Evaluating the Accuracy of ChatGPT in Assessing Writing and Speaking: A Verification Study Using ICNALE GRA

Uchida, Satoru

(Citation)
Learner Corpus Studies in Asia and the World,6:1-12

(Issue Date)
2024-03-20

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100487710

(URL)
https://hdl.handle.net/20.500.14094/0100487710

Evaluating the Accuracy of ChatGPT in Assessing Writing and Speaking —A Verification Study Using ICNALE GRA—

UCHIDA, Satoru (Kyushu University)

Abstract

Since the emergence of ChatGPT at the end of 2022, extensive applied research has been conducted on Large Language Models (LLMs). This study investigates the extent to which ChatGPT can accurately assess learners' writing and speaking, utilizing the ICNALE GRA. The ICNALE GRA comprises 140 instances each of writing and speaking data, rated by 80 evaluators. The results revealed that the highest correlation between the overall writing scores and the scores given by ChatGPT 3.5 was 0.801, while for ChatGPT 4.0, it was 0.888. In contrast, the highest correlation with the overall speaking scores was lower, at 0.341 for ChatGPT 3.5 and 0.561 for ChatGPT 4.0. However, it was shown that this correlation could be increased to 0.641 with prompt engineering.

Keywords

learner corpus, ICNALE GRA, automatic scoring, ChatGPT, prompt engineering

1. Introduction

The assessment of English writing and speaking has historically been a labor-intensive task for educators, but the emergence of generative AIs, epitomized by ChatGPT, is beginning to change the landscape. Many generative AIs, accepting inputs in a chat format and offering user-friendly interfaces, are now accessible for learners to use directly.

There has been substantial research into the application of generative AIs in education and research. For instance, studies have discussed its relationship with Data Driven Learning (Crosthwaite & Baisa, 2023), its potential as a writing assistant (Su et al., 2023), its utility in identifying language genres (Kuzman et al., 2023), and its connection with corpus linguistics (Uchida, 2024), among other topics.

There has been extensive research on the automated evaluation of English writing (see Uto, 2021; Ramesh & Sanampudi, 2022; Ding & Zou, 2024 for extensive reviews), with recent studies also focusing on the accuracy of ChatGPT. Mizumoto and Eguchi (2023) analyzed 12,100 essays from the ETS Corpus using the GPT-3 text-davinci-003 model, a

precursor to ChatGPT, and reported high accuracy results, especially when combining traditional linguistic features. Furthermore, Pfau et al. (2023) evaluated the grammatical accuracy of writing with ChatGPT, finding some discrepancies in data from beginner learners but generally high agreement with human evaluators across most levels. Additionally, Yancey et al. (2023) assessed the CEFR levels of writing with ChatGPT versions 3.5 and 4.0, demonstrating that presenting certain examples to ChatGPT could achieve accuracy close to that of human evaluators.

Given this background, the present paper examines the extent to which ChatGPT's evaluations align with a relatively new dataset, the ICNALE GRA (cf. Ishikawa, 2020, 2023), which includes evaluation data from 80 raters. The rationale for using ChatGPT includes its status as one of the most widely used generative AIs at present, with ChatGPT 3.5 being a pioneering generative Large Language Model (LLM) that is available for free. Previous studies have predominantly focused on writing, with less emphasis on speaking. This paper employs the spoken interaction data from the ICNALE GRA to verify the accuracy with which ChatGPT can evaluate speaking proficiency.

2. Data and methodology

2.1 ICNALE GRA

ICNALE GRA stands for the International Corpus Network of Asian Learners of English, Global Rating Archive, which was released in October 2023 as a learner corpus equipped with evaluation data. It includes 140 instances each of writing and speaking (conversation) data, which have been carefully sampled from the ICNALE to ensure balance. Each piece of data has been evaluated by 80 raters who underwent rigorous training. This corpus is unparalleled worldwide for the extensive number of evaluations attached to learner data. Notable features of this dataset include its inclusion of outputs from various Asian learners, native speaker data, writing data with corrections, and both speaking and writing data.

The 80 raters were required to conduct two types of assessments. Initially, they performed a holistic scoring to determine what score, out of 100, the given essay or speaking conversation would receive. Subsequently, they were presented with the following ten items and asked to analytically rate each on a scale of up to 10 points.

【Language】intelligibility, complexity, accuracy, fluency
【content】comprehensibility, logicality, sophistication, purposefulness
【attitude】willingness to communicate, involvement

To minimize discrepancies among evaluators, they were instructed to ensure that their average scores fell within 45-55 for the holistic scoring and within 4-6 for each item in the analytical scoring. The correlation between the holistic and analytical scores is approximately 0.9988, demonstrating a near-perfect alignment. The mean of these two scores is designated as the Overall Rating Score (ORS). In this study, the ORS is utilized as the reference score.

