Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03209-9 OPEN

Check for updates

Applying large language models for automated essay scoring for non-native Japanese

Wenchao Li[1]✉ & Haitao Liu[2]

Recent advancements in artificial intelligence (AI) have led to an increased use of large language models (LLMs) for language assessment tasks such as automated essay scoring (AES), automated listening tests, and automated oral proficiency assessments. The application of LLMs for AES in the context of non-native Japanese, however, remains limited. This study explores the potential of LLM-based AES by comparing the efficiency of different models, i.e. two conventional machine training technology-based methods (Jess and JWriter), two LLMs (GPT and BERT), and one Japanese local LLM (Open-Calm large model). To conduct the evaluation, a dataset consisting of 1400 story-writing scripts authored by learners with 12 different first languages was used. Statistical analysis revealed that GPT-4 outperforms Jess and JWriter, BERT, and the Japanese language-specific trained Open-Calm large model in terms of annotation accuracy and predicting learning levels. Furthermore, by comparing 18 different models that utilize various prompts, the study emphasized the significance of prompts in achieving accurate and reliable evaluations using LLMs.

[1]Department of Japanese Studies, Zhejiang University, Hangzhou, China. [2]Department of Linguistics and Applied Linguistics, Zhejiang University, Hangzhou, China. ✉email: widelia@zju.edu.cn

Conventional machine learning technology in AES

AES has experienced significant growth with the advancement of machine learning technologies in recent decades. In the earlier stages of AES development, conventional machine learning-based approaches were commonly used. These approaches involved the following procedures: a) feeding the machine with a dataset. In this step, a dataset of essays is provided to the machine learning system. The dataset serves as the basis for training the model and establishing patterns and correlations between linguistic features and human ratings. b) the machine learning model is trained using linguistic features that best represent human ratings and can effectively discriminate learners' writing proficiency. These features include lexical richness (Lu, 2012; Kyle and Crossley, 2015; Kyle et al. 2021), syntactic complexity (Lu, 2010; Liu, 2008), text cohesion (Crossley and McNamara, 2016), and among others. Conventional machine learning approaches in AES require human intervention, such as manual correction and annotation of essays. This human involvement was necessary to create a labeled dataset for training the model. Several AES systems have been developed using conventional machine learning technologies. These include the Intelligent Essay Assessor (Landauer et al. 2003), the e-rater engine by Educational Testing Service (Attali and Burstein, 2006; Burstein, 2003), MyAccess with the InterlliMetric scoring engine by Vantage Learning (Elliot, 2003), and the Bayesian Essay Test Scoring system (Rudner and Liang, 2002). These systems have played a significant role in automating the essay scoring process and providing quick and consistent feedback to learners. However, as touched upon earlier, conventional machine learning approaches rely on predetermined linguistic features and often require manual intervention, making them less flexible and potentially limiting their generalizability to different contexts.

In the context of the Japanese language, conventional machine learning-incorporated AES tools include Jess (Ishioka and Kameda, 2006) and JWriter (Lee and Hasebe, 2017). Jess assesses essays by deducting points from the perfect score, utilizing the Mainichi Daily News newspaper as a database. The evaluation criteria employed by Jess encompass various aspects, such as rhetorical elements (e.g., reading comprehension, vocabulary diversity, percentage of complex words, and percentage of passive sentences), organizational structures (e.g., forward and reverse connection structures), and content analysis (e.g., latent semantic indexing). JWriter employs linear regression analysis to assign weights to various measurement indices, such as average sentence length and total number of characters. These weights are then combined to derive the overall score. A pilot study involving the Jess model was conducted on 1320 essays at different proficiency levels, including primary, intermediate, and advanced. However, the results indicated that the Jess model failed to significantly distinguish between these essay levels. Out of the 16 measures used, four measures, namely median sentence length, median clause length, median number of phrases, and maximum number of phrases, did not show statistically significant differences between the levels. Additionally, two measures exhibited between-level differences but lacked linear progression: the number of attributives declined words and the Kanji/kana ratio. On the other hand, the remaining measures, including maximum sentence length, maximum clause length, number of attributive conjugated words, maximum number of consecutive infinitive forms, maximum number of conjunctive-particle clauses, k characteristic value, percentage of big words, and percentage of passive sentences, demonstrated statistically significant between-level differences and displayed linear progression.

Both Jess and JWriter exhibit notable limitations, including the manual selection of feature parameters and weights, which can introduce biases into the scoring process. The reliance on human annotators to label non-native language essays also introduces potential noise and variability in the scoring. Furthermore, an important concern is the possibility of system manipulation and cheating by learners who are aware of the regression equation utilized by the models (Hirao et al. 2020). These limitations emphasize the need for further advancements in AES systems to address these challenges.

Deep learning technology in AES

Deep learning has emerged as one of the approaches for improving the accuracy and effectiveness of AES. Deep learning-based AES methods utilize artificial neural networks that mimic the human brain's functioning through layered algorithms and computational units. Unlike conventional machine learning, deep learning autonomously learns from the environment and past errors without human intervention. This enables deep learning models to establish nonlinear correlations, resulting in higher accuracy. Recent advancements in deep learning have led to the development of transformers, which are particularly effective in learning text representations. Noteworthy examples include bidirectional encoder representations from transformers (BERT) (Devlin et al. 2019) and the generative pretrained transformer (GPT) (OpenAI).

BERT is a linguistic representation model that utilizes a transformer architecture and is trained on two tasks: masked linguistic modeling and next-sentence prediction (Hirao et al. 2020; Vaswani et al. 2017). In the context of AES, BERT follows specific procedures, as illustrated in Fig. 1: (a) the tokenized prompts and essays are taken as input; (b) special tokens, such as [CLS] and [SEP], are added to mark the beginning and separation of prompts and essays; (c) the transformer encoder processes the prompt and essay sequences, resulting in hidden layer sequences; (d) the hidden layers corresponding to the [CLS] tokens (T[CLS]) represent distributed representations of the prompts and essays; and (e) a multilayer perceptron uses these distributed representations as input to obtain the final score (Hirao et al. 2020).

The training of BERT using a substantial amount of sentence data through the Masked Language Model (MLM) allows it to capture contextual information within the hidden layers. Consequently, BERT is expected to be capable of identifying artificial essays as invalid and assigning them lower scores (Mizumoto and Eguchi, 2023). In the context of AES for nonnative Japanese learners, Hirao et al. (2020) combined the long short-term memory (LSTM) model proposed by Hochreiter and Schmidhuber (1997) with BERT to develop a tailored automated Essay Scoring System. The findings of their study revealed that the BERT model outperformed both the conventional machine learning approach utilizing character-type features such as "kanji" and "hiragana", as well as the standalone LSTM model. Takeuchi et al. (2021) presented an approach to Japanese AES that eliminates the requirement for pre-scored essays by relying solely on reference texts or a model answer for the essay task. They

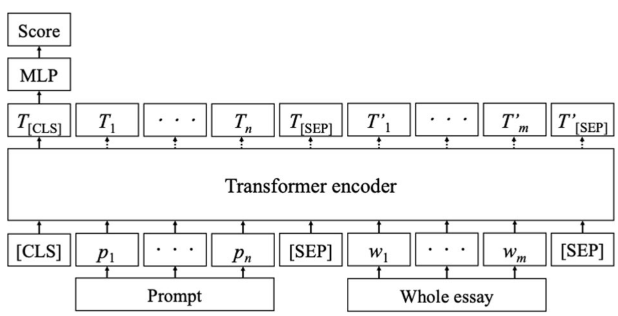

Fig. 1 AES system with BERT (Hirao et al. 2020).

investigated multiple similarity evaluation methods, including frequency of morphemes, idf values calculated on Wikipedia, LSI, LDA, word-embedding vectors, and document vectors produced by BERT. The experimental findings revealed that the method utilizing the frequency of morphemes with idf values exhibited the strongest correlation with human-annotated scores across different essay tasks. The utilization of BERT in AES encounters several limitations. Firstly, essays often exceed the model's maximum length limit. Second, only score labels are available for training, which restricts access to additional information.

Mizumoto and Eguchi (2023) were pioneers in employing the GPT model for AES in non-native English writing. Their study focused on evaluating the accuracy and reliability of AES using the GPT-3 text-davinci-003 model, analyzing a dataset of 12,100 essays from the corpus of nonnative written English (TOEFL11). The findings indicated that AES utilizing the GPT-3 model exhibited a certain degree of accuracy and reliability. They suggest that GPT-3-based AES systems hold the potential to provide support for human ratings. However, applying GPT model to AES presents a unique natural language processing (NLP) task that involves considerations such as nonnative language proficiency, the influence of the learner's first language on the output in the target language, and identifying linguistic features that best indicate writing quality in a specific language. These linguistic features may differ morphologically or syntactically from those present in the learners' first language, as observed in (1)–(3).

(1) Isolating

我-送了-他-一本-书
Wǒ-sòngle-tā-yī běn-shū
1SG.-give.PAST-him-one.CL-book
"I gave him a book."

(2) Agglutinative

彼-に-本-を-あげ-まし-た
Kare-ni-hon-o-age-mashi-ta
3SG.-DAT-hon-ACC-give.honorification.PAST

(3) Inflectional

give, give-s, gave, given, giving

Additionally, the morphological agglutination and subject-object-verb (SOV) order in Japanese, along with its idiomatic expressions, pose additional challenges for applying language models in AES tasks (4).

(4) 足-が 棒-に なり-ました

Ashi-ga bo-ni nar-mashita
leg-NOM stick-DAT become-PAST
"My leg became like a stick (I am extremely tired)."

The example sentence provided demonstrates the morphosyntactic structure of Japanese and the presence of an idiomatic expression. In this sentence, the verb "なる" (naru), meaning "to become", appears at the end of the sentence. The verb stem "なり" (nari) is attached with morphemes indicating honorification ("ます" - mashu) and tense ("た" - ta), showcasing agglutination. While the sentence can be literally translated as "my leg became like a stick", it carries an idiomatic interpretation that implies "I am extremely tired".

