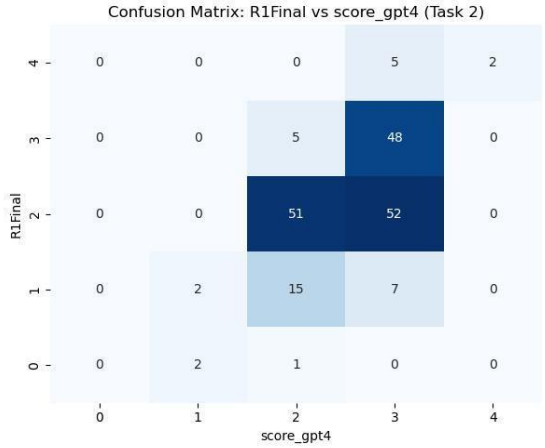

Figure 4: RQ2: Confusion Matrix for Task 2 (Human Rater 1 vs GPT-4).

agreement for scores of 2 and 3 (n = 51 and 48 respectively), with fewer deviations but a visible tendency to avoid the extremes—GPT-4 rarely assigns scores of 0 or 4. These patterns suggest that while both R2 and GPT-4 align with R1 in the mid-range, GPT-4 displays a narrower scoring range with a conservative bias.

4.3 RQ 3: Difference between Task 1 and Task 2

To evaluate differences in scoring agreement across task types, we compared R1–R2 and R1–GPT-4 agreement statistics for both Task 1 and Task 2, using the common sample (N = 166). As shown in Table 4, GPT-4 is again the most lenient and consistent (M = 2.39, SD = 0.73 for Task 1; M = 2.61, SD = 0.54 for Task 2), while R1 is the harshest (M = 2.07, SD = 1.28 Task 1; M = 2.22, SD = 0.77 for Task 2). Additionally, on average, both the human raters and GPT-4 assigned higher scores to responses for Task 2 than to those for Task 1.

As shown in Table 5, for both tasks, the exact

Task	Rater	Mean	SD	Min	Max
1	Human Rater 1	2.07	1.28	0	4
	Human Rater 2	2.20	1.22	0	4
	GPT-4	2.39	0.73	0	3
2	Human Rater 1	2.22	0.77	0	4
	Human Rater 2	2.38	0.77	0	4
	GPT-4	2.61	0.54	1	4

Table 4: Descriptive Statistics of Scores by Task and Individual Rater (Sample 3: N = 166).

Task	Rater Pair	Exact	Adjacent	QWK
1	Rater 1 vs Rater 2	0.34	0.43	0.56
	Rater 1 vs GPT-4	0.27	0.52	0.41
2	Rater 1 vs Rater 2	0.49	0.46	0.57
	Rater 1 vs GPT-4	0.43	0.51	0.43

Table 5: Rater Agreement and Interrater Reliability by Task and Rater Pair (Sample 3: N = 166).

agreement between R1 and GPT-4 was lower than that between human raters (27% vs. 34% for Task 1; 43% vs. 49% for Task 2), while the adjacent agreement between R1 and GPT-4 was higher than that between human raters (52% vs. 43% for Task 1; 51% vs. 46% for Task 2). Overall, the exact and adjacent agreement rates were higher for Task 2 than Task 1, with R1–R2 increasing from 77% to 95% and R1–GPT-4 from 79% to 94%. In contrast, QWK values remained relatively stable across tasks, rising only slightly from 0.56 for Task1 to 0.57 for Task 2 for R1–R2 and from 0.41 to 0.43 for R1–GPT-4. The stable QWK coefficients for both the R1-R2 and the R1-GPT-4 pairs across tasks suggest overall reliability in scoring despite differences in task type.

Difference across tasks	Exact	Adjacent	QWK
Rater 1 vs. Rater 2	-0.15	-0.03	-0.01
Rater 1 vs. GPT-4	-0.16	0.01	-0.02

Table 6: Difference in Rater Agreement and Interrater Reliability Across Tasks (N = 166).

