Journal of Educational Sciences
Journal homepage: https://jes.ejournal.unri.ac.id/index.php/JES

P-ISSN 2581-1657
E-ISSN 2581-2203

A Rubric-Integrated Assessment System Using a Large Language Model for Automated Essay Evaluation in Secondary Vocational Schools

Rochman Bambang Eko Saputro, Rizki Hikmawan*
Information System and Technology Education, Universitas Pendidikan Indonesia, Bandung, 40154, Indonesia

ARTICLE INFO

Article history:

Received: 11 Dec 2025
Accepted: 28 April 2026
Available Online: 05 May 2026

Keywords:

Artificial Intelligence,
Large Language Model,
Vocational Education,
Automated Assessment,
Competency-Based Rubric

* *Corresponding author:*

E-mail: hikmariz@upi.edu

Article Doi:

https://doi.org/10.31258/jes.10.5.p.11-23

A B S T R A C T

This study aims to develop and evaluate an artificial intelligence-based essay assessment system integrated with the National Work Competency Standard (SKKNI) in vocational Informatics education. The system was designed using the ADDIE model, featuring two core functions: (1) automated rubric generation based on indicators, criteria, and weights provided by teachers, and (2) automated essay scoring by a multimodal large language model capable of processing both image and text inputs. Evaluation was conducted on ten student essays assessed in parallel by three experienced teachers and the AI system. Reliability analysis using the Intraclass Correlation Coefficient (ICC) revealed a score of 0.975, classified as "excellent," indicating strong agreement between AI-generated scores and human ratings. Qualitative findings from teacher interviews confirmed that the system not only reduces administrative burden but also reinforces the teacher's role as a pedagogical curator, ensuring assessment remains aligned with learning objectives. The key conclusion is that integrating artificial intelligence within a pedagogically centered process (rather than replacing educators) yields a reliable, valid, and sustainable approach to automated essay scoring in vocational education.

1. Introduction

Essay-based assessment is a rich form of evaluation that captures students' higher-order thinking skills; however, it traditionally faces significant challenges related to teachers' limited time, variability in grading subjectivity, and delayed feedback (Legi et al., 2024; Matsukawa & Iwasaki, 2024; Baihaqi, 2021). The digitalization of education has fundamentally changed teachers' work habits, requiring technological adaptation to support professionalism while also creating new challenges related to workload that need to be managed efficiently (Hilhamsyah et al., 2024). These challenges constitute a major barrier to the sustainability of

assessment practices, particularly in vocational education settings burdened by high administrative loads and limited resources. In this context, the adoption of Large Language Models (LLMs) is expected to serve as a strategic solution to alleviate instructor workload and support more sustainable assessment practices (Agostini & Picasso, 2024; Fagbohun et al., 2024). Nevertheless, despite advances in AI technology that have opened avenues for automation, most existing systems remain fragmented, focusing solely on numerical score computation without integrating the pedagogical frameworks that underpin teacher assessment, such as indicators, criteria, and national competency standards (Durak & Onan, 2025; Hammad et al., 2024).

Evaluations of LLM effectiveness in essay assessment reveal inconsistent outcomes. Some studies report only moderate correlations between AI scores and human raters, suggesting that while AI can produce scores that appear valid, it remains vulnerable to bias, inconsistency, and a lack of transparency in its justification of judgments (Kooli & Yusuf, 2024; Rony et al., 2025). The study by Beaulieu-Jones et al. (2024) demonstrated that GPT-4 generates different responses when queried repeatedly, highlighting long-term stability issues that threaten trust in automated systems. Conversely, research by Rony et al. (2025) revealed that although GPT-4o produces scores aligned with human raters, the qualitative reasoning behind those scores is often inconsistent, indicating that numerical outputs do not always reflect a defensible assessment process. The validity and reliability of LLM-based assessment are highly dependent on the alignment between the assessment instrument (rubric) and the linguistic and pedagogical context of the learners, as emphasized by Pack et al. (2024) in the context of assessing English language learners' writing.

In vocational education, the use of AI to support competency-based assessment remains in its early stages (Çela et al., 2025). Efforts to automate rubric generation by Dockens and Shelton (2025) and Sholapurapu and Sayed (2025) have not systematically linked the rubric generation process to official competency standards such as the Indonesian National Work Competency Standard (SKKNI), resulting in outputs that are often irrelevant to local curricular needs. A systematic review by Yan et al. (2023) identified that the adoption of LLMs in education continues to be hindered by practical challenges (including low technological readiness, insufficient transparency, and privacy risks) and recommends a human-centered approach as a prerequisite for sustainability. Furthermore, Hammad et al. (2024), in their systematic review, found that only a small fraction of AI-based assessment studies involves teachers in the rubric design cycle, and almost none integrate rubric generation with essay assessment within a unified platform. This gap creates a disconnection between technological potential and classroom reality, where teachers remain the sole assessors capable of holistically integrating context, criteria, and competencies. Recent studies also indicate that AI integration in education is most effective when it is supported by appropriate pedagogical design, contextual adaptation, and digital literacy development (Wardhani et al., 2025; Yozaga et al., 2026; Sahputra et al., 2025).

