OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 14, Number 1 (2025)
https://doi.org/10.29140/vli.v14n1.2078

Castledown

Generative AI and CEFR Levels: Evaluating the Accuracy of Text Generation with ChatGPT-4o Through Textual Features

Satoru Uchida
Kyushu University

Abstract

Since its emergence, generative AI has significantly impacted various fields, including English language education. Numerous academic studies have investigated its capabilities in grammar correction, writing evaluation, and dynamics of user interaction. However, there have been insufficient investigations into whether texts generated by such AI align appropriately with CEFR proficiency levels. This study addresses this gap by exploring the applicability of generative AI to CEFR standards. Multiple texts were generated using ChatGPT-4o with specified CEFR levels and analyzed using a vocabulary level analyzer (CVLA) to evaluate text features. The findings revealed discrepancies between AI-generated texts and textbook standards, significant divergences between levels below B1 and above B2, and a noticeable topic bias. Although AI-generated texts seem to differ by level, they require careful evaluation before being applied to CEFR-based education.

Background

English education is undergoing a significant transformation, largely driven by the advent of generative AI tools such as ChatGPT. These AI systems can produce fluent English, excel in translation, and serve as valuable alliances for English teachers and

learners. Numerous studies have demonstrated its utility by examining the accuracy of writing assessments (Mizumoto & Eguchi, 2023; Uchida, 2024a; Yamashita, 2024), grammar correction (Mizumoto et al., 2024; Schmidt-Fajlik, 2023), and learning methodologies (Crosthwaite & Baisa, 2023; Mizumoto, 2023). In addition, detailed review articles on AI use in educational contexts, such as those by Lo (2023) and Law (2024), have emerged.

Although generative AI has shown promise in various educational applications, research on its relationship to the Common European Framework of Reference for Languages (CEFR)—a global standard of language education—remains limited. One valuable attempt was by Yancey et al. (2023), who evaluated the effectiveness of the GPT-4 in rating short essays written by English learners on the CEFR scale. GPT-4's performance was compared to GPT-3.5 and existing automated writing evaluation (AWE) systems. When calibration examples were provided, GPT-4 nearly matched the accuracy of modern AWE methods, although its agreement with human ratings varied depending on the test-taker's native language (L1). Ramadhani et al. (2023) employed a systemic functional linguistic approach, examining lexical density, grammatical intricacy, and lexical variation to analyze the readability and complexity of ChatGPT-generated texts. Their analysis, focusing on whether these texts aligned with CEFR standards, revealed that text complexity does not consistently match CEFR levels. They found that text length correlates with perceived complexity, suggesting that ChatGPT struggles to control linguistic features appropriately across different CEFR levels, leading to a potential misalignment with student proficiency.

This study expands upon the research conducted by Ramadhani et al. (2023) by utilizing the latest AI model, ChatGPT-4o, and analyzing it through a CEFR-based Vocabulary Level Analyzer (CVLA), providing a fresh perspective on the alignment of AI-generated texts with CEFR levels. This approach is expected to objectively reveal the extent to which the current AI aligns with CEFR levels. The research questions addressed in this study are as follows:

RQ1: To what extent can ChatGPT-4o generate texts that align with CEFR levels?
RQ2: Do the topics of texts generated by ChatGPT-4o vary according to CEFR levels?
RQ3: Does providing examples improve the results of text generation by ChatGPT-4o?

Methods

Text Generation Using ChatGPT

This study utilizes ChatGPT, one of the most prominent generative AI models. Released in late November 2022, ChatGPT-3.5 marked the beginning of generative AI, leaving a significant impact on many users. The model is built upon Transformer technology, a deep learning architecture known for its self-attention mechanism, which allows it to process and generate sequences of text more efficiently than previous models. It learns from large-scale datasets and is further refined through reinforcement learning from

human feedback (RLHF), enabling it to perform a variety of tasks with high accuracy. This evolution continued with the release of ChatGPT-4.0, in March 2023. As of August 2024, the latest version was ChatGPT-4o (ver. 2024-08-06), which supports multimodal inputs. This study employs this model via API. To ensure randomness, the temperature was set to 0.8, with higher values increasing the randomness.

