Journal of Informatics and Vocational Education (JOIVE)
Vol.9, No.2, July 2026, pp. 13~22
ISSN: 2746-7813

Human–GenAI Score Alignment in Rubric-Constrained Essay Assessment: Procedural Convergence Without Pedagogical Equivalence

Muhamad Akda Fathul Barri[1], Rizki Hikmawan[2*]
[1,2] Department of Information System and Technology Education, Indonesia University of Education, Indonesia

Article Info

Article history:

Received Feb 23, 2026
Revised Mar 4, 2026
Accepted Jul 1, 2026

Corresponding Author:

Rizki Hikmawan, Information System and Technology Education,
Indonesia University of Education
Jl. Dr. Setiabudi No. 229, Isola, Kota Bandung, Jawa Barat 40154 , Indonesia.
Email: hikmariz@upi.edu

ABSTRACT

The rapid integration of Generative Artificial Intelligence (GenAI) into educational contexts has intensified scrutiny regarding its reliability in rubric-based assessment of open-ended responses. This exploratory study examines Human–AI score alignment in junior secondary Informatics essay assessment using a parallel scoring design. Seventeen student responses (n = 17) were independently evaluated by a certified teacher and ChatGPT under an identical analytic rubric structured across progressively increasing cognitive levels. Inter-rater agreement was analyzed using a two-way mixed-effects Intraclass Correlation Coefficient (ICC) with absolute agreement. Results demonstrate very high alignment (ICC = 0.985), indicating strong procedural consistency in rubric application between human and AI scoring. However, qualitative comparison of feedback reveals substantive pedagogical divergence. While GenAI feedback exhibits structural clarity and explicit rubric referencing, teacher feedback reflects contextual sensitivity, instructional intent, and diagnosis of student misconceptions. These findings suggest that rubric-constrained GenAI scoring can approximate teacher-generated scores in structured assessment settings, particularly for formative purposes. Nevertheless, the limited sample size restricts generalizability, and score alignment should not be interpreted as pedagogical equivalence. GenAI is therefore best positioned as an augmentative decision-support tool operating under teacher oversight. This study contributes empirical evidence to ongoing discourse on responsible Human–AI assessment integration and underscores the mediating role of rubric design in alignment outcomes.

Keywords: Generative Artificial Intelligence, Rubric-Based Assessment, Formative Assessment, Human–AI Alignment, Secondary Education

1. INTRODUCTION

Driven by the urgency of post-pandemic recovery, the rapid acceleration of digital transformation in education has intensified longstanding concerns regarding assessment quality in classroom settings. While instructional delivery has increasingly incorporated digital tools, the evaluation of open-ended student responses remains heavily dependent on teachers' professional judgment. In formative assessment contexts, teachers are expected to ensure fairness, consistency, and pedagogical relevance while simultaneously managing increasing workload demands. These pressures raise important questions about how assessment reliability and instructional integrity can be maintained within evolving technological environments.

Assessment is no longer understood solely as a mechanism for documenting learning outcomes, but as a central component of learning quality assurance and instructional decision-making (Bearman et al., 2024; Sortwell et al., 2024). Contemporary evaluation frameworks emphasize construct validity, reliability, transparency, and credibility in authentic classroom contexts. Within this expanded role, analytic rubrics have been widely adopted to standardize evaluative criteria and clarify performance expectations. By translating curriculum standards into operational descriptors, rubrics aim to reduce ambiguity and support consistency in teacher judgment.

However, rubric-based assessment does not eliminate subjectivity. Even when identical rubrics are applied, inter-rater discrepancies frequently emerge due to differences in interpretation, calibration, and contextual judgment

(Martin et al., 2025). Manual rubric implementation requires sustained cognitive effort, including descriptor interpretation, cross-criteria comparison, and maintenance of internal scoring consistency over time (Ling, 2024; Panadero et al., 2023). Such demands increase vulnerability to interpretative drift and internal bias, particularly in the evaluation of open-ended responses that require nuanced professional reasoning. These inconsistencies extend beyond numerical scores and affect the coherence, specificity, and formative orientation of feedback, with potential implications for fairness and assessment-for-learning practices (Ajjawi et al., 2023; Williams, 2024).

The emergence of Generative Artificial Intelligence (GenAI) introduces new possibilities for addressing these procedural challenges. Large language model–based systems such as ChatGPT differ from earlier automated scoring approaches that relied primarily on predefined patterns or rigid statistical modeling. Contemporary systems demonstrate advanced semantic processing capabilities, enabling them to interpret rubric descriptors, evaluate complex student responses, and generate structured feedback within constrained evaluative frameworks (Atasoy & Moslemi Nezhad Arani, 2025; Kizilcec et al., 2024). These affordances raise the possibility that GenAI may approximate elements of human evaluative application under clearly defined rubric conditions.

