

A Comparative Study of Human and AI Assessment in Japanese Sakubun Writing

Jeni Putra^{1*}, Masashi Umemoto², Susi Widiyanti³, Novia Hayati⁴, Muthi Afifah⁵

^{1,3,4,5}Universitas Pendidikan Indonesia, Bandung, Indonesia

²Nagasaki University of Foreign Studies, Nagasaki, Japan

Received: 20-08-2026; Revised: 01-10-2026; Accepted: 09-10-2026; Available Online: 17-10-2026
Published: 31-10-2026

Abstract

This study compares the assessment of Japanese *sakubun* by a native Japanese teacher and two large language model based AI tools, Gemini AI (3.1) and ChatGPT (5.5), and examines the extent to which AI generated scores align with human assessment. The data comprised eight *sakubun* texts on the theme of *meiwaku* (inconvenience) written by second year students with approximately JLPT N4 level Japanese proficiency. Each *sakubun* was assessed using the same rubric covering four dimensions: content, organization, expression, and accuracy, as well as an overall score. The scores were descriptively compared across the three evaluators at both the criterion and overall levels. The results showed that the Japanese teacher assigned the highest average overall score of 16.25, followed by Gemini AI with 14.25 and ChatGPT with 13.75. Of the two AI tools, Gemini AI demonstrated closer overall alignment with the teacher's assessment. Both AI systems also exhibited a score compression tendency. Higher quality *sakubuns* were generally evaluated lower than by the teacher, whereas some lower quality *sakubuns* received relatively higher scores. The most substantial discrepancies were observed in content and expression, indicating that AI assessment was less sensitive to idea development, contextual appropriateness, naturalness of expression, and subtle linguistic nuances. These findings suggest that although Gemini AI and ChatGPT can provide useful supplementary assessment, human evaluator judgment remains essential, particularly for evaluating qualitative and context sensitive aspects of Japanese writing.

Keywords: Japanese writing assessment; artificial intelligence; automated writing assessment; human–AI agreement; Japanese language education

How to cite (APA): Putra, J., Umemoto, M., Widiyanti, S., Hayati, N., & Afifah, M. (2026). A Comparative Study of Human and AI Assessment in Japanese Sakubun Writing. *KIRYOKU*, 10(2), 710-723. <https://doi.org/10.14710/kiryoku.v10i2.710-723>

DOI: <https://doi.org/10.14710/kiryoku.v10i2.710-723>

1. Introduction

Writing assessment is a fundamental component of second language education because it provides evidence of learners' ability to develop ideas, organize discourse, and use linguistic resources appropriately. While current studies have reported the effectiveness of several

¹Corresponding Author. E-mail: bungjeni@upi.edu

Telp: +62 821-1615-3011

Copyright ©2026, The authors. Published by Kiryoku: Jurnal Studi Kejepangan. This open access article is distributed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

e-ISSN: 2581-0960, p-ISSN: 2599-0497

methods in *sakubun* courses (Utari et al., 2024; Wahani, 2025), assessing written *texts* manually requires considerable time and effort, particularly when teachers must evaluate multiple aspects of writing across a large number of learners (Ramesh & Sanampudi, 2022). In the context of Japanese as a foreign language (JFL), this challenge is compounded by the inherently interpretive nature of writing assessment. Lee and Hasebe (2020) describe the evaluation of learner writing as an isolated task in which teachers often have to rely on their own linguistic knowledge and intuition, making the standardization of assessment difficult. These limitations have encouraged the development of automated essay scoring (AES) systems as tools for improving the efficiency and consistency of writing assessment. Earlier applications have included machine learning and deep learning approaches to Japanese essay grading (Oktaviani et al., 2021), while more recent developments have been characterized by the rapid integration of Artificial Intelligence (AI) and Large Language Models (LLM) into writing, feedback, and assessment practices (Johnson & Zhang, 2024; Zahira et al., 2025). The emergence of generative AI therefore offers new possibilities for reducing teachers' assessment workload, but it also raises questions about whether AI can evaluate learner writing in ways that are comparable to professional human judgment.

Recent research demonstrates both the potential and the limitations of automated and AI-based writing assessment. In Japanese language education, Hirao et al. (2020) developed an AES (Advanced Encryption Standard) system for nonnative Japanese learners using holistic and multiple-trait scores, with BERT (Bidirectional Encoder Representations from Transformers) showing promising performance, while Lee and Hasebe (2020) developed jWriter to estimate JFL learners' writing proficiency and provide diagnostic information based on linguistic features. More recently, Li and Liu (2024) compared conventional feature-based systems with BERT, GPT-4, and a Japanese-specific large language model and reported that GPT-4 achieved the strongest overall agreement with human ratings, while also demonstrating that prompt design substantially affected scoring performance. Similar developments have been reported in other L2 writing contexts. Pack et al. (2024) found that GPT-4 demonstrated relatively strong validity and intrarater reliability in scoring English learner essays, although model performance could fluctuate across occasions, while Mansour et al. (2024) found that ChatGPT and Llama could perform both holistic and trait-level essay evaluation but still showed a performance gap compared with state-of-the-art AES models. Research has also demonstrated that incorporating grammatical features can improve AES performance (Doi et al., 2024), whereas ChatGPT may be less successful in evaluating higher-order aspects such as coherence and cohesion, sometimes producing generic feedback based on superficial linguistic features (Yoon et al., 2023). Consequently, despite the increasing sophistication of AI-based assessment, its validity, reliability, fairness, and interpretability remain subjects of continuing debate (Johnson & Zhang, 2024; Lazendic, 2023).