Text generation was conducted using zero-shot learning to assess the default behavior of ChatGPT. The prompts were as follows:

> You are a proficient English writer. Generate a passage suitable for CEFR level {level} reading. When creating the passage, pay attention to the following points.
>
> (1) Ensure that the vocabulary and grammar used in the sentences are appropriate for {level}.
> (2) Keep the length of the passage suitable for {level}.
> (3) Choose a topic that is relevant and appropriate for {level}.
> (4) Ensure that the passage is designed for English language learners.
> (5) Return the output in the following JSON format:
> {"topic": "xxx", "title": "yyy", "content": ["sentence1", "sentence2", ...]}

Using this prompt, 30 passages for each level, from A1 to C2 (180 passages), were generated and saved as text files.

Next, a one-shot learning approach was employed, in which a sample for each level was provided before the texts were generated. By supplying an example, it is anticipated that the output would adhere more accurately to the specified CEFR levels owing to the increased reference information. The prompts were as follows:

> You are a proficient English writer. Generate a passage suitable for CEFR level {level} reading, using the provided sample as a reference. When creating the passage, pay attention to the following points:
>
> (1) Ensure that the vocabulary and grammar used in the sentences are appropriate for {level}.
> (2) Keep the length of the passage suitable for {level}.
> (3) Choose a topic that is relevant and appropriate for {level}.
> (4) Ensure that the passage is designed for English language learners.
> (5) Return the output in the following JSON format:
> {"topic": "xxx", "title": "yyy", "content": ["sentence1", "sentence2", ...]}
> <sample>
> {sample_passage}
> </sample>

The sample passages provided by the Council of Europe (COE) were used as illustrative reading tasks (https://www.coe.int/en/web/common-european-framework-reference-languages/reading-comprehension). These included materials from sources

such as Aptis (an English proficiency test conducted by the British Council) and the Cambridge English Language Assessment. However, because only one sample was available for the C levels, they were excluded from this study. Thirty texts were generated for each of the A1, A2, B1, and B2 levels, and the results were analyzed accordingly.

Vocabulary Level Analysis Using Software

For the vocabulary level analysis, Version 2.0 of the CVLA (https://cvla.langedu.jp/), an online tool developed by Uchida and Negishi (2018) to estimate the CEFR-J levels of reading and listening texts, was utilized. This tool sets four statistical indicators for each level based on textbooks compiled according to CEFR levels and estimates the level using a regression model. The four indicators are: Automated Readability Index (ARI), which measures text complexity; VperSent, which represents the average number of verbs per sentence and serves as an indicator of grammatical complexity within sentences; AvrDiff, which calculates the average difficulty of words based on the CEFR-J Wordlist (assigning a value of 1 for A1 content words, 2 for A2, 3 for B1, and 4 for B2); and BperA, which indicates the proportion of B-level content words relative to A-level words. Table 1 shows the average scores of these indicators, which were used to estimate the CEFR-J level. A notable characteristic of these indicators is that they increase proportionally with the level.

Table 1 *Average Values of Each Indicator by CEFR Level*

CEFR	ARI	VperSent	AvrDiff	BperA
A1	5.73	1.49	1.31	0.08
A2	7.03	1.82	1.41	0.12
B1	10.00	2.37	1.57	0.18
B2	12.33	2.88	1.71	0.26

Uchida and Negishi (2021) validated the accuracy of the CVLA using sample texts from the COE, where each input consisted of multiple passages combined into a single text to represent each level. Table 2 presents the results of the study. As shown in the table, the estimated levels generally corresponded well with the actual levels, except for the B2 level, which was estimated as C1. In the case of B2, the final estimated value (4.68) fell between adjacent levels (B1 and C1), indicating that the tool effectively captures the hierarchical structure of the levels.