Yet the pedagogical credibility of such systems remains empirically uncertain. Existing research often prioritizes technical performance indicators such as correlation coefficients and agreement indices (Bui & Barrot, 2025), potentially reducing assessment quality to numerical convergence. However, score similarity does not necessarily imply alignment in evaluative reasoning or pedagogical intent. Feedback plays a central role in formative learning processes by supporting conceptual development, self-regulation, and instructional adjustment. Automated systems may demonstrate high statistical agreement while diverging in how feedback is framed, contextualized, or instructionally oriented. Without careful examination, such divergence risks weakening the formative function of classroom assessment.

Recent scholarship increasingly recognizes that human-AI assessment alignment is multidimensional and contingent upon rubric precision and operational clarity (Jackaria et al., 2024; Pecuchova et al., 2025; Shen et al., 2025). Nevertheless, empirical evidence remains concentrated in higher education and large-scale testing environments. These contexts differ substantially from compulsory schooling, where assessment practices are deeply contextualized and intertwined with students' developmental trajectories. Classroom-based evidence examining rubric-constrained GenAI application in lower secondary education is still limited (Xing et al., 2024).

In lower secondary Informatics instruction, open-ended assessment tasks present distinctive challenges. Students are required to articulate conceptual and procedural often algorithmic reasoning in written form, while teachers must provide consistent and pedagogically meaningful evaluation under time constraints (Hidayatullah et al., 2024; Messer et al., 2024; Ukkonen et al., 2025). These conditions create an authentic context for examining whether GenAI can align with teacher-generated scoring and feedback when operating under a teacher-developed analytic rubric.

Addressing these gaps, the present study critically evaluates Human-AI alignment in rubric-based assessment by positioning teachers' professional judgment as the primary benchmark. Rather than treating numerical score agreement as the sole indicator of quality, this study examines both scoring alignment and the substantive characteristics of feedback generated by ChatGPT under identical rubric conditions. By integrating quantitative agreement analysis with qualitative feedback comparison, the study seeks to extend current understandings of assessment alignment beyond statistical equivalence toward pedagogical coherence within authentic classroom contexts.

Based on this background, the study addresses the following research questions:

1. To what extent are the scores and feedback substance generated by ChatGPT aligned with a teacher's manual assessment when the same teacher-developed assessment rubric is applied?
2. How does the teacher perceive the quality, fairness, and pedagogical usefulness of assessment outcomes generated by ChatGPT when used in conjunction with a teacher-developed rubric?

2. RESEARCH METHOD

2.1. RESEARCH DESIGN

This study employed an exploratory mixed-method design integrating qualitative inquiry with embedded quantitative reliability analysis. The design aimed to examine the alignment between teacher-based assessment and Generative Artificial Intelligence (GenAI)-based assessment, specifically ChatGPT (GPT-5.2, Free version), within an authentic lower secondary classroom context.

The methodological framework is conceptually grounded in the principles of assessment for learning, which position assessment as a formative mechanism for improving instructional quality rather than merely a summative evaluative tool (Ajjawi et al., 2023; Bearman et al., 2024).Within this perspective, GenAI is examined as an

augmentative assessment instrument rather than a substitutive replacement of professional teacher judgment (Nazaretsky et al., 2022; Sortwell et al., 2024).

Operationally, a parallel independent scoring procedure was implemented. Identical student response artifacts were evaluated separately by a professional teacher and ChatGPT using the same standardized rubric. The design enabled direct examination of numerical score agreement, feedback substance alignment, and teacher perception of AI-generated assessment outputs.

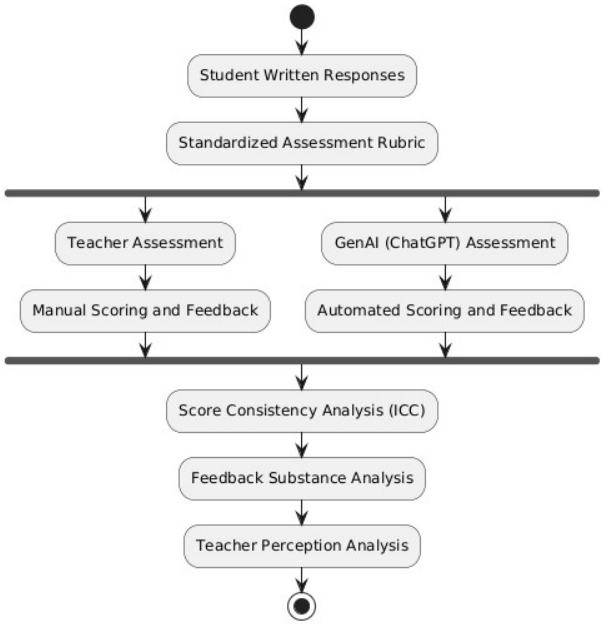

Figure 1. Methodological workflow of parallel assessment between teacher and Generative Artificial Intelligence using a standardized rubric.