Despite these advances, an important research gap remains in the application of general-purpose LLMs to the assessment of Japanese as a foreign language (JFL) writing. Li and Liu (2024) note that research on LLM-based automated essay scoring (AES) for non-native Japanese writing remains limited. Existing research on Japanese AES has largely focused on developing computational models, estimating proficiency levels, extracting linguistic features, and evaluating model performance using relatively large learner corpora. More broadly, Li and Ng (2024) argue that recent AES research has concentrated heavily on improving predictive performance on benchmark datasets, leaving broader questions concerning the interpretation and educational use of automated scores in need of further investigation. In particular, comparatively little is known about how commercially available, general-purpose

AI systems evaluate the same Japanese learner *sakubun* when instructed to follow an identical analytic rubric, how their judgments differ from those of a native Japanese teacher across specific dimensions of writing, and whether systematic scoring patterns emerge among learners with relatively low Japanese proficiency. To address this gap, the present study compares the scores assigned to eight Japanese *sakubun* texts, written by learners with approximately JLPT N5 and N4 level proficiency on the theme of *meiwaku*, by a native Japanese teacher, Gemini (3.1 version), and ChatGPT (GPT-5.5 version). The eight *sakubun* comprised four pairs of compositions, with each pair written by the same student, allowing for within-student comparisons rather than comparisons of unrelated texts. The four students were purposively selected based on their final course grades to capture variation in achievement: one from the high-achieving group, two from the middle-achieving group, and one from the low-achieving group. This focused selection enabled an in-depth comparison of the paired texts and the scores assigned to them; the study does not aim to generalize its findings to the wider student population.

Previous research has highlighted both the strengths and limitations of AI-assisted writing assessment. Bouziane and Bouziane (2024) reported that ChatGPT performed particularly well in evaluating grammar, spelling, sentence structure, relevance, and supporting evidence, whereas human evaluators were more effective in assessing thematic consistency. In the context of Japanese writing assessment, Li and Liu (2024) suggested that GPT-based systems have the potential to reduce rater bias and subjectivity, lessen the time, workload, and cost of assessment, and provide constructive feedback that supports independent learning. However, these studies leave open the question of how general-purpose LLMs compare with a native Japanese teacher when assessing the same learner texts using an identical analytic rubric. To address this gap, the present study compares the scores assigned to eight Japanese *sakubun* texts by a native Japanese teacher, Gemini, and ChatGPT. It examines four assessment dimensions, namely content, organization, expression, and accuracy, as well as the overall score, to identify patterns of agreement and discrepancy among the assessors. The study contributes a criterion-level comparison in a JFL classroom context, providing evidence about the potential and limitations of AI as a supplementary tool for Japanese writing assessment.

2. Methods

This study employed a descriptive comparative approach to examine similarities and differences between human evaluator and AI assessments of Japanese learner writing. The sample consisted of eight *sakubun* texts written by four students enrolled in the *Shouchuukyuu Hyouki Sakubun* course during the odd semester of the 2025–2026 academic year. The four students were purposively selected from the class based on their relative writing proficiency, as indicated by their performance on previous daily assignments in the course.

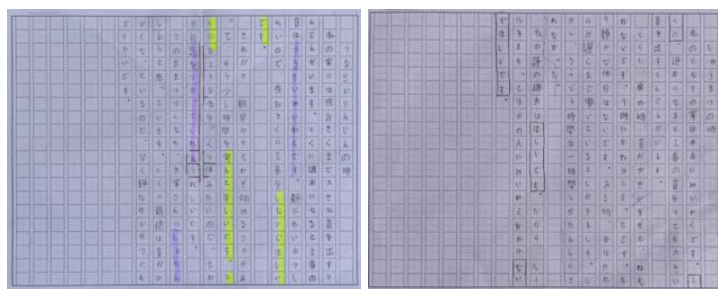


Figure 1. Student Sakubun by AI Assistance Condition

Most students in the class demonstrated Japanese language proficiency ranging from approximately JLPT N5 to N4. To capture variation in writing ability, SP (Student Participant) 1 was selected as a higher-proficiency learner, SP2 and SP3 as intermediate-proficiency learners, and SP4 as a lower-proficiency learner. Each student contributed two *sakubun* texts, resulting in four pairs of compositions and eight texts in total. This selection enabled the study to examine patterns of agreement and discrepancy in human and AI scoring across different levels of learner writing proficiency. Because the sample was purposively selected and limited to four students, the findings are interpreted as descriptive and are not intended to be generalized to the wider student population. Figure 1 presents examples of the *sakubun* texts produced by the four student participants.

As part of the data collection procedure, participants completed two *sakubun* writing tasks on the theme of *meiwaku* (inconvenience or annoyance), students were encouraged to use the expressions *～てほしい* (*~tehoshii*) and/or *～ないでほしい* (*~naidehoshii*) to emphasize their sense of discomfort and their expectations regarding others' actions. Each *sakubun* was required to follow a complete structure consisting of an introduction (*hajimeni*), body (*naiyō*), and conclusion (*musubi*).