2.2 Statistical features using CVLA

To elucidate the relationship between the ICNALE GRA and traditionally used statistical metrics, the CEFR-based Vocabulary Level Analyzer (CVLA, ver. 2.0; https://cvla.langedu.jp/) is utilized. Developed by Uchida and Negishi (2018), this online application estimates the CEFR-J level of a submitted text based on four statistical metrics. These metrics are the Automated Readability Index (ARI), which represents the text's readability; VperSent, the average number of verbs per sentence; AvrDiff, the average lexical level of content words; and BperA, the ratio of level B content words to level A content words. ARI and VperSent indicate the complexity of sentence structure, while AvrDiff and BperA reflect lexical complexity. It is important to note that CVLA was designed for estimating the levels of reading texts and listening scripts and does not accommodate learner outputs (writing and speaking). Therefore, in this study, the CEFR-J levels estimated by CVLA are not considered; instead, only the numerical values are used.

2.3 ChatGPT settings

Access to ChatGPT was conducted via a Python script through the API. The models used were ChatGPT 3.5 as "gpt-3.5-turbo-1106" and ChatGPT 4.0 as "gpt-4-1106-preview," with the outputs set to be in JSON format. Additionally, *max_tokens* was set to 3,000 and *temperature* to 0.

The prompt specified for the system input was as follows, with the ICNALE essays and scripts entered as the user input.

[Holistic]
Your task is to rate a passage written [utterances spoken] by an English learner on a 100-point scale.
The output should be a JSON file with the key "rating."

[Analytical]

Your task is to rate a passage written [utterances spoken] by an English learner on a 10-point scale in terms of the following points:

Intelligibility (Language), Complexity (Language), Accuracy (Language), Fluency (Language), Comprehensibility (Content), Logicality (Content), Sophistication (Content), Purposefulness (Content), Willingness to Communicate (Attitude), Involvement (Attitude)

The output should be a JSON file with the keys "Intelligibility," "Complexity," "Accuracy," "Fluency," "Comprehensibility," "Logicality," "Sophistication," "Purposefulness," "Willingness_to_Communicate," and "Involvement."

Furthermore, during the evaluation of speaking data, the system prompt was appended with "Note that the rating should be given only to learner's utterances marked with '[S]:', not to the entire conversation." to ensure that the examiner's dialogue was excluded.

3. Results

3.1 CVLA

Table 1 presents the correlation between the CVLA scores and the Overall Rating Score (ORS) for writing in the ICNALE GRA. The CEFR scores in the table represent the final scores from CVLA, which are the averages of the estimates from the four scores. From this table, it is evident that ARI and VperSent, which indicate sentence complexity, do not show a significant correlation with the ORS. Conversely, AvrDiff and BperA, representing lexical aspects, exhibit significant correlations. While the final CEFR score is statistically significant, the effect size is small.

Table1

Correlation matrix of CVLA features and the ICNALE GRA writing ORS

	ARI	VperSent	AvrDiff	BperA	CEFR	ORS
ARI	1.000					
VperSent	**0.870**	1.000				
AvrDiff	**0.323**	0.109	1.000			
BperA	**0.331**	0.090	**0.910**	1.000		
CEFR	**0.917**	**0.837**	**0.599**	**0.598**	1.000	
ORS	0.098	0.004	**0.450**	**0.451**	**0.242**	1.000

Note: Bold items indicate statistically significant correlations ($p \leq 0.01$), calculated using scipy.stats' pearsonr in Python. The same applies to subsequent tables.

Table 2 displays the correlation between the CVLA scores and the ORS for speaking in the ICNALE GRA. In this experiment, only the learner's utterances were used for the speaking data, excluding the examiner's speech. Although the results indicate statistically significant correlations across all metrics for this data, the correlation coefficients are below 0.5, indicating a moderate effect size.

Table 2
Correlation matrix of CVLA features and the ICNALE GRA speaking ORS

	ARI	VperSent	AvrDiff	BperA	CEFR	ORS
ARI	1.000					
VperSent	**0.925**	1.000				
AvrDiff	**0.311**	0.197	1.000			
BperA	**0.367**	**0.290**	**0.830**	1.000		
CEFR	**0.943**	**0.932**	**0.509**	**0.584**	1.000	
ORS	**0.428**	**0.378**	**0.395**	**0.392**	**0.476**	1.000