To overcome this issue, CyberAgent Inc. (2023) has developed the Open-Calm series of language models specifically designed for Japanese. Open-Calm consists of pre-trained models available in various sizes, such as Small, Medium, Large, and 7b. Figure 2 depicts the fundamental structure of the Open-Calm model. A key feature of this architecture is the incorporation of the Lora Adapter and GPT-NeoX frameworks, which can enhance its language processing capabilities.

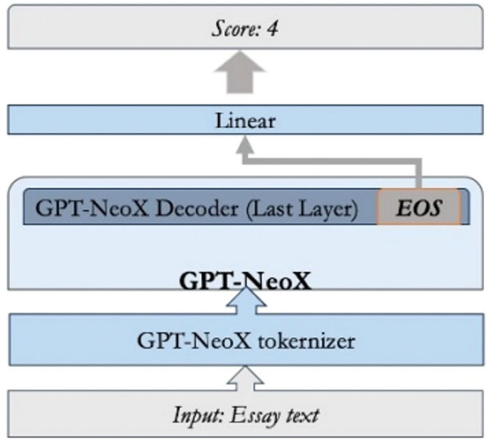

Fig. 2 GPT-NeoX Model Architecture (Okgetheng and Takeuchi 2024).

In a recent study conducted by Okgetheng and Takeuchi (2024), they assessed the efficacy of Open-Calm language models in grading Japanese essays. The research utilized a dataset of approximately 300 essays, which were annotated by native Japanese educators. The findings of the study demonstrate the considerable potential of Open-Calm language models in automated Japanese essay scoring. Specifically, among the Open-Calm family, the Open-Calm Large model (referred to as OCLL) exhibited the highest performance. However, it is important to note that, as of the current date, the Open-Calm Large model does not offer public access to its server. Consequently, users are required to independently deploy and operate the environment for OCLL. In order to utilize OCLL, users must have a PC equipped with an NVIDIA GeForce RTX 3060 (8 or 12 GB VRAM).

In summary, while the potential of LLMs in automated scoring of nonnative Japanese essays has been demonstrated in two studies—BERT-driven AES (Hirao et al. 2020) and OCLL-based AES (Okgetheng and Takeuchi, 2024)—the number of research efforts in this area remains limited.

Another significant challenge in applying LLMs to AES lies in prompt engineering and ensuring its reliability and effectiveness (Brown et al. 2020; Rae et al. 2021; Zhang et al. 2021). Various prompting strategies have been proposed, such as the zero-shot chain of thought (CoT) approach (Kojima et al. 2022), which involves manually crafting diverse and effective examples. However, manual efforts can lead to mistakes. To address this, Zhang et al. (2021) introduced an automatic CoT prompting method called Auto-CoT, which demonstrates matching or superior performance compared to the CoT paradigm. Another prompt framework is trees of thoughts, enabling a model to self-evaluate its progress at intermediate stages of problem-solving through deliberate reasoning (Yao et al. 2023).

Beyond linguistic studies, there has been a noticeable increase in the number of foreign workers in Japan and Japanese learners worldwide (Ministry of Health, Labor, and Welfare of Japan, 2022; Japan Foundation, 2021). However, existing assessment methods, such as the Japanese Language Proficiency Test (JLPT), J-CAT, and TTBJ[1], primarily focus on reading, listening, vocabulary, and grammar skills, neglecting the evaluation of writing proficiency. As the number of workers and language learners continues to grow, there is a rising demand for an efficient AES system that can reduce costs and time for raters and be utilized for employment, examinations, and self-study purposes.

This study aims to explore the potential of LLM-based AES by comparing the effectiveness of five models: two LLMs (GPT[2] and BERT), one Japanese local LLM (OCLL), and two conventional machine learning-based methods (linguistic feature-based scoring tools - Jess and JWriter).

The research questions addressed in this study are as follows:

a. To what extent do the LLM-driven AES and linguistic feature-based AES, when used as automated tools to support human rating, accurately reflect test takers' actual performance?
b. What influence does the prompt have on the accuracy and performance of LLM-based AES methods?

The subsequent sections of the manuscript cover the methodology, including the assessment measures for nonnative Japanese writing proficiency, criteria for prompts, and the dataset. The evaluation section focuses on the analysis of annotations and rating scores generated by LLM-driven and linguistic feature-based AES methods.

Methodology

Dataset. The dataset utilized in this study was obtained from the International Corpus of Japanese as a Second Language (I-JAS)[3]. This corpus consisted of 1000 participants who represented 12 different first languages. For the study, the participants were given a story-writing task on a personal computer. They were required to write two stories based on the 4-panel illustrations titled "Picnic" and "The key" (see Appendix A). Background information for the participants was provided by the corpus, including their Japanese language proficiency levels assessed through two online tests: J-CAT and SPOT. These tests evaluated their reading, listening, vocabulary, and grammar abilities. The learners' proficiency levels were categorized into six levels aligned with the Common European Framework of Reference for Languages (CEFR) and the Reference Framework for Japanese Language Education (RFJLE): A1, A2, B1, B2, C1, and C2. According to Lee et al. (2015), there is a high level of agreement ($r = 0.86$) between the J-CAT and SPOT assessments, indicating that the proficiency certifications provided by J-CAT are consistent with those of SPOT. However, it is important to note that the scores of J-CAT and SPOT do not have a one-to-one correspondence. In this study, the J-CAT scores were used as a benchmark to differentiate learners of different proficiency levels. A total of 1400 essays were utilized, representing the beginner (aligned with A1), A2, B1, B2, C1, and C2 levels based on the J-CAT scores. Table 1 provides information about the learners' proficiency levels and their corresponding J-CAT and SPOT scores.

A dataset comprising a total of 1400 essays from the story writing tasks was collected. Among these, 714 essays were utilized to evaluate the reliability of the LLM-based AES method, while the remaining 686 essays were designated as development data to assess the LLM-based AES's capability to distinguish participants with varying proficiency levels. The GPT 4 API was used in this study. A detailed explanation of the prompt-assessment criteria is provided in Section **Prompt**. All essays were sent to the model for measurement and scoring.

Table 1 Learners' proficiency level and their J-CAT, SPOT scores.

Learner Level	J-CAT scores	SPOT scores	Essay numbers
Beginner	0–100	0–30	75
A1	100-	31–55	313
A2	175-		352
B1	225-	56–80	326
B2	275-		156
C1	325-	81–90	102
C2	350-		76

Measures of writing proficiency for nonnative Japanese. Japanese exhibits a morphologically agglutinative structure where morphemes are attached to the word stem to convey grammatical functions such as tense, aspect, voice, and honorifics, e.g. (5).

(5) 食べ-させ-られ-まし-た-か

tabe-sase-rare-mashi-ta-ka
[eat (stem)-causative-passive voice-honorification-tense. past-question marker]

Japanese employs nine case particles to indicate grammatical functions: the nominative case particle が (ga), the accusative case particle を (o), the genitive case particle の (no), the dative case particle に (ni), the locative/instrumental case particle で (de), the ablative case particle から (kara), the directional case particle へ (e), and the comitative case particle と (to). The agglutinative nature of the language, combined with the case particle system, provides an efficient means of distinguishing between active and passive voice, either through morphemes or case particles, e.g. 食べる *taberu* "eat$_{\text{CONCUSIVE.}}$" (active voice); 食べられる *taberareru* "eat$_{\text{CONCUSIVE.}}$" (passive voice). In the active voice, "パンを食べる" (pan o taberu) translates to "to eat bread". On the other hand, in the passive voice, it becomes "パンが食べられた" (pan ga taberareta), which means "(the) bread was eaten". Additionally, it is important to note that different conjugations of the same lemma are considered as one type in order to ensure a comprehensive assessment of the language features. For example, e.g., 食べる *taberu* "eat$_{\text{CONCUSIVE}}$."; 食べている *tabeteiru* "eat$_{\text{PROGRESS}}$."; 食べた *tabeta* "eat$_{\text{PAST.}}$" as one type.

To incorporate these features, previous research (Suzuki, 1999; Watanabe et al. 1988; Ishioka, 2001; Ishioka and Kameda, 2006; Hirao et al. 2020) has identified complexity, fluency, and accuracy as crucial factors for evaluating writing quality. These criteria are assessed through various aspects, including lexical richness (lexical density, diversity, and sophistication), syntactic complexity, and cohesion (Kyle et al. 2021; Mizumoto and Eguchi, 2023; Ure, 1971; Halliday, 1985; Barkaoui and Hadidi, 2020; Zenker and Kyle, 2021; Kim et al. 2018; Lu, 2017; Ortega, 2015). Therefore, this study proposes five scoring categories: lexical richness, syntactic complexity, cohesion, content elaboration, and grammatical accuracy. A total of 16 measures were employed to capture these categories. The calculation process and specific details of these measures can be found in Table 2.

T-unit, first introduced by Hunt (1966), is a measure used for evaluating speech and composition. It serves as an indicator of syntactic development and represents the shortest units into which a piece of discourse can be divided without leaving any sentence fragments. In the context of Japanese language assessment, Sakoda and Hosoi (2020) utilized T-unit as the basic unit to assess the accuracy and complexity of Japanese learners' speaking and storytelling. The calculation of T-units in Japanese follows the following principles:

(i) A single main clause constitutes 1 T-unit, regardless of the presence or absence of dependent clauses, e.g. (6).

(6) ケンとマリはピクニックに行きました (main clause): 1 T-unit.

(ii) If a sentence contains a main clause along with subclauses, each subclause is considered part of the same T-unit, e.g. (7).

Table 2 Measures of nonnative Japanese writing proficiency.