Topic Analysis Method

For the topic analysis, the AI was instructed to output both the topic and title during text generation. These topic tags were used as classification labels. However, owing to inconsistencies in the labels (e.g., pets, my-pet-cat; travel, travel-and-accommodation; cultural-events, cultural-festivals), a manual review was conducted to standardize them into simpler categories (e.g., pets, travel, and culture). Topics that did not display any notable patterns were categorized under "others."

Table 2 *Accuracy of CVLA (adapted from Uchida & Negishi, 2021)*

CEFR	ARI	VperSent	AvrDiff	BperA	Final
COE_A1	1.83	2.08	1.16	0.05	0.65
	PreA1	A2.2	PreA1	A1.1	A1.1
COE_A2	6.88	2.15	1.52	0.12	2.18
	A2.1	B1.1	B1.1	A2.1	A2.2
COE_B1	7.32	2.79	1.58	0.19	2.95
	A2.1	B2.1	B1.2	B1.1	B1.1
COE_B2	10.54	3.67	1.80	0.30	4.68
	B1.2	C2	C1	C1	C1
COE_C1	11.19	4.72	1.90	0.39	5.24
	B2.1	C2	C1	C2	C1

Results

This section presents the results of the level estimation using CVLA and topic analysis of the texts generated by ChatGPT-4o. In this generation process, word count was not specified for either zero-shot or one-shot scenarios because it is challenging to accurately count words owing to the unique tokenization method used in large language models (LLMs). The prompt specified, "Keep the length of the passage suitable for {level}," so the AI was expected to output passages of an appropriate length based on its estimation. Table 3 presents the average word count and standard deviation for each level.

As shown in the table, the word count increased with the level in both the zero-shot and one-shot learning scenarios. However, because longer texts are typically expected at higher levels, there may be room for discussion as to whether the generated length is appropriate for those levels.

Table 3 *Average Word Count (SD) for Each CEFR Level*

CEFR Level	Zero Shot	One Shot
A1	63.97(10.90)	78.67(19.38)
A2	92.20(15.15)	110.90(18.66)
B1	148.90(23.68)	185.57(85.36)
B2	177.63(22.93)	239.43(37.50)
C1	224.30(27.65)	NA
C2	240.90(25.66)	NA

Table 4 *Specified Levels (column) and CVLA Analysis Results (row) for Texts Generated with Zero-Shot Learning*

	A1	A2	B1	B2	C1	C2	Total
PreA1	30	29					59
A1.1			3				3
A1.2			2				2
A1.3		1	3				4
A2.1			5				5
A2.2			8				8
B1.1			**5**				5
B1.2			**4**				4
C1				5			5
C2				25	30	30	85
Total	30	30	30	30	30	30	180

CEFR Level Estimation

As mentioned previously, the CVLA is designed to estimate CEFR-J levels with a certain degree of accuracy. In this analysis, sublevels such as A1.1, A1.2, and A1.3 were grouped as A1. The CEFR levels specified to ChatGPT were compared with the analysis results obtained from CVLA for the texts generated by ChatGPT. Table 4 presents the results for zero-shot learning.

The results show that only 5% (9 out of 180 texts) matched the specified CEFR levels according to the CVLA analysis, indicating very low accuracy. Specifically, many texts intended for the A1 and A2 levels were classified as pre-A1. By contrast, texts intended for the B2, C1, and C2 levels were often classified as C2. These findings suggest a tendency for ChatGPT to generate texts that are either excessively simple at lower levels or overly complex at higher levels. Additionally, for the B1 level, there was noticeable variability in the level of the generated texts.

Table 5 compares the texts generated by ChatGPT after being provided with sample passages at the specified levels with the CVLA analysis results. The data show that when using one-shot learning, where examples are provided, the accuracy slightly improves to 12.5% (15 out of 120 texts). However, the overall trend remains consistent, with similar discrepancies in level classification.