3.2 Writing

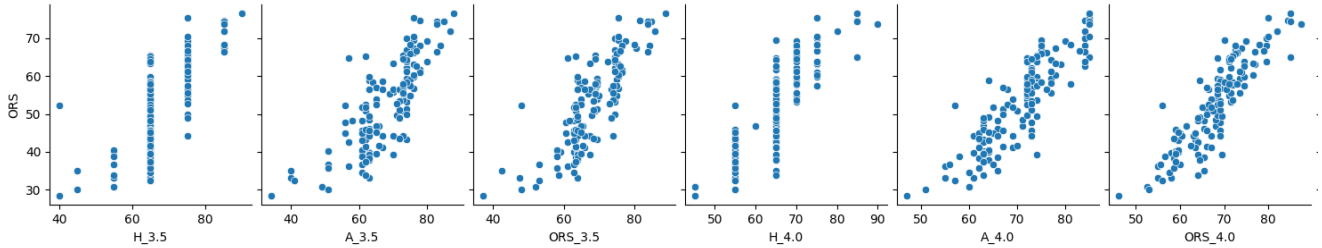
Next, the correlation between ChatGPT evaluation scores and the ICNALE GRA writing scores will be examined. Evaluations were conducted using two versions of ChatGPT, 3.5 and 4.0, and two types of scores were calculated: holistic scoring (denoted as H_) and analytical scoring (denoted as A_). In addition, the correlation with the average of these scores (denoted as ORS_) was also examined. Table 3 shows the correlation table and Figure 1 shows the scatter plot of ChatGPT scores against the ICNALE GRA ORS.

Table 3
Correlation matrix of ChatGPT scores and ICNALE GRA writing ORS

	H_3.5	A_3.5	ORS_3.5	H_4.0	A_4.0	ORS_4.0	ORS
H_3.5	1.000						
A_3.5	**0.789**	1.000					
ORS_3.5	**0.939**	**0.952**	1.000				
H_4.0	**0.717**	**0.731**	**0.766**	1.000			
A_4.0	**0.732**	**0.790**	**0.806**	**0.836**	1.000		
ORS_4.0	**0.756**	**0.794**	**0.820**	**0.958**	**0.958**	1.000	
ORS	**0.722**	**0.790**	**0.801**	**0.828**	**0.873**	**0.888**	1.000

Figure 1

Scatterplot of ChatGPT scores and the ICNALE GRA writing ORS



All ChatGPT scores showed a high correlation with the ICNALE GRA, the highest being 0.888 for ChatGPT 4.0's ORS (average of holistic and analytical scoring). One point of note, as seen in the scatter plot, is the scoring intervals in the holistic scoring, which were calculated in increments of 5 points, such as 55, 60, and 65. As a result, there's a possibility that the correlations are unnaturally inflated. While this could potentially be changed by the prompt, it has become clear that simply instructing the system to score out of 100 is not sufficient.

3.3 Speaking

The next step is the evaluation of the speech data. As mentioned above, the learner's utterances are marked with "[S]:", and the entire conversation is entered for evaluation to include the context. Table 4 shows the correlation between ChatGPT scores and the ICNALE GRA speaking ORS, and Figure 2 shows the scatter plot.

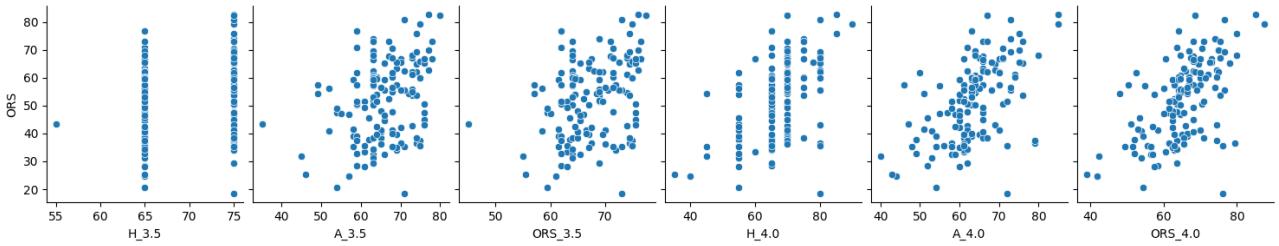
Table 4

Correlation matrix of ChatGPT scores and the ICNALE GRA speaking ORS

	H_3.5	A_3.5	ORS_3.5	H_4.0	A_4.0	ORS_4.0	ORS
H_3.5	1.000						
A_3.5	**0.611**	1.000					
ORS_3.5	**0.853**	**0.935**	1.000				
H_4.0	**0.494**	**0.620**	**0.631**	1.000			
A_4.0	**0.532**	**0.682**	**0.689**	**0.859**	1.000		
ORS_4.0	**0.531**	**0.673**	**0.683**	**0.969**	**0.959**	1.000	
ORS	**0.243**	**0.341**	**0.334**	**0.536**	**0.547**	**0.561**	1.000

Figure 2

Scatterplot of ChatGPT scores and the ICNALE GRA speaking ORS



Compared to writing, the correlations are generally lower for speaking. Moreover, examining the scatter plot reveals a particularly severe bias in the holistic scores given by ChatGPT 3.5. The highest value recorded was 0.561 for the ORS of ChatGPT 4.0.

