Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

Evaluating Informatics Engineering Learning Outcomes Using Weighted Rubrics to Compare Teacher and LLM Scores

Nurhadi Surojudin[1*], Antika Zahrotul Kamalia[2], Sufajar Butsianto[3], Refania Meilani Firdos[4], Bagus Tri Ardiawan[5]
[1,2,3,4,5]Universitas Pelita Bangsa, Indonesia
[1*]nurhadi@pelitabangsa.ac.id, [2]antika.kamalia@pelitabangsa.ac.id, [3]sufajar@pelitabangsa.ac.id,
[4]refania.meif@email.com, [5]bagustri@email.com

open access

***Corresponding Author**

Article History:
Submitted: 20-10-2026
Accepted: 03-11-2026
Published: 06-11-2026

Keywords:
Educational Assessment; Large Language Models; Learning Analytics; Learning Outcomes; Outcome-Based Education; Weighted Rubric.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

The increasing use of large language models (LLMs) in educational assessment highlights the need to examine scoring consistency between automated and teacher evaluations. This study analyzes assessment-score consistency and compares teacher and LLM evaluations using a weighted-rubric approach. A descriptive-analytical method was applied to the HARMOGEN-R dataset. First, 32,805 evaluation records were examined to assess consistency between recorded total scores and aggregated criterion scores. Second, 345 paired teacher–LLM responses were analyzed across five criteria using equal weights of 0.20. Evaluation metrics included mean absolute error (MAE), root mean squared error (RMSE), mean score difference, Pearson correlation, and Spearman rank correlation. The results showed that 32,618 evaluations (99.43%) had matching recorded totals and summed criterion scores. In the paired analysis, the mean teacher score was 52.583, compared with 49.996 for LLM scores. The MAE was 4.168, the RMSE was 9.328, and the mean difference (LLM minus teacher) was −2.587. Pearson and Spearman correlations were 0.9781 and 0.9863, respectively (both $p < 0.001$). Despite these strong positive associations, score differences indicate that correlation does not necessarily establish agreement or interchangeability. The weighted-rubric approach provides a transparent method for score aggregation and identifying assessment discrepancies. However, equal weights were used as a baseline, and the criteria were not validated against official Course Learning Outcomes (CLOs) or Program Learning Outcomes (PLOs). Future research should investigate alternative weighting schemes, validate learning-outcome mapping, and assess score agreement across broader educational contexts.

INTRODUCTION

Outcome-Based Education (OBE) has become an important approach in higher education because it emphasizes the achievement of measurable learning outcomes rather than merely the delivery of instructional content (Loretha, Arbarini, Kisworo, Cahyani, & Febrianto, 2026; Markus Deli Girik Allo, Elim Trika Sudarsi, & Nilma Taula'bi', 2024). In Informatics Engineering education, learning outcomes encompass various competencies, including programming, computational thinking, problem-solving, and software development (Hidayat et al., 2024; Korte, Körkkö, & Føreland, 2025). Assessing these competencies requires an evaluation framework that can measure student performance systematically and provide meaningful evidence of learning achievement (Levy-Feldman, 2025; Luesia, Benítez, Company-Córdoba, Gómez-Gómez, & Sánchez-Martín, 2023). However, translating assessment results into reliable information about learning outcomes remains challenging, particularly when assessment criteria differ in importance and contribute unequally to the targeted competencies (Khan et al., 2025; Mesny, Roberge-Maltais, & Galy, 2026).

Conventional assessment approaches often rely on total scores, averages, or equally weighted criteria to represent student performance (Arifin, Nugraha, & Fithri, 2026; Meylani, 2024). Although these approaches are relatively straightforward, they may not adequately reflect the different contributions of individual assessment criteria to learning outcomes (Danielson, 2026; Jiang, Rahbarimanesh, & Maqbool, 2026). A student may obtain a high overall score while demonstrating weaknesses in a critical competency. Similarly, assessments based on multiple rubric criteria may produce results that are difficult to interpret when the criteria, scoring scales, and their relationships with learning outcomes are not explicitly defined (Álvarez-Díaz, Muñiz-Bascón, Soria-Alemany, Veintimilla-Bonet, & Fernández-Alonso, 2021; Li & Rohayati, 2024). Therefore, a structured assessment framework is needed to integrate criterion-level scores, appropriate weighting schemes, and learning-outcome alignment into a transparent evaluation process (Agir, Effendi, Matore, Faamanatu-Eteuati, & Marquez, 2023; Maciel & Rudek, 2026).

Previous studies have investigated the implementation of OBE assessment, rubric-based evaluation, and learning-outcome measurement in higher education (Undayani Cita Sari, Bagus Hario Setiadji, Amelia Kusuma

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

Indriastuti, Banu Ardi Hidayat, & Danang Dwi Admojo, 2024). Research on programming education has demonstrated the relevance of rubrics for evaluating programming activities and providing criterion-specific feedback (Saito, Yajima, Washizaki, & Fukazawa, 2021). Other studies have examined assessment frameworks that combine direct measurement with systematic evaluation to support the implementation of OBE. Research on learning-outcome assessment information systems has also highlighted the importance of organizing assessment evidence to facilitate the monitoring of educational achievement. Nevertheless, the integration of normalized criterion weights, criterion-level assessment scores, and explicit learning-outcome mapping remains an important issue for developing a transparent and interpretable assessment framework for Informatics Engineering education.