Therefore, this study designs an AI-based system that integratively (1) assists teachers in automatically generating assessment rubrics based on indicators, criteria, weights, and direct mapping to national competency standards, and (2) uses teacher-validated rubrics as the foundation for evaluating student essays and delivering relevant, constructive personalized feedback. This direction is also supported by recent studies showing that AI in education becomes more meaningful when it is embedded in structured instructional design and adapted to specific learning contexts, including vocational settings (Rahayu et al., 2025; Sukma & Hikmawan, 2026). The objective of this research is to design an integrated prototype system and evaluate its effectiveness through a comparison of assessment outcomes between the AI system and three human raters using the same rubric on a set of student essays. This study aims to provide a technological solution that not only reduces teachers' administrative burden (Fagbohun et al., 2024; Hilhamsyah et al., 2024) but also strengthens the consistency, transparency, and pedagogical validity of competency-based assessment, while maintaining the teacher's role as the ultimate curator of the assessment process.

2. Methodology

This study employs a Research and Development (R&D) approach based on the ADDIE model (Analysis, Design, Development, Implementation, Evaluation). In the Analysis phase, the needs of Informatics teachers were identified regarding time constraints, assessment consistency, and alignment with the Indonesian National Work Competency Standard (SKKNI) for the Informatics field. The Design phase encompassed the design of the system architecture, user interface, and the integration workflow between teacher inputs, automated rubric generation, and multimodal essay assessment. In the Development phase, a prototype system was developed with two core functions: (1) assisting teachers in automatically generating assessment rubrics based on indicators, criteria, weights, and mapping to SKKNI, which can then be reviewed and adjusted by the teacher; and (2) automatically assessing student essay responses using the multimodal large language model GPT-4. Prior to main implementation, the system underwent pilot testing involving two teachers and five student essay samples. The pilot testing results demonstrated that the system functioned effectively: text extraction from documents and images operated with high accuracy (>95%), the rubric generation process accurately reflected the input SKKNI criteria, and the AI-generated feedback was relevant to the specified indicators. Based on these results, the system was deemed ready for main implementation.

In the Implementation phase, ten students uploaded their digital case analysis assignment responses in file formats (PDF, Word, or image). The assessment process began when the system validated the file format and size, followed by storage of the file on the server. Subsequently, the system retrieved the assessment rubric from the database, constructed a prompt containing the rubric and assessment instructions, and embedded the student's response image into the prompt (in base64 or URL format). This prompt was then sent to the GPT-4 API, a multimodal model capable of processing both text and images simultaneously. GPT-4 processed the

student's response based on the rubric, generating a score and personalized feedback, which was then received by the system, stored in the database, and displayed on the student and teacher dashboard, with the process illustrated in Figure 1.