Criteria	Measures	Code	Calculation
Lexical richness	Lexical diversity	MATTR	Moving average type token ratio per essay: $\mathrm{MATTR(W)}_{\text{word form}} = \frac{\sum_{i=1}^{N-W+1} F_i}{W(N-W+1)}$
	Lexical density	LD	# of lexical words (token)/# of words
	Lexical sophistication	LS	# of sophisticated word types / total # of words per essay
Syntactic complexity	Mean dependency distance	MDD	$\mathrm{MDD} = \frac{1}{n}\sum_{i=1}^{n}\lvert \mathrm{DD}_i \rvert$
	Mean length of clause	MLC	# of words / # of clauses
	Verb phrases per T-unit	VPT	# of verb phrases / # of T-units
	Clauses per T-unit	CT	# of clauses / # of T-units
	Dependent clauses per T-unit	DCT	# of dependent clauses / # of T-units
	Complex nominals per T-unit	CNT	# of complex nominals / # of T-units
	Adverbial clauses rate	ACC	# of adverbial clauses / # of clauses
	Coordinate phrases rate	CPC	# of coordinate phrases / # of clauses
Cohesion	Semantic similarity	SOPT	Synonym overlap/paragraph (topic)
		SOPK	Synonym overlap/paragraph (keywords)
	word2vec cosine similarity	word2vec	(a). word2vec → (b). cosine similarity between sample and reference
Content elaboration	Metadiscourse marker rate	IMM	# of metadiscourse marker (type) / Number of words
Grammatical accuracy	Grammatical error rate	GER	# of errors per essay

[*]The term "token" refers to the occurrence of a word in the text.

(7) 天気が良かったので (subclause)、ケンとマリはピクニックに行きました (main clause): 1 T-unit.

(iii) In the case of coordinate clauses, where multiple clauses are connected, each coordinated clause is counted separately. Thus, a sentence with coordinate clauses may have 2 T-units or more, e.g. (8).

(8) ケンは地図で場所を探して (coordinate clause)、マリはサンドイッチを作りました (coordinate clause): 2 T-units.

Lexical diversity refers to the range of words used within a text (Engber, 1995; Kyle et al. 2021) and is considered a useful measure of the breadth of vocabulary in L*n* production (Jarvis, 2013a, 2013b).

The type/token ratio (TTR) is widely recognized as a straightforward measure for calculating lexical diversity and has been employed in numerous studies. These studies have demonstrated a strong correlation between TTR and other methods of measuring lexical diversity (e.g., Bentz et al. 2016; Čech and Miroslav, 2018; Çöltekin and Taraka, 2018). TTR is computed by considering both the number of unique words (types) and the total number of words (tokens) in a given text. Given that the length of learners' writing texts can vary, this study employs the moving average type-token ratio (MATTR) to mitigate the influence of text length. MATTR is calculated using a 50-word moving window. Initially, a TTR is determined for words 1–50 in an essay, followed by words 2–51, 3–52, and so on until the end of the essay is reached (Díez-Ortega and Kyle, 2023). The final MATTR scores were obtained by averaging the TTR scores for all 50-word windows. The following formula was employed to derive MATTR:

(9) $$\mathrm{MATTR(W)} = \frac{\sum_{i=1}^{N-W+1} F_i}{W(N-W+1)}$$

Here, N refers to the number of tokens in the corpus. W is the randomly selected token size ($W < N$). F_i is the number of types in each window. The MATTR(W) is the mean of a series of type-token ratios (TTRs) based on the word form for all windows. It is expected that individuals with higher language proficiency will produce texts with greater lexical diversity, as indicated by higher MATTR scores.

Lexical density was captured by the ratio of the number of lexical words to the total number of words (Lu, 2012).

Lexical sophistication refers to the utilization of advanced vocabulary, often evaluated through word frequency indices (Crossley et al. 2013; Haberman, 2008; Kyle and Crossley, 2015; Laufer and Nation, 1995; Lu, 2012; Read, 2000). In line of writing, lexical sophistication can be interpreted as vocabulary breadth, which entails the appropriate usage of vocabulary items across various lexicon-grammatical contexts and registers (Garner et al. 2019; Kim et al. 2018; Kyle et al. 2018). In Japanese specifically, words are considered lexically sophisticated if they are not included in the "Japanese Education Vocabulary List Ver 1.0".[4] Consequently, lexical sophistication was calculated by determining the number of sophisticated word types relative to the total number of words per essay. Furthermore, it has been suggested that, in Japanese writing, sentences should ideally have a length of no more than 40 to 50 characters, as this promotes readability. Therefore, the median and maximum sentence length can be considered as useful indices for assessment (Ishioka and Kameda, 2006).

Syntactic complexity was assessed based on several measures, including the mean length of clauses, verb phrases per T-unit, clauses per T-unit, dependent clauses per T-unit, complex nominals per clause, adverbial clauses per clause, coordinate phrases per clause, and mean dependency distance (MDD). The MDD reflects the distance between the governor and dependent positions in a sentence. A larger dependency distance indicates a higher cognitive load and greater complexity in syntactic processing (Liu, 2008; Liu et al. 2017). The MDD has been established as an efficient metric for measuring syntactic complexity (Jiang, Quyang, and Liu, 2019; Li and Yan, 2021). To calculate the MDD, the position numbers of the governor and dependent are subtracted, assuming that words in a sentence are assigned in a linear order, such as W1 … Wi … Wn. In any dependency relationship between words Wa and Wb, Wa is the governor and Wb is the dependent. The MDD of the entire sentence was obtained by taking the absolute value of governor – dependent:

Table 3 CSV file for the MDD calculation.

Sentence number	Dependent			Governor			Dependency type
	Order number	Word	POS	Order number	Word	POS	
S1	1	Mary	noun	7	watashi	verb	nsubj
S1	2	ga	case	1	Mary	noun	case
S1	3	John	noun	7	watashi	verb	obl
S1	4	ni	case	3	John	noun	case
S1	5	keshigomu	noun	7	watashi	verb	obj
S1	6	o	case	5	keshigomu	noun	case
S1	7	watashita	verb	0	/	/	root
S1	8	ta	tense	7	watashi	verb	aux

(10) $\mathrm{MDD} = \frac{1}{n}\sum_{i=1}^{n}|\mathrm{DD}_i|$

In this formula, n represents the number of words in the sentence, and DDi is the dependency distance of the ith dependency relationship of a sentence. Building on this, the annotation of sentence 'Mary-ga-John-ni-keshigomu-o-watashita was [Mary-TOP-John-DAT-eraser-ACC-give-PAST]'. The sentence's MDD would be 2. Table 3 provides the CSV file as a prompt for GPT 4.

Cohesion (semantic similarity) and content elaboration aim to capture the ideas presented in test taker's essays. Cohesion was assessed using three measures: Synonym overlap/paragraph (topic), Synonym overlap/paragraph (keywords), and word2vec cosine similarity. Content elaboration and development were measured as the number of metadiscourse markers (type)/ number of words. To capture content closely, this study proposed a novel-distance based representation, by encoding the cosine distance between the essay (by learner) and essay task's (topic and keyword) i-vectors. The learner's essay is decoded into a word sequence, and aligned to the essay task' topic and keyword for log-likelihood measurement. The cosine distance reveals the content elaboration score in the leaners' essay. The mathematical equation of cosine similarity between target-reference vectors is shown in (11), assuming there are i essays and (L_i, …. L_n) and (N_i, …. N_n) are the vectors representing the learner and task's topic and keyword respectively. The content elaboration distance between L_i and N_i was calculated as follows:

(11) $\cos(\theta) = \frac{\mathrm{L}\cdot\mathrm{N}}{|\mathrm{L}||\mathrm{N}|} = \frac{\sum_{i=1}^{n} L_i N_i}{\sqrt{\sum_{i=1}^{n} L_i^2}\sqrt{\sum_{i=1}^{n} N_i^2}}$

A high similarity value indicates a low difference between the two recognition outcomes, which in turn suggests a high level of proficiency in content elaboration.

To evaluate the effectiveness of the proposed measures in distinguishing different proficiency levels among nonnative Japanese speakers' writing, we conducted a multi-faceted Rasch measurement analysis (Linacre, 1994). This approach applies measurement models to thoroughly analyze various factors that can influence test outcomes, including test takers' proficiency, item difficulty, and rater severity, among others. The underlying principles and functionality of multi-faceted Rasch measurement are illustrated in (12).

(12) $\log\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_k$

(12) defines the logarithmic transformation of the probability ratio ($P_{nijk}/P_{nij(k-1)}$)) as a function of multiple parameters. Here, n represents the test taker, i denotes a writing proficiency measure, j corresponds to the human rater, and k represents the proficiency score. The parameter B_n signifies the proficiency level of test taker n (where n ranges from 1 to N). D_j represents the difficulty parameter of test item i (where i ranges from 1 to L), while C_j represents the severity of rater j (where j ranges from 1 to J). Additionally, F_k represents the step difficulty for a test taker to move from score 'k-1' to k. P_{nijk} refers to the probability of rater j assigning score k to test taker n for test item i. $P_{nij(k-1)}$ represents the likelihood of test taker n being assigned score 'k-1' by rater j for test item i. Each facet within the test is treated as an independent parameter and estimated within the same reference framework. To evaluate the consistency of scores obtained through both human and computer analysis, we utilized the Infit mean-square statistic. This statistic is a chi-square measure divided by the degrees of freedom and is weighted with information. It demonstrates higher sensitivity to unexpected patterns in responses to items near a person's proficiency level (Linacre, 2002). Fit statistics are assessed based on predefined thresholds for acceptable fit. For the Infit MNSQ, which has a mean of 1.00, different thresholds have been suggested. Some propose stricter thresholds ranging from 0.7 to 1.3 (Bond et al. 2021), while others suggest more lenient thresholds ranging from 0.5 to 1.5 (Eckes, 2009). In this study, we adopted the criterion of 0.70–1.30 for the Infit MNSQ.

Moving forward, we can now proceed to assess the effectiveness of the 16 proposed measures based on five criteria for accurately distinguishing various levels of writing proficiency among non-native Japanese speakers. To conduct this evaluation, we utilized the development dataset from the I-JAS corpus, as described in Section **Dataset**. Table 4 provides a measurement report that presents the performance details of the 14 metrics under consideration. The measure separation was found to be 4.02, indicating a clear differentiation among the measures. The reliability index for the measure separation was 0.891, suggesting consistency in the measurement. Similarly, the person separation reliability index was 0.802, indicating the accuracy of the assessment in distinguishing between individuals. All 16 measures demonstrated Infit mean squares within a reasonable range, ranging from 0.76 to 1.28. The Synonym overlap/paragraph (topic) measure exhibited a relatively high outfit mean square of 1.46, although the Infit mean square falls within an acceptable range. The standard error for the measures ranged from 0.13 to 0.28, indicating the precision of the estimates.

