Evaluating the Reliability and Agreement of Rubric-Guided LLM Scoring Versus Human Grading Across Three University Courses

Howard Kim , Sung-Tae Lee , Jongwon Lee *

Posted Date: 19 May 2026

doi: 10.20944/preprints202605.1284.v1

Keywords: large language models; rubric-based grading; educational assessment; human–AI agreement; intraclass correlation coefficient; Bland–Altman analysis; text-response scoring; co-grading

 doi:10.20944/preprints202605.1284.v1

Article

Evaluating the Reliability and Agreement of Rubric-Guided LLM Scoring Versus Human Grading Across Three University Courses

Howard Kim [1,2], Sung-Tae Lee [3] and Jongwon Lee [3,*]

[1] Department of AI Creation, College of AI Convergence, Seoul Cyber University, Seoul 01133, Republic of Korea
[2] Department of Management of Technology, General Graduate School, Sungkyunkwan University, Suwon 16419, Republic of Korea
[3] Department of Computer Science and Engineering, Seoul Cyber University, Seoul 01133, Republic of Korea
* Correspondence: jlee@iscu.ac.kr

Abstract

This study examines whether a rubric-guided large language model (LLM) can approximate local human grading practice for text-based responses in three university courses. A total of 930 student responses from Prompt Engineering, Photoshop Design, and AI Video Production were scored by two human instructors and by ChatGPT using the same five-criterion analytic rubric (Accuracy, Logical Flow, Specificity, Quality, and Originality; 0.0–3.0 each; Total 0–15). Human consensus (HC) was defined as the mean of the two human scores and was treated as a pragmatic reference rather than ground truth. Pairwise agreement among H1, H2, AI, and HC was evaluated using ICC(3,1), Pearson correlations, mean absolute error (MAE), and Bland–Altman bias and limits of agreement (LoA); a course-specific held-out calibration analysis was additionally conducted. On the Total score, human–human agreement was strong (ICC = 0.819 [0.797, 0.839]). AI–H1 and AI–H2 Total-score agreement were ICC = 0.700 [0.666, 0.732] and 0.767 [0.739, 0.792], respectively, while AI–HC agreement was ICC = 0.763 [0.735, 0.789], with MAE = 1.603 and LoA = [−4.246, 4.045]. At the trait level, AI–HC ICCs exceeded H1–H2 ICCs for all five rubric dimensions, although Quality remained weakly defined in the human baseline. On a 70/30 held-out test split, a course-specific linear calibration modestly improved Total-score ICC from 0.774 to 0.782 and reduced MAE from 1.624 to 1.215, narrowing the LoA from [−4.290, 4.188] to [−3.157, 3.329]. However, threshold-adjacent agreement remained imperfect after calibration. The findings concern written responses only and support a conservative conclusion: rubric-guided LLM scoring can assist human grading under fixed local rubrics, but the current evidence supports calibrated human–AI co-grading rather than unsupervised replacement.

Keywords: large language models; rubric-based grading; educational assessment; human–AI agreement; intraclass correlation coefficient; Bland–Altman analysis; text-response scoring; co-grading

1. Introduction

Grading open-ended student work consistently and fairly remains a persistent challenge in higher education. Analytic rubrics can improve transparency, promote alignment between learning outcomes and assessment, and make feedback more actionable. Even so, inter-rater variability is common, particularly when criteria involve holistic judgment, creativity, or perceived quality. Reliability research therefore recommends explicit model choice and careful interpretation when intraclass correlation coefficients (ICC) are used to support claims about scoring consistency [1–3].

Automated essay scoring (AES) systems predate large language models (LLMs) by decades. Classical systems such as e-rater demonstrated meaningful human–machine associations under constrained prompts and operationally standardized settings, while also highlighting persistent issues of construct coverage, transparency, and governance [4,5]. The recent rise of LLMs has renewed interest

 doi:10.20944/preprints202605.1284.v1

in rubric-guided scoring because models can respond directly to natural-language criteria rather than only to hand-engineered features. Early evidence suggests that LLMs can align reasonably well with human raters when rubrics are explicit and prompt structure is stable, but published studies also warn that agreement may deteriorate without calibration and that case-level substitution often remains difficult [6–10].

For this reason, agreement-oriented evaluation is more informative than correlation alone. Correlation indicates whether higher human scores tend to correspond to higher AI scores, but it does not show whether the two methods are close enough to be operationally interchangeable. ICC provides a reliability-oriented summary of agreement under a specified rater design, whereas Bland–Altman analysis makes systematic bias and the likely range of pairwise differences visible on the original score scale [1,11]. In studies of AI-assisted grading, design fairness is also critical: agreement with the mean of two human raters can appear stronger than agreement with a single human because averaging reduces random error.

The present study investigates whether a rubric-guided LLM can approximate human scoring of text-based responses collected in three university courses at one institution, rather than whether an LLM can generally replace grading in higher education. This distinction is important because the dataset includes written answers about Photoshop workflows, prompt-design principles, and AI-video tools rather than the students' final design or multimedia artifacts. The paper should therefore be read as a study of agreement on rubric scoring of written responses under a fixed scoring prompt.

Four research questions are addressed. *RQ1*: How do AI–H1, AI–H2, and AI–HC agreement compare with the observed human–human baseline across rubric traits and the Total score? *RQ2*: How strongly are AI and human scores associated overall and by course? *RQ3*: What kinds of systematic bias, individual-level disagreement, and threshold-adjacent error remain after aggregate agreement is taken into account? *RQ4*: How much can simple course-specific held-out calibration improve operational fit, and what conservative recommendations for human–AI co-grading follow from these results?