The writing task was conducted in two stages. In the first stage, participants were permitted to use AI assistance only to check the suitability of vocabulary (*goi*) and discuss expressions (*hyougen*) appropriate to the context of their *sakubun*. These limits were explicitly explained to the students before they began writing. The use of AI to generate sentences, paragraphs, or an entire composition, or to develop or organize the content, was prohibited. The researcher closely supervised the students while they completed the task to monitor their adherence to these instructions. Three weeks later, as part of the final examination, the same participants completed a second *sakubun* task on the same topic and with the same structural requirements, but without AI assistance. Keeping the topic and structural requirements constant made the two texts from each student more directly comparable, while the second task provided an opportunity to examine their independent writing on the same topic. The comparison was intended to identify descriptive differences between writing completed with limited AI assistance and writing completed without it. However, these differences cannot be attributed solely to AI use because the tasks were completed three weeks apart and under different writing conditions.

The *sakubun* were independently assessed by three evaluators: one native Japanese teacher as a human evaluator, Gemini AI, and ChatGPT. The human evaluator held a master's degree and had eight years of experience teaching Japanese. A single human rater was used because the study focused on an in-depth comparison between expert teacher judgment and the scores generated by two general-purpose LLMs, rather than on estimating inter-rater reliability among human assessors. The selected teacher served as an expert reference point because of her Japanese language background, teaching experience, and familiarity with learner writing in the course context. Using one human rater also ensured that all human scores were based on a consistent interpretation of the rubric and avoided introducing variation resulting from differences among multiple human raters. To ensure comparability across the three evaluators, the same analytic assessment rubric was used for all *sakubun*. The rubric was adopted from Tajima (2010) and consisted of four dimensions: content (*naiyou*/内容), organization (*kousei*/構成), expression (*hyougen*表現), and accuracy (*seikakusa*正確さ). Each dimension included specific scoring criteria. The rubric developed by Tajima is presented in Figure 2.

The rubric therefore served as the main research instrument for generating comparable assessment data from the human evaluator and the two AI systems. The scores obtained from the three evaluators were then compiled according to participant, *sakubun*, assessment criterion, evaluator, and writing condition. The data were analyzed descriptively by comparing the scores assigned by the human evaluator, Gemini AI, and ChatGPT. Gemini AI and ChatGPT were selected because they are among the most widely used LLM-based AI tools in Indonesia (Prakoso et al., 2025). For the assessment process, the human evaluator was asked to evaluate each *sakubun* using the criteria specified in Tajima's (2010) rubric. However, Gemini AI and ChatGPT were unable to read the students' handwritten *sakubun* accurately when only scanned copies were attached to the prompt. Therefore, the students' *sakubuns* were transcribed into digital text before being submitted to Gemini AI and ChatGPT together with the assessment prompt. The prompt used for both AI tools is presented below.

あなたは日本人の日本語教師です。以下の基準(得点)を用いて、上の学生（日本語能力はN4程度）が書いた作文を採点してください。…

Anata wa Nihonjin no Nihongo kyoushi desu. Ika no kijun (tokuten) o mochiite, ue no gakusei (Nihongo nouryoku wa N4 teido) ga kaita sakubun o saiten shite kudasai. ...

For both Gemini AI and ChatGPT, the assessment procedure was conducted in the following sequence: first, the transcribed student *sakubun* was provided; second, the assessment instruction prompt presented above was entered; and third, the Tajima (2010) rubric was reproduced in the prompt.

In this study, the student-written *sakubun* texts served as the primary data or writing samples to be evaluated, while the students who produced the *sakubuns* constituted the participants and sources of the written data. The *sakubuns* were assessed by three evaluators—one human evaluator, Gemini AI, and ChatGPT—who functioned as raters or assessment agents. To ensure that all evaluators applied the same assessment standards, Tajima's (2010) rubric was used as the scoring instrument. The criterion-level and overall scores assigned by the three evaluators constituted the assessment-score data, which were subsequently compared to examine the extent to which the AI-generated assessments aligned with the human evaluator's evaluation.

表3 「内容」に関する基準		表4 「構成」に関する基準	
得点	基準	得点	基準
6	具体的な内容が秩序立てられ明確に述べられている。詳しく説明され、量的にも十分である。	6	適切な段落分けがなされ、段落間のつながりも整合性があり、バランスの良い構成になっている。書き出しや終わりの部分が適切である。
5	具体的な内容が明確に述べられている。特に際立ってはいるが、十分な説明がある。	5	適切な段落分けがなされているが、行頭の1字あけなどの形式面での不備が見られる。段落間のつながりや全体のまとまりに問題はないが、あっても軽微なものである。書き出しや終わりの部分がやや自然であったりする。
4	具体的な内容がおおむね明確に述べられている。部分的に不十分を感じても、全体的な内容として課題に対しての十分な説明になっている。	4	段落分けがなされているが、形式面での不備が、意味内容とのずれなど、不適切な部分がある。書き出しや終わりの部分が1文字しかなかったり、本論部分の段落分けがなかったりする。
3	内容の具体性、記述の明確さにやや劣る点が見られる。説明不足の部分や記述の偏り、内容のわかりにくい部分が見られる。	3	段落分けがなされているが、不適切な部分があって、全体のまとまりに影響している。段落間でのつながりの悪い箇所がある。あるいは、段落分けはなされていないが、文章の流れに問題はなくまとまった記述にはなっている。
2	内容の具体性や量に不足が認められる。あるいは、記述の明確さにおいて劣っているなど、問題点がある。	2	段落分けがなされていない。形式面での不備が、意味内容とのずれなど、まとまりに問題となっている。あるいは、段落間のつながりが悪い。文章の流れや全体のまとまりに問題がある。
1	内容の具体性や量が乏しい。あるいは、記述が明確ではなく、内容がわかりにくい。	1	構成意識がない。形式的にも（形式段落）意味的にも（意味内容のまとまり）段落に相当するものがない。文と文のつながりが悪い部分が目立ち、全体がまとまっていない。

