ISSN 1392-0340 (Print)
ISSN 2029-0551 (Online)
https://doi.org/10.15823/p.2025.160.2
Pedagogika / Pedagogy
2025, t. 160, Nr. 4, p. 23–39 / Vol. 160 No. 4, pp. 23–39, 2025

VYTAUTO DIDŽIOJO
UNIVERSITETO
ŠVIETIMO
AKADEMIJA

AI-Assisted vs. Teacher-Based Assessment of EFL Students' Writing: A Comparative Study

İbrahim Tekman[1], Tülay Kaya Tekman[2]

[1] Cyprus International University, School of Foreign Languages, English Preparatory Program, Haspolat Campus, 99258, Nicosia, Northern Cyprus, itekman@ciu.edu.tr
[2] Near East University, Faculty of Education, Turkish Language Teaching Department, Near East Boulevard, 99138, Nicosia, Northern Cyprus, tulay.kaya@neu.edu.tr

Annotation. Artificial intelligence (AI) plays a growing role in foreign language learning, especially in assessing learners' writing. This study examined 11 EFL compositions, evaluated by both an AI program and three teachers. Findings show AI offers speed, consistency, and objectivity, but struggles with content and organization. Thus, AI should be used as a supportive tool, complementing teachers rather than replacing them in EFL writing assessment.

Keywords: *artificial intelligence, EFL, writing, assessment.*

Introduction

Artificial intelligence (AI) plays a significant role in many areas of our lives. Parallel to the advancements in technology, AI has affected every aspect of life, including education. The use of AI has created various changes in education and foreign language learning and teaching is one of the fields that has experienced innovations and advancements through the integration of AI in recent years.

Many researchers described writing as a complex and critical skill in the foreign language learning and teaching process (Anwar & Ahmed, 2016; Baharudin et al.,

2023; Pandya & Saiyad, 2025). Writing not only includes and requires grammar and vocabulary knowledge, but it also requires the ability to organise and support ideas and communicate successfully.

Students' writing skills are traditionally assessed by teachers, which requires significant time and effort. The integration of AI in the assessment procedure of writing products has attracted the interest of many researchers, as it might help teachers to improve the assessment process (Ke & Ng, 2019; Wang, 2022; Bevilacqua et al., 2023; Almegren et al., 2024; Bucol & Sangkawong, 2024; Li et al., 2024; Mirzaeian, 2025).

The use of AI in the assessment of writing compositions might reduce teachers' workload and effort and provide instant feedback. Thus, it is believed to be important to analyze and evaluate the potential and effectiveness of AI in the assessment of writing skill. The present study aims to compare AI-assisted scoring with teacher scoring in evaluating writing tasks of students learning English as a foreign language and to identify the areas where similarities and differences occur. Specifically, the study seeks to compare AI-assisted and teacher-based assessments and to identify the strengths and weaknesses of AI-assisted assessment in relation to teacher evaluation. Understanding how teacher assessment and AI assessment differ can help educators make more informed decisions about integrating technology into the evaluation of writing skills. In this context, this study aimed to find answers to the following questions:

In composition assessments:

1. Are there significant differences between AI-assisted assessment and teacher assessment?
2. What are the strengths and weaknesses of AI-assisted assessment compared to teacher assessment?

Literature Review

In recent years, research on the integration of AI in foreign language learning, teaching and assessment has gained growing interest. EFL writing assessment is one of the areas that attracted the most attention. In order to be sufficient and effective in writing, EFL learners need continuing practice and reliable feedback and evaluation.

Researchers highlight that AI tools provide instant and personal feedback on grammar and vocabulary use, spelling, and syntactic complexity (Ranalli, 2018). These AI tools guide learners to identify their errors and revise their drafts in an effective way, and they also reduce teachers' workload as they provide automated scoring and enable the assessment of writing compositions consistently (Zawacki-Richter et al., 2019; Holmes et al., 2021). Studies suggest that integrating AI into EFL writing assessment has pedagogical benefits and effectiveness.

The potential effectiveness of AI has made it a significant and debated area in language testing. Therefore, conducting this study and analysing the strengths and weaknesses of AI integration in EFL writing assessment is critical for broader discussion and identifying the gaps on this topic. In order to better understand the benefits and challenges of AI in practice, it is necessary to examine the existing studies that evaluated the AI applications in the assessment of EFL writing. When the literature is reviewed, it can be found that various studies have investigated the use and integration of AI in EFL writing assessment and revealed the advantages and limitations.