The contributions of the study are fourfold. First, a multi-metric evaluation of rubric-guided LLM scoring is provided using agreement statistics that are directly relevant to educational deployment: ICC, Pearson association, MAE, and Bland–Altman bias and limits of agreement. Second, AI performance is interpreted against the observed human–human baseline using not only AI–HC but also AI–H1 and AI–H2 comparisons, thereby making the averaging effect of HC explicit. Third, threshold-oriented operational checks and a held-out course-specific calibration analysis are reported rather than leaving calibration as a purely future concern. Fourth, the empirical findings are translated into a conservative workflow proposal for calibrated human–AI co-grading in text-response settings, emphasizing oversight for low-stability traits and decision-critical cases. The remainder of the paper reviews related work, summarizes the study design and analysis strategy, reports the empirical results, and discusses implications, limitations, ethics, and data availability.

2. Related Work

2.1. Automated Scoring Before LLMs

Automated essay scoring has a long history in educational measurement. Earlier AES systems showed that machine-generated scores could achieve useful correlations with human ratings in constrained writing tasks, particularly when prompts, feature spaces, and scoring rubrics were tightly controlled [4,5]. At the same time, this literature consistently emphasized that operational success depends on clear construct definition, validation against local scoring practice, and human governance over use. These concerns remain relevant even though contemporary LLM-based scorers differ substantially from older feature-engineered systems.

 doi:10.20944/preprints202605.1284.v1

2.2. Rubric-Guided LLM Scoring

LLMs make rubric-guided scoring appealing because they can read natural-language instructions, attend to multiple criteria simultaneously, and generate structured outputs without bespoke feature engineering. Recent educational studies report encouraging alignment between LLM and human ratings under explicit scoring instructions [6–8]. However, domain-specific investigations also caution that LLM scorers can display systematic over- or under-scoring, limited individual-level agreement, and sensitivity to how constructs are operationalized in the rubric [9,10]. The implication is not that LLM scoring is unusable, but that it requires context-specific validation rather than broad generalization.

2.3. Agreement, Validity, and Human–AI Support

Methodologically, the distinction between association and agreement is especially important in AI-assisted assessment. A system may preserve rank order well enough to produce a strong correlation while still disagreeing with human raters by operationally meaningful margins. ICC and Bland–Altman analyses are therefore better aligned with the question of whether AI scores are close enough to human scores for practical use [1,11]. In addition, prior work on hybrid assessment workflows recommends first-pass AI scoring, discrepancy-triggered review, and explicit governance policies instead of unconstrained automation [12–14].

2.4. Positioning of the Present Study

Within this literature, the present study contributes a conservative agreement analysis of rubric-guided LLM scoring on 930 de-identified university responses drawn from three courses at one institution. The study is narrower than many public discussions of AI grading: it evaluates agreement on written responses under a fixed rubric, not full automation of higher-education grading across diverse artifacts, disciplines, or institutional contexts. This narrower positioning aligns the study's claims with the evidence provided by the dataset and analyses.

3. Materials and Methods

3.1. Study Context and Dataset

The dataset comprised 930 student assessment records from three university courses: Photoshop Design ($n = 421$), Prompt Engineering ($n = 405$), and AI Video Production ($n = 104$). The integrated score structure was formed by matching student records across two human-scoring spreadsheets and one AI-scoring spreadsheet by course sheet and student identifier. Each matched record included five rubric scores and a Total score from each scorer.

The three courses involved different text-response contexts. Prompt Engineering responses asked students to explain principles or design strategies for generative-AI prompting. Photoshop Design responses required students to describe production workflows, technical decisions, and design-quality considerations for visual tasks. AI Video Production responses focused on features, uses, and workflow reasoning related to AI-assisted video creation tools. The broader curricular context of these courses is described in Kim *et al.* [15].

A common analytic rubric was used across the dataset. The rubric dimensions were Accuracy, Logical Flow, Specificity, Quality, and Originality, each scored from 0.0 to 3.0 with one-decimal precision. The Total score ranged from 0 to 15. A crucial boundary condition is that the evaluated input for the present study was the student's *written response*. In the Photoshop Design and AI Video Production courses, the AI grader did not inspect the students' final visual or multimedia artifacts. The findings should therefore be interpreted as evidence about text-response scoring, not as validation of AI grading for authentic multimedia products.

For reproducible case matching, the unit of analysis was one unique course-by-student case. Records were matched across the two human workbooks and the AI workbook by course identifier and student identifier. When duplicate rows existed within a course, the submitted row was retained; if no submitted row existed, the non-submission row was retained as a zero-score case according to

 doi:10.20944/preprints202605.1284.v1

the workbook convention. In the AI workbook, four non-submission rows with blank rubric scores were set to zero, and one malformed Total entry was reconstructed from the five component scores. After cleaning, the final matched dataset contained 930 cases with no missing analytic score cells.

As an additional integrity check, the recorded AI Total score was compared with the arithmetic sum of the five AI rubric traits. In 74 of 930 cases (8.0%), the recorded AI Total did not equal the sum of the component scores. A sensitivity analysis using recomputed Totals produced nearly identical Total-score conclusions (AI–HC ICC = 0.763 using the recorded Total versus 0.760 using the recomputed Total), so the substantive interpretation of the study was unchanged.

3.2. Human and AI Scoring Procedure

Two course instructors, who were not authors of this study, independently scored each response using the same rubric. For analysis, human consensus (HC) was defined as the arithmetic mean of the two human scores for each trait and for the Total score. HC was treated as a pragmatic reference rather than a gold-standard label. This distinction matters because averaging two human raters reduces random error and can make AI agreement appear stronger than comparison with a single rater would suggest.

The AI scorer was ChatGPT (o4-mini-high), used with a fixed prompt template and stable inference settings. Scoring was conducted under deterministic settings, including a temperature of zero, and the model was instructed to output numeric rubric scores aligned with the common analytic rubric. The scoring protocol did not report subject-specific few-shot exemplars or fine-tuning procedures. The fixed-prompt design improved procedural consistency, but the present study does not establish robustness to prompt edits, model changes, or alternative output constraints.

3.3. Analysis Framework

The analysis centered on four complementary metrics: ICC, Pearson correlation, MAE, and Bland–Altman mean bias with 95% limits of agreement (LoA). Human–human agreement (H1–H2) served as the local baseline, and AI–HC performance was evaluated using the same framework. Course-level Total-score analyses were used to identify domain-dependent changes in agreement.