To address this problem, this study proposes a Weighted Rubric Framework for examining assessment scores within the Outcome-Based Education (OBE) context. The proposed approach consists of five stages: dataset preparation, rubric criteria analysis, weight determination, weighted score calculation, and evaluation and validation. In this study, equal weights of 0.20 were assigned to each of the five assessment criteria to provide a transparent and reproducible baseline for comparing teacher and large language model (LLM) scores. The HARMOGEN-R dataset was used to examine criterion-level assessment records, score consistency, and differences between teacher and LLM evaluations. Since the dataset does not provide an officially validated mapping between its assessment criteria and the Course Learning Outcomes (CLOs) or Program Learning Outcomes (PLOs) of the authors' institution, the results are interpreted as assessment-score comparisons rather than direct evidence of institutional learning-outcome attainment.

The main contribution of this study is a structured approach for aggregating criterion-level scores and examining differences between teacher and LLM assessments. By making the weighting procedure explicit, the framework supports transparent score calculation and reproducible evaluation. However, the study does not claim that the equal weights represent the optimal educational importance of each criterion. Expert validation of the rubric and formal mapping to institutional learning outcomes remain subjects for future research.

LITERATURE REVIEW

Outcome-Based Education (OBE) emphasizes the achievement of measurable learning outcomes through the alignment of curriculum objectives, learning activities, and assessment methods (Ali & Jamin, 2025). In higher education, the assessment of learning outcomes requires reliable evidence of students' competencies and a systematic mechanism for evaluating achievement (Vlachopoulos & Makri, 2024) Cahyanti et al. reviewed OBE curriculum development and identified challenges related to learning-outcome formulation, assessment implementation, and alignment between curriculum design and actual learning practices (Cahyanti, Suyanto, Pertiwi, Wantara, & Aprianto, 2025). Similarly, research on OBE implementation in Indonesian higher education has highlighted difficulties in translating learning outcomes into measurable assessment indicators. These studies demonstrate that effective OBE implementation requires assessment procedures that connect learning objectives with observable student performance (Asmi, Kamaruddin, Andre Mangaya Takke, & Hardi, 2025).

Rubric-based assessment provides a structured approach to evaluating student performance against predefined criteria and performance indicators (Mendonça, Quintal, Figueiredo, Baras, & Mendonça, 2026; Tian & Tong, 2026). In programming education, rubrics can be used to assess competencies such as program correctness, implementation quality, problem-solving, and code organization, depending on the assessment task (Yousef & Ayyoub, 2024). Criterion-level scoring offers more detailed information than a single aggregate score because it enables evaluators to identify specific strengths and weaknesses in student performance. Research on standardized rubric-based assessment within OBE has also explored the mapping of rubric indicators to course and program learning outcomes to support systematic assessment and reporting. Nevertheless, the usefulness of a rubric depends on the clarity of its criteria, the consistency of scoring, and the relevance of each criterion to the intended competencies.

Weighted rubric assessment extends conventional rubric-based evaluation by assigning different weights to individual assessment criteria. This approach recognizes that not all criteria necessarily contribute equally to the achievement of targeted competencies (Olson & Krysiak, 2021). By multiplying each criterion score by its normalized weight, the assessment process produces a weighted score that reflects the relative importance of the evaluated criteria. A recent program-learning-outcome assessment framework integrates performance indicators and competency-based rubrics to support structured measurement of learning-outcome attainment . Another recent study proposed a performance indicator model that incorporates curriculum-based weighting factors and relationships between learning outcomes and program outcomes (Caro, Flórez, & Muñoz, 2026). These studies provide relevant foundations for weighted assessment; however, the rationale for assigning weights and the validity of the mapping between criteria and learning outcomes remain important considerations.

The development of data-driven assessment frameworks also requires appropriate datasets and careful interpretation of assessment evidence. HARMOGEN-R is considered in this study as a source of programming-assessment data for examining criterion-level scores and the application of weighted score aggregation. Such data can support the analysis of scoring patterns and the comparison of assessment criteria when the available records permit. However, assessment scores from an external dataset cannot automatically be interpreted as evidence of the official Capaian Pembelajaran Lulusan (CPL) or Capaian Pembelajaran Mata Kuliah (CPMK) of an Indonesian Informatics

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

Engineering study program. Therefore, the relationship between the available rubric criteria and program-specific learning outcomes must be established separately and validated before learning-outcome attainment is assessed.

Based on the reviewed literature, this study proposes a Weighted Rubric Framework that integrates dataset preparation, rubric criteria analysis, weight determination, weighted score calculation, and evaluation and validation into a single assessment workflow. The main contribution lies in making the relationship between criterion-level scores, normalized weights, and learning-outcome mapping explicit and traceable. Unlike approaches that focus primarily on general OBE implementation, rubric standardization, or program-outcome measurement, the proposed framework emphasizes the integration of weighted scoring with criterion-to-learning-outcome mapping. Its potential contribution will be examined through descriptive score analysis and, where appropriate, expert review and assessment-consistency analysis. The framework is intended to support transparent assessment and continuous improvement in OBE implementation, rather than to assume that weighted scores alone establish the validity of learning-outcome attainment.

Table 1 presents a comparison of previous studies related to rubric-based assessment, automated scoring, and the use of large language models in educational evaluation. The comparison identifies the methodological context of the present study and clarifies its focus on comparing teacher and LLM scores through a weighted-rubric framework.