Accordingly, the methodological intention of this study is not to validate GenAI assessment performance in a normative sense, but to explore epistemic alignment between human professional judgment and algorithmic assessment logic within a bounded classroom context.

2.2. RESEARCH CONTEXT, PARTICIPANTS, AND DATA SOURCES

2.2.1. RESEARCH CONTEXT AND PARTICIPANTS

The study was conducted in a lower secondary school (SMP) within the context of Informatics instruction. The focus was placed on the assessment of open-ended tasks requiring students to articulate conceptual and procedural reasoning in written form. Such tasks are considered appropriate for rubric-based evaluation because they allow systematic mapping of students' cognitive performance (Atasoy & Moslemi Nezhad Arani, 2025).

One Informatics teacher was selected through purposive sampling to ensure ecological validity of manual assessment. The teacher has more than five years of pedagogical experience and active involvement in rubric development and calibration. This selection aligns with prior research emphasizing the role of expert judgment as a benchmark in exploratory studies of AI-supported assessment (Kizilcec et al., 2024). ChatGPT (GPT-5.2, Free version) was positioned as a synthetic rater whose scoring characteristics and feedback outputs were compared to authentic teacher assessment.

The data corpus consisted of 17 written student response artifacts (n = 17) collected from regular classroom assignments. All responses were anonymized prior to analysis. Although the sample size does not aim for statistical representativeness, it is methodologically acceptable for exploratory alignment analysis using fixed-rater reliability estimation models (Bui & Barrot, 2025; Shen et al., 2025).

2.2.2. RESEARCH INSTRUMENTS

The research employed five primary instruments to ensure contextual equivalence and traceability of analysis.

Table 1. Research Instruments and Their Methodological Functions

No	Instrument	Function In The Study
1	Teacher assessment rubric	Primary reference for scoring and construct validity
2	ChatGPT assessment prompt	ChatGPT assessment prompt
3	Assessment scores	Numerical consistency analysis
4	Written feedback	Analysis of feedback quality and characteristics
5	Teacher perception instrument	Exploration of AI acceptance and pedagogical utility

The teacher-developed assessment rubric was structurally informed by the national Informatics curriculum guidelines for lower secondary education and aligned with specific learning outcome indicators. Furthermore, the teacher perception instrument was adapted from validated technology acceptance and assessment utility frameworks (e.g., UTAUT and evaluative judgement models) to ensure construct validity. This rubric served as the evaluative anchor ensuring equivalence of scoring logic between teacher and AI (Atasoy & Moslemi Nezhad Arani, 2025).

2.3. ASSESSMENT PROCEDURE AND GENAI PROMPT DESIGN

2.3.1. PARALLEL ASSESSMENT PROCEDURE

A parallel–segregated assessment mechanism was implemented to minimize sequence dependency bias. The teacher first evaluated all student artifacts using the standardized rubric and produced both numerical scores and written feedback. Subsequently, the identical artifact set was evaluated by ChatGPT.

During AI assessment, a strict non-intervention protocol was applied. No human-in-the-loop correction, regeneration, or iterative refinement was conducted. Each artifact was processed once (single-pass execution) to control procedural variability. This design ensured that any identified differences in scores or feedback characteristics could be attributed to assessment mechanism differences rather than procedural inconsistencies.

2.3.2. CHATGPT ASSESSMENT PROMPT DESIGN

Prompt engineering was treated as a controlled methodological instrument rather than an experimental manipulation variable. The prompt was designed according to educational prompt structuring principles emphasizing clarity of role, evaluative boundaries, and output specification (Bhatti, 2026; Korzynski et al., 2023; Shen et al., 2025).

The prompt integrated four structured components:

1. Role specification (ChatGPT as rubric-based assessor)
2. Unified presentation of rubric and student responses
3. Restriction to rubric indicators without external assumptions
4. Output specification (numerical scores and constructive feedback)

All artifacts were evaluated using an identical prompt configuration to ensure replicability and isolate algorithmic scoring characteristics (Bui & Barrot, 2025).