表5 「表現」に関する基準		表6 「正確さ」に関する基準	
得点	基準	得点	基準
6	論述試験の答案として適切な表現・語彙が用いられて、十分に説明されている。語彙や表現の多様性が希である。	6	誤りがなく、意味が明確に伝わる文章である。
5	論述試験の答案として適切な表現・語彙が用いられているが、不適切なものが1、2箇所目立っている。多様な語彙や表現が使われている。	5	誤りが1、2箇所あるが、軽微なものである。意味がわかりにくい部分はない。
4	書き言葉の表現・語彙が用いられているが、話し言葉や簡約的な不適切なものが若干目立っている。語彙や表現は特に多様ではないが、劣ってはいない。	4	誤りが数箇所あるが、軽微なものである。意味がわかりにくい部分はないが、あっても全体に影響はない。
3	書き言葉の表現・語彙が用いられているが、話し言葉や簡約的な不適切なものが単純な間違い程度ではなく目立っている。語彙や表現にやや乏しきや単語が認められる。	3	誤りが数箇所あったり、意味のわからない部分があったりして、全体の正確さに影響がある。
2	表現・語彙が不適切なものがあり、乏しさや単語が認められる。漢字が使われるべきところで使われていないなど、漢字力の問題が認められる。	2	誤りが数箇所あったり、意味のわからない部分があったりして、全体の正確さに問題がある。
1	表現・語彙が乏しく、記述が不十分になっている。全体の文章が稚拙である。漢字の使用が少ないなど漢字力が劣っている。	1	誤りが目立ち、意味がわからない部分があったりして、不正確な印象がある。

Figure 2. Tajima's (2010, pp. 92–94) Assessment Rubric

Copyright ©2026, The authors. Published by Kiryoku: Jurnal Studi Kejepangan. This open access article is distributed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

The data were analyzed using descriptive comparative analysis. First, the mean overall score assigned by each evaluator was calculated to identify general scoring tendencies. Next, score differences between the teacher and each AI system were examined to determine the magnitude and direction of discrepancies, including whether the AI systems tended to assign lower or higher scores than the teacher. A criterion-level comparison was then conducted across the four dimensions of the rubric: content, organization, expression, and accuracy, to identify the dimensions showing the largest differences among evaluators. In addition, scoring patterns for *sakubuns* produced with limited AI assistance and those produced without AI assistance were compared to explore whether AI use during the writing process was associated with differences in the evaluations assigned by the human evaluator and the AI systems. Through these descriptive comparisons, the study examined the extent to which AI-based assessments corresponded with human evaluator judgment and identified areas in which the evaluators' scores converged or diverged.

3. Result and Discussion

This section presents the assessment results for eight *sakubun* texts written by four student participants. It first compares the scores assigned to the same texts by a human evaluator, Gemini, and ChatGPT, then examines the teacher's scores for each student's two writing conditions. These descriptive comparisons provide a basis for identifying patterns across assessors and across the paired texts. Figure 3 presents the scores assigned to the same student *sakubun* texts by the teacher, Gemini, and ChatGPT, illustrating similarities and differences between human and AI assessments.

Table 1 details the human evaluator's scores for the four student participants (SP1–SP4). It compares texts written without AI assistance (“none”) with texts written with limited AI support (“limited”) across four criteria: content, organization, expression, and accuracy.

以下のように採点します。			作文①		採点結果	
評価項目	得点	理由	内容	1・2・3・4・5・6	評価項目	得点
A. 内容	4	個人の語彙で通文に固まっていること、後述のないこと、ルールを守ってほしいことが書かれており、内容はだいたい伝わります。ただし、「母は彼らが遅くまで働いていると死なました」など、状況が少しわかりにくい部分があります。	構成	1・2・3・4・5・6	A. 内容	4
B. 構成	4	はじめ・理由・まとめの流れはありますが、段落分けもできていますが、同じ「特に」が続いたり、語のつながりが少し不自然なところがあります。	表現	1・2・3・4・5・6	B. 構成	4
C. 表現	3	言いたいことは伝わりますが、「週末の時」「個人の言」「大きい過ぎて」「静の週末」など、不自然な表現がいくつかあります。語彙もやや単語です。	正確さ	1・2・3・4・5・6	C. 表現	3
D. 正確さ	3	文法や語彙の誤りが数か所あり、意味がわかりにくい文もあります。ただし、全体のテーマは理解できます。	コメント	「この「まず」が、文脈でつながりません。文法は正しいですが、意味が通じません。また、この「まず」は、文脈でつながりません。文法は正しいですが、意味が通じません。」		