Agreement and error were defined on the observed score scale. MAE was computed as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|AI_i - HC_i|, \tag{1}$$

and Bland–Altman summaries were reported using

$$\Delta = \frac{1}{N}\sum_{i=1}^{N}(AI_i - HC_i), \qquad \mathrm{LoA} = \Delta \pm 1.96\,SD(AI - HC). \tag{2}$$

Here, positive bias indicates that the AI scored higher than HC on average, and negative bias indicates lower AI scores. Pearson correlations are reported as supporting descriptors of association rather than as substitutes for agreement. Confidence intervals for the Pearson coefficients in the main tables were computed using Fisher's z-transformation.

The ICC results are interpreted as agreement-oriented summary indices following standard reporting guidance [1–3]. In the revised analysis, four pairings were evaluated on the observed score scale: H1–H2, AI–H1, AI–H2, and AI–HC. For each pairing, ICC(3,1) with 95% confidence intervals, Pearson correlation, MAE, mean bias, and 95% Bland–Altman limits of agreement are reported.

Because aggregate agreement does not guarantee operational interchangeability, threshold-level consistency was additionally evaluated on the Total score at 9.0 and 12.0 points using overall agreement and adjacent-zone agreement, with the adjacent zone defined as cases within ± 1.5 points of the HC threshold. To examine whether simple post-hoc correction could reduce systematic course-dependent bias, a locked hold-out calibration analysis was also conducted. Within each course, cases were split

 doi:10.20944/preprints202605.1284.v1

into 70% development and 30% test subsets, a linear calibration model of the form $HC = a + b \times AI$ was estimated on the development subset, and the fixed parameters were then applied unchanged to the hold-out test subset.

3.4. Interpretive Guardrails

Three interpretive guardrails are important for this study. First, AI–HC agreement is not equivalent to agreement with an external gold standard. Second, the study concerns written responses only and does not validate AI grading of final creative artifacts. Third, even when aggregate agreement appears strong, individual-level interchangeability must still be evaluated against the observed error range. These guardrails shape how the results are interpreted in the following sections. Figure 1 summarizes the overall study design, including data sources, raters, the analytic rubric, and the agreement-oriented analysis framework.

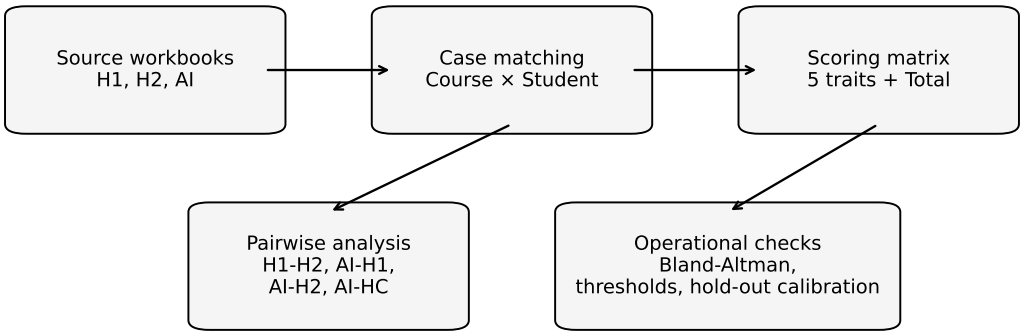

Figure 1. Overview of study design: data sources, raters, rubric structure, and analysis framework.

4. Results

4.1. Human–Human Reliability Baseline

The two human instructors showed strong agreement on the Total score (ICC = 0.819 [0.797, 0.839], $r = 0.832$, MAE = 1.353), indicating that the aggregate score was comparatively stable in the local grading process. At the trait level, human–human agreement was strongest for Logical Flow (ICC = 0.676), followed by Specificity (0.634), Accuracy (0.583), and Originality (0.486), whereas Quality remained effectively unreliable (ICC = −0.042). This revised pattern indicates that the local human baseline was stable for the Total score and several analytic traits, but not for Quality.

In practical terms, the human baseline now supports a more differentiated interpretation than in the earlier draft. The Total score remains the most defensible aggregate outcome, and several rubric traits show moderate agreement between human raters, but Quality remains too weakly constrained to support strong interchangeability claims.

4.2. Overall Pairwise Agreement and Association

Relative to the two individual human raters, the AI showed asymmetrical agreement on the Total score (Table 1). AI–H1 agreement was ICC = 0.700 [0.666, 0.732] with MAE = 1.815, whereas AI–H2 agreement was higher at ICC = 0.767 [0.739, 0.792] with MAE = 1.642. Against the averaged human reference, AI–HC Total-score agreement was ICC = 0.763 [0.735, 0.789], $r = 0.789$, MAE = 1.603, mean bias = −0.101, and 95% LoA = [−4.246, 4.045]. This pattern indicates that the AI was closer to one human rater than the other and that part of the apparent AI–HC agreement reflects the variance-reducing effect of score averaging in HC.

 doi:10.20944/preprints202605.1284.v1

At the trait level, AI–HC ICCs exceeded the corresponding H1–H2 ICCs for all five rubric dimensions (Table 2; Figure 2): Accuracy (0.724 vs. 0.583), Logical Flow (0.682 vs. 0.676), Specificity (0.768 vs. 0.634), Quality (0.537 vs. −0.042), and Originality (0.515 vs. 0.486). Pearson correlations told a similar but not identical story: alignment was strongest for the Total score ($r = 0.789$), followed by Specificity ($r = 0.782$) and Accuracy ($r = 0.735$). Logical Flow was somewhat lower ($r = 0.685$), while Quality ($r = 0.553$) and Originality ($r = 0.537$) remained the weakest dimensions. Trait-level pairwise analyses also showed that the AI was generally closer to H2 than H1 for Accuracy, Logical Flow, Specificity, Originality, and the Total score, whereas Quality showed the reverse pattern. These results reinforce the need to report AI–H1 and AI–H2 separately rather than relying on AI–HC alone.