Table 2. Prompt Components and Functional Descriptions

Prompt Component	Function In The Study
Role specification	GenAI as a rubric-based assessment assistant
Input Data	Student responses and assessment rubric
Assesment Process	Evaluation per rubric dimension
Output	Numerical scores and written feedback
Language Style	Objective, constructive, and pedagogical

This prompt design approach aims to enhance alignment between human and AI assessment logic, as recommended in literature on human–AI agreement in rubric-based assessment (Jackaria et al., 2024). By constraining AI interpretive space and positioning the rubric as the sole evaluative reference, the prompt is expected to produce assessment outputs that are not only score-consistent but also pedagogically meaningful, thereby warranting further evaluation through consistency and teacher perception analyses. All automated assessments were conducted using ChatGPT (GPT-5.2 free version), accessed in January 2026. Default system parameters were retained without manual adjustment of sampling temperature or generation constraints to preserve ecological authenticity of platform usage.

2.4. DATA ANALYSIS TECHNIQUES

2.4.1. SCORING CONSISTENCY ANALYSIS

To examine scoring agreement between teacher and ChatGPT, the Intraclass Correlation Coefficient (ICC) was calculated using a two-way mixed-effects model with absolute agreement definition and single measures, denoted as ICC (3,1).

This model was selected because:

1. The raters (teacher and ChatGPT) were fixed rather than randomly sampled.
2. The objective was to measure absolute score agreement rather than relative consistency.

All analyses were conducted using IBM SPSS Statistics (Version 25). Statistical significance was set at $\alpha = .05$, and 95% confidence intervals were reported to indicate reliability precision. Although the sample size ($n = 17$) is modest, the narrow 95% confidence interval for the ICC (0.939–0.995) suggests stable agreement estimates within this bounded dataset. The analysis emphasizes within-sample calibration rather than population-level generalization, consistent with exploratory classroom validation research. Interpretation followed established thresholds (Aydın et al., 2025; Hikmawan et al., 2025):

Table 3. ICC Interpretation Criteria

ICC Value	Interpretation
< 0.50	Poor consistency
0.50–0.75	Moderate consistency
0.75–0.90	Good consistency
> 0.90	Excellent consistency

To detect potential systematic mean differences between raters, a paired-samples t-test was conducted. Prior to conducting the ICC and parametric t-test analyses, the assumption of normality was assessed using the Shapiro-Wilk test, which is appropriate for small sample sizes. The score distributions for both human and GenAI assessments were found to adequately meet the normality assumption, justifying the use of these statistical models. Effect size was calculated using Cohen's d for dependent samples to quantify magnitude of difference (Atasoy & Moslemi Nezhad Arani, 2025; Quah et al., 2024).

2.4.2. FEEDBACK SUBSTANCE ANALYSIS

Written feedback was analyzed using deductive thematic analysis informed by effective feedback frameworks emphasizing clarity of goals, identification of performance gaps, and actionable improvement guidance (Ajjawi et al., 2023; Ling, 2024). Coding categories were derived from rubric dimensions and feedback quality constructs. Each feedback instance (teacher and ChatGPT) served as a unit of analysis. Coding was conducted by the researcher, with systematic documentation through an audit trail to ensure analytical transparency.

2.4.3. TEACHER PERCEPTION ANALYSIS

In response to the second research question, teachers' cognitive reception of algorithmic assessment outputs was explored through methodological triangulation combining semi-structured interviews and written reflection protocols. These complementary data sources were used to capture both articulated professional reasoning and reflective judgment regarding the use of GenAI in rubric-based assessment contexts.

The qualitative data corpus was analyzed using a deductive thematic analysis, guided by three a priori evaluative dimensions derived from assessment and educational technology literature: (1) perceived validity of assessment quality, (2) sense of assessment fairness, and (3) pragmatic pedagogical usefulness (Bearman et al., 2024; Kizilcec et al., 2024; Pitpit et al., 2025). Interview transcripts and written reflections were initially analyzed separately and subsequently compared to identify convergent and divergent thematic patterns across data sources.

Interview and reflection data were analyzed separately and then compared to identify convergent and divergent themes. Memoing and cross-referencing with raw data excerpts were employed to strengthen interpretive credibility.

2.4.4. TRUSTWORTHINESS OF QUALITATIVE ANALYSIS

Trustworthiness was ensured through audit trail documentation, methodological triangulation, and systematic memoing (Nazaretsky et al., 2022). Credibility was strengthened through cross-verification between interview and reflection data. Dependability and confirmability were maintained by sequential documentation of coding decisions. Transferability was addressed through detailed contextual description of instructional setting and rubric framework.