Figure 3. Sample *Sakubun* Scores by Human, Gemini AI and ChatGPT

Table 1. Human evaluator Evaluation Scores

Sample	AI Assistance	Human evaluator				Total
		Content	Organization	Expression	Accuracy	
SP1	none	6	5	4	4	19
	none	6	5	5	4	20
SP2	limited	5	5	5	3	18
	none	4	5	1	1	11
SP3	limited	4	4	2	3	13
	none	4	5	3	2	14
SP4	limited	6	4	5	5	20
	none	5	4	3	3	15

Table 2 presents the scores assigned by Gemini to the student *sakubun* texts to provide an overview of assessments across drafts produced with and without AI assistance. Meanwhile, table 3 summarizes ChatGPT's evaluation of the same texts, showing the scores assigned under the limited-AI and no-AI assistance conditions.

Table 4 presents the overall scoring tendencies of the human evaluator, Gemini AI, and ChatGPT. The human evaluator assigned the highest mean total score of 16.25, whereas Gemini AI and ChatGPT produced lower mean scores of 14.25 and 13.75, respectively. Relative to the teacher, the mean score difference was -2.00 points for Gemini and -2.50 points for ChatGPT, indicating that both AI systems generally tended to assign lower scores than the human evaluator.

Table 2. Gemini AI Sakubun Scores

Sample	AI Assistance	Gemini				Total
		Content	Organization	Expression	Accuracy	
SP1	none	4	4	3	3	14
	none	5	5	3	3	16
SP2	limited	5	5	4	4	18
	none	4	4	2	2	12
SP3	limited	4	4	3	3	14
	none	3	4	2	2	11
SP4	limited	5	5	4	3	17
	none	4	4	2	2	12

Table 3. ChatGPT Sakubun Scores

Sample	AI Assistance	Chat GPT				Total
		Content	Organization	Expression	Accuracy	
SP1	none	4	3	3	2	12
	none	4	4	3	3	14
SP2	limited	4	4	3	3	14
	none	4	4	3	2	13
SP3	limited	4	4	3	2	13
	none	4	4	3	3	14
SP4	limited	5	5	4	4	18
	none	4	3	3	2	12

Table 4. Overall scoring tendencies

Indicator	Human Evaluator	Gemini AI	ChatGPT
Mean total score	16.25	14.25	13.75
Mean difference from teacher	-	-2.00	-2.50
Mean absolute error from teacher	-	2.50	3.00
Score range	11-20	11-18	12-18
Exact agreement with teacher	-	1/8	2/

The mean absolute error further indicates that Gemini was slightly closer to the teacher's scoring than ChatGPT. Gemini had a mean absolute error of 2.50 points, compared with 3.00 points for ChatGPT. Interestingly, ChatGPT produced more exact score matches with the teacher, namely two out of eight *sakubuns*, while Gemini produced only one exact match. This result demonstrates that exact score agreement alone is insufficient for determining which AI system is generally closer to human assessment. An evaluator may produce several identical scores while still showing relatively large discrepancies on other *sakubuns*. Therefore, the magnitude and pattern of score differences, together with the criteria underlying the judgments, need to be examined rather than relying solely on exact score agreement. This interpretation is consistent with Saricaoglu & Bilki (2025), who emphasize that agreement between AI and human scores alone provides limited information about whether the system adequately represents the relevant dimensions of writing proficiency.

Another notable difference concerns the score ranges. The teacher used a wider range from 11 to 20, whereas Gemini scores ranged from 11 to 18 and ChatGPT scores from 12 to 18. This narrower distribution suggests a score compression tendency, in which AI assessments were more concentrated around the middle of the scoring scale. The phenomenon was particularly visible at the upper end, where the AI systems did not assign scores as high as those given by the teacher. Thus, although the AI systems followed the general direction of the rubric, they appeared less likely to reproduce the full range of distinctions made by the human evaluator. This pattern may indicate that the models were more cautious when assigning extreme scores, especially high scores that require clear evidence of consistently strong performance across multiple rubric criteria. As a result, writing of relatively high quality may have been evaluated more conservatively by the AI systems. Previous research likewise suggests that LLM-based assessment can achieve relatively close overall alignment with human scoring, while human raters may still provide more nuanced judgments, particularly for finer characteristics of writing (Zhang et al., 2025).

3.1 Differences According to Individual Learners

As shown in Table 5, the largest discrepancy was found for SP1, who represented the highest-proficiency learner in the sample. The teacher assigned SP1 a mean score of 19.5, whereas Gemini assigned 15.0 and ChatGPT only 13.0. The resulting differences were -4.5 points for Gemini and -6.5 points for ChatGPT. By contrast, the differences were considerably smaller for SP2 and SP3. Gemini scored SP2 only 0.5 points above the teacher, while ChatGPT scored SP3 exactly the same as the teacher. These results indicate that the discrepancy between AI and human assessment was not uniform across learners.