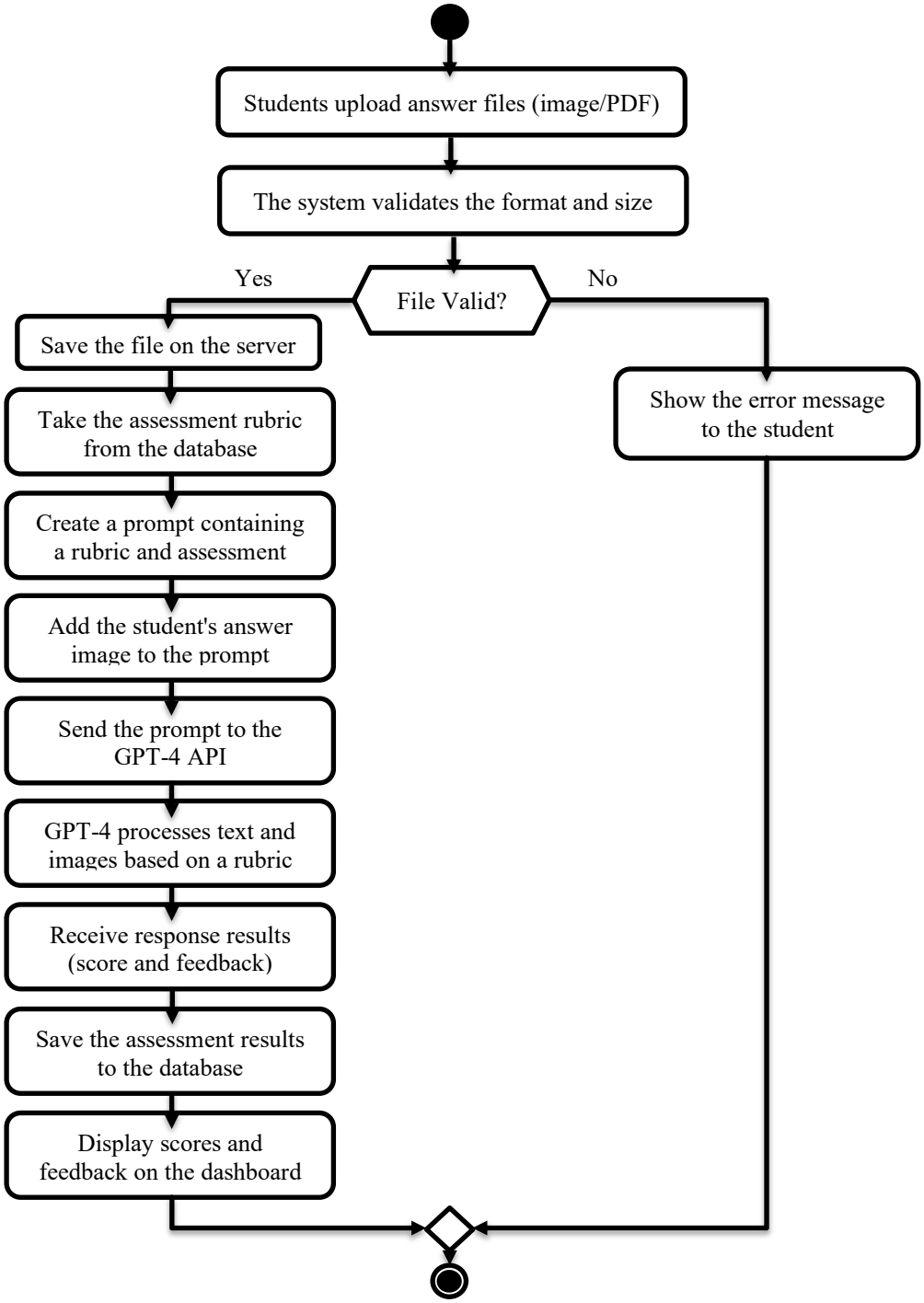

Figure 1. Workflow of the Essay Assessment System

To evaluate the impact of the rubric on assessment consistency, this study implemented two assessment conditions. Condition A (with rubric): three teachers and the AI system assessed ten essays using a rubric previously agreed upon and

mapped to SKKNI. Condition B (without rubric): the same three teachers and the AI system assessed the same ten essays without using a rubric, relying solely on a general instruction: "Grade this essay holistically on a scale of 0–100." In total, 50 assessments were conducted. The selection of the GPT-4 model was based on the findings of Pack et al. (2024), which demonstrated that GPT-4 exhibited the best performance in assessing language learners' essays, with high validity and interrater reliability. To ensure traceability between the experimental design and the reported findings, the two conditions were applied to the same set of ten essays and involved the same three teachers and the AI system. This parallel design allowed direct comparison of rating consistency across rubric-guided and non-rubric scoring contexts while minimizing variation caused by differences in essay content or assessor composition.

In the Evaluation phase, the system's effectiveness was assessed both quantitatively and qualitatively. Quantitative analysis focused on interrater reliability between the AI system and the average scores of the three teachers under both conditions. The Intraclass Correlation Coefficient (ICC) was calculated using SPSS software, employing a two-way random effects model with consistency type, in accordance with current best practices in reliability research (Liljequist et al., 2019). ICC values were interpreted according to the categories developed by Koo and Li (2016), as presented in Table 1. In addition, qualitative data were collected through semi-structured interviews with the three teachers to evaluate their perceptions regarding the clarity, consistency, usefulness, and trustworthiness of the system in supporting daily assessment practices, including a comparison between assessment with and without the rubric.

Table 1. ICC Score Categories

ICC Score Range	Category
< 0.50	Poor
0.50 – 0.75	Moderate
0.75 – 0.90	Good
> 0.90	Excellent

3. Results and Discussion

This study aimed to develop and evaluate an artificial intelligence-based assessment system integrated with rubrics in the context of vocational Informatics education. The system was designed using the ADDIE model, with two core functions: (1) assisting teachers in automatically generating assessment rubrics based on the Indonesian National Work Competency Standard (SKKNI), and (2) automatically evaluating student essay responses using GPT-4. Evaluation was conducted on ten student essays assessed in parallel by three human raters and the AI system under two conditions: (1) with a validated rubric and (2) without a rubric, relying solely on a general holistic instruction ("Score this essay on a scale of 0–100"). The results are presented first, followed by a comprehensive analysis that contextualizes the findings within current literature and pedagogical theory.