Table 1. Overall Total-score agreement across pairwise comparisons.

Pair	ICC(3,1) [95% CI]	Pearson r [95% CI]	MAE	Mean Bias	95% LoA
H1–H2	0.819 [0.797, 0.839]	0.832 [0.812, 0.851]	1.353	−0.651	[−3.929, 2.627]
AI–H1	0.700 [0.666, 0.732]	0.734 [0.703, 0.762]	1.815	+0.225	[−4.356, 4.806]
AI–H2	0.767 [0.739, 0.792]	0.774 [0.746, 0.798]	1.642	−0.426	[−4.757, 3.905]
AI–HC	0.763 [0.735, 0.789]	0.789 [0.764, 0.812]	1.603	−0.101	[−4.246, 4.045]

HC denotes the arithmetic mean of the two human raters. Positive mean bias indicates that AI scored higher than the comparator on average. LoA denotes the 95% Bland–Altman limits of agreement. Pearson confidence intervals were computed using Fisher's z-transformation.

Table 2. Trait-level agreement between the human baseline (H1–H2) and AI versus human consensus (AI–HC).

Trait	H1–H2 ICC [95% CI]	AI–HC ICC [95% CI]	AI–HC Pearson r [95% CI]	MAE	Mean Bias
Accuracy	0.583 [0.539, 0.624]	0.724 [0.691, 0.753]	0.735 [0.704, 0.763]	0.387	+0.072
Logical Flow	0.676 [0.639, 0.709]	0.682 [0.646, 0.715]	0.685 [0.649, 0.717]	0.384	+0.008
Specificity	0.634 [0.594, 0.671]	0.768 [0.740, 0.793]	0.782 [0.755, 0.806]	0.355	−0.151
Quality	−0.042 [−0.106, 0.022]	0.537 [0.489, 0.581]	0.553 [0.507, 0.597]	0.525	+0.198
Originality	0.486 [0.435, 0.533]	0.515 [0.466, 0.561]	0.537 [0.489, 0.581]	0.465	−0.207

Full pairwise trait-level results for AI–H1 and AI–H2 are available from the authors on request. The Pearson 95% confidence intervals were computed using Fisher's z-transformation with $N = 930$. The numerical coincidence between the Quality ICC 95% CI ([0.489, 0.581]) and the Originality Pearson r 95% CI ([0.489, 0.581]) reflects the shared point estimate of 0.537 across the two statistics under a large sample size; the values were independently computed and verified.

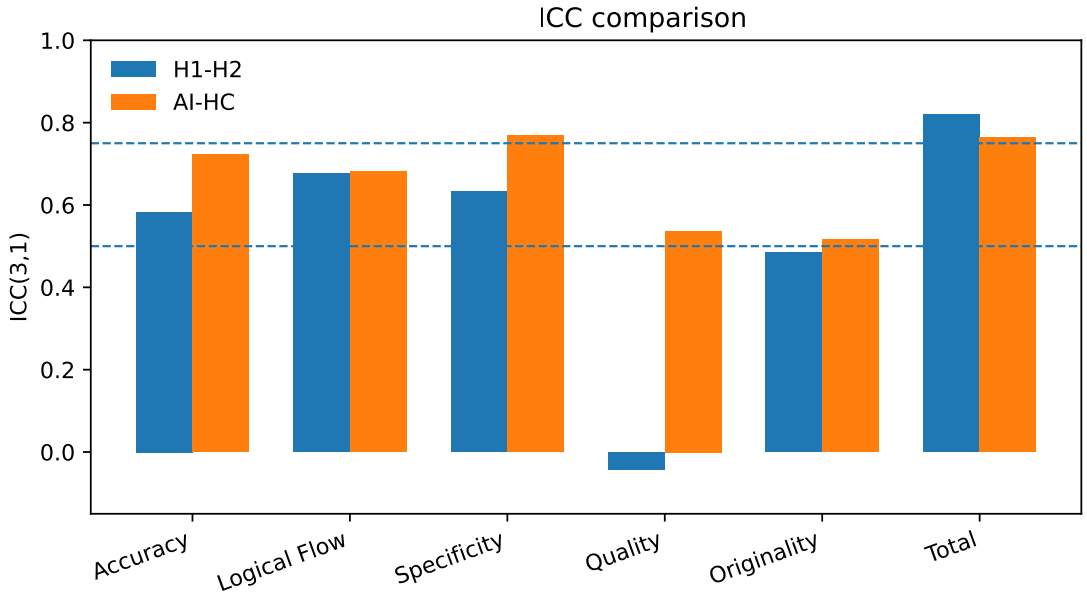

Figure 2. ICC(3,1) comparison between the revised human–human baseline (H1–H2) and AI–human consensus (AI–HC) across rubric traits and the Total score. Dashed lines indicate conventional thresholds (Good = 0.75; Moderate = 0.50).

 doi:10.20944/preprints202605.1284.v1

4.3. Bias, Error Magnitude, and Individual-Level Disagreement

Table 1 shows that good aggregate agreement did not eliminate case-level disagreement. On the 0–15 Total scale, AI–HC MAE was 1.603 and the 95% LoA were [−4.246, 4.045]. This means that although the AI tracked the overall human scoring pattern, the AI and HC could still differ by approximately four points in either direction for individual responses. If an institution were to treat a discrepancy of roughly ±1.5 points as an acceptable operational band on the 15-point scale, the observed Total-score LoA would still substantially exceed that tolerance. Figure 3 visualizes this Total-score Bland–Altman pattern, with the dashed lines indicating the 95% LoA and the dash-dot line indicating the proportional-bias regression.