The SP1 result is particularly important because it suggests that AI may have difficulty fully recognizing characteristics that distinguish a highly rated *sakubun* from an average one. A human evaluator may respond not only to explicitly observable linguistic features but also to the maturity of idea development, appropriateness of expression, contextual relevance, naturalness, and subtle relationships among ideas. LLMs, in contrast, may apply rubric criteria relatively systematically while being less sensitive to such fine-grained distinctions. Zhang et al. (2025) discuss similar concerns, noting evidence that AI-generated assessments can approach human scoring standards while human judgments remain more nuanced. Their study also emphasizes the continued importance of human involvement in language assessment.

The narrower AI score ranges in Table 5 and the large downward adjustment for SP1 further indicate a tendency toward the middle of the scale. In other words, the AI systems tended to reduce differences among *sakubuns* rather than reproduce the wider differentiation in teacher scoring. This does not mean that AI always assigns average scores; rather, AI scores in this dataset were less dispersed than the teacher's. The effect was particularly pronounced for high-scoring writing, which both AI systems evaluated substantially lower. Given the present study's sample of only eight *sakubuns*, however, this pattern should be interpreted as a descriptive tendency within this dataset, not as a general property of LLM-based assessment.

Participant proficiency categorization also does not necessarily determine performance on each writing task. Although SP4 was categorized as the lower-proficiency learner based on general Japanese proficiency, the teacher assigned SP4 a relatively high mean *sakubun* score of 17.5. This suggests that general language proficiency and performance on a specific writing task are related but distinct constructs.

3.2 Differences Across Rubric Dimensions

Among the four dimensions, content showed the largest mean discrepancy between the teacher and both AI systems. Gemini scored content 0.75 points lower than the teacher, whereas ChatGPT scored it 0.87 points lower. This pattern suggests that AI assessment may have been more conservative in recognizing the quality of idea development and the appropriateness or sufficiency of content. Such higher-order aspects require an evaluator to interpret not only what information is explicitly present but also how effectively ideas respond to the writing task and how meaningful their development is within the context of the *sakubun*.

Differences were also observed in organization. Gemini's mean score of 4.38 was relatively close to the teacher's 4.63, producing a difference of only -0.25. ChatGPT, however, assigned a mean score of 3.88, resulting in a difference of -0.75. This indicates that ChatGPT tended to evaluate paragraph development and logical progression more strictly than the teacher, whereas Gemini's assessment of organization was substantially closer to the human evaluation.

The results for expression require a slightly different interpretation. Based solely on the signed mean difference, ChatGPT appeared closer to the teacher, with a difference of -0.37 compared with -0.62 for Gemini. However, the essay-level absolute differences showed relatively large discrepancies in expression for both AI systems. This distinction is important because positive and negative discrepancies can cancel each other out when only mean signed differences are considered. The results therefore suggest that aspects such as naturalness, appropriateness of expression, and subtle linguistic choices may remain difficult for AI to evaluate in precisely the same way as a human evaluator.

Table 5. Score differences by learner

Learner	Human evaluator	Gemini AI	ChatGPT	Gemini - Teacher	ChatGPT-Teacher
SP1	19.5	15.0	13.0	-4.5	-6.5
SP2	14.5	15.0	13.5	+0.5	-1.0
SP3	13.5	12.5	13.5	-1.0	0.0
SP4	17.5	14.5	15.0	-3.0	-2.5

This finding is relevant to Saricaoglu and Bilki (2025), who found that ChatGPT-4 could accurately identify many strengths and weaknesses across multiple dimensions of L2 writing but differed in the specificity of its feedback across dimensions. In particular, feedback was more general for Task Response and Coherence and Cohesion than for some linguistic dimensions. Their findings suggest that the ability to identify an issue does not necessarily mean that an LLM interprets and explains complex discourse-level qualities with the same specificity as a human evaluator. Thus, the representation and evaluation of coherence, especially how relationships among ideas contribute to the overall quality of a *sakubun*, remain important areas for further investigation.

For accuracy, the differences were more moderate. Gemini scored 0.38 points below the teacher, while ChatGPT scored 0.50 points below. Compared with content and some aspects of expression, linguistic accuracy may provide relatively explicit textual evidence, such as grammatical forms, word choice, and syntactic constructions, that an LLM can detect more directly. Nevertheless, even linguistic assessment is not entirely straightforward because the appropriateness of a form may depend on context and intended meaning.

Taken together, the criterion-level findings suggest that the degree of agreement between AI and human assessment varies depending on what aspect of writing is being evaluated. The results therefore reinforce the importance of examining individual rubric dimensions rather than considering only the total score. Two evaluators may assign similar total scores while reaching those scores through different judgments concerning content, organization, expression, and accuracy.

3.3 Differences According to the Use of AI During Writing

Table 7 compares the assessment results according to whether AI assistance was used during the writing process. The participants were allowed to decide whether to use AI when completing the writing tasks. Although SP1 had the highest Japanese proficiency among the four participants, this learner chose to write both *sakubuns* without AI assistance. In contrast, the other three participants contributed *sakubuns* written with and without AI assistance. Consequently, the dataset consisted of three AI-assisted *sakubuns* and five non-AI-assisted *sakubuns*.