As observed in Table 6, for the R1–R2 pair from Task 1 to Task 2, the exact agreement increased by 15%, and the adjacent agreement increased by 3%. However, the comparison of

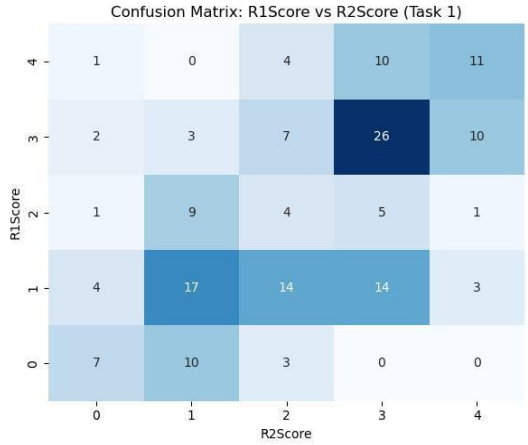

Figure 5: RQ3: Confusion Matrix for Task 1 (Human Rater 1 vs Human Rater 2).

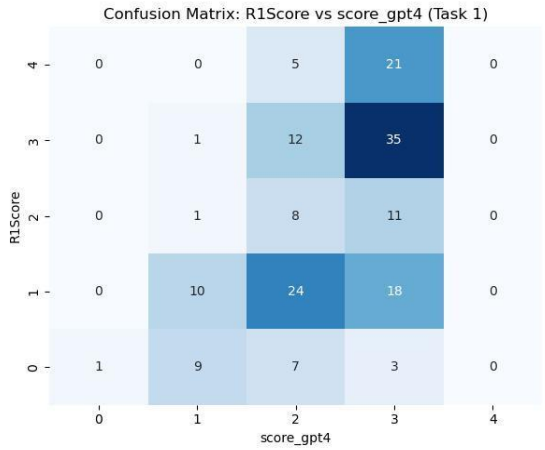

Figure 6: RQ3: Confusion Matrix for Task 1 (Human Rater 1 vs GPT-4).

QWK statistics indicates nearly no difference in agreement between the two tasks (1%). For the R1–GPT-4 pair from Task 1 to Task 2, the exact agreement increased by 16, and the adjacent agreement decreased by 1%. The comparison of QWK statistics shows a slight increase in agreement (2%) from Task 1 to Task. While the QWK coefficients remained largely stable, the increases in exact agreement indicate more consistent scoring on the drawing task.

We also generated confusion matrices for all scoring comparisons and tasks based on Sample 3, comparing R1's scores with those from R2 and GPT-4 on Task 1 (see Figures 5 and 6). The patterns are similar to those results in RQ 1. The R1–R2 matrix indicates moderate agreement, with the highest concentration along the diagonal, particularly at score 3 (n=26). However, off-diagonal cells reveal notable adjacent and distant mismatches, such as R1 assigning a point of 4 when R2 gave a 1 or 2, suggesting some inconsistency at higher scores. In contrast, the R1–GPT-4 matrix shows a narrower distribution, with GPT-4 scores concentrated at 2 and 3 and no scores assigned at the extremes (0 or 4). Although GPT-4 shows alignment with R1 in the mid-range (e.g., R1 = 3 most often matched GPT-4 = 3), it avoids the use of the full score scale, pointing to a compressed and conservative scoring tendency.

Figures 7 and 8 show confusion matrices comparing R1's scores with those from R2 and GPT-4 on Task 2 using the subsample for RQ 3. The patterns are similar to those results in RQ 2. The R1–R2 matrix shows strong agreement at score 2 (n = 62) and moderate alignment at score 3 (n = 26), though some off-diagonal variation appears—particularly when R1 assigned a 3 but R2 gave a 2 or 4, suggesting some upper-end disagreement. The R1–GPT-4 matrix reveals a narrower scoring pattern, with GPT-4 clustering scores tightly around 2 and 3, and avoiding extreme values (0 and 4). While GPT-4 shows high agreement with R1 in the mid-range (e.g., R1 = 2 and GPT-4 = 2 or 3), its reluctance to assign the highest score may reflect a conservative or compressed scoring pattern.