An audit trail was maintained throughout the coding process to document analytic decisions and category refinement. Reflexive memos were recorded to mitigate interpretative bias and preserve analytic transparency.

2.5. ETHICAL CONSIDERATIONS AND RESEARCH VALIDITY

This study was conducted in accordance with established ethical principles of educational research. All data collection procedures were carried out following formal approval from the participating school and professional consent from the Informatics teacher involved in the study. Student participation was indirect and limited to the use of learning artifacts generated during regular instructional activities, without additional intervention beyond standard classroom practice.

To protect student privacy, all response data were anonymized prior to analysis, with personal identifiers and potentially identifying information removed. Data were used exclusively for academic research purposes and analyzed in aggregate form. No individual assessment results were utilized for formal academic evaluation or administrative decision-making.

To minimize assessment bias, the study employed standardized assessment rubrics, parallel assessment procedures between teacher and Generative Artificial Intelligence (GenAI), and excluded researcher intervention during evaluation of student learning outcomes. This approach was designed to preserve data integrity, analytical objectivity, and ecological validity of the research findings (Kizilcec et al., 2024).

3. RESULT AND ANALYSIS

3.1. RESULT

This section reports the empirical results of a parallel rubric-based assessment conducted by an Informatics teacher and a Generative Artificial Intelligence (GenAI) system (ChatGPT) on students' open-ended responses. The results are organized into three main components: (1) descriptive score distribution, (2) scoring consistency analysis, and (3) general characteristics of written feedback.

3.1.1. RESEARCH CONTEXT AND PARTICIPANTS

A total of 17 student response artifacts (N = 17) were included in the analysis, with no missing or excluded cases. Each artifact was independently evaluated by the teacher and ChatGPT using an identical rubric with a total score range of 0–16.

Table 4. Descriptive Statistics of Teacher and ChatGPT Assessment Scores

Asessor	Mean Score	Standart Deviation	Minimum Score	Maximum Score
Teacher	11.71	4.09	3	16
ChatGPT	12.12	4.33	3	16

Overall, the score distributions generated by ChatGPT closely approximate those of the teacher in terms of central tendency, dispersion, and score range. Both assessors produced identical minimum and maximum scores, indicating comparable sensitivity in identifying extreme performance levels.

However, ChatGPT produced a slightly higher mean score (M = 12.12) compared to the teacher (M = 11.71), suggesting a potential systematic tendency toward marginally more generous scoring. At the individual level, most score differences were limited to ±1 point, with exact score matches observed in the majority of cases. Discrepancies were primarily observed in responses requiring implicit reasoning or reflective justification, indicating subtle interpretive variation rather than structural misalignment in rubric application.

3.1.2. SCORING CONSISTENCY BETWEEN TEACHER AND CHATGPT

To examine the degree of absolute agreement between teacher and ChatGPT scores, an Intraclass Correlation Coefficient (ICC) analysis was conducted using a two-way mixed-effects model with absolute agreement and single measures, denoted as ICC (3,1). This model is appropriate when the raters are fixed and the objective is to measure absolute scoring agreement (Atasoy & Moslemi Nezhad Arani, 2025).

Table 5. ICC Analysis Results between Teacher and ChatGPT Assessments

Parameter	Value
ICC (Single Measures)	0.985
95% Confidence Level	0.939 – 0.995
F-value	184.462

df1 / df2	16 / 16
Significance (p-value)	< .001
Consistency Interpretation	Excellent

The ICC value of 0.985 indicates excellent absolute agreement, exceeding established thresholds for high reliability in rubric-based evaluation contexts (Martin et al., 2025; Quah et al., 2024). Given the two-rater fixed design, the high single-measure ICC suggests minimal variance attributable to rater differences relative to between-student score variability. The narrow confidence interval further demonstrates the stability of the observed agreement across the dataset. These findings suggest strong procedural alignment between teacher scoring and ChatGPT outputs when operating under explicit rubric constraints.

3.1.3. MEAN DIFFERENCE ANALYSIS

Although agreement was exceptionally high, agreement does not preclude systematic differences in mean scoring. To detect potential bias, a paired-samples t-test was conducted. The analysis revealed a statistically significant mean difference between teacher and ChatGPT scores:

$$t(16) = -2.746,\ p = .014.$$

This pattern reflects a systematic calibration tendency rather than random scoring variation, indicating structured interpretative alignment under rubric constraints. The mean difference was 0.41 points (ChatGPT higher), indicating a modest but statistically significant calibration tendency. Effect size was calculated using Cohen's d for dependent samples:

$$d = 0.67.$$

According to conventional benchmarks, this represents a moderate effect size. These results indicate that while absolute agreement is very high, ChatGPT demonstrates a small but statistically meaningful tendency toward higher scoring. Therefore, the alignment observed through ICC reflects strong relative agreement in scoring patterns, but not perfect equivalence in central tendency. This distinction is methodologically important: high reliability does not eliminate the possibility of systematic calibration differences between human and AI assessors.