To conduct a more detailed analysis, Tables 6 and 7 present the average values and standard deviations (in parentheses) for the four indicators calculated by the CVLA across 30 texts for each level. In both the zero-shot and one-shot learning scenarios, it is evident that the values deviate from the benchmarks listed in Table 1. Focusing on ARI, the largest difference is observed between B1 and B2, indicating that text readability significantly increases at the B2 level. Regarding VperSent, the C1 and

Table 5 *Specified Levels (column) and CVLA Analysis Results (row) for Texts Generated with One-Shot Learning*

	A1	A2	B1	B2	Total
PreA1	29	14			43
A1.1	**1**	12			13
A1.2		3			3
A2.1		**1**	3		4
A2.2			5		5
B1.1			**7**		7
B1.2			**6**		6
B2.1			7		7
B2.2			1		1
C1			1	4	5
C2				26	26
Total	30	30	30	30	120

C2 levels exceed 3 in zero-shot learning, whereas the B2 level exceeds 3 in one-shot learning, suggesting that grammatical complexity is substantially higher than that of the corresponding CEFR texts. When examining AvrDiff and BperA, which represent vocabulary levels, there is minimal difference between A1 and A2 in both scenarios, indicating that ChatGPT finds it challenging to differentiate between these two levels. Additionally, there is a sharp increase in values at the B2 level in both cases, suggesting the use of a more advanced vocabulary. Furthermore, compared with Table 1, it is evident that the levels generated by ChatGPT deviate significantly from the textbook measurements. For instance, for BperA, the values for A1 are 0.08 in textbooks and 0.04 in ChatGPT outputs; for A2, 0.12 in textbooks and 0.04 in ChatGPT; for B1, 0.18 in textbooks and 0.15 in ChatGPT; and for B2, 0.26 in textbooks and 0.76 in ChatGPT. A similar trend is observed across other indicators, where ChatGPT tends to produce extremely low values for lower levels and extremely high values for upper levels.

Topic Frequency

Next, the topics of the texts generated by ChatGPT were examined. The topics were manually reviewed based on the labels initially provided by ChatGPT. Although brief, this information provides insight into the content of the texts. Table 8 presents the results for zero-shot learning, and Table 9 presents the results for one-shot learning.

The one-shot learning approach displays greater topic variation; however, certain biases related to the level were evident in both scenarios. At the A1 and A2 levels, topics often revolve around themes related to daily life, such as family, beach,

Table 6 *Average Values (SD) of Text Levels and Indicators in Zero-Shot Learning*

CEFR	ARI	VperSent	AvrDiff	BperA	CEFR score
A1	-2.81(1.03)	1.16(0.08)	1.09(0.04)	0.04(0.02)	-0.51(0.19)
A2	0.62(1.45)	1.41(0.16)	1.15(0.07)	0.04(0.02)	0.10(0.33)
B1	4.94(1.32)	2.16(0.44)	1.51(0.15)	0.15(0.08)	2.08(0.80)
B2	12.21(1.35)	2.79(0.3)	2.23(0.17)	0.76(0.23)	6.97(1.30)
C1	17.93(1.70)	3.56(0.53)	2.46(0.11)	1.09(0.22)	9.78(1.19)
C2	20.17(1.74)	3.89(0.61)	2.62(0.12)	1.33(0.37)	11.46(1.76)

Table 7 *Average Values (SD) of Text Levels and Indicators in One-Shot Learning*

CEFR	ARI	VperSent	AvrDiff	BperA	CEFR score
A1	–1.39(0.89)	1.2(0.13)	1.12(0.09)	0.04(0.04)	–0.30(0.34)
A2	1.68(0.76)	1.62(0.24)	1.23(0.08)	0.05(0.03)	0.52(0.37)
B1	7.32(1.54)	2.56(0.48)	1.62(0.17)	0.22(0.10)	3.00(0.82)
B2	14.14(1.57)	3.38(0.44)	2.17(0.21)	0.65(0.24)	6.96(1.40)

Table 8 *Text Levels and Topics in Zero-Shot Learning*

Topic	A1	A2	B1	B2	C1	C2	Total
AI					1	19	20
Climate_change				2	7	4	13
Culture			1	1			2
Daily_life	28	28					56
Environment				13	6		19
Family	2						2
Health			12				12
Hobbies		2					2
Others				4		3	7
SNS					3		3
Sustainability				6	9	1	16
Technology				4	4	3	11
Travel			17				17
Total	30	30	30	30	30	30	180

hobbies, and parks. At the B1 level, there is a noticeable focus on health and travel in zero-shot learning and on restaurants and travel in one-shot learning. For B2, environment and sustainability are commonly recurring themes in both cases, with culture also frequently appearing in one-shot learning. The C levels, which were only examined in zero-shot learning, show that sustainability and climate change are common topics at the C1 level, while AI-related themes are prevalent at the C2 level.