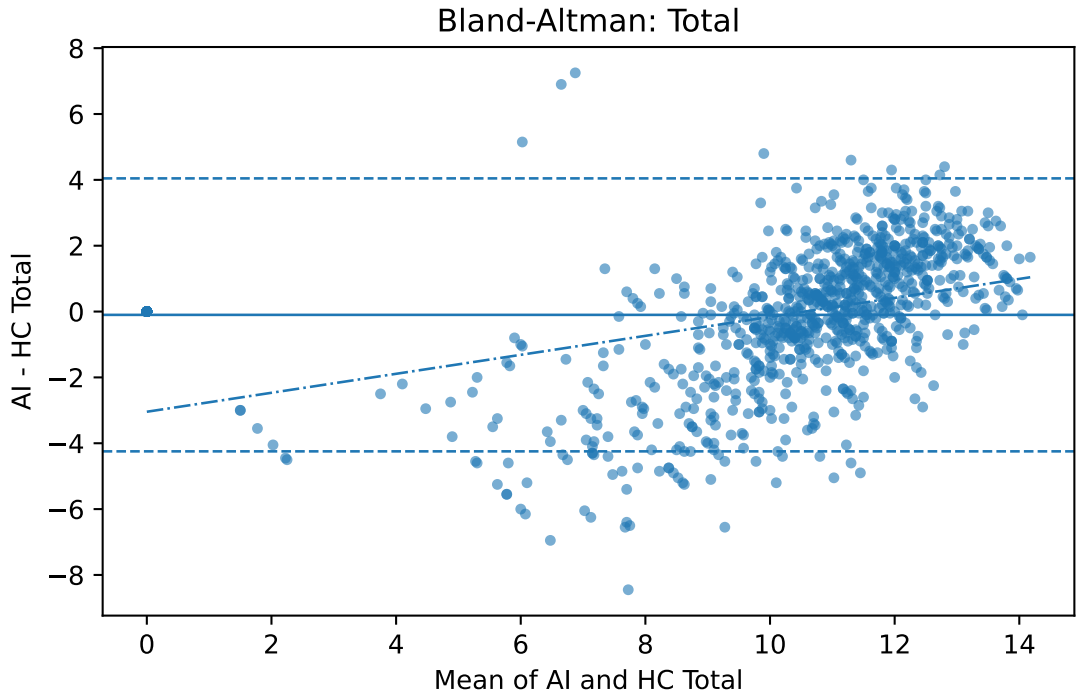

Figure 3. Bland–Altman plot for the AI–HC Total score (0–15). Solid line: mean bias (−0.101); dashed lines: 95% limits of agreement [−4.25, 4.04]; dash-dot: proportional-bias regression.

Trait-level disagreement was narrower but still meaningful. Accuracy showed only a small positive bias (+0.072), Logical Flow was essentially unbiased (+0.008), Specificity showed a modest negative bias (−0.151), Quality a modest positive bias (+0.198), and Originality the largest negative bias among the rubric traits (−0.207). The LoA for rubric traits were roughly on the order of one point in each direction on a 0–3 scale. In other words, the AI's average bias by trait was small, but the potential difference for an individual response could still represent a substantial proportion of the rubric range. Figure 4 presents the trait-level Bland–Altman plots that correspond to these biases and LoA values.

An additional descriptive analysis indicated a moderate positive proportional-bias pattern for the Total score (difference–mean correlation ≈ 0.393). Interpreted conservatively, this suggests that the AI tended to be relatively harsher at the lower end of the Total scale and relatively more generous at the higher end. Because Table 1 emphasizes the primary agreement metrics, this pattern is treated as descriptive support for prospective calibration rather than as a stand-alone central claim.

 doi:10.20944/preprints202605.1284.v1

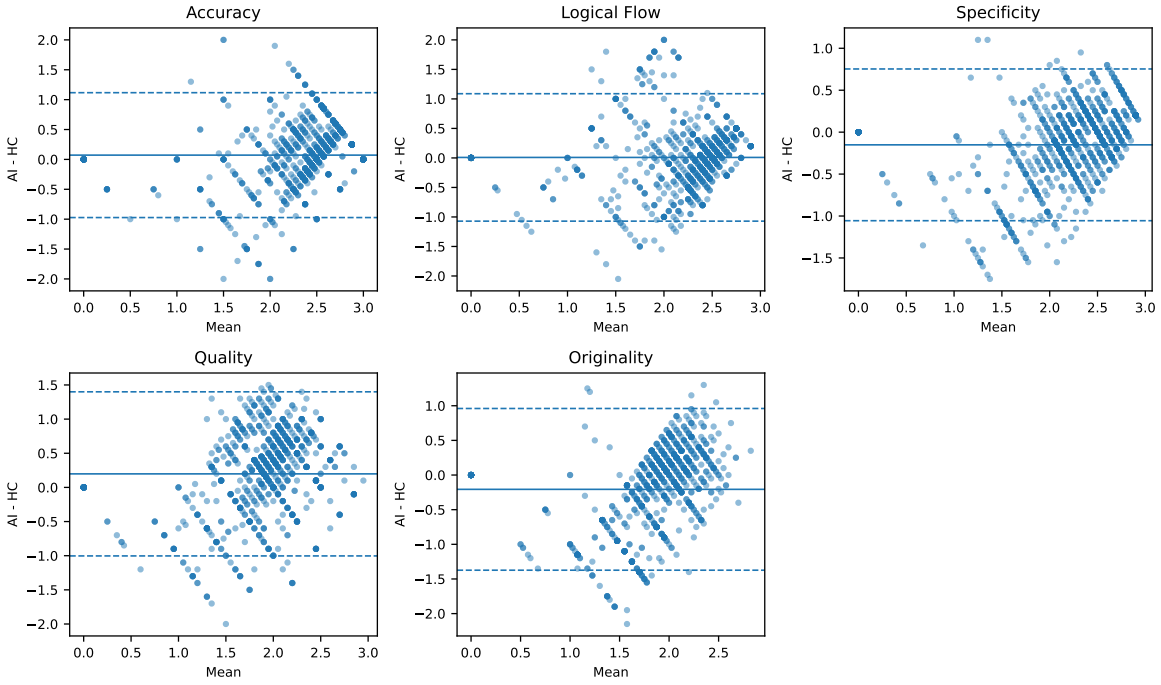

Figure 4. Bland–Altman plots by rubric trait: (**a**) Accuracy, (**b**) Logical Flow, (**c**) Specificity, (**d**) Quality, (**e**) Originality. Each panel shows mean bias (solid line) and 95% LoA (dashed lines).