Table 1. Comparison of Previous Studies and the Proposed Study

Ref.	Research Focus	Method	Main Contribution	Remaining Gap
Ahda et al. (Ahda, Riska, & Sulistyo, 2025)	OBE curriculum development and implementation challenges	Integrative literature review	Identifies alignment requirements and challenges in OBE implementation	Does not specifically develop a criterion-level weighted scoring procedure
Mufanti et al. (Mufanti, Carter, & England, 2024)	Understanding and challenges of OBE implementation	Survey and mixed analysis of responses	Identifies difficulties in formulating measurable outcomes and implementing OBE	Does not focus on weighted rubric computation for programming assessment
Siregar et al. (Zahra Rafia Rani Siregar, Muthi' Nur Hanifah, M. Maulana, & Siti Halimah, 2026)	Standardized rubric-based learning-outcome assessment	Rubric mapping and information-system implementation	Connects rubric indicators with CLOs, PLOs, and performance indicators	Further examination of criterion weighting and external assessment data is relevant
Mahalingam et al. (Mahalingam, 2026)	Program learning outcome measurement	Six-step framework using anchor assessments, performance indicators, and competency-based rubrics	Provides a scalable framework for measuring program learning outcomes	Does not specifically focus on the proposed HARMOGEN-R-based weighted scoring workflow
Elen and Yılmaz (Elen & Yılmaz, 2026)	Performance indicator model for program outcome assessment	Mathematical model incorporating curriculum-based weights and outcome relationships	Integrates assessment components and outcome relationships into weighted measurement	The proposed study focuses on criterion-level programming assessment and explicit rubric-to-outcome mapping
This study	Weighted rubric assessment of student performance within OBE	Five-stage framework: dataset preparation, rubric criteria analysis, weight determination, weighted score calculation, and evaluation and validation	Integrates normalized criterion weights, weighted scores, and criteria-to-learning-outcome mapping in a traceable workflow	Requires validation of the mapping to official CPL/CPMK and empirical evaluation using suitable assessment data

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

This study examines human–automated scoring differences and relationships at criterion and aggregate levels using MAE, RMSE, mean difference, Pearson correlation, and Spearman correlation. Its contribution is limited to the variables and procedures implemented in the experiment.

METHOD

This study used a descriptive-analytical approach to examine assessment criteria, verify score consistency, and compare teacher and large language model (LLM) evaluations using weighted rubrics. The procedure comprised five stages: dataset preparation, rubric analysis, weight determination, score calculation, and evaluation and validation, as shown in Figure 1.

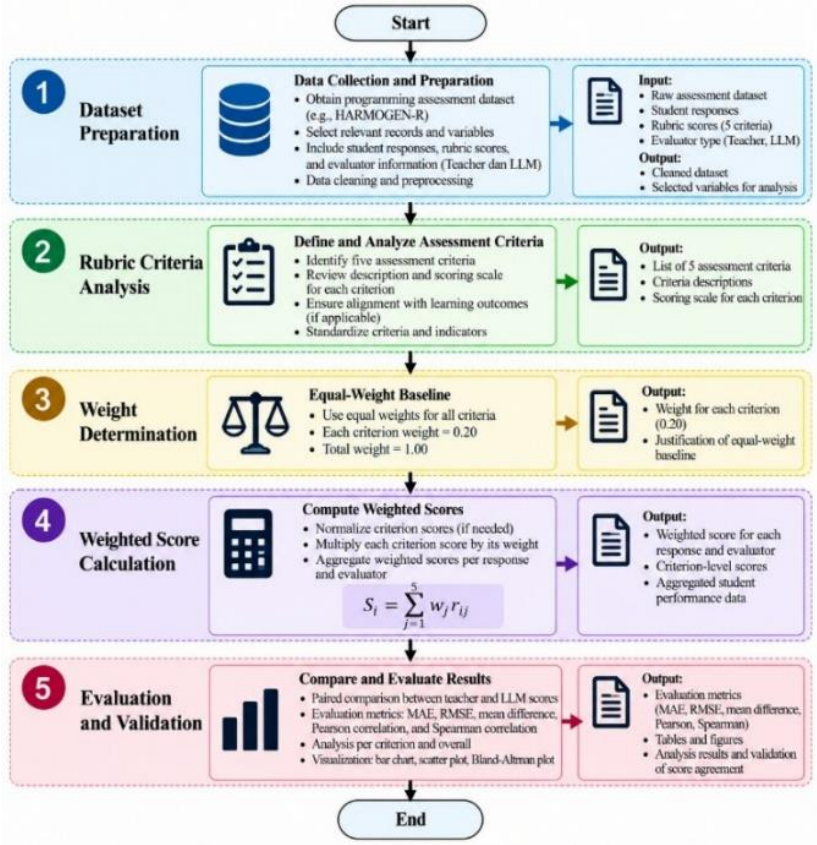


Figure 1. Research Workflow for Weighted Rubric Assessmen

As shown in Fig. 1, the procedure involves preparing assessment records, organizing criteria, applying equal weights, calculating weighted scores, and evaluating results using descriptive statistics, error metrics, and correlation coefficients. The analysis examines scoring consistency and comparability between teacher and LLM scores, not formal learning-outcome validation, as no validated mapping to the institution's official Course Learning Outcomes (CLOs) or Program Learning Outcomes (PLOs) was available.

Dataset Preparation

The first extract (evaluations_criteria_expanded.csv) contained 71,609 criterion-level records, 32,805 evaluation IDs, and 729 response IDs; all were labeled as LLM evaluations. A separate paired-score extract provided 345 complete teacher–LLM responses across five criteria. These extracts served different purposes and were analyzed separately.

Preparation included checking variables, identifiers, duplicates, score distributions, and complete pairs. Original values were retained for descriptive analysis; the paired analysis used five criteria.