In a study by Song & Song (2023), the effectiveness of ChatGPT, an AI tool, in enhancing academic writing skills and motivation of EFL students were investigated. They conducted a study with control and experimental groups, and they found that the students who received AI-assisted instruction showed a significant improvement in their various aspects of writing and motivation. Marzuki et al. (2023) conducted a study to examine the effect of AI writing tools on the content and organization in students' writing, as perceived by EFL teachers. They stated that using AI writing tools was effective and beneficial in evaluating the quality of EFL students' writing. In another study, Tsai et al. (2024) examined the practicality and challenges of an AI tool, and they found that the AI tool was effective at providing instant feedback, but they also stated the potential for unfairness in writing assessment.

Xiao (2024) also carried out a study where it was concluded that an AI tool was effective in improving EFL learners' confidence and proficiency. However, it was suggested to avoid depending on it excessively. In another study conducted by Etaat (2024) AI-based written corrective feedback was found to be helpful and effective in improving the students' writing skills, especially in lexis and mechanics, followed by grammar. More recently, Wetzler et al. (2025) compared the essay grading by human instructors and AI. They concluded that AI has the potential to support the grading process, but human instructors and AI differ in scoring. Aytekin (2025) investigated the reliability and validity of AI-based grading systems, and she found that AI cannot replace human assessment, but it has high potential for consistency and speed.

Overall, studies underlined that AI has advantages in surface-level writing evaluations, but it has weaknesses in evaluating higher-level writing skills. However, research on this topic is still limited in comparing human grading and AI-based grading using the same rubric, especially for intermediate-level EFL learners. This study aims to address this gap by comparing teacher grading and AI grading in B1-level EFL learner compositions using the same rubric and thus contributes to the understanding of AI's role in EFL writing assessment.

Methodology

Research Design

Document analysis was used in this study. It is one of the qualitative research methods that are commonly used to examine written records (Bowen, 2009). Document analysis was deemed appropriate, as the aim of the research was to compare the assessment of student compositions across two different evaluative approaches, which are AI-assisted assessment and teacher-based assessment. Through this design, the study sought to identify both the consistencies and the divergences in evaluation outcomes, thereby providing insights into the potential of AI as a supplementary tool in language assessment.

Study Group

The study included 11 students learning English as a foreign language and 11 compositions written by these students. The participants consisted of B1-level students enrolled in the English Preparatory Program at a private university in the Turkish Republic of Northern Cyprus (TRNC). This level was chosen because intermediate learners are expected to demonstrate sufficient linguistic knowledge to construct coherent texts, while still presenting challenges in accuracy and organization that make assessment meaningful.

Assessment Process

In the study, the compositions written by 11 students were evaluated independently by both an AI system (DeepSeek) and three experienced English language teachers. To ensure consistency, the same analytic writing rubric was applied across all evaluations. The same writing rubric was used in all evaluations. The assessment criteria determined by the researcher were as follows: 25% content, 20% organization, 20% vocabulary use, 30% language use, and 5% mechanics. The same rubric was provided to both teachers and the DeepSeek AI program, and they were instructed to score the compositions according to these criteria. The scores obtained from each evaluation were then compared and analyzed.

AI-Assisted Assessment Procedure

The AI-assisted grading was conducted using the DeepSeek AI program. DeepSeek is a large language model (LLM) developed by DeepSeek Company, built on a transformer-based architecture and trained on a massive multilingual dataset of text and code. This training enables it to perform complex natural language processing tasks, including automated essay scoring. To ensure the AI was aligned with the study's specific rubric, a precise configuration was implemented. The primary method of

configuration was through detailed prompting. DeepSeek was instructed to base its evaluations on a designed writing rubric. This rubric included five main criteria – Grammar, Vocabulary, Organization, Mechanics, and Content – along with a corresponding scoring scale for each. Following these instructions, DeepSeek's evaluation algorithm scored the writings in accordance with the criteria specified in the rubric.