4.4. Held-Out Calibration and Threshold Agreement

Because the Bland–Altman results indicated operationally meaningful case-level error, whether simple post-hoc calibration could improve out-of-sample performance was tested. Table 3 summarizes the three calibration models evaluated on the hold-out test set. In the pooled hold-out test set, uncalibrated AI–HC Total-score agreement was ICC = 0.774 [0.722, 0.817], $r = 0.799$, MAE = 1.624, mean bias = −0.051, and 95% LoA = [−4.290, 4.188]. A course-specific linear calibration modestly improved ICC to 0.782 [0.732, 0.823], reduced MAE to 1.215, and narrowed the LoA to [−3.157, 3.329]. A simpler bias-only correction produced only a small MAE reduction (1.591) and did not narrow the upper LoA. In practical terms, the linear calibration improved aggregate fit out of sample, but it did not eliminate threshold-sensitive error.

Threshold-level analyses led to the same conclusion. On the hold-out test set at the 9.0-point threshold, overall agreement improved from 0.879 to 0.900 after course-specific linear calibration, but adjacent-zone agreement for cases within ±1.5 points of the HC threshold remained 0.810. At the 12.0-point threshold, overall agreement improved from 0.683 to 0.762 and adjacent-zone agreement improved from 0.608 to 0.688. These results support calibration as a useful error-reduction step while still favoring a triage-oriented co-grading workflow for threshold-adjacent cases.

Table 3. Held-out Total-score calibration and threshold agreement.

Model (hold-out test)	ICC(3,1) [95% CI]	Pearson r [95% CI]	MAE	Mean Bias	95% LoA	Comment
Uncalibrated	0.774 [0.722, 0.817]	0.799 [0.752, 0.837]	1.624	−0.051	[−4.290, 4.188]	Baseline AI–HC
Bias-only calibration	0.782 [0.732, 0.824]	0.803 [0.758, 0.841]	1.591	+0.075	[−4.037, 4.188]	Small error reduction
Linear course-specific calibration	0.782 [0.732, 0.823]	0.806 [0.760, 0.843]	1.215	+0.086	[−3.157, 3.329]	Best MAE and narrowest LoA

Operational threshold checks on the same hold-out set showed that linear course-specific calibration improved overall agreement from 0.879 to 0.900 at the 9-point threshold and from 0.683 to 0.762 at the 12-point threshold. Adjacent-zone agreement remained at 0.810 at the 9-point threshold and improved from 0.608 to 0.688 at the 12-point threshold, but did not reach the level of confident case-level interchangeability.

4.5. Course-Level Total-Score Performance

Agreement varied across courses (Table 4; Figure 5). AI–HC Pearson correlations on the Total score were highest in Prompt Engineering ($r = 0.823$), followed by Photoshop Design ($r = 0.793$), and

 doi:10.20944/preprints202605.1284.v1

lower in AI Video Production ($r = 0.711$). The corresponding AI–HC Total ICCs were 0.795, 0.770, and 0.685, respectively. The corresponding human–human Total ICCs were 0.861 [0.834, 0.884], 0.883 [0.860, 0.902], and 0.679 [0.561, 0.771], respectively, indicating the strongest local human baseline in Photoshop Design rather than Prompt Engineering.

Pairwise comparisons showed that the AI was closer to H2 than H1 across all three courses on the Total score. AI–H1 Total ICCs were 0.737 in Prompt Engineering, 0.690 in Photoshop Design, and 0.552 in AI Video Production, whereas AI–H2 Total ICCs were 0.809, 0.805, and 0.708, respectively. Bias direction remained course-dependent: AI–HC bias was positive in Prompt Engineering (+0.438) but negative in Photoshop Design (−0.499) and AI Video Production (−0.587). These results reinforce the conclusion that deployment assumptions should be made at the course or task level, not at the abstract level of "AI grading" in general.

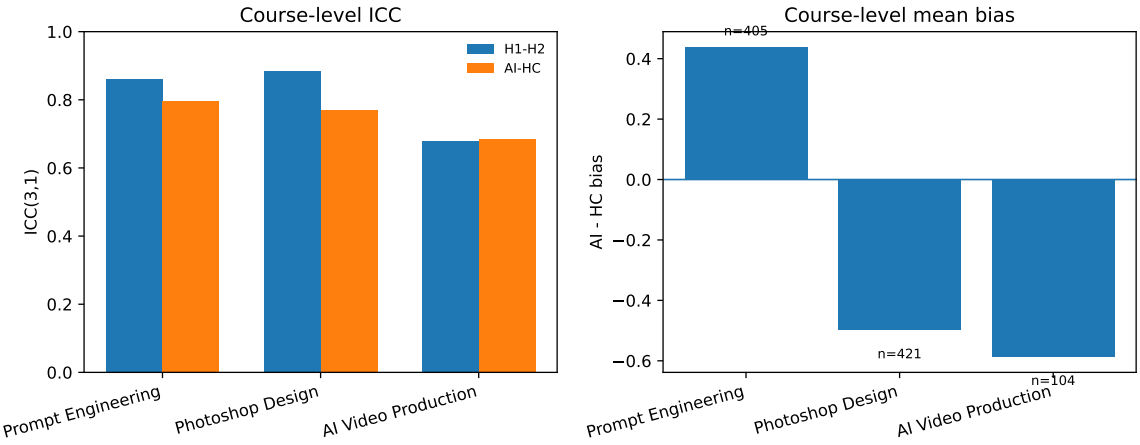

Figure 5. Course-level Total-score performance: (**a**) ICC comparison between H1–H2 and AI–HC by course; (**b**) mean bias (AI − HC) by course with sample sizes.

Table 4. Course-level Total-score agreement.