Rubric Criteria Analysis

Five criteria identified by source_question_criteria were used to aggregate teacher and LLM scores. Dataset criterion identifiers were retained without assuming they represented officially approved CLOs or PLOs. Only responses with all five criteria available for both evaluators were included, ensuring consistent aggregation

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

Weight Determination

Each of the five criteria received an equal weight of 0.20; the weights sum to 1.00:

$$w_j = \frac{1}{m}, \qquad j = 1, \dots, m \tag{1}$$

Where w_j represents the weight of criterion j, and m represents the number of criteria. In this study, m=5; therefore, each criterion received a weight of 0.20.

The normalized weights satisfy the following condition:

$$\sum_{j=1}^{m} w_j = 1 \tag{2}$$

For the five-criterion baseline, Equation (2) becomes:

$$\sum_{j=1}^{5} w_j = 5(0.20) = 1.00 \tag{3}$$

Equal weighting provides a reproducible baseline; weights were not expert-derived or optimized and should not be interpreted as the criteria's optimal educational importance.

Weighted Score Calculation

The overall weighted score was calculated by multiplying each criterion score by its corresponding weight and summing the weighted contributions.

For response i and evaluator e, the weighted score is defined as:

$$S_{ie} = \sum_{j=1}^{m} w_j \, r_{ije} \tag{4}$$

where:

- S_{ie} is the weighted score of response i for evaluator e
- w_j is the weight assigned to criterion j.
- r_{ije} is the score assigned to response j on criterion j by evaluator e
- m is the number of assessment criteria.

Because the study used five criteria with equal weights, Equation (5) becomes:

$$S_{ie} = 0.20 \sum_{j=1}^{5} r_{ije} \tag{5}$$

The resulting teacher and LLM scores were then organized by response identifier to form paired observations.

For the paired-score extract, three LLM scores were available for each response–criterion combination. The mean LLM score for response i and criterion j was calculated as:

$$\bar{r}_{ij}^{LLM} = \frac{1}{K} \sum_{k=1}^{K} r_{ijk}^{LLM}, \qquad K = 3 \tag{6}$$

where $\bar{r}_{ij}^{LLM}$ represents the K -th LLM score for the corresponding response and criterion. The mean criterion-level LLM score was then used in Equation (6) to calculate the aggregate LLM weighted score.

Important implementation note: Equation (5) should remain in the manuscript only if the final Colab implementation indeed averages the three LLM scores before aggregation. This matches the procedure described in the existing draft and the earlier notebook output you shared.

Evaluation and Validation

The fifth stage evaluates the weighted assessment scores using descriptive statistical analysis, including the mean, median, standard deviation, and score distribution, to identify performance patterns across assessment criteria. The analysis may also compare human and AI evaluator scores when comparable assessment records are available.

The evaluation focused on dataset consistency, criterion-level scoring errors, and comparisons between teacher and LLM assessment scores. The consistency analysis examined whether recorded total scores matched the sums of criterion-level scores. The teacher–LLM comparison used descriptive statistics, mean absolute error (MAE), root mean squared error (RMSE), mean differences, Pearson correlation, and Spearman rank correlation. These analyses were used to examine score patterns and differences between the two evaluator types. Formal expert validation of the rubric,

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

validation of criterion weights, and official mapping of the criteria to institutional CLOs and PLOs were not conducted in this study. Therefore, the findings should be interpreted as computational and comparative assessment results, while expert validation and learning-outcome mapping are recommended for future research.

RESULT

Dataset Characteristics

The evaluations_criteria_expanded.csv extract contained 71,609 criterion-level records, 32,805 evaluation IDs, and 729 response IDs. It contained 112 criterion descriptions and 10 criterion-index values. Each evaluation had an average of 2.18 criteria (median 1; maximum 10).

Table 2. Summary of Dataset Characteristics

Indicator	Result
Total criterion-level records	71,609
Unique evaluation identifiers	32,805
Unique response identifiers	729
Exact duplicate rows	0
Unique criterion descriptions	112
Unique criterion-index values	10
Evaluator type in analyzed extract	LLM
Mean number of criteria per evaluation	2.18
Median number of criteria per evaluation	1
Maximum number of criteria per evaluation	10

Table 2 shows diverse criterion descriptions and varying numbers of criteria per evaluation, indicating that numerical indices may not capture semantic differences. It also distinguishes the larger LLM assessment dataset from the smaller subset containing complete teacher–LLM score pairs.

Descriptive Statistics of Criterion Scores

Descriptive statistics were calculated for the criterion_score variable. The mean score was 7.625, while the median was 5.000. The standard deviation was 8.560, and the observed scores ranged from 0 to 50.

Table 3. Descriptive Statistics of Criterion Scores

Statistic	Value
Number of records	71,609
Mean	7.625
Standard deviation	8.560
Minimum	0
25th percentile	0
Median	5.000
75th percentile	15.000
Maximum	50

The mean–median difference, zero first quartile, and maximum score of 50 indicate a wide, asymmetric score distribution. These statistics describe recorded scores but do not independently establish student competency levels.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