Teacher-Based Assessment Procedure

Before the assessment, teachers were provided with detailed instructions to ensure clarity in scoring. The rubric descriptors were carefully explained, and teachers were encouraged to follow the scoring guidelines closely to maintain consistency.

Findings

In this section, tables presenting the evaluations of the compositions based on the criteria of content, organization, vocabulary use, language use, and mechanics by both teachers and the DeepSeek program were provided.

Table 1
EFL Students' Writing Scores by Teacher and AI Evaluations (%)

Participant	1. teacher's evaluation (%)	2. teacher's evaluation (%)	3. teacher's evaluation (%)	AI (DeepSeek) evaluation (%)
1. Student	65	62	65	68
2. Student	62	59	56	76
3. Student	61	54	50	57
4. Student	74	64	66	71
5. Student	70	62	65	66
6. Student	62	59	58	73
7. Student	51	59	48	82
8. Student	72	65	66	81
9. Student	63	52	51	61
10. Student	67	68	60	82
11. Student	66	72	64	84

Based on the findings in Table 1, it was found that the first student's composition was evaluated as 65% by the 1st and 3rd teachers, 62% by the 2nd teacher, and 68% by DeepSeek. For the second student's composition, the evaluations were 62% by the 1st teacher, 59% by the 2nd teacher, 56% by the 3rd teacher, and 76% by DeepSeek. It was

also found that the third student's composition was evaluated as 61% by the 1st teacher, 54% by the 2nd teacher, 50% by the 3rd teacher, and 57% by DeepSeek. The fourth student's composition was assessed as 74% by the 1st teacher, 64% by the 2nd teacher, 66% by the 3rd teacher, and 71% by DeepSeek. In addition, it can be seen that the fifth student's composition was evaluated as 70% by the 1st teacher, 62% by the 2nd teacher, 65% by the 3rd teacher, and 66% by DeepSeek. The sixth student's composition was assessed as 62% by the 1st teacher, 59% by the 2nd teacher, 58% by the 3rd teacher, and 73% by DeepSeek. For the seventh student's composition, the evaluations were 51% by the 1st teacher, 59% by the 2nd teacher, 48% by the 3rd teacher, and 82% by DeepSeek. It was determined that the eighth student's composition was evaluated as 72% by the 1st teacher, 65% by the 2nd teacher, 66% by the 3rd teacher, and 81% by DeepSeek. For the ninth student's composition, the evaluations were 63% by the 1st teacher, 52% by the 2nd teacher, 51% by the 3rd teacher, and 61% by DeepSeek. Looking at the evaluation results for the tenth student's composition, it was found that the 1st teacher gave 67%, the 2nd teacher 68%, the 3rd teacher 60%, and DeepSeek 82%. Finally, it was determined that the eleventh student's composition was evaluated as 66% by the 1st teacher, 72% by the 2nd teacher, 64% by the 3rd teacher, and 84% by DeepSeek. Based on the findings, the overall assessment results of teachers and DeepSeek were largely similar, with DeepSeek giving higher scores specifically for the compositions of the 2nd , 7th , 10th, and 11th students.

Table 2

Findings Related to the Evaluation of the 1st Student's Writing Task

1. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total
1. Teacher	19	15	11	17	3	65
2. Teacher	20	15	13	12	2	62
3. Teacher	15	15	12	21	2	65
AI (DeepSeek)	18	14	13	20	3	68

As shown in Table 2, it was found that the 1st teacher graded 19% for content, 15% for organization, 11% for vocabulary use, 17% for language use, and 3% for mechanics. The 2nd teacher's scores were 20% for content, 15% for organization, 13% for vocabulary use, 12% for language use, and 2% for mechanics. The 3rd teacher graded content as 15%, organization as 15%, vocabulary use as 12%, language use as 21%, and mechanics as 2%. Finally, DeepSeek scored content as 18%, organization as 14%, vocabulary use as 13%, language use as 20%, and mechanics as 3%. Based on these findings, it can be seen that the teacher assessments and the AI assessment for the 1st student's writing task were very close to each other.