3.1.4. GENERAL CHARACTERISTICS OF WRITTEN FEEDBACK

Beyond numerical scores, qualitative inspection of written feedback revealed distinct but complementary characteristics between teacher and ChatGPT outputs. ChatGPT-generated feedback demonstrated strong structural consistency, typically characterized by:

1. Explicit reference to rubric indicators,
2. Transparent justification of assigned scores, and
3. Neutral, criteria-oriented language emphasizing performance alignment.

This pattern supports traceability and auditability, which are central to fairness-oriented assessment frameworks (Ajjawi et al., 2023; Bearman et al., 2024). In contrast, teacher-generated feedback exhibited greater contextual nuance and pedagogical depth. Teachers frequently incorporated references to student misconceptions, prior instructional experiences, and broader learning trajectories. For instance, where ChatGPT provided structurally rigid feedback such as, *"The response correctly identifies the algorithm steps but lacks a detailed explanation of the iteration process,"* the teacher provided instructionally anchored feedback: *"You correctly identified the algorithm, but remember our discussion last week about loops. If the condition is never met, what happens to the program? Try to trace the loop step-by-step."* Feedback often extended beyond rubric descriptors to include interpretive commentary grounded in classroom context.

3.2. ANALYSIS

3.2.1. HUMAN–AI ALIGNMENT IN RUBRIC-BASED SCORING

The exceptionally high ICC (0.985) indicates strong procedural alignment between teacher and ChatGPT scoring under a shared rubric framework. Nevertheless, the stability of this reliability estimate should be interpreted with caution. Given the small sample size and single-teacher context, ICC coefficients derived from bounded datasets may overestimate agreement when generalized to broader instructional settings.

The present findings therefore reflect alignment within a specific evaluative configuration rather than evidence of universal scoring equivalence. When evaluative criteria are explicitly articulated and operationalized, Generative

Artificial Intelligence systems appear capable of consistently applying formal assessment structures in a manner that closely mirrors professional scoring outcomes.

From a theoretical perspective, this finding reinforces the role of rubrics as epistemic scaffolds that constrain interpretive space and reduce ambiguity in open-ended assessment contexts (Bearman et al., 2024; Panadero et al., 2023). Rather than functioning solely as scoring tools, rubrics operate as boundary objects that mediate evaluative reasoning for both human and algorithmic assessors ((Taşçı, 2025).

However, the statistically significant mean difference demonstrates that alignment does not imply calibration equivalence. The moderate effect size suggests that ChatGPT exhibits a consistent scoring bias toward slightly higher scores. This pattern may reflect differences in interpretive strictness, tolerance for partial correctness, or probabilistic weighting of rubric descriptors.

Thus, the observed agreement is more accurately conceptualized as constraint-induced procedural alignment rather than convergence in evaluative cognition. Importantly, this study evaluates cross-agent scoring alignment and does not establish construct validity equivalence between human and AI-based assessment. High agreement reflects similarity in score assignment under shared rubric constraints, but does not confirm that underlying evaluative constructs are interpreted identically. Teacher judgment remains grounded in contextual knowledge, pedagogical experience, and implicit instructional priorities, whereas ChatGPT relies on probabilistic semantic mapping against rubric indicators.

3.2.2. PEDAGOGICAL FIDELITY AND FEEDBACK QUALITY

Although ChatGPT demonstrated strong structural compliance with rubric descriptors, qualitative differences in feedback indicate limitations in pedagogical fidelity. Within assessment-for-learning frameworks, effective feedback must not only describe performance gaps but also support self-regulation and conceptual development (Sortwell et al., 2024). While ChatGPT-generated feedback satisfied criteria of clarity and transparency, it did not consistently embed evaluative commentary within a broader instructional narrative.

Teacher-generated feedback, by contrast, frequently integrated contextual sensitivity and anticipatory instructional intent. Such feedback extends beyond formal rubric compliance and reflects professional discretion shaped by classroom experience (Nazaretsky et al., 2022). Accordingly, high numerical agreement should not be conflated with equivalence in pedagogical reasoning. Procedural consistency in scoring does not guarantee transformative formative impact in feedback delivery.