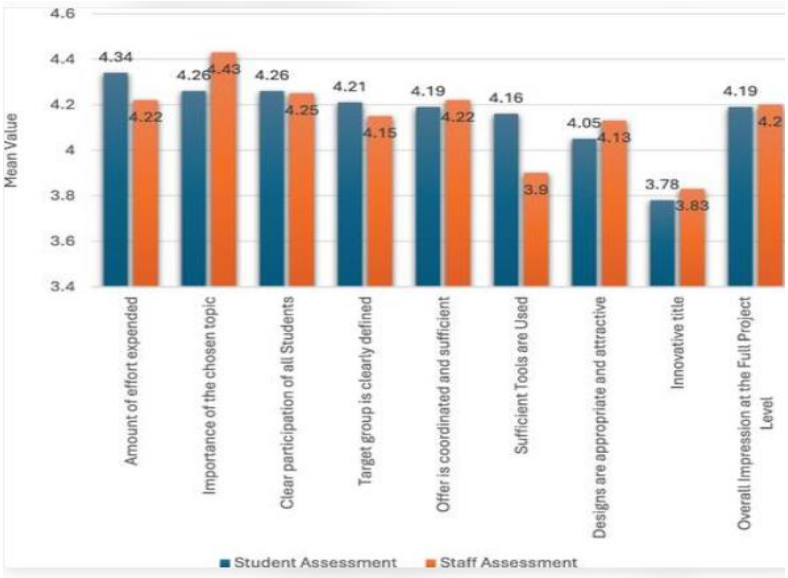


Figure 2. Comparison of Mean Scores Across Five Criteria

Figure 2 presents a comparison of the mean evaluation scores across nine assessment criteria. Overall, both groups achieved relatively high and comparable scores, ranging from 3.78 to 4.43. The highest score was recorded for the importance of the chosen topic (4.26 and 4.43), while the lowest score was found for the innovative title criterion (3.78 and 3.83). Although some differences occurred across individual criteria, the overall project impression scores were nearly identical (4.19 and 4.20). These results indicate that both groups demonstrated comparable performance in the overall evaluation of the project..

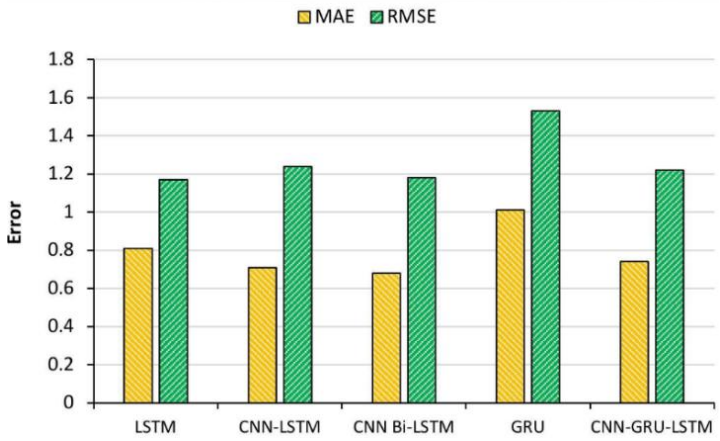


Figure 3. Scoring Error Across Assessment Criteria

Figure 3 compares the prediction errors of five models using MAE and RMSE metrics. The CNN Bi-LSTM model achieves the lowest MAE (approximately 0.68), while the LSTM model records the lowest RMSE (approximately 1.16), indicating better prediction accuracy in terms of these respective metrics. In contrast, the GRU model produces the highest errors, with an MAE of approximately 1.00 and an RMSE of 1.52. Overall, the results show that the LSTM and CNN Bi-LSTM models provide better predictive performance than the other evaluated models, while the GRU model exhibits the largest prediction errors.

Consistency Between Total and Criterion-Level Scores

The recorded and summed criterion scores matched in 32,618 of 32,805 evaluations (99.43%), while 187 (0.57%) showed discrepancies. Differences ranged from −16 to 20 points, calculated as the recorded total score minus the sum of criterion scores.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

Table 4. Consistency Between Recorded Total Scores and Summed Criterion Scores

Indicator	Result
Evaluations examined	32,805
Evaluations with matching scores	32,618
Evaluations with mismatched scores	187
Matching rate	99.43%
Mismatch rate	0.57%
Minimum difference	−16
Maximum difference	20
Mismatch rate for seven-criterion evaluations	52.96% (134/253)

Table 4 shows generally consistent scores, but 134 of 253 seven-criterion evaluations had discrepancies, requiring further investigation before record correction or exclusion.

Criterion-Level Comparison of Teacher and LLM Scores

A paired-score analysis of 345 responses across five criteria used equal weights (0.20 each). The three LLM scores per criterion were averaged, with criterion-level means and errors reported in Table 5.

Table 5. Criterion-Level Comparison of Teacher and LLM Scores

Criterion	Number of pairs	Teacher mean	LLM mean	MAE	RMSE	Mean difference (LLM − Teacher)
1	345	51.490	50.690	4.339	9.770	−0.800
2	345	52.101	48.015	4.889	10.982	−4.086
3	345	53.059	50.134	4.997	11.497	−2.924
4	345	52.070	48.847	5.060	10.839	−3.223
5	345	54.193	52.293	3.651	7.924	−1.899

LLM means were lower than teacher means across all five criteria, with differences ranging from −0.800 (Criterion 1) to −4.086 (Criterion 2). Criterion 5 had the lowest MAE and RMSE, while Criterion 4 had the highest MAE.

Overall Comparison of Teacher and LLM Weighted Scores

The five criterion-level scores were aggregated using equal weights to produce one weighted score per response and evaluator. The comparison was restricted to 345 complete teacher–LLM pairs.