The interrater reliability between the AI system and the average scores assigned by the three human raters was analyzed using the Intraclass Correlation Coefficient (ICC) with a two-way random effects model. The results were clear and meaningful. Under the rubric-guided condition, the ICC reached 0.975, indicating an exceptionally high level of agreement between AI and human raters. In the rubric-free condition, the ICC decreased to 0.846, which remains within the "good" range but is substantially lower than the rubric-guided condition. According to the classification framework developed by Koo and Li (2016), an ICC of 0.975 is classified as "excellent," whereas 0.846 falls into the "good" category. This clear divergence underscores the significant impact of a well-structured, teacher-validated rubric on the consistency of AI-driven assessments, as summarized in Table 2.

Table 2. Comparison of ICC Scores

Assessment Condition	ICC Score	Category
With Rubric	0.975	Excellent
Without Rubric	0.846	Good

To strengthen the traceability between the experimental design and the reported ICC values, Table 3 presents the essay-level score distribution for each rater under the rubric-guided and non-rubric conditions. In the rubric-guided condition, the AI scores generally remained close to the score range assigned by the three teachers. Essay 1 showed identical scoring between the AI and all teacher raters, while Essays 2, 7, 8, 9, and 10 showed relatively small differences. By contrast, in the non-rubric condition, the AI scores were consistently higher than the teacher scores across nearly all essays, with especially visible deviations in Essays 4, 5, 6, 8, and 10. This distribution pattern provides a more concrete explanation of how the two experimental conditions produced different levels of interrater agreement. More specifically, the score distribution shown in Table 3 constitutes the empirical basis from which the ICC values were derived. The tighter clustering of scores in the rubric-guided condition, compared with the wider deviations observed in the non-rubric condition, indicates that rubric use contributed directly to greater scoring consistency across raters.

Table 3. Score Distribution of Human Raters and AI

Essay	Teacher 1	Teacher 2	Teacher 3	AI With Rubric	AI Without Rubric
1	100	100	100	100	92
2	90	85	90	88.75	98
3	95	93	94	90	99
4	88	90	87	85	97
5	92	90	91	85	98
6	84	86	83	80	97
7	97	95	96	98	99
8	80	78	79	80	96
9	93	90	92	95.5	97
10	87	88	86	85.25	96

In terms of system output, Figure 2 presents the AI-generated assessment for one student, revealing a total score of 88.75 out of 100. This score was calculated based on weighted indicators, with detailed feedback provided for each criterion: "Relevance to Topic" (score: 4/4, weight: 20), "Clarity and Coherence of Writing" (score: 3/4, weight: 15), "Logical Essay Structure" (score: 3/4, weight: 15), "Creativity and Relevance of Proposed Solutions" (score: 3/4, weight: 18.75), and "Understanding of Concepts Related to Misinformation or Cyberbullying" (score: 4/4, weight: 20). Additionally, the system delivered a holistic, constructive comment: "The essay clearly articulates the dangers of cyberbullying with practical solutions and a deep understanding of the underlying concepts. To achieve a perfect score, improve transitions between sections, deepen the analysis of impacts on victims, and explore more creative solutions with brief case examples." This demonstrates that the system does not merely generate numerical scores but also provides personalized, actionable feedback that supports student learning.

Kriteria	Skor (1–4)	Nilai Berbobot	Alasan
Kesesuaian dengan Tema	4	20	Esai secara jelas menekankan bahaya dan upaya pencegahan cyberbullying sesuai tema yang dipilih, menunjukkan pemahaman mendalam terhadap isu tersebut.
Kerapihan dan Kejelasan Tulisan	3	15	Tulisan cukup rapi dan jelas secara umum, dengan beberapa kesalahan minor yang tidak mengganggu pemahaman.
Struktur Essay yang Logis dan Sesuai	3	15	Teks memiliki pendahuluan, isi, dan penutup yang ada serta alur ide cukup terhubung, meski beberapa transisi bisa lebih menguat.
Kreativitas dan Relevansi Solusi yang Ditawarkan	3	18.75	Solusi praktis dan relevan, namun belum sangat inovatif; beberapa gagasan bisa dikembangkan lebih kreatif.
Pemahaman terhadap Konsep Informasi Hoax atau Cyber Bullying	4	20	Pemahaman sangat baik mengenai konsep hoax/cyberbullying dan dampaknya, dengan contoh-contoh yang relevan.
Total Skor		**88.75 / 100**	
Catatan Keseluruhan		Total skor menunjukkan prestasi yang sangat baik. Esai telah menguraikan bahaya cyberbullying secara jelas, dengan solusi praktis dan pemahaman mendalam tentang konsep terkait. Untuk mencapai skor 100, tingkatkan transisi antar bagian, kembangkan analisis dampak lebih dalam pada korban, dan eksplorasi ide-ide solusi yang lebih kreatif serta studi kasus singkat.	