Table 9 *Text Levels and Topics in One-Shot Learning*

Topic	A1	A2	B1	B2	total
Animals/pets	5	1	1		7
Beach		16			16
Culture				7	7
Daily_life	12				12
Environment				4	4
Family	2				2
Food	1				1
Hobbies	1		1		2
Holiday		1	3		4
Leisure			2		2
Others			3	5	8
Park	5	6			11
Restaurants			15		15
School	2				2
Shopping		1			1
Space_exploration				2	2
Sustainability				6	6
Technology				3	3
Travel			4	1	5
Volunteering				2	2
Weekends	1	2	1		4
Zoo	1	3			4
Total	30	30	30	30	120

Discussion

The results of the experiments reveal that, while the texts generated by ChatGPT may appear to vary by level on the surface, they do not adequately reflect CEFR levels upon closer examination. This finding aligns with the conclusion of Ramadhani et al. (2023) that the readability and complexity of texts issued by ChatGPT do not consistently adhere to the CEFR standards. Specifically, the A1 and A2 levels tend to be overly simplified, B1 exhibits some variability, and the B2, C1, and C2 levels are excessively complex. Yancey et al. (2023) reported that providing examples improves the accuracy of writing assessments. However, in the present study, while one-shot learning slightly increased topic variety, it did not significantly reduce bias or improve alignment with CEFR levels.

These findings indicate that, while specifying a CEFR level in ChatGPT can produce texts with appropriate vocabulary and grammar, the outputs often fail to align with CEFR standards and exhibit topic bias. To further investigate this, trigram counts (in lemma form) were performed on the zero-shot learning results, which reflected ChatGPT's default behavior more accurately than one-shot learning. This analysis was performed using Python's spaCy library (ver. 3.7.5) using the en_core_web_sm model. The results are presented below.

A1
i go to (53), with my family (37), my family i (37), go to school (28), i wake up (27), brush my tooth (27), wake up at (25), with my friend (25), up at 7 (24), i eat breakfast (24)

A2
with my family (32), at 7 o'clock (29), wake up at (26), up at 7 (26), i go to (26), brush my tooth (24), i wake up (23), wash my face (23), breakfast with my (21), after breakfast i (21)

B1
fruit and vegetable (19), to the mountain (13), it be a (12), go on a (10), a trip to (10), with my family (10), in the evening (10), the evening we (10), last summer i (9), i go on (9)

B2
for future generation (14), one of the (13), can lead to (13), lead to a (9), in recent year (9), climate change be (8), it be essential (8), play a crucial (8), a crucial role (8), of the most (7)

C1
one of the (13), in recent year (12), climate change be (10), of climate change (10), renewable energy source (9), recent year the (8), it be crucial (8), of the most (7), can lead to (7), be essential for (7)

C2
of artificial intelligence (14), the potential for (11), artificial intelligence ai (11), the advent of (10), one of the (9), intelligence ai have (9), it be imperative (8), decision make process (7), be imperative to (7), ensure that the (7)

This list reveals that, even with a certain level of randomness maintained in ChatGPT (temperature = 0.8), there are repeated expressions and topics across the generated texts. Levels A1 and A2 were particularly similar, frequently featuring scenarios involving specific times of day, going to school, and eating breakfast. At the B1 level, while there are still overlaps with the A-level trigrams, such as references to family and time of day, the topics begin to shift at B2, with a focus on future generations and environmental issues. C1 continues to emphasize environmental concerns, whereas C2 frequently discusses AI. Considering that each level comprises 30 texts, these trigrams with frequencies of 10 or more indicate that similar patterns recur across multiple texts, demonstrating a repetition of themes and expressions.