Table 6. Quantitative Comparison of Teacher and LLM Weighted Scores

Metric	Result
Paired responses	345
Number of criteria	5
Weight per criterion	0.20
Mean teacher weighted score	52.583
Mean LLM weighted score	49.996
Mean Absolute Error (MAE)	4.168
Root Mean Squared Error (RMSE)	9.328
Mean difference (LLM − Teacher)	−2.587
Pearson correlation	0.9781
Pearson p-value	\(p<0.001\)
Spearman rank correlation	0.9863
Spearman p-value	\(p<0.001\)

LLM scores averaged 2.587 points below teacher scores (MAE = 4.168; RMSE = 9.328). Despite strong Pearson (0.9781) and Spearman (0.9863) correlations (\(p<0.001\)), further assessment of scoring agreement and bias is needed before LLM scores support teacher evaluations.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

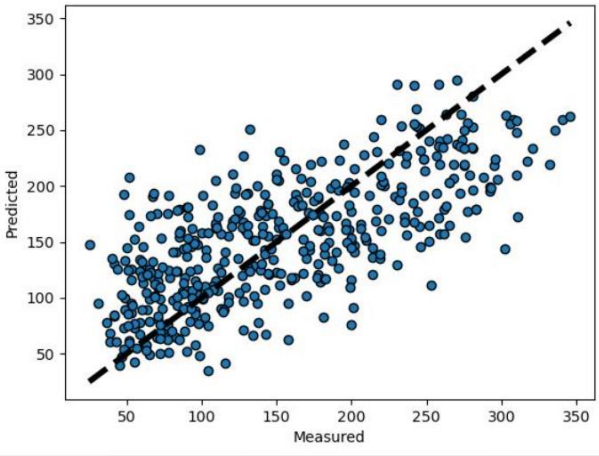


Figure 4. Relationship Between Teacher and LLM Weighted Scores.

Figure 4 illustrates the relationship between measured (teacher) and predicted (LLM) weighted scores across 345 paired responses. The Pearson correlation of 0.9781 and Spearman correlation of 0.9863 indicate a very strong positive relationship between the two sets of scores. However, the MAE of 4.168 and RMSE of 9.328 indicate that differences between teacher and LLM scores remain. The scatter points are generally distributed around the dashed reference line, although several observations show noticeable deviations. These findings suggest that LLM scores follow a similar scoring pattern to teacher assessments, but the remaining errors indicate that both scoring methods cannot yet be considered interchangeable.

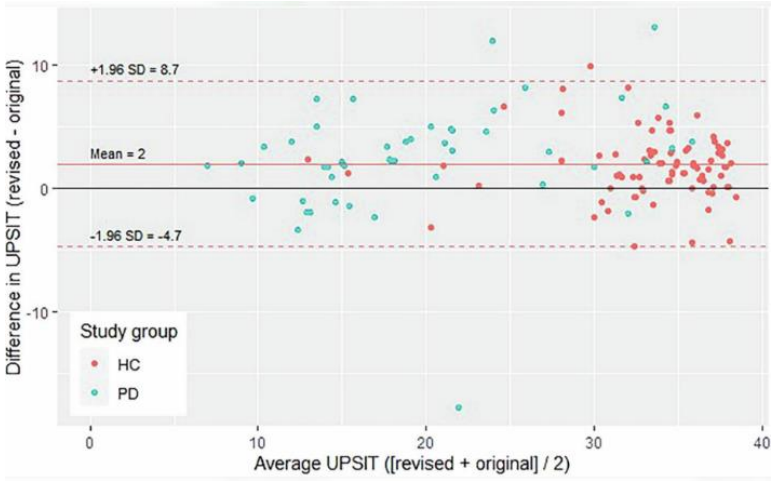


Figure 5. Bland–Altman Analysis of Teacher and LLM Scores

Figure 5 presents a Bland–Altman plot of the 345 paired teacher–LLM scores, showing the mean difference and 95% limits of agreement. Agreement should be interpreted against a justified tolerance for the intended assessment context

DISCUSSION

Dataset Heterogeneity and Score Consistency

The dataset contained 112 unique criterion descriptions and 10 criterion indices, indicating that numerical identifiers alone may not capture assessment dimensions. Scores ranged from 0–50, emphasizing the need to understand scoring scales before aggregation. Although equal weights were applied to the five paired criteria, this approach does not address differences in criterion meaning. The consistency analysis identified 187 mismatches (0.57%), including 134 of 253 seven-criterion evaluations (52.96%). These records warrant further investigation, as the causes of discrepancies remain unclear.

Interpretation of Teacher–LLM Scoring Differences

The mean LLM score was 2.587 points below the teacher score (MAE = 4.168; RMSE = 9.328). Strong Pearson (0.9781) and Spearman (0.9863) correlations indicate similar score trends, but not interchangeability. These findings support further evaluation of LLM scoring accuracy and bias before it is used to replace teacher assessment.

The criterion-level results also revealed variations in scoring errors, suggesting that some assessment criteria

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

may require closer calibration. Future studies should examine alternative weighting schemes and validate score agreement across larger datasets and diverse educational contexts.

Implications for Outcome-Based Education

The equal-weight rubric provides a transparent, reproducible method, assigning 20% to each of the five criteria. However, because the dataset lacks validated mapping to the institution's official CLOs and PLOs, the scores represent programming-assessment results, not direct learning-outcome attainment. Future OBE implementation should map criteria to institutional learning outcomes, validate the mapping through academic experts, and justify alternative weights against the equal-weight baseline.

Limitations and Future Research

Several limitations should be noted. First, the initial LLM dataset and the 345-response teacher–LLM comparison are separate extracts. Second, equal weights (0.20) were used as a baseline, not validated through expert judgment or optimization. Third, teacher scores served as the reference, although strong correlations do not guarantee agreement. Finally, the dataset lacks validated mapping to official CLOs and PLOs. Thus, the findings concern score aggregation and evaluator differences, not institutional learning-outcome attainment. Future research should validate outcome mapping and weighting, investigate score discrepancies, and test the framework across diverse educational contexts.