4. Discussion

4.1 Usefulness of AIs for assessing learners' output

Observing the results from CVLA, the significance of vocabulary complexity (AvrDiff and BperA) in both writing and speaking suggests that this aspect plays an important role in evaluating learner data. While the metrics related to sentence complexity (ARI and VperSent) showed significance in speaking, this may depend on the method of transcription. Given the difficulty in defining sentences within the speaking register, these metrics may not truly be reliable indicators.

Regarding ChatGPT's outcomes, higher versions yielded higher scores, and the ORS, which averages the holistic and analytical scores, was found to be the highest. This suggests that further improvements in AI versions could lead to even more accurate assessments.

Comparing writing and speaking, writing showed higher correlations. This could be attributed to the predominance of written data over spoken data in the training material for LLMs. As training with spoken data progresses, it is expected that the accuracy of speaking scores will also improve.

4.2 Prompt engineering

In writing, the highest correlation coefficient reached 0.888, which can be considered sufficiently accurate for practical use. On the other hand, speaking remained at 0.561, leaving room for improvement. Therefore, this section experiments with how accuracy changes through simple prompt engineering.

First, an attempt was made to give ChatGPT a clear role by instructing it with "You

are an experienced English teacher." This is expected to provide the system with a deeper understanding of what task it is supposed to perform and to clarify the context of the task. This experiment was conducted using ChatGPT 4.0, which had shown the highest accuracy, with prompts for analytical scoring.

Next, in addition to clarifying the role, detailed instructions for scoring criteria were provided during analytical scoring. For example, for Intelligibility, the following explanation was given based on Ishikawa (2020):

Intelligibility: To what extent can you "decode," namely, verbally understand what is said/written? In speech evaluation, factors such as pronunciation and intonation will influence the degree of intelligibility. In essay evaluation, factors such as spellings and sentence structures may influence it. Please note that intelligibility, which concerns the understandability of the language, should be discriminated from comprehensibility, which concerns the understandability of the content. You may sometimes find a speech/essay that is intelligible but not comprehensible such as a logical but nonsensical statement. Meanwhile, you do not usually find a speech/essay that is comprehensible but unintelligible because if the text cannot be decoded, its content cannot be conveyed.

Similar explanations were provided for the other nine evaluation criteria, followed by instructions to evaluate speaking. Table 5 presents the results of this approach. When the role of the teacher was specified, the correlation increased to 0.586, marking an improvement of 0.039. Furthermore, when combining the teacher role with detailed explanations, the correlation coefficient increased to 0.641, showing an increase of 0.094.

Table 5
Results of prompt engineering with speaking data

	A_4.0	A_4.0_teacher	A_4.0_teacher+explanation
ORS	**0.547**	**0.586**	**0.641**

Although the increase is modest, it can be said that the possibility of improving results through prompt engineering has been demonstrated. Additionally, referencing the profiles of ICNALE GRA evaluators to create various personas could also be viable, and comparing the outcomes of these different approaches would likely be interesting.

5. Conclusion

This study verified the accuracy of ChatGPT scores using ICNALE GRA data. The results showed a maximum correlation coefficient of 0.888 for writing and up to 0.641 for speaking after prompt engineering. These results suggest that generative AI scores have a certain level of reliability. In particular, it was found that higher correlations are achieved in written language, with larger LLM parameters (newer models), and when roles and evaluation criteria are clearly defined.

By its very nature, generative AIs involve a degree of randomness and are constantly updated. This suggests that the results presented in this study may not always be replicable, highlighting a limitation of research using LLMs such as ChatGPT. Furthermore, this study did not employ few-shot learning or fine-tuning with evaluation data, which could potentially alter accuracy. Addressing this issue is a task for future research.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP22H00677, JP 22K12326, and JP 21H03564.

Use of AIs

This study examined the capability of ChatGPT in assessing learners' writing and speaking and thus extensively used it for the experiments. In addition, during the preparation of this work, the author used ChatGPT 4.0 in order to improve, proofread, and translate the writing. After using this tool, the author reviewed and edited the content as required and took full responsibility for the content of the publication.

Bibliography

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics, 3*(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12402-3

Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of Asian learners' L2 English essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World, 5*, 121-144.

Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. https://arxiv.org/abs/2303.03953

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics, 2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

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

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics, 4*(1), 100089. https://doi.org/10.1016/j.acorp.2024.100089

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In *Proceedings of Asia Pacific Corpus Linguistics Conference, 4*, 463-467.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika, 48*, 459-484. https://doi.org/10.1007/s41237-021-00142-y

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 576-584. https://doi.org/10.18653/v1/2023.bea-1.49