Table 6. Score differences across rubric dimensions

Dimension	Human evaluator	Gemini	ChatGPT	Gemini – Human evaluator	ChatGPT – Human Evaluator
Content	5.00	4.25	4.13	-0.75	-0.87
Organization	4.63	4.38	3.88	-0.25	-0.75
Expression	3.50	2.88	3.13	-0.62	-0.37
Accuracy	3.13	2.75	2.63	-0.38	-0.50

Table 7. Score differences according to AI use during writing

Writing condition	n	Human evaluator	Gemini AI	Chat GPT	Gemini – Human evaluator	ChatGPT – Human evaluator
AI-assisted	3	17.00	16.33	15.00	-0.67	-2.00
Non-AI-assisted	5	15.80	13.00	13.00	-2.80	-2.80

The results show that the score differences between the human and the two AI systems were smaller for AI-assisted *sakubuns*. For the three AI-assisted *sakubuns*, the mean score assigned by the teacher was 17.00, compared with 16.33 for Gemini and 15.00 for ChatGPT. Gemini therefore differed from the human evaluator by only -0.67 points, while the difference for ChatGPT was -2.00 points. In contrast, for the five *sakubuns* written without AI assistance, both Gemini and ChatGPT produced a mean score of 13.00, compared with the human's mean score of 15.80, resulting in a difference of -2.80 points for both AI systems. Descriptively, therefore, the evaluations of AI-assisted *sakubuns* showed greater convergence between the human evaluator and the AI evaluators, particularly in the case of Gemini.

In this sample, the scores assigned by the human evaluator, Gemini, and ChatGPT showed smaller discrepancies for the AI-assisted *sakubun* than for the non-AI-assisted *sakubun*, indicating closer agreement among the assessors for the AI-assisted texts. Saricaoglu and Bilki (2025) found that ChatGPT-4 could identify strengths and weaknesses across several dimensions of L2 writing, although the specificity of its feedback varied by dimension. Their findings provide context for interpreting LLM-based writing assessment but do not explain the difference in scoring agreement observed in the present study. Because the linguistic features of the *sakubun* and the reasons underlying individual scores were not systematically analysed, this study cannot establish why the assessors agreed more closely on the AI-assisted *sakubun* texts. Nor can the difference be attributed solely to AI assistance, as the two tasks were completed three weeks apart under different writing conditions.

A particularly important factor is the distribution of the participants across the two writing conditions. SP1, who had the highest Japanese proficiency, voluntarily chose not to use AI in either *sakubun* despite being permitted to do so. This decision may reflect greater confidence in independent Japanese writing and a lower perceived need for technological assistance. More importantly for the present analysis, SP1 was also the participant for whom the largest discrepancies between human and AI scoring were observed. The teacher assigned SP1 a mean score of 19.5, whereas Gemini and ChatGPT assigned mean scores of 15.0 and 13.0, respectively. Thus, two of the five *sakubuns* in the non-AI-assisted group were produced by the learner whose writing generated the greatest divergence between human and AI evaluation.

This distribution has an important methodological implication. The larger human-AI difference observed in the non-AI-assisted condition cannot be attributed solely to the absence of AI assistance. It may also reflect characteristics of SP1's higher-level writing that were valued strongly by the human evaluator but were not recognized to the same extent by the AI systems. In other words, AI-use condition and learner characteristics are not fully independent in the present dataset. The smaller score difference for AI-assisted *sakubuns* may therefore reflect both the linguistic characteristics of AI-assisted writing and differences among the learners who produced the *sakubuns*.

This interpretation is also consistent with the learner-level results presented in Table 5. While the AI systems tended to produce scores concentrated within a narrower range, the human evaluator gave particularly high scores to SP1. The substantial downward adjustment of SP1's *sakubuns* by both AI systems suggests that LLM-based evaluation may be less sensitive to some qualities of relatively advanced learner writing, such as nuanced expression, depth of content, natural development of ideas, or contextually appropriate language use.

Therefore, the apparent convergence between human and AI assessment in the AI-assisted condition should be interpreted cautiously.

Overall, the findings indicate that AI-assisted *sakubuns* in this dataset received more similar scores from the teacher and the AI evaluators than non-AI-assisted *sakubuns*. However, given the small and unequal group sizes, with three AI-assisted and five non-AI-assisted *sakubuns*, as well as the fact that the highest-proficiency participant contributed exclusively to the non-AI-assisted condition, the findings should be regarded as descriptive rather than causal. Further research with a larger sample and a more balanced research design is required to determine whether AI assistance itself systematically increases agreement between human and LLM-based writing assessment.

4. Conclusion

This study compared the assessments of eight Japanese *sakubun* texts by a native Japanese teacher, Gemini AI, and ChatGPT, using the same analytic rubric covering content, organization, expression, accuracy, and overall score. Both AI systems generally assigned lower scores than the teacher, although Gemini showed closer overall alignment based on the mean score difference and mean absolute error. The AI assessments also showed a tendency toward score compression, particularly for higher-quality *sakubun*, with the largest discrepancy observed for SP1. At the rubric level, differences were most pronounced in content; expression also showed notable variation when absolute differences were considered.