Course	n	H1–H2 ICC [95% CI]	AI–H1 ICC	AI–H2 ICC	AI–HC ICC [95% CI]	AI–HC Mean Bias
Prompt Engineering	405	0.861 [0.834, 0.884]	0.737	0.809	0.795 [0.757, 0.829]	+0.438
Photoshop Design	421	0.883 [0.860, 0.902]	0.690	0.805	0.770 [0.728, 0.806]	−0.499
AI Video Production	104	0.679 [0.561, 0.771]	0.552	0.708	0.685 [0.567, 0.775]	−0.587

Course-level results are shown for the Total score. Full course-level Pearson, MAE, bias, and LoA values are available from the authors on request.

5. Discussion

5.1. What the Results Support—And What They Do Not

In this study, the phrase "human-level" should be interpreted narrowly and locally. The revised pairwise analysis shows that AI–HC Total-score agreement (ICC = 0.763) was close to AI–H2 agreement (0.767), lower than the human–human baseline (0.819), and higher than AI–H1 agreement (0.700). This pattern indicates that the model approximated one human rater more closely than the other and that agreement with HC partly reflects the variance-reducing effect of score averaging.

Accordingly, the strongest defensible claim is one of constrained aggregate proximity to local scoring practice under a fixed rubric and fixed prompt, not equivalence to human raters in general and not agreement with an external gold standard. The Bland–Altman results and threshold analyses further show that individual-level and threshold-adjacent disagreements remain operationally meaningful. The present evidence therefore supports a calibrated co-grading model rather than unsupervised replacement.

 doi:10.20944/preprints202605.1284.v1

5.2. Trait Differences and Construct Clarity

The revised trait pattern is informative because it shifts the construct interpretation. Under transparent preprocessing, Quality remained the only trait with effectively absent human–human reliability, whereas Originality showed moderate human agreement rather than near-zero agreement. This means that the main construct-instability problem in the current rubric is concentrated in Quality, not uniformly across all subjectively framed traits.

The AI–H1 and AI–H2 asymmetries further support this interpretation. For Originality, AI–H2 agreement was much stronger than AI–H1 agreement, suggesting that part of the observed variation reflects rater-specific interpretation rather than a single stable latent construct. By contrast, Quality remained unstable across human and AI comparisons, so agreement results involving Quality should be treated cautiously and not as evidence of strong construct validity.

5.3. Scope Boundary and Construct Validity

An important interpretive issue in this study is construct validity, especially for courses that are naturally associated with creative or multimedia outputs. To address this issue, the scope boundary is defined explicitly: the study evaluates written responses only. In Photoshop Design and AI Video Production, the AI did not evaluate actual posters, edited images, or final video products. It evaluated students' explanatory descriptions of design reasoning, workflow steps, or tool knowledge.

This boundary does not make the study unimportant, but it materially changes what the findings mean. The results are relevant to text-based assessment tasks embedded within design-related courses. They should not be generalized as evidence that an LLM can evaluate authentic creative products in those domains without additional multimodal validation.

5.4. Implications for Human–AI Co-Grading

The most useful practical implication of the study is still a conservative co-grading model rather than a replacement model, but the revised results now allow this recommendation to be stated more concretely. First, the pairwise results show that AI agreement depends on which human rater is used as the reference, so operational systems should not be validated against an averaged benchmark alone. Second, the held-out calibration analysis shows that simple local calibration can reduce absolute error and narrow the disagreement range. Third, the threshold analysis shows that even calibrated systems remain vulnerable in threshold-adjacent cases.

A defensible deployment pattern is therefore a calibrated triage workflow. AI can be used for first-pass scoring, batch standardization, and feedback drafting on relatively well-anchored dimensions, while cases near grade boundaries, cases involving Quality, and cases with policy-sensitive consequences should be routed to human adjudication. The threshold evidence is particularly important here: even after calibration, adjacent-zone agreement remained at 0.810 at the 9-point threshold and reached only 0.688 (up from 0.608) at the 12-point threshold. Table 5 outlines the corresponding deployment patterns. This aligns with prior human–AI collaborative assessment work and with broader guidance on educational AI governance [12–14].

 doi:10.20944/preprints202605.1284.v1

Table 5. Recommended conservative human–AI grading workflow.

Condition	Recommended Handling	Rationale
Analytic traits with established agreement (Accuracy, Specificity, part of Total)	Use AI as a first-pass scorer or third reader, followed by batch auditing.	These traits showed the strongest alignment with HC and the clearest rubric anchors.
Low-stability or interpretively ambiguous cases (especially Quality)	Require routine human review or second-reader confirmation.	Quality showed effectively absent human–human reliability, so AI agreement should not be treated as interchangeable evidence.
Large disagreement, threshold-adjacent cases, or policy-sensitive decisions	Route to human adjudication.	Hold-out threshold checks show that adjacent-zone disagreements remain operationally meaningful even after calibration.
Course migration, rubric revision, or model update	Re-calibrate and re-validate locally before operational use.	Agreement varied by course and may shift with construct wording or model behavior.

5.5. Relation to Prior LLM-Scoring Studies

The present findings are broadly consistent with recent studies reporting that rubric-guided LLM scoring can approximate human ratings under controlled conditions [6–8]. They are also consistent with the cautionary side of the literature: agreement is stronger for well-anchored, text-based criteria than for subjective or construct-ambiguous judgments, and aggregate alignment does not guarantee case-level interchangeability [9,10]. What this study adds is a clearer comparison against the local human–human baseline and an explicit narrowing of claims to text-response scoring in three courses at one institution.

5.6. Limitations and Future Work

Several limitations remain. First, the study is based on one institution and three courses only, so external validity across institutions, grading cultures, and disciplines remains untested. Second, the evaluated inputs were written responses rather than the students' final multimedia artifacts. Third, the hold-out calibration analysis was conducted within one random split and one linear correction family; future work should test repeated resampling, alternative calibration strategies, and robustness across model updates and prompt perturbations.