Table 3

Findings Related to the Evaluation of the 2nd Student's Writing Task

2. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	15	13	13	19	2	62
2. Teacher	16	12	12	16	3	59
3. Teacher	14	12	11	17	2	56
AI (DeepSeek)	20	17	15	20	4	76

As can be seen in Table 3, it was determined that Teacher 1 graded 15% for content, 13% for organization, 13% for vocabulary use, 19% for language use, and 2% for mechanics. Teacher 2 gave 16% for content, 12% for organization, 12% for vocabulary use, 16% for language use, and 3% for mechanics. Teacher 3, on the other hand, graded content as 14%, organization as 12%, vocabulary use as 11%, language use as 17%, and mechanics as 2%. Lastly, it was found that DeepSeek AI evaluated content as 20%, organization as 17%, vocabulary use as 15%, language use as 20%, and mechanics as 4%. Based on these findings, it can be observed that the AI gave higher scores for content and organization in Student 2's writing compared to teachers' evaluations.

Table 4

Findings Related to the Evaluation of the 3rd Student's Writing Task

3. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	15	15	12	16	3	61
2. Teacher	16	12	10	14	2	54
3. Teacher	13	11	11	12	3	50
AI (DeepSeek)	20	12	11	15	3	57

As seen in Table 4, for the 3rd student's writing, it was found that Teacher 1 gave the following scores: 15% for content, 15% for organization, 12% for vocabulary use, 16% for language use, and 3% for mechanics. Teacher 2 gave 16% for content, 12% for organization, 10% for vocabulary use, 14% for language use, and 2% for mechanics. Teacher 3, on the other hand, evaluated content as 13%, organization as 11%, vocabulary use as 11%, language use as 12%, and mechanics as 3%. Lastly, it was found that DeepSeek AI evaluated content as 20%, organization as 12%, vocabulary use as 11%, language use as 15%, and mechanics as 3%. Based on these findings, it can be seen that teachers' evaluations and the AI evaluation for the 3rd student's writing are very similar.

Table 5

Findings Related to the Evaluation of the 4th Student's Writing Task

4. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	19	16	15	20	4	74
2. Teacher	18	15	15	14	2	64
3. Teacher	15	13	13	22	3	66
AI (DeepSeek)	19	15	14	20	3	71

As can be seen in Table 5, the findings reveal that teacher 1 evaluated the writing with 19% for content, 16% for organization, 15% for vocabulary use, 20% for language use, and 4% for mechanics for the 4th student's writing task. Teacher 2 evaluated the writing with 18% for content, 15% for organization, 15% for vocabulary use, 14% for language use, and 2% for mechanics. Teacher 3, on the other hand, evaluated content as 15%, organization as 13%, vocabulary use as 13%, language use as 22%, and mechanics as 3%. Lastly, it was found that DeepSeek AI evaluated content as 19%, organization as 15%, vocabulary use as 14%, language use as 20%, and mechanics as 3%. Based on these findings, it can be concluded that teachers' evaluations and the AI evaluation for the 4th student's writing are quite similar.

Table 6

Findings Related to the Evaluation of the 5th Student's Writing Task

5. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	19	15	14	19	3	70
2. Teacher	20	13	13	14	2	62
3. Teacher	15	15	12	21	2	65
AI (DeepSeek)	17	14	13	18	4	66

Table 6 shows that, for the 5th student writing, teacher 1 graded 19% for content, 15% for organization, 14% for vocabulary use, 19% for language use, and 3% for mechanics. Teacher 2 gave 20% for content, 13% for organization, 13% for vocabulary use, 14% for language use, and 2% for mechanics. Teacher 3, on the other hand, graded content as 15%, organization as 15%, vocabulary use as 12%, language use as 21%, and mechanics as 2%. Lastly, it was found that DeepSeek evaluated content as 17%, organization as 14%, vocabulary use as 13%, language use as 18%, and mechanics as 4%. Based on these findings, it can once again be said that the teacher and AI evaluations are very close to each other.

Table 7

Findings Related to the Evaluation of the 6th Student's Writing Task

6. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	16	15	12	17	2	62
2. Teacher	18	13	12	14	2	59
3. Teacher	15	12	12	17	2	58
AI (DeepSeek)	20	18	14	18	3	73

Based on the findings in Table 7, it was determined that teacher 1 gave 16% for content, 15% for organization, 12% for vocabulary use, 17% for language use, and 2% for mechanics for the 6th student's writing. Teacher 2 was found to have given 18% for content, 13% for organization, 12% for vocabulary use, 14% for language use, and 2% for mechanics. Teacher 3 evaluated the writing with 15% for content, 12% for organization, 12% for vocabulary use, 17% for language use, and 2% for mechanics. Lastly, it was found that DeepSeek AI evaluated content as 20%, organization as 18%, vocabulary use as 14%, language use as 18%, and mechanics as 3%.

