International Journal of Educational Studies and Policy (IJESP)

Volume: 6, Issue: 2, November 2025

Automated Assessment of Students' Critical Writing Skills with ChatGPT

Serdar Tekin[1], Şeyhmus Aydoğdu[2]

ABSTRACT

Critical writing, a subskill of critical thinking, is a crucial skill for students to obtain in their education life. Since these skills require high-level cognitive skills such as analysis and evaluation, open-ended questions are used to evaluate students. Automated essay scoring (AES) tools can be used to overcome the difficulties in evaluating open-ended questions. This study aims to investigate the reliability of ChatGPT 3.5 as an AES tool for evaluating critical writing. It examines variations in average scores between a human rater and ChatGPT across diverse critical writing criteria, utilizing 59 essays from tertiary-level students majoring in teaching English as a foreign language. Reliability between raters was determined by intraclass correlation coefficients and the average score difference between raters was determined by Repeated Measures ANOVA. The findings indicate that ChatGPT, as an AES tool, demonstrates low reliability in assessing critical writing skills, suggesting its current role as a supplementary tool rather than a replacement for human raters. It was also found that ChatGPT tends to give higher scores than the human rater. The discussion aligns the results with existing literature, proposing future research avenues to leverage ChatGPT's potential as a supplementary tool for enhancing critical writing skills.

Keywords: Critical writing skills, Automated essay scoring (AES), ChatGPT, Artificial intelligence

DOI: https://doi.org/10.63612/ijesp.1735968

Article Info:

Received: 06.07.2025 **Accepted:** 31.08.2025 **Article Type:** Research Article

Cite as: Tekin, S & Aydoğdu, Ş. (2025). Automated assessment of students' critical writing skills with ChatGPT. *International Journal of Educational Studies and Policy, 6*(2), 343-359.

[1]Corresponding Author: Serdar Tekin, Tallinn University, stekin@tlu.ee, ORCID: 0000-0003-4625-4324

[2]Şeyhmus Aydoğdu, University of Nottingham, aydogduseyhmus@gmail.com, ORCID: 0000-0002-9075-8055

Introduction

Critical thinking is a multifaceted concept that encompasses a broad set of cognitive skills such as analyzing, evaluating, inferring, and interpreting (Nosich, 2022). As a subset of critical thinking, critical writing is the practical application of these skills into writing. In addition to the analysis and evaluation of information, critical writing involves individuals expressing their thoughts, arguments, and analyses in a written format. Therefore, it necessitates careful consideration and meticulous organization of arguments and assumptions, justified reasons, logical structures, and effective communication of opinions (Barnet *et al.,* 2017). Since critical writing provides individuals with several advantages, including enabling a deep understanding of complex issues, constructing persuasive arguments, and empowering them to challenge established knowledge, it is a crucial skill for students to possess to excel academically (Nosich, 2022). Rather than mere description or summarization, it demands a meticulous analysis of the underlying assumptions, justifications, and implications of arguments both in favor of and against these assumptions (Graff & Birkenstein, 2018). The assessment of such a meticulous piece of writing requires expertise and excessive time (Hussein *et al.*, 2019).

Although several tools and pedagogies have been developed to improve writing skills (Moore *et al.*, 2016; Strobl *et al.*, 2019), evaluating this skill is a laborious endeavor. The evaluation of critical writing could be argued to pose a number of challenges for teachers. The most evident one is probably the subjectivity of evaluators (Liu *et al.*, 2014). Unlike objective tests, the evaluation of a piece of critical writing is subjective, as it requires the interpretation of clarity, depth and breadth, and the level of persuasiveness. In this regard, it is challenging to establish consistent evaluation criteria that work smoothly. Moreover, although it is possible to create a number of criteria to evaluate critical writing, it is difficult to develop standardized assessment tools that take into consideration a great diversity of ways of writing. Additionally, since it requires thorough reading and meticulous analysis of each assignment one by one, evaluation of critical writing essays usually takes excessive time, especially in large classes. Thus, automated essay scoring (AES) tools have been emerging as a practical solution recently to overcome these challenges.

AES relies on computers performing the traditional human scoring process (Shermis & Burstein, 2003). A crucial concern in automatic scoring revolves around the validity of scores generated by computational systems (Attali, 2013). To validate these automated assessments, numerous studies explore the consistency between human evaluators and automated raters (Almusharraf & Alotaibi, 2023; Hoang & Kunnan, 2016). Despite the increasing volume of research in this domain, there exists a significant gap in utilizing artificial intelligence (AI) as automated raters (Mizumoto & Eguchi, 2023), emphasizing the need for an inquiry into the reliability of these tools.