The topic bias observed in this study can be partially justified by considering CEFR can-do statements (https://www.coe.int/en/web/common-european-framework-reference-languages/table-2-cefr-3.3-common-reference-levels-self-assessment-grid). For example, in the self-assessment grid for reading, the A2 level mentions simple everyday materials, such as advertisements, prospectuses, menus, and timetables, as well as short, simple personal letters emphasizing actions related to daily life. In contrast, the B1 level refers to job-related language and the B2 level includes mentions of contemporary problems. To some extent, these references align with the topic bias observed in ChatGPT outputs. Additionally, the expression bias revealed through trigram analysis may stem from underlying topic bias. While such biases could be considered natural to some extent, users of generative AI should be fully aware of these tendencies, particularly when specifying CEFR levels, to ensure that the generated content aligns with the intended goals.

Conclusion

The conclusions of this study are summarized in response to the research questions as follows:

RQ1: To what extent can ChatGPT-4o generate texts that align with CEFR levels?
A: Analysis using CVLA indicates that it is difficult to determine whether ChatGPT-4o can generate texts that consistently align with the specified CEFR levels. There is a tendency for lower levels to become overly simple, and more advanced levels to become overly difficult.

RQ2: Do the topics of texts generated by ChatGPT-4o vary according to CEFR levels?
A: There is a noticeable bias in topics, with A-level texts focusing on daily life, B-level texts on travel and environmental issues, and C-level texts on sustainability and AI. This bias can be considered natural based on CEFR can-do statements. However, users should be aware of this tendency when working with AI-generated text.

RQ3: Does providing examples improve the results of text generation by ChatGPT-4o?
A: Providing examples slightly improves alignment with CEFR levels and slightly increases the variety of generated topics but does not lead to dramatic improvements.

The findings emphasize the importance of checking readability, word difficulty, and topic relevance and being cautious when using ChatGPT to generate texts for CEFR levels. Although this study used the latest version of ChatGPT4o available at the time of writing, the model is expected to continue being updated, and the reproducibility of the results presented here cannot be guaranteed. As noted by Uchida (2024b), this is one of the limitations of research on generative AI. Additionally, because this study tested only one generative AI, the findings cannot be generalized to all generative AI models. Future directions may include more technically advanced approaches, such as fine-tuning generative AI, which could be a promising area of exploration. It is also necessary to investigate how human experts assess the CEFR levels of texts generated by ChatGPT. This could provide valuable insights into the alignment between AI-generated outputs and human judgment.

Declaration of Use of AI

This study explored the use of generative AI in the generation of texts at each CEFR level with the extensive use of ChatGPT in the experiments. In addition, ChatGPT 4o was used to improve, proofread, and translate the writing during the preparation of this study. The author carefully reviewed and edited the content and took full responsibility for the final publication.

Acknowledgment

This study was supported by the JSPS KAKENHI (grant number JP22H00677).

References

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

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6*(100174), https://doi.org/10.1016/j.caeo.2024.100174

Lo, C. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics, 3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

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

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Ramadhani, R., Aulawi, H., & Ulfa, R. L. (2023). Readability of reading texts as authentic materials issued by ChatGPT: A systemic functional perspective. *Indonesian Journal of English Language Teaching and Applied Linguistics, 8*(2), 149–168.

Schmidt-Fajlik, R. (2023). ChatGPT as a grammar checker for Japanese English Language Learners: A comparison with Grammarly and ProWritingAid. *AsiaCALL Online Journal, 14*(1), 105–119. https://doi.org/10.54855/acoj.231417

Uchida, S. (2024a). Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification study using ICNALE GRA. *Learner Corpus Studies in Asia and the World, 6*, 1–12.

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

Uchida, S. (2024b). 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. (2021). Estimating the CEFR levels of English reading materials: Evaluation of CVLA. *Journal of Corpus-Based Lexicology Studies, 3*, 1–14. https://doi.org/10.24546/81012537

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics, 3*(3), 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. *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