From these findings, it can be said that in evaluating the 6th student's writing, the AI assigned higher scores for content and organization compared to teachers. When scoring the content and organization of the writing, teachers tend to place greater emphasis on content, expecting students to provide more explanations and examples to support their ideas. They also give more importance to the organization of ideas and paragraphs. On the other hand, in AI evaluation, it is considered sufficient for the content and paragraphs to be meaningful and fluent.

Table 8

Findings Related to the Evaluation of the 7th Student's Writing Task

7. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	12	8	10	19	2	51
2. Teacher	19	10	13	15	2	59
3. Teacher	13	6	10	16	3	48
AI (DeepSeek)	22	19	17	21	3	82

Based on the findings in Table 8, it was determined that teacher 1 gave 12% for content, 8% for organization, 10% for vocabulary use, 19% for language use, and 2% for mechanics. Teacher 2 was found to have given 19% for content, 10% for organization, 13% for vocabulary use, 15% for language use, and 2% for mechanics. Teacher 3, on the other hand, was found to have evaluated content as 13%, organization as 6%, vocabulary use as 10%, language use as 16%, and mechanics as 3%. Lastly, it was

observed that DeepSeek AI evaluated content as 22%, organization as 19%, vocabulary use as 17%, language use as 21%, and mechanics as 3%.

From these findings, it can be said that in the evaluation of the 7th student's writing, the AI once again assigned higher scores for content and organization compared to teachers.

Table 9

Findings Related to the Evaluation of the 8th Student's Writing Task

8. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	19	16	14	20	3	72
2. Teacher	18	13	15	17	2	65
3. Teacher	18	14	12	20	2	66
AI (DeepSeek)	21	19	16	22	3	81

Based on the findings in Table 9, for the 8th student's writing, it was determined that teacher 1 graded the content as 19%, organization as 16%, vocabulary use as 14%, language use as 20%, and mechanics as 3%. Teacher 2 was found to have given 18% for content, 13% for organization, 15% for vocabulary use, 17% for language use, and 2% for mechanics. Teacher 3, on the other hand, was found to have evaluated content as 18%, organization as 14%, vocabulary use as 12%, language use as 20%, and mechanics as 2%. Lastly, it was found that DeepSeek AI evaluated content as 21%, organization as 19%, vocabulary use as 16%, language use as 22%, and mechanics as 3%.

From these findings, when the evaluations of the 8th student's writing are examined, it can be observed that, once again, the AI assigned higher scores for organization compared to teachers.

Table 10

Findings Related to the Evaluation of the 9th Student's Writing Task

9. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	15	14	15	16	3	63
2. Teacher	13	12	12	13	2	52
3. Teacher	12	11	13	13	2	51
AI (DeepSeek)	17	14	12	16	2	61

Based on the findings in Table 10, it was found that teacher 1 evaluated the 9th student's writing as followed: 15% for content, 14% for organization, 15% for vocabulary use, 16% for language use, and 3% for mechanics. Teacher 2 gave 13% for content, 12% for organization, 12% for vocabulary use, 13% for language use, and 2% for mechanics.

Teacher 3, on the other hand, evaluated content as 12%, organization as 11%, vocabulary use as 13%, language use as 13%, and mechanics as 2%. Lastly, it was found that DeepSeek AI evaluated content as 17%, organization as 14%, vocabulary use as 12%, language use as 16%, and mechanics as 2%.

As shown in Table 10, the evaluations of the 9th student's writing indicate that the scores given by teachers and the AI are quite similar.