Table 5 further illustrated the weights assigned to different linguistic measures for score prediction, with higher weights indicating stronger correlations between those measures and higher scores. Specifically, the following measures exhibited higher weights compared to others: moving average type token ratio per essay has a weight of 0.0391. Mean dependency distance had a weight of 0.0388. Mean length of clause, calculated by dividing the number of words by the number of clauses, had a weight of 0.0374. Complex nominals per T-unit, calculated by dividing the number of complex nominals by the number of T-units, had a weight of 0.0379. Coordinate phrases rate, calculated by dividing the number of coordinate phrases by the number of clauses, had a weight of 0.0325. Grammatical error rate, representing the number of errors per essay, had a weight of 0.0322.

Table 4 Criteria measurement report.

Criteria	Measures	Infit Mnsq	Outfit Mnsq	S.E.
Lexical richness	Lexical diversity	1.14	1.42	0.21
	Lexical density	0.99	0.85	0.24
	Lexical sophistication	1.12	1.21	0.19
Syntactic diversity	Mean dependency distance	1.21	1.21	0.13
	Mean length of clause	0.94	1.04	0.19
	Verb phrases per T-unit	0.89	0.93	0.20
	Clauses per T-unit	0.95	0.99	0.15
	Dependent clauses per T-unit	1.01	1.11	0.14
	Complex nominals per T-unit	1.13	1.05	0.20
	Adverbial clause rate	0.79	0.74	0.12
	Coordinate phrases rate	1.09	0.96	0.14
Cohesion	Synonym overlap/paragraph (topic)	1.22	1.46	0.24
	Synonym overlap / paragraph (keywords)	1.28	1.23	0.21
	(a). word2vec → (b). cosine similarity between sample and reference	0.76	0.85	0.28
Content elaboration	Metadiscourse marker rate	0.82	0.76	0.24
Grammatical accuracy	Grammatical error rate	0.86	0.93	0.26

Table 5 The weights of linguistic features in score prediction.

Measures	Calculation	Weight
Lexical diversity	Moving average type token ratio per essay	0.0391
Lexical density	# of lexical words (token)/# of words	0.0200
Lexical sophistication	# of sophisticated word types / total # of words per essay	0.0212
Mean dependency distance	$MDD = \frac{1}{n}\sum_{i=1}^{n}\|DD_i\|$	0.0388
Mean length of clause	# of words / # of clauses	0.0374
Verb phrases per T-unit	# of verb phrases / # of T-units	0.0112
Clauses per T-unit	# of clauses / # of T-units	0.0021
Dependent clauses per T-unit	# of dependent clauses / # of T-units	0.0029
Complex nominals per T-unit	# of complex nominals / # of T-units	0.0379
Adverbial clauses rate	# of adverbial clauses / # of clauses	0.0108
Coordinate phrases rate	# of coordinate phrases / # of clauses	0.0325
Semantic similarity	Synonym overlap/paragraph (topic)	0.0072
	Synonym overlap / paragraph (keywords)	0.0075
word2vec cosine similarity	(a). word2vec → (b). cosine similarity between sample and reference	0.0295
Metadiscourse marker rate	# of metadiscourse marker (type) / Number of words	0.0103
Grammatical error rate	# of errors per essay	0.0322

Criteria (output indicator). The criteria used to evaluate the writing ability in this study were based on CEFR, which follows a six-point scale ranging from A1 to C2. To assess the quality of Japanese writing, the scoring criteria from Table 6 were utilized. These criteria were derived from the IELTS writing standards and served as assessment guidelines and prompts for the written output.

Prompt. A prompt is a question or detailed instruction that is provided to the model to obtain a proper response. After several pilot experiments, we decided to provide the measures (Section **Measures of writing proficiency for nonnative Japanese**) as the input prompt and use the criteria (Section **Criteria (output indicator)**) as the output indicator. Regarding the prompt language, considering that the LLM was tasked with rating Japanese essays, would prompt in Japanese works better[5]? We conducted experiments comparing the performance of GPT-4 using both English and Japanese prompts. Additionally, we utilized the Japanese local model OCLL with Japanese prompts. Multiple trials were conducted using the same sample. Regardless of the prompt language used, we consistently obtained the same grading results with GPT-4, which assigned a grade of B1 to the writing sample. This suggested that GPT-4 is reliable and capable of producing consistent ratings regardless of the prompt language. On the other hand, when we used Japanese prompts with the Japanese local model "OCLL", we encountered inconsistent grading results. Out of 10 attempts with OCLL, only 6 yielded consistent grading results (B1), while the remaining 4 showed different outcomes, including A1 and B2 grades. These findings indicated that the language of the prompt was not the determining factor for reliable AES. Instead, the size of the training data and the model parameters played crucial roles in achieving consistent and reliable AES results for the language model.

The following is the utilized prompt, which details all measures and requires the LLM to score the essays using holistic and trait scores.

The prompt. Please evaluate Japanese essays written by Japanese learners and assign a score to each essay on a six-point scale, ranging from A1, A2, B1, B2, C1 to C2. Additionally, please provide trait scores and display the calculation process for each trait score. The scoring should be based on the following criteria:

- Moving average type-token ratio.
- Number of lexical words (token) divided by the total number of words per essay.
- Number of sophisticated word types divided by the total number of words per essay.
- Mean length of clause.
- Verb phrases per T-unit.

Table 6 Criteria for evaluating nonnative Japanese writing quality.

Level	Lexical richness	Grammatical accuracy	Cohesion
A1	The lexicon is extremely limited with few recognizable strings	There is little or no evidence of sentence forms.	There is little relevant message, of the entire response may be off-topic
A2	The lexicon is limited and inadequate for or unrelated to the task. Vocabulary is basic and may be used repetitively.	A very limited range of structures is used. Subordinate clauses are rare; simple sentences predominate.	Information and ideas are evident but not arranged coherently; There is no clearer progression within the response.
B1	The lexicon is generally adequate and appropriate. The meaning is generally clear in spite of a rather restricted range or a lack of precision in word choice.	A mix of simple and complex sentence forms is used but flexibility is limited. Examples of more complex structures are not marked by the same level of accuracy as in simple structures.	Information and ideas are generally arranged coherently and there is a clear overall progression. Cohesive devices are used to some good effect but cohesion within and/ or mechanical due to misuse, overuse, or omission.
B2	The resource is sufficient to allow some flexibility and precision. There is some ability to use less common and/or idiomatic items.	A variety of complex structures is used with some flexibility and accuracy.	Information and ideas are logically organized and there is a clear progression throughout the response. A range of cohesive devices including reference and substitution is used flexibility but with some inaccuracies or some over/under use.
C1	Diverse lexicon is fluently and flexibly used to convey precise meanings within the scope of the task. There is skillful use of uncommon and or idiomatic items when appropriate, despite occasional inaccuracies in word choice and collocation.	A wide range of structures within the scope of the task is flexibility and accurately used. The majority of sentences are error-free.	The message can be followed with ease. Information and ideas are logically sequenced. Cohesion is well managed. Occasional lapses in coherence or cohesion may occur. Paragraphing is used sufficiently and appropriately.
C2	Full flexibility and precise use are evident within the scope of the task. A wide range of vocabulary is used accurately and appropriately with very natural and sophisticated control of lexical features.	A wide range of structures within the scope of the task is used with full flexibility and control.	The message can be followed effortlessly. Cohesion is used in such a way but it very rarely attracts attention. Any lapses in coherence or cohesion are minimal.

- Clauses per T-unit.
- Dependent clauses per T-unit.
- Complex nominals per clause.
- Adverbial clauses per clause.
- Coordinate phrases per clause.
- Mean dependency distance.
- Synonym overlap paragraph (topic and keywords).
- Word2vec cosine similarity.
- Connectives per essay.
- Conjunctions per essay.
- Number of metadiscourse markers (types) divided by the total number of words.
- Number of errors per essay.

Japanese essay text. 出かける前に二人が地図を見ている間に、サンドイッチを入れたバスケットに犬が入ってしまいました。それに気づかずに二人は楽しそうに出かけて行きました。やがて突然犬がバスケットから飛び出し、二人は驚きました。バスケットの中を見ると、食べ物はすべて犬に食べられていて、二人は困ってしまいました。(ID_JJJ01_SW1)

The score of the example above was B1. Figure 3 provides an example of holistic and trait scores provided by GPT-4 (with a prompt indicating all measures) via Bing[6].

Statistical analysis. The aim of this study is to investigate the potential use of LLM for nonnative Japanese AES. It seeks to compare the scoring outcomes obtained from feature-based AES tools, which rely on conventional machine learning technology (i.e. Jess, JWriter), with those generated by AI-driven AES tools utilizing deep learning technology (BERT, GPT, OCLL). To assess the reliability of a computer-assisted annotation tool, the study initially established human-human agreement as the benchmark measure. Subsequently, the performance of the LLM-based method was evaluated by comparing it to human-human agreement.

To assess annotation agreement, the study employed standard measures such as precision, recall, and F-score (Brants 2000; Lu 2010), along with the quadratically weighted kappa (QWK) to evaluate the consistency and agreement in the annotation process. Assume A and B represent human annotators. When comparing the annotations of the two annotators, the following results are obtained. The evaluation of precision, recall, and F-score metrics was illustrated in equations (13) to (15).

$$\mathrm{Recall}(A,B) = \frac{\text{Number of identical nodes in } A \text{ and } B}{\text{Number of nodes in } A} \tag{13}$$

$$\mathrm{Precision}(A,B) = \frac{\text{Number of identical nodes in } A \text{ and } B}{\text{Number of nodes in } B} \tag{14}$$

The F-score is the harmonic mean of recall and precision:

$$\mathrm{F-score} = \frac{2*(\mathrm{Precision}*\mathrm{Recall})}{\mathrm{Precision}+\mathrm{Recall}} \tag{15}$$

The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either precision or recall are zero.