Figure 2. AI-generated Assessment Output for a Single Student

To clarify how teacher oversight was maintained in the system, the finalized rubric is presented in Figure 3. Figure 3 illustrates the rubric generated by the AI and subsequently reviewed and validated by the teachers before being used in the scoring process. The rubric includes five core indicators: (1) content relevance, (2) clarity and organization of writing, (3) structural coherence, (4) creativity and relevance of responses, and (5) understanding of the topic. The teachers reviewed and calibrated the performance descriptors for each level (e.g., "Excellent," "Good," "Sufficient," "Poor") and adjusted the weightings of each indicator. This review process confirms that the system functions not as a replacement, but as an assistant, ensuring that the final rubric retains strong pedagogical validity and alignment with local curriculum standards.

The qualitative insights from semi-structured interviews with the three teachers added an invaluable human dimension to these quantitative findings. Prior to adopting the system, teachers described essay assessment as an exhausting and often inequitable administrative burden. "We teach multiple classes with hundreds of students. Correcting essays one by one takes hours, often until late at night. Sometimes, due to fatigue or time constraints, we simply look at the length of the writing or whether the student attended class regularly, and give a safe score," one teacher explained. Although well-intentioned, this practice risks overlooking the

depth of student thinking and reinforces subjective bias, a practice fundamentally at odds with the principles of competency-based assessment.

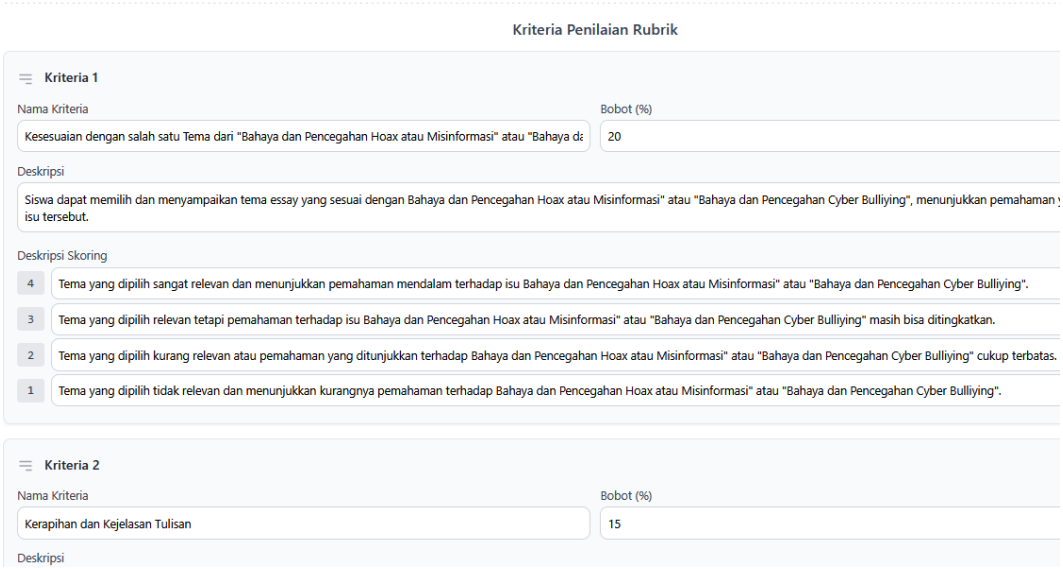

Figure 3. AI-Generated Rubric Reviewed and Validated by Teachers

The central finding of this study that interrater reliability increased from ICC = 0.846 (without rubric) to ICC = 0.975 (with rubric) strongly supports the hypothesis that a clear, structured, and teacher-validated rubric is a critical determinant of high consistency in LLM-based assessment. The 0.128 difference in ICC values demonstrates that even highly capable models like GPT-4 remain vulnerable to variability when not anchored in an explicit pedagogical framework. This aligns with the findings of Seo et al. (2025), who found that the reliability of LLM evaluators is strongly influenced by the level of agreement among human raters; when human raters themselves lack consistency, the reliability of AI evaluations tends to decline. In other words, the quality of human judgment serves as a prerequisite for successful AI integration, not merely its target.