There are some noticeable differences between Task 1 and Task 2 in both R1–R2 and R1–GPT-4 scoring patterns:

- Human-human consistency: in both tasks, R1 and R2 show the strongest agreement at score 2, but Task 2 generally shows tighter clustering along the diagonal—especially at scores 2 and 3, suggesting slightly higher inter-rater consistency than Task 1. In Task 1, disagreements are more spread out, including more extreme mismatches (e.g., R1 = 4 vs R2 = 0 or 1), whereas in Task 2, disagreements tend to stay within adjacent scores.
- GPT-4 scoring behavior: GPT-4 is conservative across both tasks, avoiding extreme scores, but this effect is more pronounced in Task 2. In Task 2, GPT-4

rarely assigns a score of 4 and leans heavily toward scores of 2 and 3, even when R1 gave higher scores. In Task 1, while GPT-4 still avoids extremes, the spread is slightly broader, especially around score 3.

In general, GPT-4 is capable of approximating human scoring with high adjacent agreement (>0.5) in both text-based and drawing-based tasks. However, the QWK statistics for the R1–GPT-4

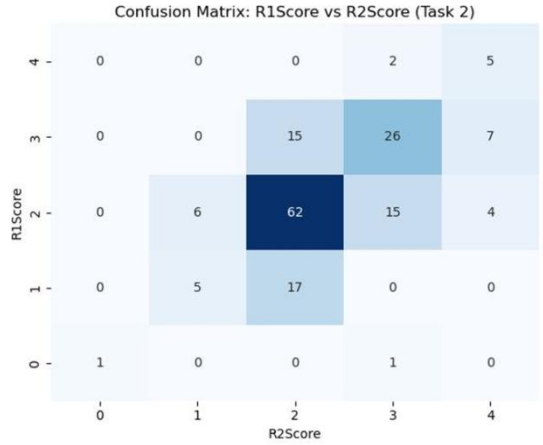

Figure 7: RQ3: Confusion Matrix for Task 2 (Human Rater 1 vs Human Rater 2).

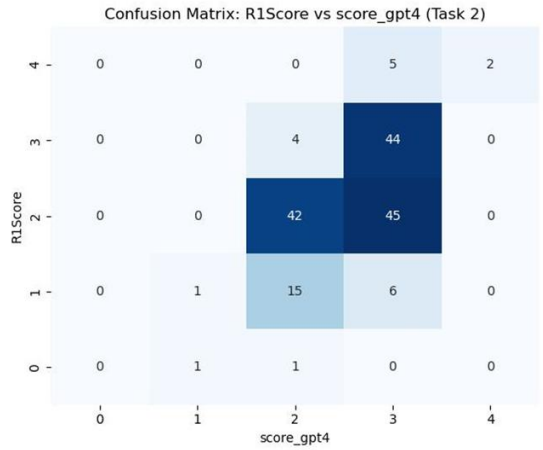

Figure 8: RQ3: Confusion Matrix for Task 2 (Human Rater 1 vs GPT-4).

pair indicate that GPT-4 does not replicate the full scoring pattern of human raters, particularly at extremes (i.e., 0 or 4). The agreement patterns between tasks are similar, but Task 2 (drawing) shows slightly stronger alignment in both adjacent and overall agreement, possibly due to fewer extreme scores in that task.

5 Discussion

In the text-based task, GPT-4 achieved comparable adjacent agreement with human raters but demonstrated a narrower scoring range and a reluctance to assign extreme scores. This centralizing tendency results in lower QWK values, reflecting a more conservative scoring pattern. In the drawing-based task, although adjacent agreement remained high, GPT-4's performance was slightly weaker, particularly in its interpretation of complex visual elements such as depth, composition, and emotional nuance.