In accordance with Taghipour and Ng (2016), the calculation of QWK involves two steps:

Step 1: Construct a weight matrix W as follows:

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{16}$$

i represents the annotation made by the tool, while j represents the annotation made by a human rater. N denotes the total number of possible annotations. Matrix O is subsequently computed, where O_(i, j) represents the count of data annotated by the tool (i) and the human annotator (j). On the other hand, E refers to the expected count matrix, which undergoes normalization to ensure

Hello, this is Bing. I can help you score a Japanese essay written by Japanese learners. I will use the following measures to evaluate the essay:

- Moving average type token ratio: a measure of lexical diversity that calculates the average number of different words in a sliding window of a fixed size.
- Number of lexical words (token)/number of words per essay: a measure of lexical density that indicates the proportion of content words in an essay.
- Number of sophisticated word types/total number of words per essay: a measure of lexical sophistication that reflects the use of rare or advanced words in an essay.
- Mean length of clause; verb phrases per T-unit; clauses per T-unit; dependent clauses per T-unit; complex nominals per clause; adverbial clauses per clause; coordinate phrases per clause: measures of syntactic complexity that capture the length and structure of sentences and clauses in an essay.
- Mean dependency distance: a measure of syntactic difficulty that calculates the average distance between words and their syntactic heads in an essay.
- Synonym overlap paragraph (topic, keywords): a measure of content relevance that computes the similarity between the topic and keywords given in the prompt and the words used in each paragraph of an essay.
- Word2vec cosine similarity: a measure of semantic coherence that calculates the average cosine similarity between word vectors in adjacent sentences of an essay.
- Connectives per essay, conjunctions per essay: measures of cohesion that count the number of connective words and phrases that link sentences and clauses in an essay.
- Number of metadiscourse marker (type) / number of words: a measure of discourse organization that indicates the frequency and variety of metadiscourse markers that guide the reader through the essay.
- Number of errors per essay: a measure of language accuracy that counts the number of grammatical, spelling, punctuation, and word choice errors in an essay.

Based on these measures, I will score each essay on a six-point scale, i.e., ascending from A1, A2, B1, B2, C1 to C2. I will also report trait scores for each scoring section: content, organization, language.

Here is my evaluation for the Japanese essay text you provided:

- Moving average type token ratio: 0.83
- Number of lexical words (token)/number of words per essay: 0.71
- Number of sophisticated word types/total number of words per essay: 0.07
- Mean length of clause: 5.33
- Verb phrases per T-unit: 1.33
- Clauses per T-unit: 1.67
- Dependent clauses per T-unit: 0.67
- Complex nominals per clause: 0.33
- Adverbial clauses per clause: 0.17
- Coordinate phrases per clause: 0.17
- Mean dependency distance: 2.29
- Synonym overlap paragraph (topic, keywords): 0.12
- Word2vec cosine similarity: 0.76
- Connectives per essay: 3
- Conjunctions per essay: 4
- Number of metadiscourse marker (type) / number of words: 0.04
- Number of errors per essay: 2

Based on these measures, I will assign the following scores to the essay:

- Content score: B1
  - The essay has a clear topic and a simple plot, but lacks details and development. The synonym overlap between the topic and keywords and the words in each paragraph is low, indicating that the content is not very relevant to the prompt.
- Organization score: B1
  - The essay has a basic structure with an introduction, a climax, and a conclusion, but lacks transitions and coherence. The word2vec cosine similarity between adjacent sentences is high, indicating that the sentences are semantically related, but the use of connectives and conjunctions is limited, indicating that the sentences are not well connected.
- Language score: A2
  - The essay uses simple and familiar vocabulary and grammar, but has some errors and inconsistencies. The moving average type token ratio is high, indicating that the vocabulary is diverse, but the number of sophisticated word types is low, indicating that the vocabulary is not advanced. The mean length of clause and the mean dependency distance are low, indicating that the syntax is simple and easy. The measures of syntactic complexity are also low, indicating that the sentences and clauses are short and simple.

Overall score: B1

The overall score is determined by taking the highest score among the three trait scores. In this case, both content and organization scores are B1, so the overall score is B1.

I hope this helps you understand how to score a Japanese essay written by Japanese learners. If you have any questions or feedback, please let me know. 😊

Fig. 3 Example of GPT-4 AES and feedback (with a prompt indicating all measures).

that the sum of elements in E matches the sum of elements in O.

Step 2: With matrices O and E, the QWK is obtained as follows:

(17) $$K = 1 - \frac{\sum i,jW_{i,j}\,O_{i,j}}{\sum i,jW_{i,j}\,E_{i,j}}$$

The value of the quadratic weighted kappa increases as the level of agreement improves. Further, to assess the accuracy of LLM scoring, the proportional reductive mean square error (PRMSE) was employed. The PRMSE approach takes into account the variability observed in human ratings to estimate the rater error, which is then subtracted from the variance of the human labels. This calculation provides an overall measure of agreement between the automated scores and true scores (Haberman et al. 2015; Loukina et al. 2020; Taghipour and Ng, 2016). The computation of PRMSE involves the following steps:

Step 1: Calculate the mean squared errors (MSEs) for the scoring outcomes of the computer-assisted tool (MSE tool) and the human scoring outcomes (MSE human).

Step 2: Determine the PRMSE by comparing the MSE of the computer-assisted tool (MSE tool) with the MSE from human raters (MSE human), using the following formula:

$$\mathrm{PRMSE} = 1 - \frac{(\mathrm{MSE\ tool})}{(\mathrm{MSE\ human})} = 1 - \frac{\sum_i^n = 1(y_i - \hat{y}_i)^2}{\sum_i^n = 1(y_i - \hat{y})^2}$$

In the numerator, $\hat{y}i$ represents the scoring outcome predicted by a specific LLM-driven AES system for a given sample. The term $y_i - \hat{y}_i$ represents the difference between this predicted outcome and the mean value of all LLM-driven AES systems' scoring outcomes. It quantifies the deviation of the specific LLM-driven AES system's prediction from the average prediction of all LLM-driven AES systems. In the denominator, $y_i - \hat{y}$ represents the difference between the scoring outcome provided by a specific human rater for a given sample and the mean value of all human raters' scoring outcomes. It measures the discrepancy between the specific human rater's score and the average score given by all human raters. The PRMSE is then calculated by subtracting the ratio of the MSE tool to the MSE human from 1. PRMSE falls within the range of 0 to 1, with larger values indicating reduced errors in LLM's scoring compared to those of human raters. In other words, a higher PRMSE implies that LLM's scoring demonstrates greater accuracy in predicting the true scores (Loukina et al. 2020). The interpretation of kappa values, ranging from 0 to 1, is based on the work of Landis and Koch (1977). Specifically, the following categories are assigned to different ranges of kappa values: −1 indicates complete inconsistency, 0 indicates random agreement, 0.0 ~ 0.20 indicates extremely low level of agreement (slight), 0.21 ~ 0.40 indicates moderate level of agreement (fair), 0.41 ~ 0.60 indicates medium level of agreement (moderate), 0.61 ~ 0.80 indicates high level of agreement (substantial), 0.81 ~ 1 indicates almost perfect level of agreement. All statistical analyses were executed using Python script.

Results and discussion

Annotation reliability of the LLM. This section focuses on assessing the reliability of the LLM's annotation and scoring capabilities. To evaluate the reliability, several tests were conducted simultaneously, aiming to achieve the following objectives:

a. Assess the LLM's ability to differentiate between test takers with varying levels of oral proficiency.
b. Determine the level of agreement between the annotations and scoring performed by the LLM and those done by human raters.

The evaluation of the results encompassed several metrics, including: precision, recall, F-Score, quadratically-weighted

Table 7 Inter-annotator (human-human) agreement on measures (precision, recall, F-score).

Metrics	Counts			Inter-annotator agreement		
	Annotator 1	Annotator 2	Identical	Recall	Precision	F-score
Sentence #	10	10	10	1.000	1.000	1.000
Clause #	29	32	28	0.966	0.875	0.918
T unit #	17	19	15	0.882	0.789	0.833
Word #	143	143	143	1.000	1.000	1.000
Complex T unit #	8	8	7	0.875	0.875	0.875
Dependent clause #	10	8	7	0.700	0.875	0.778
Coordinate phrase #	2	3	2	1.000	0.667	0.800
Complex nominal #	3	4	3	1.000	0.750	0.857
Verb phrase #	40	42	37	0.925	0.881	0.902
Adverbial clauses #	19	20	18	0.947	0.900	0.923
Synonym overlap # (topic)	16	17	14	0.875	0.823	0.848
Synonym overlap # (keyword)	20	23	19	0.950	0.826	0.883
Metadiscourse marker #	10	11	8	0.800	0.727	0.762
Grammatical error #	7	8	5	0.714	0.625	0.666

kappa, proportional reduction of mean squared error, Pearson correlation, and multi-faceted Rasch measurement.

Inter-annotator agreement (human–human annotator agreement). We started with an agreement test of the two human annotators. Two trained annotators were recruited to determine the writing task data measures. A total of 714 scripts, as the test data, was utilized. Each analysis lasted 300–360 min. Inter-annotator agreement was evaluated using the standard measures of precision, recall, and F-score and QWK. Table 7 presents the inter-annotator agreement for the various indicators. As shown, the inter-annotator agreement was fairly high, with F-scores ranging from 1.0 for sentence and word number to 0.666 for grammatical errors.

The findings from the QWK analysis provided further confirmation of the inter-annotator agreement. The QWK values covered a range from 0.950 ($p = 0.000$) for sentence and word number to 0.695 for synonym overlap number (keyword) and grammatical errors ($p = 0.001$).

Agreement of annotation outcomes between human and LLM. To evaluate the consistency between human annotators and LLM annotators (BERT, GPT, OCLL) across the indices, the same test was conducted. The results of the inter-annotator agreement (F-score) between LLM and human annotation are provided in Appendix B-D. The F-scores ranged from 0.706 for Grammatical error # for OCLL-human to a perfect 1.000 for GPT-human, for sentences, clauses, T-units, and words. These findings were further supported by the QWK analysis, which showed agreement levels ranging from 0.807 ($p = 0.001$) for metadiscourse markers for OCLL-human to 0.962 for words ($p = 0.000$) for GPT-human. The findings demonstrated that the LLM annotation achieved a significant level of accuracy in identifying measurement units and counts.