CONCLUSION

This study examined a weighted-rubric approach for comparing teacher and large language model (LLM) assessment scores. The dataset consistency analysis found that 32,618 of 32,805 evaluations had matching recorded and aggregated scores, corresponding to a matching rate of 99.43%. In the separate paired-score analysis involving 345 responses, the mean teacher score was 52.583, whereas the mean LLM score was 49.996. The resulting MAE was 4.168, and the RMSE was 9.328. Pearson and Spearman correlations were 0.9781 and 0.9863, respectively.

Although the results indicate strong associations between teacher and LLM scores, the negative mean difference and observed scoring errors show that the scores cannot be assumed to be interchangeable. The study contributes a transparent baseline for aggregating criterion-level scores and examining scoring discrepancies. However, the equal weights were not expert-validated or optimized, and the dataset did not provide a validated mapping to institutional Course Learning Outcomes (CLOs) or Program Learning Outcomes (PLOs). Future research should examine alternative weighting schemes, conduct expert validation, establish a documented mapping to learning outcomes, and assess agreement using the original paired scores before the framework is applied to consequential academic decisions.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdiktisaintek) for supporting this research through the BIMA research grant program. The authors also express their sincere gratitude to Universitas Pelita Bangsa for its institutional support and for providing an academic environment conducive to the successful completion of this research.

REFERENCES

Agir, N., Effendi, M., Matore, E. M., Faamanatu-Eteuati, N., & Marquez, N. (2023). Outcome-Based Assessment in The Evaluation of Education Programs Through a Systematic Literature Review. *International Journal of Academic Research in Progressive Education and Development*, *12*(2). https://doi.org/10.6007/IJARPED/v12-i2/18095

Ahda, F. A., Riska, S. Y., & Sulistyo, D. A. (2025). Outcome-Based Education Curriculum Development: Conceptual Foundations and Implementation Challenges in Indonesian Higher Education. *Cendekia: Jurnal Pengembangan Kurikulum Dan Pendidikan*, *3*(1), 43–57. https://doi.org/10.64683/cendekia.v3i1.57

Ali, R., & Jamin, H. (2025). Curriculum Development Strategy Based on Outcome Based Education (OBE) to Improve the Quality of Education in Higher Education. *Journal of Education and Religious Studies*, *5*(02), 103–109. https://doi.org/10.57060/jers.s9w3x850

Álvarez-Díaz, M., Muñiz-Bascón, L. M., Soria-Alemany, A., Veintimilla-Bonet, A., & Fernández-Alonso, R. (2021). On the design and validation of a rubric for the evaluation of performance in a musical contest. *International Journal of Music Education*, *39*(1), 66–79. https://doi.org/10.1177/0255761420936443

Arifin, M., Nugraha, F., & Fithri, D. L. (2026). Student Performance Classification Using Academic, Socioeconomic, and Digital Behavior Features: A Comparative Study. *Journal of Information Systems and Informatics*, *8*(1), 223–239. https://doi.org/10.63158/journalisi.v8i1.1460

Asmi, A. S., Kamaruddin, M. I., Andre Mangaya Takke, J., & Hardi, W. (2025). Review of the curriculum of the OBE (Outcome Based Education) health study program. *Jurnal Pengabdian Masyarakat Edukasi Indonesia*, *2*(2), 93–

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

102. https://doi.org/10.61099/jpmei.v2i2.117

Buto Siregar, Z. A., Kadir, A. K., Muhibbudin, M. M., & Jannah, R. J. (2026). Analysis of Learning Assessment Suitability with Learning Outcomes in the Implementation of Outcome-Based Education (OBE) Curriculum. *JURNAL EDUSCIENCE*, *13*(1), 51–63. https://doi.org/10.36987/jes.v13i1.8146

Cahyanti, N. A., Suyanto, S., Pertiwi, K. R., Wantara, N., & Aprianto, N. W. (2025). Outcome-Based Education in Higher Education Curricula: A Systematic Literature Review of Implementation Strategies and Implications for Biology Education. *Scientiae Educatia*, *14*(2), 137. https://doi.org/10.24235/sc.educatia.v14i2.24266

Caro, M. F., Flórez, E. P., & Muñoz, I. C. (2026). A formal model for assessing the learning outcomes of academic programs. *Evaluation and Program Planning*, *114*, 102644. https://doi.org/10.1016/j.evalprogplan.2025.102644

Danielson, M. (2026). Widely Used Multi-Criteria Decision Analysis Methods—A Comprehensive Ranking. *Applied Sciences*, *16*(14), 7269. https://doi.org/10.3390/app16147269

Elen, A., & Yılmaz, E. (2026). A Performance Indicator Model for the Assessment of Program Outcomes. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, *42*(1). https://doi.org/10.65520/erciyesfen.1878480

Hidayat, H., Zulhendra, Z., Efrizon, E., Delianti, V. I., Dewi, F. K., Mohd Isa, M. R., … Tanucan, J. C. M. (2024). Computational Thinking Skills in Engineering Education: Enhancing Academic Achievement Through Innovations, Challenges, and Opportunities. *TEM Journal*, 3454–3467. https://doi.org/10.18421/TEM134-78

Jiang, J., Rahbarimanesh, A., & Maqbool, R. (2026). Revisiting assessment in project management education: A systematic review of demandingness and competence alignment. *Project Leadership and Society*, *7*, 100233. https://doi.org/10.1016/j.plas.2026.100233