Table 11

Findings Related to the Evaluation of the 10th Student's Writing Task

10. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	18	12	12	21	4	67
2. Teacher	20	13	15	18	2	68
3. Teacher	16	11	12	18	3	60
AI (DeepSeek)	22	18	16	22	4	82

Based on the findings in Table 11, it can be seen that teacher 1 gave 18% for content, 12% for organization, 12% for vocabulary use, 21% for language use, and 4% for mechanics of the 10th student's writing. Teacher 2 gave 20% for content, 13% for organization, 15% for vocabulary use, 18% for language use, and 2% for mechanics. Teacher 3 provided 16% for content, 11% for organization, 12% for vocabulary use, 18% for language use, and 3% for mechanics. Lastly, it was found that DeepSeek AI evaluated content as 22%, organization as 18%, vocabulary use as 16%, language use as 22%, and mechanics as 4%.

Based on the findings from the 10th student's writing, AI considers the logical progression of the composition to be sufficient. It can be said that AI does not pay much attention to paragraphs and transitions. For AI, it is enough that the sentences are logical and fluent. If the composition progresses logically and coherently, it may overlook transitions between paragraphs. It is seen that according to AI, consistency between sentences in the composition is considered sufficient.

Table 12

Findings Related to the Evaluation of the 11th Student's Writing Task

11. Participant	Content (%25)	Organization (%20)	Vocabulary (%20)	Language Use (%30)	Mechanics (%5)	Total (%100)
1. Teacher	18	10	15	19	4	66
2. Teacher	20	14	17	18	3	72
3. Teacher	17	13	14	16	4	64
AI (DeepSeek)	23	19	17	21	4	84

Based on the findings in Table 12, for the 11th student's writing, it was found that teacher 1 evaluated the writing as followed: 18% for content, 10% for organization, 15% for vocabulary use, 19% for language use, and 4% for mechanics. Teacher 2 gave 20% for content, 14% for organization, 17% for vocabulary use, 18% for language use, and 3% for mechanics. Teacher 3, on the other hand, evaluated content as 17%, organization as 13%, vocabulary use as 14%, language use as 16%, and mechanics as 4%. Lastly, DeepSeek AI evaluated content as 23%, organization as 19%, vocabulary use as 17%, language use as 21%, and mechanics as 4%.

Based on Table 12, when the composition of the 11th student is examined, it can be concluded that, similar to several other students, the AI assigned higher scores for organization compared to teachers. Teachers tend to place greater emphasis on the organization of ideas and paragraphs.

Discussion

Based on the findings of this study, meaningful results have been reached. First of all, it can be stated that AI is sufficiently effective, fast, and consistent in evaluating grammar, vocabulary usage, word repetition, sentence structure, spelling, and basic writing conventions. Similarly, Bouziane & Bouziane (2024) noted in their study that AI demonstrated superior performance in assessing areas such as grammar, spelling, sentence structure, and punctuation, aligning with the findings of the present research. Likewise, Harunasari (2023) highlighted that AI is effective in identifying and correcting grammar and spelling errors. Xiao et al. (2025) also found that AI scoring was successful in scoring grammar use. This similarity among studies shows that AI has a high potential in recognizing surface-level errors.

Another finding of the study is that teacher evaluations are more effective than AI in assessing content and organization. When scoring the content and organization of a composition, teachers place greater emphasis on content than AI and expect students to provide more explanations and examples to support their ideas. In addition, teachers give more importance to the organization of ideas and paragraphs. However, if the content and paragraphs are meaningful and fluent, AI considers it sufficient. When evaluating compositions, AI tends to pay less attention to details. Based on this, it can be said that if the writing is logical and fluent, AI does not focus on the details. For example, according to the findings from the 10th student's writing sample, AI considers a composition acceptable as long as it progresses logically. It does not pay close attention to paragraphing or transitions. For AI, sentence-level consistency in the composition was sufficient. This difference illustrates an important pedagogical concern: AI is able to evaluate surface-level features, but it remains limited to interpret higher-order features such as argumentation, developing ideas, and rhetorical structure.

In terms of educational perspective, this shows the need for human teachers who are able to evaluate writing in communicative purpose and context, which are limited in AI's analytical capacity.

Geçkin et al. (2023) found a similar result in their study, in which they observed that teachers' evaluation was more successful in assessing the organization of writing. Likewise, in a study by Sun (2023), it was stated that AI was limited in the deep evaluation of the content. Yamashita (2020) also underlined that while AI is effective in error detection, it is insufficient in evaluating argumentation, coherence, and rhetorical appropriateness, which are higher order skills. These results indicate that AI is effective in controlled and form focused assessments, but it has weaknesses in holistic and communicative writing assessments.