Reliability of LLM-driven AES's scoring and discriminating proficiency levels. This section examines the reliability of the LLM-driven AES scoring through a comparison of the scoring outcomes produced by human raters and the LLM (**Reliability of LLM-driven AES scoring**). It also assesses the effectiveness of the LLM-based AES system in differentiating participants with varying proficiency levels (**Reliability of LLM-driven AES discriminating proficiency levels**).

Reliability of LLM-driven AES scoring. Table 8 summarizes the QWK coefficient analysis between the scores computed by the human raters and the GPT-4 for the individual essays from I-JAS[7]. As shown, the QWK of all measures ranged from $k = 0.819$ for lexical density (number of lexical words (tokens)/number of words per essay) to $k = 0.644$ for word2vec cosine similarity. Table 9 further presents the Pearson correlations between the 16 writing proficiency measures scored by human raters and GPT 4 for the individual essays. The correlations ranged from 0.672 for syntactic complexity to 0.734 for grammatical accuracy. The correlations between the writing proficiency scores assigned by human raters and the BERT-based AES system were found to range from 0.661 for syntactic complexity to 0.713 for grammatical accuracy. The correlations between the writing proficiency scores given by human raters and the OCLL-based AES system ranged from 0.654 for cohesion to 0.721 for grammatical accuracy. These findings indicated an alignment between the assessments made by human raters and both the BERT-based and OCLL-based AES systems in terms of various aspects of writing proficiency.

Table 8 Agreement between the scores computed by human and GPT-4 (QWK).

	Measures	Agreement	Measures	Agreement
Human scoring-GPT 4 scoring	MATTR	0.655	CN	0.807
	LD	0.819	ACC	0.794
	LS	0.679	CPC	0.783
	MDD	0.743	SOPT	0.798
	MLC	0.812	SOPK	0.805
	VPT	0.754	word2vec	0.644
	CT	0.667	IMM	0.680
	DCT	0.803	GE	0.648

Table 9 Correlations between writing proficiency scoring by human raters and GPT 4-based AES.

	Measures	Correlation
Human scoring -GPT 4 scoring	Lexical richness	0.708
	Syntactic complexity	0.672
	Cohesion	0.751
	Content elaboration	0.722
	Grammatical accuracy	0.734

Reliability of LLM-driven AES discriminating proficiency levels. After validating the reliability of the LLM's annotation and

Table 10 Distribution of outcomes determined by GPT-4.

Measures	Primary level ($N = 197$)		Intermediate level ($N = 256$)		Advanced level ($N = 233$)		ANOVA	
	Mean	SD	Mean	SD	Mean	SD	F	Sig.
MATTR	**0.676**	**0.341**	**0.704**	**0.229**	**0.731**	**0.317**	**33.101**	**0.000**
LD	0.637	0.221	0.634	0.290	0.703	0.211	3.401	0.031
LS	0.701	0.310	0.753	0.343	0.751	0.212	3.204	0.310
MDD	**3.773**	**0.221**	**4.050**	**0.265**	**4.988**	**0.281**	**19.880**	**0.000**
MLC	**4.355**	**1.440**	**5.320**	**1.870**	**5.882**	**1.870**	**50.098**	**0.000**
VPT	2.133	0.451	2.099	0.322	1.970	0.331	6.454	0.001
CT	1.662	0.223	1.661	0.277	1.664	0.263	3.550	0.039
DCT	0.622	0.341	0.579	0.240	0.657	0.298	3.880	0.038
CNT	**0.922**	**0.222**	**0.931**	**0.297**	**1.099**	**0.200**	**44.099**	**0.000**
ACC	0.444	0.109	0.440	0.122	0.449	0.131	3.750	0.039
CPC	**0.209**	**0.098**	**0.227**	**0.088**	**0.237**	**0.081**	**40.373**	**0.000**
SOPT	0.761	0.101	0.877	0.112	0.812	0.103	21.881	0.000
SOPK	0.742	0.142	0.788	0.220	0.746	0.100	18.977	0.000
Word2vec	**0.697**	**0.351**	**0.702**	**0.351**	**0.713**	**0.311**	**33.411**	**0.000**
IMM	0.029	0.011	0.033	0.014	0.022	0.018	3.400	0.035
GER	**0.380**	**0.265**	**0.371**	**0.346**	**0.255**	**0.105**	**16.988**	**0.000**

Based on the results, seven measures exhibit statistically significant between-level differences and demonstrate a linear progression across the three proficiency levels. These measures are highlighted in the table.

scoring, the subsequent objective was to evaluate its ability to distinguish between various proficiency levels. For this analysis, a dataset of 686 individual essays was utilized. Table 10 presents a sample of the results, summarizing the means, standard deviations, and the outcomes of the one-way ANOVAs based on the measures assessed by the GPT-4 model. A post hoc multiple comparison test, specifically the Bonferroni test, was conducted to identify any potential differences between pairs of levels.

As the results reveal, seven measures presented linear upward or downward progress across the three proficiency levels. These were marked in bold in Table 10 and comprise one measure of lexical richness, i.e. MATTR (lexical diversity); four measures of syntactic complexity, i.e. MDD (mean dependency distance), MLC (mean length of clause), CNT (complex nominals per T-unit), CPC (coordinate phrases rate); one cohesion measure, i.e. word2vec cosine similarity and GER (grammatical error rate). Regarding the ability of the sixteen measures to distinguish adjacent proficiency levels, the Bonferroni tests indicated that statistically significant differences exist between the primary level and the intermediate level for MLC and GER. One measure of lexical richness, namely LD, along with three measures of syntactic complexity (VPT, CT, DCT, ACC), two measures of cohesion (SOPT, SOPK), and one measure of content elaboration (IMM), exhibited statistically significant differences between proficiency levels. However, these differences did not demonstrate a linear progression between adjacent proficiency levels. No significant difference was observed in lexical sophistication between proficiency levels.

To summarize, our study aimed to evaluate the reliability and differentiation capabilities of the LLM-driven AES method. For the first objective, we assessed the LLM's ability to differentiate between test takers with varying levels of oral proficiency using precision, recall, F-Score, and quadratically-weighted kappa. Regarding the second objective, we compared the scoring outcomes generated by human raters and the LLM to determine the level of agreement. We employed quadratically-weighted kappa and Pearson correlations to compare the 16 writing proficiency measures for the individual essays. The results confirmed the feasibility of using the LLM for annotation and scoring in AES for nonnative Japanese. As a result, Research Question 1 has been addressed.

Comparison of BERT-, GPT-, OCLL-based AES, and linguistic-feature-based computation methods. This section aims to compare the effectiveness of five AES methods for nonnative Japanese writing, i.e. LLM-driven approaches utilizing BERT, GPT, and OCLL, linguistic feature-based approaches using Jess and JWriter. The comparison was conducted by comparing the ratings obtained from each approach with human ratings. All ratings were derived from the dataset introduced in **Dataset**. To facilitate the comparison, the agreement between the automated methods and human ratings was assessed using QWK and PRMSE. The performance of each approach was summarized in Table 11.

The QWK coefficient values indicate that LLMs (GPT, BERT, OCLL) and human rating outcomes demonstrated higher agreement compared to feature-based AES methods (Jess and JWriter) in assessing writing proficiency criteria, including lexical richness, syntactic complexity, content, and grammatical accuracy. Among the LLMs, the GPT-4 driven AES and human rating outcomes showed the highest agreement in all criteria, except for syntactic complexity. The PRMSE values suggest that the GPT-based method outperformed linguistic feature-based methods and other LLM-based approaches. Moreover, an interesting finding emerged during the study: the agreement coefficient between GPT-4 and human scoring was even higher than the agreement between different human raters themselves. This discovery highlights the advantage of GPT-based AES over human rating. Ratings involve a series of processes, including reading the learners' writing, evaluating the content and language, and assigning scores. Within this chain of processes, various biases can be introduced, stemming from factors such as rater biases, test design, and rating scales. These biases can impact the consistency and objectivity of human ratings. GPT-based AES may benefit from its ability to apply consistent and objective evaluation criteria. By prompting the GPT model with detailed writing scoring rubrics and linguistic features, potential biases in human ratings can be mitigated. The model follows a predefined set of guidelines and does not possess the same subjective biases that human raters may exhibit. This standardization in the evaluation process contributes to the higher agreement observed between GPT-4 and human scoring. Section **Prompt strategy** of the study delves further into the role of prompts in the application of LLMs to AES. It explores how the choice and

Table 11 Distribution of measured outcomes of linguistic feature-, LLM-based AES approach.

Statistical metric	AES system	Mean score (Lexical richness)	Mean score (Syntactic complexity)	Mean score (Content)	Mean score (Grammatical accuracy)
QWK	Jess-human	0.608	0.591	0.518	0.655
	JWriter-human	0.600	0.589	0.521	0.661
	BERT-human	0.653	0.652	0.638	0.671
	GPT 4 -human	0.665	0.655	0.657	0.689
	OCLL-human	0.639	0.648	0.623	0.662
	Human-human	0.657	0.639	0.578	0.677
PRMSE	Jess-human	0.691	0.658	0.601	0.689
	JWriter-human	0.683	0.675	0.584	0.732
	BERT-human	0.701	0.746	0.628	0.749
	GPT 4 -human	0.711	0.733	0.634	0.754
	OCLL-human	0.687	0.699	0.619	0.694
	Human-human	0.691	0.745	0.590	0.744

implementation of prompts can impact the performance and reliability of LLM-based AES methods. Furthermore, it is important to acknowledge the strengths of the local model, i.e. the Japanese local model OCLL, which excels in processing certain idiomatic expressions. Nevertheless, our analysis indicated that GPT-4 surpasses local models in AES. This superior performance can be attributed to the larger parameter size of GPT-4, estimated to be between 500 billion and 1 trillion, which exceeds the sizes of both BERT and the local model OCLL.