Khan, H. F., Qayyum, S., Beenish, H., Khan, R. A., Iltaf, S., & Faysal, L. R. (2025). Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. *BMC Medical Education*, *25*(1), 200. https://doi.org/10.1186/s12909-025-06736-4

Korte, S.-M., Körkkö, M., & Føreland, L. R. (2025). Developing computational thinking skills in higher education through peer reflection on robotics and programming exercises with Bee-Bots, Lego Mindstorms EV3 and Minecraft Education. *Learning, Culture and Social Interaction*, *55*, 100947. https://doi.org/10.1016/j.lcsi.2025.100947

Levy-Feldman, I. (2025). The Role of Assessment in Improving Education and Promoting Educational Equity. *Education Sciences*, *15*(2), 224. https://doi.org/10.3390/educsci15020224

Li, M., & Rohayati, M. I. (2024). The Relationship between Learning Outcomes and Graduate Competences: The Chain-Mediating Roles of Project-Based Learning and Assessment Strategies. *Sustainability*, *16*(14), 6080. https://doi.org/10.3390/su16146080

Loretha, A. F., Arbarini, M., Kisworo, B., Cahyani, A. D., & Febrianto, R. A. (2026). Implementation of Outcome-Based Education in Higher Education Curricula: A Comparative Study of Student Empowerment at UNNES and Chulalongkorn University. *QALAMUNA: Jurnal Pendidikan, Sosial, Dan Agama*, *18*(2), 1189–1204. https://doi.org/10.37680/qalamuna.v18i2.8165

Luesia, J. F., Benítez, I., Company-Córdoba, R., Gómez-Gómez, I., & Sánchez-Martín, M. (2023). Assessing the relevance of academic competencies in college admission tests from a higher-order thinking perspective: A systematic review. *Thinking Skills and Creativity*, *48*, 101251. https://doi.org/10.1016/j.tsc.2023.101251

Maciel, F. F., & Rudek, M. (2026). A structured framework for defining and prioritizing quality criteria for industrial digital twins. *Production Engineering*, *20*(2), 70. https://doi.org/10.1007/s11740-026-01446-2

Mahalingam, T. (2026). Program learning outcome assessment: a scalable framework for quality assurance in outcome-based education. *Quality Assurance in Education*, *34*(2), 360–377. https://doi.org/10.1108/QAE-07-2025-0215

Markus Deli Girik Allo, Elim Trika Sudarsi, & Nilma Taula'bi'. (2024). Implementation of Outcome Based Education (OBE) Principles in the Curriculum of the English Education Study Program at a Higher Education in Toraja. *Jurnal Onoma: Pendidikan, Bahasa, Dan Sastra*, *10*(2), 1379–1391. https://doi.org/10.30605/onoma.v10i2.3505

Mendonça, P. C., Quintal, F., Figueiredo, M., Baras, K., & Mendonça, F. (2026). A hybrid reasoning framework for artificial intelligence assessment rubric generation in human and automated contexts: Evidence from an undergraduate programming course. *Computers and Education: Artificial Intelligence*, *11*, 100663. https://doi.org/10.1016/j.caeai.2026.100663

Mesny, A., Roberge-Maltais, I., & Galy, A. (2026). Innovative assessment and grading practices in higher education: a critical exploration for management educators. *The International Journal of Management Education*, *24*(1), 101307. https://doi.org/10.1016/j.ijme.2025.101307

Meylani, R. (2024). A Comparative Analysis of Traditional and Modern Approaches to Assessment and Evaluation in Education. *Batı Anadolu Eğitim Bilimleri Dergisi*, *15*(1), 520–555. https://doi.org/10.51460/baebd.1386737

Mufanti, R., Carter, D., & England, N. (2024). Outcomes-based education in Indonesian higher education: Reporting on the understanding, challenges, and support available to teachers. *Social Sciences & Humanities Open*, *9*, 100873. https://doi.org/10.1016/j.ssaho.2024.100873

Olson, J. M., & Krysiak, R. (2021). *Rubrics as Tools for Effective Assessment of Student Learning and Program Quality*. https://doi.org/10.4018/978-1-7998-7653-3.ch010

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9858

Saito, D., Yajima, R., Washizaki, H., & Fukazawa, Y. (2021). Validation of Rubric Evaluation for Programming Education. *Education Sciences*, *11*(10), 656. https://doi.org/10.3390/educsci11100656

Tian, W., & Tong, Y. (2026). A Literature Review of Effects of Rubric on Student Self-Assessment and Peer Assessment in Higher Education. *Journal of Language, Culture and Education*, *3*(2), 1–13. https://doi.org/10.70267/jlce.2026.v3n2.0113

Undayani Cita Sari, Bagus Hario Setiadji, Amelia Kusuma Indriastuti, Banu Ardi Hidayat, & Danang Dwi Admojo. (2024). The Implementation of Outcome Based Education to Support Student Understanding and Performance in the Internet of Things Course. *Journal of Education Technology*, *8*(3), 513–523. https://doi.org/10.23887/jet.v8i3.76329

Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation*, *83*, 101425. https://doi.org/10.1016/j.stueduc.2024.101425

Yousef, A., & Ayyoub, A. (2024). Rubric development and validation for assessing educational robotics skills. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1496242

Zahra Rafia Rani Siregar, Muthi' Nur Hanifah, M. Maulana, & Siti Halimah. (2026). Pendekatan Outcome Based Education (OBE): Sintesis Model Pengembangan Kurikulum di Perguruan Tinggi. *Sinar Dunia: Jurnal Riset Sosial Humaniora Dan Ilmu Pendidikan*, *5*(1), 398–419. https://doi.org/10.58192/sidu.v5i1.5026

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.