It is also important to note that the evaluation process takes a significant amount of time for teachers, whereas AI-assisted assessment takes a short time. Sun (2023) and Kemal & Liman-Kaban (2025) reported in their studies that AI is very fast in assessing the writing. These results have clear implications, especially for large classes in which providing feedback and assessing writing are challenging tasks. Integrating AI-based assessments could help teachers to provide feedback and personalized instruction, and therefore improve the learning process.

However, the present study has the limitation that it examined a small sample size. As the data was collected from only 11 EFL students in one institution, the findings are limited in generalizability to wide EFL populations and different levels of proficiency. This limitation restricts the external validity of the results, and the conclusions should therefore be interpreted with caution. Nevertheless, working with a small group allowed for a more controlled comparison between teacher and AI grading and provided rich qualitative insights into their evaluative differences. Future research with larger and more diverse samples is recommended to verify and extend these findings across various educational contexts.

These findings are significant for EFL writing teaching and learning. Using AI-based assessment can improve efficiency, consistency and objectivity in assessing surface-level writing elements, enabling teachers to focus more on qualitative aspects like content, creativity and organization. AI should be used as an additional assisting tool to support and improve human evaluation, instead of substituting human teachers. This corresponds with contemporary views that support human and AI collaboration in education, fostering fair, balanced and valid assessment practices (Zawacki-Richter et al., 2019; Holmes et al., 2021). On the whole, the current findings enhance the literature on AI in EFL education by providing evidence for both the strengths and limitations of AI-assisted assessments.

Conclusion

In conclusion, this study has shown that AI-assisted assessment is adequately efficient, fast and consistent in assessing grammar, vocabulary, and mechanics, whereas it still has limitations in evaluating content and organization. These results suggest that AI tools could be incorporated as supportive resources instead of evaluation systems. Likewise, numerous researchers have indicated that although AI can be used as a supportive tool, its constraints should also be considered (Bucol & Sangkawong, 2023; Song & Song, 2023; Sun, 2023; Aytekin, 2025; Wetzler, 2025).

The capability of AI to provide instant and personalized feedback on surface-level language aspects is beneficial in large classroom environments in which teachers are under time pressure. Nevertheless, human assessment is crucial for identifying deeper aspects of writing, such as argumentation, coherence, and critical engagement. Therefore, teacher assessment remains essential for comprehensive evaluation.

The findings of this research indicate that AI has significant potential to improve the efficiency, consistency, and objectivity of EFL writing evaluation. Nevertheless, teachers' interpretation and pedagogy are irreplaceable. Thus, AI's analytical capability and human knowledge could be combined to create fairer, more comprehensive, and educationally relevant evaluations.

Although the small sample size of this research restricts the generalizability of its results, it offers an important basis for future studies. Future studies should involve larger and more varied samples to examine consistency across different educational settings and proficiency levels. Additionally, future studies could investigate AI's effectiveness in evaluating more complex writing genres, including creative, academic, and critical writing, to gain a deeper insight into its educational possibilities and limitations.

References

Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16.

Anwar, M. N., & Ahmed, N. (2016). Students' difficulties in learning writing skills in second language. *Science International*, *28*(4), 735–739.

Aytekin, S. A. (2025). Evaluation of artificial intelligence based automated scoring systems in terms of reliability and Validity. *Turan: Stratejik Arastirmalar Merkezi*, *17*(66), 296–302.

Baharudin, F., Ramli, N. L., Habali, A. M., Azmi, A. A., & Rahmat, N. H. (2023). Process of writing: The challenges in writing skill among ESL learners. *International Journal of Academic Research in Business and Social Sciences*, *13*(10), 33–52.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74.

Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, *3*(1), 201.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40.

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, 1–16.

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. *Journal of Educational Technology & Online Learning*, *6*(4), 1096–1108.

Etaat, F. (2024). *The effect of AI-based applications on EFL writing skill development: An Inquiry into Integration of AI into Language Learning.* [Master's thesis, The Arctic University of Norway]. The Arctic University of Norway. https://nva.sikt.no/registration/0198f4700bc0-df948fcf-37fa-4bb1-81e0-8636b9d220f8

Harunasari, S. Y. (2023). Examining the effectiveness of AI-integrated approach in EFL writing: A case of ChatGPT. *International Journal of Progressive Sciences and Technology (IJPSAT)*, *39*(2), 357–368.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning.* Center for Curriculum Redesign.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 6300–6308.