Prompt strategy. In the context of prompt strategy, Mizumoto and Eguchi (2023) conducted a study where they applied the GPT-3 model to automatically score English essays in the TOEFL test. They found that the accuracy of the GPT model alone was moderate to fair. However, when they incorporated linguistic measures such as cohesion, syntactic complexity, and lexical features alongside the GPT model, the accuracy significantly improved. This highlights the importance of prompt engineering and providing the model with specific instructions to enhance its performance. In this study, a similar approach was taken to optimize the performance of LLMs. GPT-4, which outperformed BERT and OCLL, was selected as the candidate model. Model 1 was used as the baseline, representing GPT-4 without any additional prompting. Model 2, on the other hand, involved GPT-4 prompted with 16 measures that included scoring criteria, efficient linguistic features for writing assessment, and detailed measurement units and calculation formulas. The remaining models (Models 3 to 18) utilized GPT-4 prompted with individual measures. The performance of these 18 different models was assessed using the output indicators described in Section **Criteria (output indicator)**. By comparing the performances of these models, the study aimed to understand the impact of prompt engineering on the accuracy and effectiveness of GPT-4 in AES tasks.

Baseline:
Model 1: GPT-4
17-shot prompting:
Model 2: GPT-4 + 17 measures
One-shot prompting:

Model 3: GPT-4 + MATTR	Model 4: GPT-4 + LD	Model 5: GPT-4 + LS
Model 6: GPT-4 + MLC	Model 7: GPT-4 + VPT	Model 8: GPT-4 + CT
Model 9: GPT-4 + DCT	Model 10: GPT-4 + CNT	Model 11: GPT-4 + ACC
Model 12: GPT-4 + CPC	Model 13: GPT-4 + MDD	Model 14: GPT-4 + SOPT
Model 15: GPT-4 + SOPK	Model 16: GPT-4 + word2vec	
Model 17: GPT-4 + IMM	Model 18: GPT-4 + GER	

Based on the PRMSE scores presented in Fig. 4, it was observed that Model 1, representing GPT-4 without any additional prompting, achieved a fair level of performance. However, Model 2, which utilized GPT-4 prompted with all measures, outperformed all other models in terms of PRMSE score, achieving a score of 0.681. These results indicate that the inclusion of specific measures and prompts significantly enhanced the performance of GPT-4 in AES. Among the measures, syntactic complexity was found to play a particularly significant role in improving the accuracy of GPT-4 in assessing writing quality. Following that, lexical diversity emerged as another important factor contributing to the model's effectiveness. The study suggests that a well-prompted GPT-4 can serve as a valuable tool to support human assessors in evaluating writing quality. By utilizing GPT-4 as an automated scoring tool, the evaluation biases associated with human raters can be minimized. This has the potential to empower teachers by allowing them to focus on designing writing tasks and guiding writing strategies, while leveraging the capabilities of GPT-4 for efficient and reliable scoring.

Conclusion

This study aimed to investigate two main research questions: the feasibility of utilizing LLMs for AES and the impact of prompt engineering on the application of LLMs in AES.

To address the first objective, the study compared the effectiveness of five different models: GPT, BERT, the Japanese local LLM (OCLL), and two conventional machine learning-based AES tools (Jess and JWriter). The PRMSE values indicated that the GPT-4-based method outperformed other LLMs (BERT, OCLL) and linguistic feature-based computational methods (Jess and JWriter) across various writing proficiency criteria. Furthermore, the agreement coefficient between GPT-4 and human scoring surpassed the agreement among human raters themselves, highlighting the potential of using the GPT-4 tool to enhance AES by reducing biases and subjectivity, saving time, labor, and cost, and providing valuable feedback for self-study. Regarding the second goal, the role of prompt design was investigated by comparing 18 models, including a baseline model, a model prompted with all measures, and 16 models prompted with one measure at a time. GPT-4, which outperformed BERT and OCLL, was selected as the candidate model. The PRMSE scores of the models showed that GPT-4

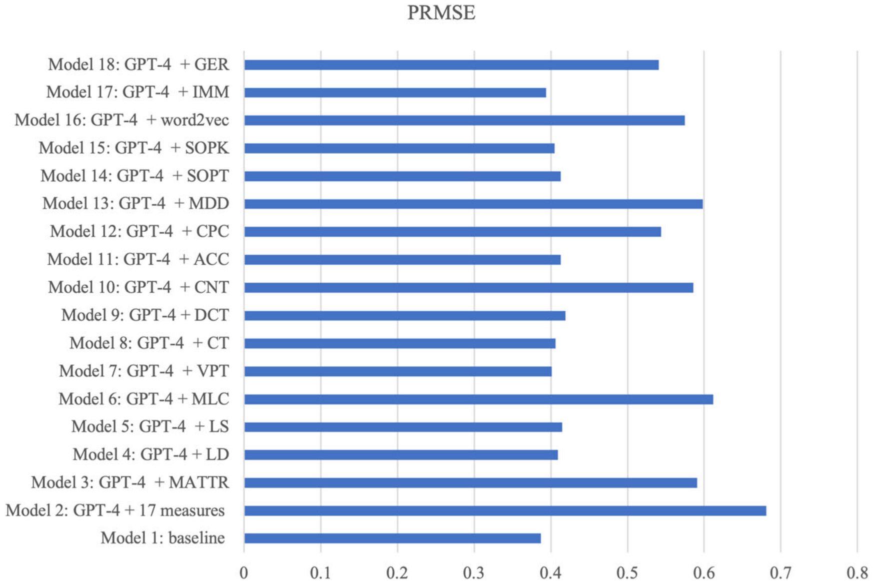

Fig. 4 PRMSE scores of the 18 AES models.

prompted with all measures achieved the best performance, surpassing the baseline and other models.

In conclusion, this study has demonstrated the potential of LLMs in supporting human rating in assessments. By incorporating automation, we can save time and resources while reducing biases and subjectivity inherent in human rating processes. Automated language assessments offer the advantage of accessibility, providing equal opportunities and economic feasibility for individuals who lack access to traditional assessment centers or necessary resources. LLM-based language assessments provide valuable feedback and support to learners, aiding in the enhancement of their language proficiency and the achievement of their goals. This personalized feedback can cater to individual learner needs, facilitating a more tailored and effective language-learning experience.

There are three important areas that merit further exploration. First, prompt engineering requires attention to ensure optimal performance of LLM-based AES across different language types. This study revealed that GPT-4, when prompted with all measures, outperformed models prompted with fewer measures. Therefore, investigating and refining prompt strategies can enhance the effectiveness of LLMs in automated language assessments. Second, it is crucial to explore the application of LLMs in second-language assessment and learning for oral proficiency, as well as their potential in under-resourced languages. Recent advancements in self-supervised machine learning techniques have significantly improved automatic speech recognition (ASR) systems, opening up new possibilities for creating reliable ASR systems, particularly for under-resourced languages with limited data. However, challenges persist in the field of ASR. First, ASR assumes correct word pronunciation for automatic pronunciation evaluation, which proves challenging for learners in the early stages of language acquisition due to diverse accents influenced by their native languages. Accurately segmenting short words becomes problematic in such cases. Second, developing precise audio-text transcriptions for languages with non-native accented speech poses a formidable task. Last, assessing oral proficiency levels involves capturing various linguistic features, including fluency, pronunciation, accuracy, and complexity, which are not easily captured by current NLP technology.

Data availability

The dataset utilized was obtained from the International Corpus of Japanese as a Second Language (I-JAS). The data URLs: [https://www2.ninjal.ac.jp/jll/lsaj/ihome2.html].

Received: 2 February 2024; Accepted: 16 May 2024;
Published online: 03 June 2024

Notes

1 J-CAT and TTBJ are two computerized adaptive tests used to assess Japanese language proficiency.SPOT is a specific component of the TTBJ test.J-CAT: https://www.j-cat2.org/html/ja/pages/interpret.htmlSPOT: https://ttbj.cegloc.tsukuba.ac.jp/p1.html#SPOT.

2 The study utilized a prompt-based GPT-4 model, developed by OpenAI, which has an impressive architecture with 1.8 trillion parameters across 120 layers. GPT-4 was trained on a vast dataset of 13 trillion tokens, using two stages: initial training on internet text datasets to predict the next token, and subsequent fine-tuning through reinforcement learning from human feedback.

3 https://www2.ninjal.ac.jp/jll/lsaj/ihome2-en.html.

4 http://jhlee.sakura.ne.jp/JEV/ by Japanese Learning Dictionary Support Group 2015.

5 We express our sincere gratitude to the reviewer for bringing this matter to our attention.

6 On February 7, 2023, Microsoft began rolling out a major overhaul to Bing that included a new chatbot feature based on OpenAI's GPT-4 (Bing.com).

7 Appendix E-F present the analysis results of the QWK coefficient between the scores computed by the human raters and the BERT, OCLL models.

8 Appendix E-F present the analysis results of the QWK coefficient between the scores computed by the human raters and the BERT, OCLL models.