Furthermore, the observed patterns in the confusion matrices suggest that while human raters showed moderate consistency across the full score range, discrepancies, particularly at higher scores, highlighted the subjective nature of human scoring. In contrast, GPT-4 demonstrated strong alignment with human raters in mid-range scores but consistently avoided extreme ratings, indicating a conservative bias. These findings imply that while GPT-4 may be a reliable tool for scoring typical responses, caution is needed when using it to evaluate very high- or low-quality work, as it may underrepresent performance extremes and affect decisions tied to those score ranges.

A consistent trend across both tasks is that GPT-4 tended to avoid assigning the lowest and highest score points, which has implications for high-stakes assessments where performance extremes are critical. Additionally, human raters showed greater variability in scoring, especially at higher performance levels, which may reflect the inherent subjectivity in assessing creative work—a dimension GPT-4 is currently limited in replicating.

These findings suggest that while GPT-4 can serve as a reliable supplemental tool in scoring student work, it should not yet be considered a full substitute for human judgment in art assessment. The performance of GPT-4 was task-sensitive, further emphasizing the need for content-specific prompt tuning and calibration.

6 Limitations

While this study provides valuable insights into the comparability of GPT-4 and human scoring in visual arts assessments, several limitations should be acknowledged. First, the analysis focused on fourth-grade student responses from a single

assessment year and context (SCAAP 2016), which may limit the generalizability of the findings to other grade levels, content areas, and assessment frameworks. Second, the scoring outcomes reflect a single implementation of GPT-4 using a specific prompt engineering approach; different prompt designs or fine-tuned model configurations may produce different results. Third, although GPT-4 was instructed to incorporate augmentation scores (e.g., +0.33 or –0.33), it did not apply this rule during scoring. This unexpected behavior highlights a potentially important limitation in the model's ability to consistently follow nuanced scoring rules, and future research should investigate the causes and implications of this issue. Finally, human-human agreement was not high, and only R1 was randomly selected as the baseline rater for all comparisons, which may constrain generalizability. One possible reason for the moderate human-human agreement is that Task 1 was new and first administered in 2016. In addition, Task 2's prompt was changed in 2015, and 2016 was only the second year it was used. Therefore, the raters were not yet sufficiently familiar with the new tasks and the rubric. However, this study is strengthened by the inclusion of cross-comparisons across two tasks and repeated analyses on multiple subsamples, thereby providing a multidimensional assessment of scoring consistency from diverse perspectives.

7 Future Research

Future research should explore the reasoning behind GPT-4's scoring decisions through qualitative content analysis. Specifically, analyzing GPT-4's rationale in comparison to the scoring rubric and underlying pedagogical goals may help illuminate how well the model interprets key assessment criteria. Additionally, it would be valuable to examine why GPT-4 consistently avoided assigning extreme scores—whether due to probabilistic constraints in its language modeling, uncertainty in interpreting creative responses, or an overly cautious alignment with prompt wording. Understanding these issues may help refine prompt engineering or model tuning for better alignment with human evaluative standards.

References

Eisner, E. W. (1999). The national assessment in the visual arts. Arts Education Policy Review, 100(6), 16-20.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2014). Validity issues for performance-based tests scored with computer-automated scoring systems. In Advances in Computerized Scoring of Complex Item Formats (pp. 413-432). Routledge.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220. https://doi.org/10.1037/h0026256.

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A Comprehensive Review of Automated Essay Scoring (AES) Research and Development. Pertanika Journal of Science and Technology, 29(3). https://doi.org/10.47836/pjst.29.3.27.

Perlman, C. C. (2003). Performance Assessment: Designing Appropriate Performance Tasks and Scoring Rubrics.

Vetrivel, S. C., Arun, V. P., Ambikapathi, R., & Saravanan, T. P. (2025). Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments. In *Adopting Artificial Intelligence Tools in Higher Education* (pp. 21-61). CRC Press.