The ICC of 0.975 under the rubric-guided condition is consistent with other studies employing similar methodologies. Hackl et al. (2023) reported ICC values between 0.94 and 0.99 for GPT-4 in evaluating university-level macroeconomics responses, attributing this high reliability to the use of clearly structured prompts and explicit evaluation criteria, reinforcing the notion that technical precision in LLMs depends heavily on the clarity of the assessment framework. Similarly, Yavuz et al. (2025) found an ICC of 0.972 for a fine-tuned ChatGPT model assessing EFL essays using an analytical rubric, demonstrating that consistent results can be replicated across different models when grounded in sound pedagogy. This study extends these findings to the Indonesian vocational education context, where linguistic diversity and heterogeneous learning styles often pose significant challenges to manual assessment. Furthermore, studies in medical education (Sreedhar et al., 2025) and physiology (Teckwani et al., 2024) confirm that LLMs can accurately assess student learning outcomes, not because of inherent technical superiority, but because they

are guided by rubrics that are explicitly aligned with domain-specific competencies and validated by subject-matter experts.

However, this study does not overlook the well-documented challenges of LLM consistency. Pack et al. (2024) found that although GPT-4 exhibits excellent intrarater reliability, its interrater reliability slightly declines across repeated sessions, suggesting that LLM outputs are not fully deterministic. Beaulieu-Jones et al. (2024) corroborated this, showing that GPT-4 produced different responses to identical queries in 36.4% of cases, highlighting a significant risk to trustworthiness in high-stakes contexts. Rony et al. (2025) added a critical nuance: while quantitative scores from models like GPT-4o may align closely with human raters, the qualitative reasoning behind those scores often lacks consistency, a finding directly relevant to our context. In our study, minor deviations between AI and teacher scores likely reflect differing evaluation paradigms: teachers may tolerate structural imperfections if the argument is strong and contextually rich, whereas the AI, even when guided by a rubric, relies on linguistic patterns learned from training data and may not fully grasp pedagogical nuance. Nevertheless, because these deviations were statistically insignificant and remained within the "excellent" reliability range, the system remains a reliable assistant, not a replacement.

The system's performance is further supported by its hybrid workflow, which positions the teacher as the final curator. This approach aligns with the human-in-the-loop principle in Artificial Intelligence in Education (AIED), where AI augments rather than replaces professional expertise (Çela et al., 2025; Nugraha & Harsono, 2024). By automating the labor-intensive tasks of scoring and initial feedback generation, the system frees teachers to focus on deeper pedagogical activities: providing nuanced feedback, supporting struggling learners, and designing personalized learning interventions consistent with the human-centered recommendations of Yan et al. (2023). The qualitative data from teacher interviews reinforce this: before the system, correcting essays was a time-consuming chore, often leading to score inflation based on non-academic factors like length or attendance. With the system, the process shifted from individual essay correction to rubric review, reducing assessment time from several hours to just 15–20 minutes per task. More importantly, teachers reported that the system enabled them to deliver personalized feedback to all students, something previously impossible due to scale. This allows educators to focus on understanding individual student needs and tailoring instruction, rather than merely assigning scores. This finding is also consistent with recent studies, showing that AI contributes most meaningfully to education when it supports instructional design, teacher facilitation, and adaptive learning resources rather than functioning as an autonomous substitute for educators (Rahayu et al., 2025; Sahputra et al., 2025).

Nevertheless, a high ICC does not guarantee full construct validity. As Yan et al. (2023) caution, LLMs still face challenges related to transparency, bias, and sustainability. AI may recognize "what is correct" but often fails to understand "why it is correct", a limitation only mitigated through human oversight. Furthermore, Rony et al. (2025) and Seo et al. (2025) remind us that not all LLMs, even within the same family, demonstrate equal consistency. Model selection must

be deliberate, and systems must be designed to allow for ongoing audit and review. Fagbohun et al. (2024) emphasize that without human supervision, AI systems risk reinforcing undetected biases, such as rewarding verbose or keyword-heavy responses over those that demonstrate deeper, more original thinking.

Despite the promising findings, the results should be interpreted with caution because the study involved only ten student essays. Such a small sample may have produced limited variability in essay quality and score dispersion, which increases the risk of overestimating agreement and may inflate the observed ICC values. In addition, the statistical stability of the reliability estimates has not yet been tested across larger or more heterogeneous samples. Therefore, the present findings should be interpreted as preliminary evidence rather than definitive proof of system reliability, and they remain subject to limited generalizability across different schools, subjects, grade levels, and assessment tasks.