3.2.3. CONCEPTUAL IMPLICATIONS FOR *ASSESSMENT FOR LEARNING*

Within the assessment-for-learning paradigm, the primary objective of assessment is to enhance learning regulation rather than merely produce summative judgments (Panadero et al., 2023; Sortwell et al., 2024). The findings suggest that ChatGPT holds potential as a consistency-enhancing and efficiency-supporting assessment assistant, particularly in open-ended tasks that impose substantial cognitive and temporal demands on teachers. When deployed under explicit rubric constraints and guided by teacher oversight, GenAI can reduce scoring variability while preserving professional authority.

However, positioning GenAI as an autonomous assessor risks conflating structural alignment with pedagogical adequacy (Vallejo Blanxart & Nicolas Sans, 2025). Overreliance on context-neutral feedback may narrow students' interpretive understanding of learning quality. Therefore, this study advances an integrative conceptualization of GenAI as an augmentative support instrument embedded within human-centered assessment ecosystems, consistent with ethical and pedagogical frameworks in educational technology innovation (Nazaretsky et al., 2022).

4. CONCLUSION

Under a shared analytic rubric, teacher and ChatGPT (GPT-5.2) scores demonstrated very high absolute agreement (ICC = 0.985), indicating that clearly operationalized criteria enable strong cross-agent scoring consistency. In structured conditions, AI can reproduce teacher scoring patterns with substantial precision. Yet reliability should not be conflated with equivalence. Conceptually, this study reframes Human-AI scoring alignment as rubric-mediated procedural convergence rather than cognitive equivalence. The significant mean difference, with AI assigning slightly higher scores, signals a systematic calibration tendency rather than random fluctuation. Procedural alignment, therefore, may coexist with subtle normative divergence in evaluative standards.

Differences become more evident at the feedback level. AI responses are structurally explicit and tightly anchored to rubric descriptors, whereas teacher feedback reflects contextual awareness and pedagogical foresight. Assessment, in this sense, involves not only scoring accuracy but also professional judgment shaped by instructional experience. These findings highlight the rubric's role as an epistemic mediator that enables alignment between human and algorithmic evaluators. At the same time, they underscore the limits of automation in formative contexts. The

evidence supports augmentation rather than substitution: AI may enhance scoring consistency and transparency, but pedagogical authority remains grounded in human expertise.

These conclusions should be interpreted within the exploratory scope of the study design and bounded classroom context. Despite the exploratory nature of this study, a major limitation is the exceptionally small sample size (n=17), alongside a single-teacher design, and one model configuration. Furthermore, the quantitative alignment in this study is based on total aggregated scores; it does not provide a criterion-level analysis to verify whether human-AI alignment is consistent across individual rubric dimensions. Broader multi-rater, cross-context investigations, and dimension-level analyses are required to examine calibration stability and long-term formative implications.

ACKNOWLEDGEMENTS

The author would like to express sincere appreciation to the Informatics teacher who contributed to the assessment process and provided professional insights throughout the implementation of this study. Gratitude is also extended to the school and students whose participation supported the data collection process.

In addition, the author acknowledges the academic support and conducive scholarly environment provided by the Department of Information Systems and Technology Education, Universitas Pendidikan Indonesia, which enabled the systematic conduct of this research. This study did not receive any specific funding from public, commercial, or non-profit funding agencies.

Furthermore, the authors declare the use of Grammarly to assist in structuring the literature review and refining the grammatical flow. However, the conceptual framework, data analysis, and conclusions were independently developed and verified by the authors.

REFERENCES

Ajjawi, R., Bearman, M., Molloy, E., & Noble, C. (2023). The role of feedback in supporting trainees who underperform in clinical environments. In *Frontiers in Medicine* (Vol. 10). Frontiers Media S.A. https://doi.org/10.3389/fmed.2023.1121602

Atasoy, A., & Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students' written texts? *Education and Information Technologies*, *30*(14), 20385–20415. https://doi.org/10.1007/s10639-025-13553-1

Aydın, B., Kışla, T., Elmas, N. T., & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, *133*. https://doi.org/10.1016/j.system.2025.103784

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment and Evaluation in Higher Education*, *49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Bhatti, A. (2026). Strategic Prompt Engineering for Enhancing AI-Generated Content in English Language Teaching Empowering EFL Contexts. *International Journal of Computer-Assisted Language Learning and Teaching*, *16*(1), 1–30. https://doi.org/10.4018/ijcallt.398504

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, *30*(2), 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Hidayatullah, E., Untari, R., & Fifardin, F. (2024). Effectiveness of AI in solving math problems at the secondary school level. *Union: Jurnal Ilmiah Pendidikan Matematika*, *12*(2), 350–360. https://doi.org/10.30738/union.v12i2.17548