Kemal, S., & Liman-Kaban, A. (2025). Comparative analysis of human graders and AI in assessing secondary school EFL journal writing. *Asian Journal of Distance Education*, *20*(1).

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, *10*(2), 2236469.

Mirzaeian, V. R. (2025). A comparative evaluation of artificial intelligence scoring versus human scoring of EFL students' essays. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *44*(1), 97–117.

Pandya, I., & Saiyad, A. (2025). From struggle to mastery: Writing as a transformative skill in second language learning. *Journal of English Language Teaching*, *67*(2), 32–37.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8–25.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843.

Sun, T. (2023). The potential use of artificial intelligence in ESL writing assessment: A case study of IELTS writing tasks. *Irish Journal of Technology Enhanced Learning*, *7*(2), 42–51.

Tsai, C. Y., Lin, Y. T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies*, *29*(17), 22427–22445.

Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course. *Library Hi Tech*, *40*(1), 80–97.

Xiao, F., Zhu, S., & Xin, W. (2025). Exploring the landscape of generative AI (ChatGPT)-powered writing instruction in English as a foreign language education: A scoping review. *ECNU Review of Education*, 20965311241310881.

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., & Wood, M. (2025). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology*, *52*(3), 298–304.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1–27.

Dirbtinio intelekto ir dėstytojo atliekamo EFL studentų rašymo vertinimo priešprieša: lyginamasis tyrimas

İbrahim Tekman[1], Tülay Kaya Tekman[2]

[1] Kipro tarptautinis universitetas, Užsienio kalbų fakultetas, Anglų kalbos parengiamoji programa, Haspolato universiteto miestelis, 99258 Nikosija, Šiaurės Kipras, itekman@ciu.edu.tr

[2] Artimųjų Rytų universitetas, Švietimo fakultetas, Turkų kalbos mokymo katedra, Artimųjų Rytų bulvaras, 99138 Nikosija, Šiaurės Kipras, tulay.kaya@neu.edu.tr

Santrauka

Dirbtinis intelektas (DI) tapo neišvengiama mūsų gyvenimo dalimi, taip pat ir švietimo sistemoje. Anglų kalbos, kaip užsienio kalbos, mokymasis (EFL) sparčiai keičiasi integruojant DI. Vienas iš svarbių užsienio kalbos mokymo aspektų yra vertinimas. DI integravimas į EFL rašymo vertinimą sulaukė susidomėjimo: toks vertinimas gali pagerinti vertinimo praktiką, sumažinti vertinimo apkrovą ir suteikti greitą grįžtamąjį ryšį.

Šiame tyrime buvo naudojamas kokybinės analizės modelis, siekiant palyginti DI grįstą vertinimą ir dėstytojų vertinimą įvertinant EFL studentų rašymo kompozicijas. Buvo įvertinta vienuolika kompozicijų, parašytų B1 kalbos lygio studentų, studijuojančių privačiame Kipro universitete, Anglų kalbos parengimo katedroje. Rašinius vertino dirbtinio intelekto programa „DeepSeek“ ir trys anglų kalbos dėstytojai, naudodami tą pačią rašymo vertinimo skalę, pagal

kurią buvo vertinamas turinys, struktūra, žodynas, kalbos vartojimas ir rašymo technika. Rezultatai parodė, kad dirbtinio intelekto atliktas vertinimas buvo veiksmingas, greitas ir nuoseklus vertinant gramatiką, žodyną ir rašymo techniką, tačiau turėjo trūkumų vertinant turinį ir struktūrą. Darytina išvada, kad dirbtiniu intelektu grįstas vertinimas gali būti integruotas kaip pagalbinė priemonė, tačiau tai negali visiškai pakeisti dėstytojų vertinimo.

Esminiai žodžiai: *dirbtinis intelektas, EFL, rašymas, vertinimas.*

Gauta 2025 08 04 / Received 04 08 2025
Priimta 2025 11 10 / Accepted 10 11 2025