This limitation also has direct implications for future research. Subsequent studies should involve a larger number of essays, more diverse student profiles, and broader disciplinary contexts in order to test the robustness of the system under more realistic classroom conditions. Future work should also examine repeated scoring sessions, cross-topic stability, and variations across different rubric structures to determine whether the consistency observed in this study remains stable over time. Such extensions would strengthen both the empirical generalizability and the practical credibility of AI-assisted assessment in vocational education. Future studies should also test the system on datasets with greater variability to determine whether the observed agreement remains stable under less homogeneous scoring conditions.

Collectively, these findings demonstrate that the integration of SKKNI, teacher-validated rubrics, and GPT-4 creates a robust, efficient, and pedagogically sound assessment workflow. This system not only reduces administrative burden (Fagbohun et al., 2024) but also enhances the consistency and transparency of competency-based assessment. These results provide empirical evidence that the success of AI in assessment does not lie in technological autonomy, but in the deliberate design of systems that preserve the teacher's central role as the guardian of pedagogical quality. Thus, this system is not merely an automation tool, it is a collaborative mechanism that restores the true meaning of assessment: not merely assigning scores, but building understanding, strengthening learning, and honoring the professional expertise of educators.

4. Conclusion

This study concludes that a rubric-integrated AI assessment system can support essay evaluation in vocational Informatics education when it is implemented within a teacher-centered pedagogical workflow. The system was designed not to replace teachers, but to assist them in generating structured rubrics, evaluating essays more consistently, and providing constructive feedback more efficiently. The findings indicate that rubric-guided assessment creates stronger alignment between AI-

assisted scoring and human judgment than holistic scoring without a shared analytical framework. This suggests that the success of AI in educational assessment depends not only on model capability, but also on the quality of the pedagogical design that guides its use. However, the present findings should be treated as preliminary because the study was conducted on a limited sample, which may have reduced score variability and restricted the generalizability of the reliability estimates. Therefore, further research is needed to test the system on broader datasets, different subjects, and varied educational settings. Overall, the study shows that AI can function responsibly and meaningfully in assessment when teachers remain the final pedagogical decision-makers.

References

Agostini, D., & Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework. *Intelligenza Artificiale, 18*(1), 121–138. https://doi.org/10.3233/ia-240033

Baihaqi, M. I. (2021). Assesmen penerapan belajar dari rumah. *JKTP: Jurnal Kajian Teknologi Pendidikan, 4*(4), 408-416.

Beaulieu-Jones, B. R., Berrigan, M. T., Shah, S., Marwaha, J. S., Lai, S.-L., & Brat, G. A. (2024). Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. *Surgery, 175*(4), 936–942. https://doi.org/10.1016/j.surg.2023.12.014

Çela, E., Vajjhala, N. R., Eappen, P., & Vedishchev, A. (2025). *Artificial intelligence in vocational education and training*. In *Transforming vocational education and training using AI*. USA: IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-8252-3.ch001

Dockens, A. L., & Shelton, K. (2025). AI for formative and summative assessment: A balanced approach. In *Emerging Trends, Global Perspectives, and Systematic Transformation in AI*. USA: IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-5102-5.ch013

Durak, H. Y., & Onan, A. (2025). A systematic review of AI-based feedback in educational settings. *Journal of Computational Social Science, 8*(4). https://doi.org/10.1007/s42001-025-00428-1

Fagbohun, O., Iduwe, N. P., Abdullahi, M., Ifaturoti, A., & Nwanna, O. M. (2024). Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *Journal of Artificial Intelligence, Machine Learning and Data Science, 2*(1), 1–8. https://doi.org/10.51219/jaimld/oluwole-fagbohun/19

Hackl, V., Müller, A. E., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education, 8*(1272229). https://doi.org/10.3389/feduc.2023.1272229

Hammad, M. M., Al-Refai, M., Musallam, W., Musleh, S., & Faouri, E. (2024). A taxonomy of AI-based assessment educational technologies. In *Proceedings of the 2024 15th International Conference on Information and Communication Systems (ICICS)*, 1–6. https://doi.org/10.1109/ICICS63486.2024.10638295

Hilhamsyah, Hidayati, D., & Imama, M. L. (2024). Teacher habits and workload in the digitalization of education. *JKTP: Jurnal Kajian Teknologi Pendidikan, 7*(4), 195-208.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Kooli, C., & Yusuf, N. (2024). Transforming Educational Assessment: Insights into the Use of ChatGPT and Large Language Models in Grading. *International Journal of Human-Computer Interaction*, 1–12. https://doi.org/10.1080/10447318.2024.2338330