Hikmawan, R., Komaru, M., Suherman, A., & Sari, A. P. (2025). *Validity and Reliability of Computational Thinking Scales for Undergraduate Students: Indonesian Adaptation*. https://jurnaldidaktika.org

Jackaria, P. M., Hajan, B. H., & Mastul, A. R. H. (2024). A Comparative Analysis of the Rating of College Students' Essays by ChatGPT versus Human Raters. *International Journal of Learning, Teaching and Educational Research*, *23*(2), 478–492. https://doi.org/10.26803/ijlter.23.2.23

Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Zeivots, S., & Raduescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100269

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(3), 25–37. https://doi.org/10.15678/EBER.2023.110302

Ling, J. H. (2024). Pedagogy: Indonesian Journal of Teaching and Learning Research A Review of Rubrics in Education: Potential and Challenges. *OPEN ACCESS JOURNAL Pedagogy: Indonesian Journal of Teaching and Learning Research*, *2*(1), 1–14.

Martin, P. P., Kranz, D., & Graulich, N. (2025). Revealing Rubric Relations: Investigating the Interdependence of a Research-Informed and a Machine Learning-Based Rubric in Assessing Student Reasoning in Chemistry. *International Journal of Artificial Intelligence in Education*, *35*(3), 1465–1503. https://doi.org/10.1007/s40593-024-00440-y

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated Grading and Feedback Tools for Programming Education: A Systematic Review. *ACM Transactions on Computing Education*, *24*(1). https://doi.org/10.1145/3636515

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, *53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of Rubrics on Academic Performance, Self-Regulated Learning, and self-Efficacy: a Meta-analytic Review. *Educational Psychology Review*, *35*(4). https://doi.org/10.1007/s10648-023-09823-4

Pecuchova, J., Benko, Ľ., & Drlik, M. (2025). Automated Grading of Open-Ended Questions in Higher Education Using GenAI Models. *International Journal of Artificial Intelligence in Education*, *35*(6), 3813–3846. https://doi.org/10.1007/s40593-025-00517-2

Pitpit, L. J., Obenza, B., & Inojales, G. Jr. (2025). Teachers' Attitudes Toward Generative AI In Assessment Planning: A UTAUT-Based Structural Equation Model. *Asia Pacific Journal of Educational Technologies, Psychology, and Social Sciences*, *1*(1), 185–202. https://doi.org/10.70847/622109

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05881-6

Shen, H., Knearem, T., Ghosh, R., Liu, M. X., Monroy-Hernández, A., Wu, T., Yang, D., Huang, Y., Mitra, T., Li, Y., & Hearst, M. (2025, April 26). Bidirectional Human-AI Alignment: Emerging Challenges and Opportunities. *Conference on Human Factors in Computing Systems - Proceedings* . https://doi.org/10.1145/3706599.3716291

Sortwell, A., Trimble, K., Ferraz, R., Geelan, D. R., Hine, G., Ramirez-Campillo, R., Carter-Thuiller, B., Gkintoni, E., & Xuan, Q. (2024). A Systematic Review of Meta-Analyses on the Impact of Formative Assessment on K-12 Students' Learning: Toward Sustainable Quality Education. In *Sustainability (Switzerland)* (Vol. 16, Number 17). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/su16177826

Taşçı, S. (2025). Human and AI Scoring of EFL Writing: The Influence of Rubrics and Genre on Reliability. *Eğitim ve Yeni Yaklaşımlar Dergisi*, *8*(2), 191–210. https://doi.org/10.52974/jena.1785369

Ukkonen, A., Pajchel, K., & Mifsud, L. (2025). Teachers' understanding of assessing computational thinking. *Computer Science Education*, *35*(4), 794–819. https://doi.org/10.1080/08993408.2024.2365566

Vallejo Blanxart, A., & Nicolas Sans, R. (2025). The role of generative AI chatbots in higher education: A student-centric conceptual analysis of benefits, ethics, and privacy concerns. *Journal of Technology and Science Education*, *15*(3), 810. https://doi.org/10.3926/jotse.3643

Williams, A. (2024). Delivering Effective Student Feedback in Higher Education: An Evaluation of the Challenges and Best Practice. *International Journal of Research in Education and Science*, *10*(2), 473–501. https://doi.org/10.46328/ijres.3404

Xing, W., Zhu, T., Wang, J., & Liu, B. (2024). A Survey on MLLMs in Education: Application and Future Directions. In *Future Internet* (Vol. 16, Number 12). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/fi16120467