These findings can be situated alongside, but not equated with, earlier research. In a study of English essay correction, Bouziane and Bouziane (2024) found that ChatGPT performed well in grammar, spelling, sentence structure, and supporting evidence, whereas teachers performed better in maintaining thematic consistency. In Japanese writing assessment, Li and Liu (2024) demonstrated the potential of large language models, particularly GPT-4, to assess writing across proficiency levels from A1 to C2 and to support human rating. Their study also identified the potential of automated assessment to reduce the time and resources required for evaluation. Against this background, the present study adds a more localized, rubric-level account of how two widely accessible AI systems assess the *same learner texts* relative to a native Japanese teacher. In particular, it identifies content-related discrepancies and a tendency toward score compression that may be obscured when attention is limited to overall agreement.

The results suggest that LLM-based assessment can apply rubric criteria and identify general strengths and weaknesses in learner writing, but it may be less sensitive than human judgment to idea development, contextual appropriateness, naturalness of expression, and cultural nuance. Score differences between the teacher and AI also tended to be smaller for AI-assisted *sakubun*, although this descriptive pattern should be interpreted cautiously given the limited number of *sakubun* texts. Gemini and ChatGPT are therefore best viewed as supplementary perspectives in Japanese writing assessment rather than replacements for teacher judgment. Future research should examine more texts, proficiency levels, topics, and task types; include multiple human raters; and investigate the consistency and reasoning of repeated AI evaluations, as well as whether characteristics of AI-assisted writing affect human–AI agreement.

Acknowledgments

This study was financially supported by the Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) Universitas Pendidikan Indonesia (UPI), to whom we express our deepest gratitude.

References

- Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, 3(1), 201. <https://doi.org/10.1007/s44217-024-00320-6>
- Doi, K., Sudoh, K., & Nakamura, S. (2024). Automated Essay Scoring Using Grammatical Variety and Errors with Multi-Task Learning and Item Response Theory. *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 316–329.
- Hirao, R., Arai, M., Shimanaka, H., Katsumata, S., & Komachi, M. (2020). Automated essay scoring system for nonnative Japanese learners. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1250–1257.
- Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports*, 14(1), 30064. <https://doi.org/10.1038/s41598-024-79208-2>
- Lazendic, G. (2023). *The validity of automated essay scoring. The Laboratory for Experimental Psychology Univerzitet u Beogradu*.
- Lee, J.-H., & Hasebe, Y. (2020). Quantitative Analysis of JFL Learners' Writing Abilities and the Development of a Computational System to Estimate Writing Proficiency. *Learner Corpus Studies in Asia and the World*, 5, 105–120.
- Li, S., & Ng, V. (2024). Automated Essay Scoring: A Reflection on the State of the Art. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17876–17888. <https://doi.org/10.18653/v1/2024.emnlp-main.991>
- Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications*, 11(1), 723. <https://doi.org/10.1057/s41599-024-03209-9>
- Mansour, W., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can Large Language Models Automatically Score Proficiency of Written Essays? *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2777–2786.
- Oktaviani, A. N., Alief, M. Z., Santiar, L., Purnamasari, P. D., & Ratna, A. A. P. (2021). Automatic Essay Grading System for Japanese Language Exam using CNN-LSTM. *2021 17th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering*, 164–169. <https://doi.org/10.1109/QIR54354.2021.9716165>
- 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. <https://doi.org/10.1016/j.caeai.2024.100234>
- Prakoso, A. A., Salsabila, K., Yahya Maulana, A., & Erza, E. K. (2025). Hubungan penggunaan teknologi artificial intelligence gemini dalam mendukung kebutuhan informasi mahasiswa. *Pustaka Karya : Jurnal Ilmiah Ilmu Perpustakaan Dan Informasi*, 13(2), 359–372. <https://doi.org/10.18592/pk.v13i2.18418>

- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Saricaoglu, A., & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. *Annual Review of Applied Linguistics*, 45, 253–273. <https://doi.org/10.1017/S0267190525100160>
- Tajima, M. (2010). 日本語の文章に対する分析的評価の信頼性に関しての検証 [Verification of the reliability of analytic assessment for Japanese writing]. *Chuo Gakuin University Journal of Humanities and Natural Sciences*, 31, 85–104. <https://cgu.repo.nii.ac.jp/records/1469>
- Utari, R., Sukmara, R., & Septiany, N. (2024). Implementation of Project-Based Learning In Improving The Creativity of Writing Japanese Essays in The Sakubun Course. *KIRYOKU*, 8(1), 139–146. <https://doi.org/10.14710/kiryoku.v8i1.139-146>
- Wahani, Y. J. (2025). Task-Based Language Teaching (TBLT) Learning Model using Worksheets to improve Japanese writing skills. *KIRYOKU*, 9(1), 40–50. <https://doi.org/10.14710/kiryoku.v9i1.40-50>
- Yoon, S., Miszoglud, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *ArXiv*.
- Zahira, N. A., Nurhafidzoh, S. U., Annur, M. K., & Saputra, D. Y. (2025). Pengaruh Penggunaan Artificial Intelligence Terhadap Kemampuan Menulis Akademik Mahasiswa. *Sindoro : Cendikia Pendidikan*, 18(2), 2661–2670.
- Zhang, T., Erlam, R., & de Magalhães, M. B. (2025). Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls. *Annual Review of Applied Linguistics*, 45, 274–293. <https://doi.org/10.1017/S0267190525000030>