Fourth, a source-data integrity issue was identified in the AI workbook: in 74 of 930 cases, the recorded AI Total did not equal the sum of the five AI rubric traits. Sensitivity analysis suggested that this inconsistency had little effect on the main Total-score conclusion, but future pipelines should enforce deterministic score aggregation at generation time. Fifth, the threshold analyses were operational checks using illustrative 9-point and 12-point cutoffs on the 15-point scale; these should not be treated as universal policy thresholds. Sixth, subgroup fairness analyses were not possible because demographic variables were unavailable in the de-identified dataset.

Future work should therefore extend the present held-out calibration analysis through repeated resampling, robustness checks across prompt and model changes, and validation in settings where actual creative artifacts—not only written explanations—are part of the assessment target. It would also be valuable to distinguish whether AI agreement improves because the model better captures the intended construct or because it more strongly regresses toward the center of the local scoring distribution.

6. Conclusions

This study shows that rubric-guided LLM scoring can approximate local human grading practice for text-based university responses, especially on the Total score and on relatively concrete analytic traits. The evidence is encouraging enough to support calibrated human–AI co-grading in narrowly defined settings, but not strong enough to justify fully automated replacement of human graders. The

 doi:10.20944/preprints202605.1284.v1

Total-score limits of agreement remain too wide for confident case-level substitution, and Quality remained unstable even for human raters.

The most defensible conclusion is therefore modest but useful: under a shared rubric and fixed scoring prompt, an LLM can serve as a structured support tool for text-response assessment across selected university courses. Its role is best understood as assisting, standardizing, calibrating, and triaging parts of the scoring workflow rather than replacing human judgment. Future work should extend the present held-out calibration analysis through repeated resampling, robustness checks across prompt and model changes, and validation in settings where actual creative artifacts—not only written explanations—are part of the assessment target.

Author Contributions: Conceptualization, H.K. and J.L.; methodology, H.K.; software, H.K.; validation, H.K., S.L., and J.L.; formal analysis, H.K.; investigation, H.K.; resources, H.K. and J.L.; data curation, H.K. and S.L.; writing—original draft preparation, H.K.; writing—review and editing, H.K., S.L., and J.L.; visualization, H.K.; supervision, J.L.; project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and received an exemption determination from the Institutional Review Board of Seoul Cyber University (Protocol No. AN01-202509-HR-001-01; determination date: 27 October 2025).

Informed Consent Statement: Written informed consent was waived for the de-identified secondary analysis of routine educational records, as approved by the Institutional Review Board.

Data Availability Statement: The de-identified score matrix used in the present analysis, analysis scripts, prompt template, and derived output tables are available from the corresponding author on reasonable request, subject to institutional privacy requirements. Raw student text responses are not publicly released because they are part of educational records. If permitted by institutional policy, a public repository version of the de-identified scoring matrix and analysis code will be deposited after publication.

Acknowledgments: The authors thank the two external course instructors who provided independent human ratings for this study (none of whom is an author of this paper) and the students whose de-identified coursework enabled the analysis. The authors are also grateful to the institutional data steward and the ethics office for guidance on de-identification and data governance. ChatGPT was used only to generate research scores for comparison and was not used to determine official course grades.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AES	Automated Essay Scoring
AI	Artificial Intelligence
H1, H2	Human Rater 1, Human Rater 2
HC	Human Consensus (mean of two human raters)
ICC	Intraclass Correlation Coefficient
LLM	Large Language Model
LoA	Limits of Agreement
MAE	Mean Absolute Error

References

1. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
2. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

 doi:10.20944/preprints202605.1284.v1

3. McGraw, K.O.; Wong, S.P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1996**, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30.
4. Attali, Y.; Burstein, J. Automated essay scoring with e-rater V.2. *J. Technol. Learn. Assess.* **2006**, *4*, 3.
5. Shermis, M.D. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assess. Writ.* **2014**, *20*, 53–76. https://doi.org/10.1016/j.asw.2014.08.002.
6. Tate, T.P.; Steiss, J.; Bailey, D.; Graham, S.; Moon, Y.; Ritchie, D.; Tseng, W.; Warschauer, M. Can AI provide useful holistic essay scoring? *Comput. Educ. Artif. Intell.* **2024**, *7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255.
7. Yavuz, F.; Celik, O.; Yavas Celik, G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *Br. J. Educ. Technol.* **2025**, *56*, 150–166. https://doi.org/10.1111/bjet.13494.
8. Aydin, B.; Kisla, T.; Elmas, N.T.; Bulut, O. Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System* **2025**, *133*, 103784. https://doi.org/10.1016/j.system.2025.103784.
9. Jade, T.; Yartsev, A. ChatGPT for automated grading of short answer questions in mechanical ventilation. *arXiv* **2025**, arXiv:2505.04645.
10. Quah, B.; Zheng, L.; Sng, T.J.H.; Yong, C.W.; Islam, I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations: A cross-sectional study. *BMC Med. Educ.* **2024**, *24*, 962. https://doi.org/10.1186/s12909-024-05881-6.
11. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *327*, 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.
12. Xiao, C.; Ma, W.; Song, Q.; Xu, S.X.; Zhang, K.; Wang, Y.; Fu, Q. Human–AI collaborative essay scoring: A dual-process framework with LLMs. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge (LAK '25)*, Dublin, Ireland, 3–7 March 2025; pp. 293–305. https://doi.org/10.1145/3706468.3706507.
13. Williamson, D.M.; Xi, X.; Breyer, F.J. A framework for evaluation and use of automated scoring. *Educ. Meas. Issues Pract.* **2012**, *31*, 2–13. https://doi.org/10.1111/j.1745-3992.2011.00273.x.
14. UNESCO. *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023. https://doi.org/10.54675/EWZM9535.
15. Kim, H.; Lee, S.M.; Yoon, J.; Cho, S.M.; Choi, J.H.; Lee, J. Design-Based Case Study of JoB+AI Model: Developing and Implementing Remote AI Competency Education Program for Disconnected Youth. *J. Future Soc.* **2025**, *16*, 192–208. https://doi.org/10.22987/jifso.2025.16.1.192.