Legi, M., Sengkey, D. F., & Sambul, A. M. (2024). Aplikasi Kecerdasan Artifisial Generatif Untuk Asesmen Pembelajaran Berdasarkan Rubrik. *Jurnal Teknik Informatika, 19*(3), 183–192. https://doi.org/10.35793/jti.v19i3.53916

Liljequist, D., Elfving, B., & Roaldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLOS ONE, 14*(7). https://doi.org/10.1371/journal.pone.0219854

Matsukawa, H., & Iwasaki, C. (2024). Development of a formative assessment system for reports using large language models and rubrics. In *Proceedings of the 2024 International Symposium on Educational Technology (ISET)*, 34–38. https://doi.org/10.1109/iset61814.2024.00016

Nugraha, C. A., & Harsono, R. J. (2024). Chemistry Storytelling with AI: A Descriptive Research of Ebook Creation for Grade 10 Students. *JKTP: Jurnal Kajian Teknologi Pendidikan, 7*(3), 149-160.

Pack, A., Barrett, A., & Escalante, J. (2024). Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Computers and Education: Artificial Intelligence, 6*, 100234–100234. https://doi.org/10.1016/j.caeai.2024.100234

Rahayu, S., Hakim, A. R., & Fitrianingsih, N. (2025). Development of Teaching Materials using Ai-Based Teachy App at Pela Elementary School. *Journal of Educational Sciences, 9*(5), 3788–3799. https://doi.org/10.31258/jes.9.5.p.3788-3799

Rony, S., Fei, T., & Arsovski, S. (2025). Educational justice. Reliability and consistency of large language models for automated essay scoring and its implications. *Journal of Applied Learning and Teaching, 8*(1). https://doi.org/10.37074/jalt.2025.8.1.21

Sahputra, D., Sari, S. M., & Syarfuni. (2025). Utilization of AI In Platform-Based Learning to Improve the Quality of Education in Banda Aceh. *Journal of Educational Sciences, 9*(5), 4461–4470.

Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., & Jung, S. (2025). Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences, 15*(2), 671. https://doi.org/10.3390/app15020671

Sholapurapu, P. K., & Sayed, M. Y. (2025). *AI-driven student feedback systems: Implementing machine learning models for personalized assessment and learning pathways*. In *Artificial Intelligence-Powered Learning Analytics and Student Feedback Mechanisms for Dynamic Curriculum Enhancement and Continuous Quality Improvement in Outcome-Based Education*. India:

RADemics Research Institute Publication. https://doi.org/10.71443/9789349552531-06

Sreedhar, R., Chang, L., Gangopadhyaya, A., Shiels, P. W., Loza, J., Chi, E., Gabel, E., & Park, Y. S. (2025). Comparing scoring consistency of large language models with faculty for formative assessments in Medical Education. *Journal of General Internal Medicine, 40*(1), 127–134. https://doi.org/10.1007/s11606-024-09050-9

Sukma, A. A., & Hikmawan, R. (2026). Development and Effectiveness of an Interactive Prompt Engineering E-Module to Enhance Vocational Students' Learning Outcomes. *Journal of Educational Sciences, 10*(3), 772–790. https://doi.org/10.31258/jes.10.3.p.772-790

Teckwani, S. H., Wong, A. H.-P., Luke, N. V., & Low, I. C. C. (2024). Accuracy and Reliability of Large Language Models in Assessing Learning Outcomes Achievement Across Cognitive Domains. *Advances in Physiology Education, 48*(4), 904–914. https://doi.org/10.1152/advan.00137.2024

Wardhani, I. S., Indrawati, I., & Asyiah, I. N. A. (2025). AI (Artificial Intelligence)-Enhanced Digital Literacy in Science Learning: A Systematic Literature Review. *Journal of Educational Sciences, 9*(4), 1957–1970. https://doi.org/10.31258/jes.9.4.p.1957-1970

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *British Journal of Educational Technology: Journal of the Council for Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology: Journal of the Council for Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Yozaga, B. T., As Zahro, M. A. N., Adianingsih, P., Murtiyasa, B., & Masduki. (2026). The use of artificial intelligence in higher education learning: A systematic review of its effectiveness and challenges in implementation. *Journal of Educational Sciences, 10*(3), 852–860. https://doi.org/10.31258/jes.10.3.p.852-860

How to cite this article:

Saputro, R. B. E., & Hikmawan, R. (2026). A Rubric-Integrated Assessment System Using a Large Language Model for Automated Essay Evaluation in Secondary Vocational Schools. *Journal of Educational Sciences*, *10(5)*, 11-23.