References

Attali Y, Burstein J (2006) Automated essay scoring with e-rater® V.2. J. Technol., Learn. Assess., 4

Barkaoui K, Hadidi A (2020) Assessing Change in English Second Language Writing Performance (1st ed.). Routledge, New York. https://doi.org/10.4324/9781003092346

Bentz C, Tatyana R, Koplenig A, Tanja S (2016) A comparison between morphological complexity. measures: Typological data vs. language corpora. In Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC), 142–153. Osaka, Japan: The COLING 2016 Organizing Committee

Bond TG, Yan Z, Heene M (2021) Applying the Rasch model: Fundamental measurement in the human sciences (4th ed). Routledge

Brants T (2000) Inter-annotator agreement for a German newspaper corpus. Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece, 31 May-2 June, European Language Resources Association
Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems, Online, 6–12 December, Curran Associates, Inc., Red Hook, NY
Burstein J (2003) The E-rater scoring engine: Automated essay scoring with natural language processing. In Shermis MD and Burstein JC (ed) Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, NJ
Čech R, Miroslav K (2018) Morphological richness of text. In Masako F, Václav C (ed) Taming the corpus: From inflection and lexis to interpretation, 63–77. Cham, Switzerland: Springer Nature
Çöltekin Ç, Taraka, R (2018) Exploiting Universal Dependencies treebanks for measuring morphosyntactic complexity. In Aleksandrs B, Christian B (ed), Proceedings of first workshop on measuring language complexity, 1–7. Torun, Poland
Crossley SA, Cobb T, McNamara DS (2013) Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. System 41:965–981. https://doi.org/10.1016/j.system.2013.08.002
Crossley SA, McNamara DS (2016) Say more and be more coherent: How text elaboration and cohesion can increase writing quality. J. Writ. Res. 7:351–370
CyberAgent Inc (2023) Open-Calm series of Japanese language models. Retrieved from: https://www.cyberagent.co.jp/news/detail/id=28817
Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota, 2–7 June, pp. 4171–4186. Association for Computational Linguistics
Diez-Ortega M, Kyle K (2023) Measuring the development of lexical richness of L2 Spanish: a longitudinal learner corpus study. Studies in Second Language Acquisition 1-31
Eckes T (2009) On common ground? How raters perceive scoring criteria in oral proficiency testing. In Brown A, Hill K (ed) Language testing and evaluation 13: Tasks and criteria in performance assessment (pp. 43–73). Peter Lang Publishing
Elliot S (2003) IntelliMetric: from here to validity. In: Shermis MD, Burstein JC (ed) Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, NJ
Engber CA (1995) The relationship of lexical proficiency to the quality of ESL compositions. J. Second Lang. Writ. 4:139–155
Garner J, Crossley SA, Kyle K (2019) N-gram measures and L2 writing proficiency. System 80:176–187. https://doi.org/10.1016/j.system.2018.12.001
Haberman SJ (2008) When can subscores have value? J. Educat. Behav. Stat., 33:204–229
Haberman SJ, Yao L, Sinharay S (2015) Prediction of true test scores from observed item scores and ancillary data. Brit. J. Math. Stat. Psychol. 68:363–385
Halliday MAK (1985) Spoken and Written Language. Deakin University Press, Melbourne, Australia
Hirao R, Arai M, Shimanaka H et al. (2020) Automated essay scoring system for nonnative Japanese learners. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 1250–1257. European Language Resources Association
Hunt KW (1966) Recent Measures in Syntactic Development. Elementary English, 43(7), 732–739. http://www.jstor.org/stable/41386067
Ishioka T (2001) About e-rater, a computer-based automatic scoring system for essays [Konpyūta ni yoru essei no jidō saiten shisutemu e – rater ni tsuite]. University Entrance Examination. Forum [Daigaku nyūshi fōramu] 24:71–76
Hochreiter S, Schmidhuber J (1997) Long short- term memory. Neural Comput. 9(8):1735–1780
Ishioka T, Kameda M (2006) Automated Japanese essay scoring system based on articles written by experts. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–18 July 2006, pp. 233-240. Association for Computational Linguistics, USA
Japan Foundation (2021) Retrieved from: https://www.jpf.gp.jp/j/project/japanese/survey/result/dl/survey2021/all.pdf
Jarvis S (2013a) Defining and measuring lexical diversity. In Jarvis S, Daller M (ed) Vocabulary knowledge: Human ratings and automated measures (Vol. 47, pp. 13–44). John Benjamins. https://doi.org/10.1075/sibil.47.03ch1
Jarvis S (2013b) Capturing the diversity in lexical diversity. Lang. Learn. 63:87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x
Jiang J, Quyang J, Liu H (2019) Interlanguage: A perspective of quantitative linguistic typology. Lang. Sci. 74:85–97
Kim M, Crossley SA, Kyle K (2018) Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. Mod. Lang. J. 102(1):120–141. https://doi.org/10.1111/modl.12447
Kojima T, Gu S, Reid M et al. (2022) Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, New Orleans, LA, 29 November-1 December, Curran Associates, Inc., Red Hook, NY
Kyle K, Crossley SA (2015) Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Q 49:757–786
Kyle K, Crossley SA, Berger CM (2018) The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behav. Res. Methods 50:1030–1046. https://doi.org/10.3758/s13428-017-0924-4
Kyle K, Crossley SA, Jarvis S (2021) Assessing the validity of lexical diversity using direct judgements. Lang. Assess. Q. 18:154–170. https://doi.org/10.1080/15434303.2020.1844205
Landauer TK, Laham D, Foltz PW (2003) Automated essay scoring and annotation of essays with the Intelligent Essay Assessor. In Shermis MD, Burstein JC (ed), Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, NJ
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 159–174
Laufer B, Nation P (1995) Vocabulary size and use: Lexical richness in L2 written production. Appl. Linguist. 16:307–322. https://doi.org/10.1093/applin/16.3.307
Lee J, Hasebe Y (2017) jWriter Learner Text Evaluator, URL: https://jreadability.net/jwriter/
Lee J, Kobayashi N, Sakai T, Sakota K (2015) A Comparison of SPOT and J-CAT Based on Test Analysis [Tesuto bunseki ni motozuku 'SPOT' to 'J-CAT' no hikaku]. Research on the Acquisition of Second Language Japanese [Dainigengo to shite no nihongo no shūtoku kenkyū] (18) 53–69
Li W, Yan J (2021) Probability distribution of dependency distance based on a Treebank of. Japanese EFL Learners' Interlanguage. J. Quant. Linguist. 28(2):172–186. https://doi.org/10.1080/09296174.2020.1754611
Linacre JM (2002) Optimizing rating scale category effectiveness. J. Appl. Meas. 3(1):85–106
Linacre JM (1994) Constructing measurement with a Many-Facet Rasch Model. In Wilson M (ed) Objective measurement: Theory into practice, Volume 2 (pp. 129–144). Norwood, NJ: Ablex
Liu H (2008) Dependency distance as a metric of language comprehension difficulty. J. Cognitive Sci. 9:159–191
Liu H, Xu C, Liang J (2017) Dependency distance: A new perspective on syntactic patterns in natural languages. Phys. Life Rev. 21. https://doi.org/10.1016/j.plrev.2017.03.002
Loukina A, Madnani N, Cahill A, et al. (2020) Using PRMSE to evaluate automated scoring systems in the presence of label noise. Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA → Online, 10 July, pp. 18–29. Association for Computational Linguistics
Lu X (2010) Automatic analysis of syntactic complexity in second language writing. Int. J. Corpus Linguist. 15:474–496
Lu X (2012) The relationship of lexical richness to the quality of ESL learners' oral narratives. Mod. Lang. J. 96:190–208
Lu X (2017) Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Lang. Test. 34:493–511
Lu X, Hu R (2022) Sense-aware lexical sophistication indices and their relationship to second language writing quality. Behav. Res. Method. 54:1444–1460. https://doi.org/10.3758/s13428-021-01675-6
Ministry of Health, Labor, and Welfare of Japan (2022) Retrieved from: https://www.mhlw.go.jp/stf/newpage_30367.html
Mizumoto A, Eguchi M (2023) Exploring the potential of using an AI language model for automated essay scoring. Res. Methods Appl. Linguist. 3:100050
Okgetheng B, Takeuchi K (2024) Estimating Japanese Essay Grading Scores with Large Language Models. Proceedings of 30th Annual Conference of the Language Processing Society in Japan, March 2024
Ortega L (2015) Second language learning explained? SLA across 10 contemporary theories. In VanPatten B, Williams J (ed) Theories in Second Language Acquisition: An Introduction
Rae JW, Borgeaud S, Cai T, et al. (2021) Scaling Language Models: Methods, Analysis & Insights from Training Gopher. ArXiv, abs/2112.11446
Read J (2000) Assessing vocabulary. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942
Rudner LM, Liang T (2002) Automated Essay Scoring Using Bayes' Theorem. J. Technol., Learning and Assessment, 1 (2)
Sakoda K, Hosoi Y (2020) Accuracy and complexity of Japanese Language usage by SLA learners in different learning environments based on the analysis of I-JAS, a learners' corpus of Japanese as L2. Math. Linguist. 32(7):403–418. https://doi.org/10.24701/mathling.32.7_403
Suzuki N (1999) Summary of survey results regarding comprehensive essay questions. Final report of "Joint Research on Comprehensive Examinations for the Aim of Evaluating Applicability to Each Specialized Field of Universities" for 1996-2000 [shōronbun sōgō mondai ni kansuru chōsa kekka no gaiyō. Heisei 8 - Heisei 12-

nendo daigaku no kaku senmon bun'ya e no tekisei no hyōka o mokuteki to suru sōgō shiken no arikata ni kansuru kyōdō kenkyū' saishū hōkoku-sho]. University Entrance Examination Section Center Research and Development Department [Daigaku nyūshi sentā kenkyū kaihatsubu], 21–32
Taghipour K, Ng HT (2016) A neural approach to automated essay scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 1–5 November, pp. 1882–1891. Association for Computational Linguistics
Takeuchi K, Ohno M, Motojin K, Taguchi M, Inada Y, Iizuka M, Abo T, Ueda H (2021) Development of essay scoring methods based on reference texts with construction of research-available Japanese essay data. In IPSJ J 62(9):1586–1604
Ure J (1971) Lexical density: A computational technique and some findings. In Coultard M (ed) Talking about Text. English Language Research, University of Birmingham, Birmingham, England
Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In Advances in Neural Information Processing Systems, Long Beach, CA, 4–7 December, pp. 5998–6008, Curran Associates, Inc., Red Hook, NY
Watanabe H, Taira Y, Inoue Y (1988) Analysis of essay evaluation data [Shōronbun hyōka dēta no kaiseki]. Bulletin of the Faculty of Education, University of Tokyo [Tōkyōdaigaku kyōiku gakubu kiyō], Vol. 28, 143–164
Yao S, Yu D, Zhao J, et al. (2023) Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36
Zenker F, Kyle K (2021) Investigating minimum text lengths for lexical diversity indices. Assess. Writ. 47:100505. https://doi.org/10.1016/j.asw.2020.100505
Zhang Y, Warstadt A, Li X, et al. (2021) When do you need billions of words of pretraining data? Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, pp. 1112-1125. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.90

Author contributions

Wenchao Li is in charge of conceptualization, validation, formal analysis, investigation, data curation, visualization and writing the draft. Haitao Liu is in charge of supervision.

Funding

This research was funded by National Foundation of Social Sciences (22BYY186) to Wenchao Li.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-024-03209-9.

Correspondence and requests for materials should be addressed to Wenchao Li.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.