This study aims to examine the rater reliability of the ChatGPT 3.5 application in the evaluation of critical writing skills. This question is important since institutions are increasingly considering AI-assisted scoring to reduce turnaround time and cost, yet any deployment must rest on empirical evidence that such systems produce dependable, fair, and interpretable judgments. While the related literature often treats ChatGPT’s roles as teacher, guide, and evaluator without distinction, this study explicitly focuses only on its function as an evaluator, excluding its teacher and guide roles as they fall outside the scope. By isolating the evaluator role, the study addresses a persistent gap in the literature, separating instructional support from measurement functions and clarifying what, exactly, is being validated. In the study, since the evaluation process of students'

writing may differ depending on the presence of sufficient contextual information, including their grades, major, and English proficiency (Ranalli, 2018), ChatGPT was used on two occasions, one including contextual details and the other without contextual information. Critical writing is deliberately chosen as a testbed because it requires argumentation, evidence integration, and audience-aware reasoning, namely, constructs that are notoriously difficult to score consistently; therefore, it provides a stringent benchmark for AI reliability. Comparing scoring with and without contextual information allows to probe whether AI models rely on surface text features alone or whether their judgments are sensitive to extratextual cues that commonly influence human raters, with potential consequences for equity and validity (Ranalli, 2018). The overarching research problem guiding this study is to what extent can ChatGPT 3.5 provide reliable and valid scoring of critical writing tasks when compared with a human rater, both with and without contextual information. To address this problem, the following research questions are examined:

1. What is the level of reliability between the human rater and ChatGPT?
2. To what extent is there a significant difference between the average scores obtained from the human rater and ChatGPT?

Theoretical Framework

Critical Writing Skills

Critical writing is an integral component of critical thinking which is a crucial 21st century skill that incorporates many sub-skills such as inquiry, investigation, examination of evidence, exploration of alternatives, argumentation, testing conclusions, rethinking assumptions, and reflecting the entire process (Spector & Ma, 2019). Critical writing is an indispensable skill for students, particularly at the tertiary level (Graff & Birkenstein, 2018). In the field of critical thinking, various vague concepts such as careful thinking and good thinking are commonly employed (Nosich, 2022; Pithers & Soden, 2000). However, within an educational context, critical thinking can be more precisely defined as the active use of several skills, including crafting justified arguments, reasoning thoughtfully and reflectively, assessing the validity and reliability of assumptions, employing inductive and deductive logic, and making logical inferences (Pithers & Soden, 2000).

As a sub-field of critical thinking, critical writing shares similar foundational principles. It involves a meticulous consideration of crucial aspects pertinent to forming a judgment on a given issue while taking into account various factors, including data, facts, thoughts, and often overlooked aspects that may have positive or negative relevance (Nosich, 2022). In the field of academic communication, critical writing assumes an important role in advancing and disseminating knowledge across diverse disciplines, serving as a cornerstone for scholars (Bean & Melzer, 2021). Furthermore, it stands as a vital and essential skill for students, enabling them to articulate their ideas effectively in formal settings such as exams and assignments. Critical writing helps to generate new ideas and a greater understanding of complicated subjects by interacting with existing information and, when needed, criticizing it. Rather than being a mere description of the topic, it necessitates a careful analysis of underlying assumptions, arguments in favor of and against these assumptions, and implications for wider society (Graff & Birkenstein, 2018). Given this, producing a strong piece of critical writing necessitates a high level of endeavor and expertise.

Critical writing incorporates a great variety of subskills, and these are also highlighted in the relevant literature. Some of these subskills include organizing the text in a logical way from the introduction to the conclusion, checking for grammar and spelling errors, analyzing and

organizing ideas and arguments coherently, supporting arguments with references and evidence, and providing a summary or recap at the end (Barnet *et al.*, 2017; Bean & Melzer, 2021; Canagarajah, 2012; Graff & Birkenstein, 2018; Huang, 2012; Nosich, 2022). These standards can be argued to be necessary to enhance a text's accuracy, clarity, and organization.

According to Bean and Melzer (2021), coherence in critical writing refers to logically ordering ideas and keeping an overt link between sentences and different sections. In a similar vein, Nosich (2022) regards it as the smooth transition between phrases and the text's flow. Another crucial part of critical writing is presenting proof and supporting the argued points. Barnet *et al.* (2017) argue that it is crucial to base different judgments on pertinent justifications and supporting data so that it would be possible to strengthen the validity of arguments. This may be accomplished by using strategies like referencing relevant sources (Nosich, 2022) and providing factual support for claims (Bean & Melzer, 2021). In contrast to descriptive writing, Graff and Birkenstein (2018) underline that critical writing requires the evaluation of competing ideas in a clear and ordered manner throughout the text. This approach enables readers to be more aware of different perspectives and make inferences based on the information presented. At this point, the text must be free of grammatical and spelling mistakes to enhance comprehension. Therefore, there are several factors to consider creating a well-structured piece of critical writing.

Automated Essay Scoring

As Shermis and Burstein (2003, p. 8) stated, "automated essay scoring is the ability of computer technology to evaluate and score written processes." AES is a multidisciplinary research domain that draws on different fields such as natural language processing and computer science (Huawei & Aryadoust, 2023) and requires the collaboration of teachers, test developers, evaluators, and computer engineers (Shermis & Burstein, 2003). It can be used as the main means of rating or to support the rater in the learning processes of AES systems. More specifically, it is used for a variety of purposes such as determining the level of academic literacy, critical thinking, plagiarism, grading, and identifying spelling errors and the quality of writing in general (Barrot, 2023).

There are several tools that are widely used for AES, including Criterion, e-rater, Grammarly, IntelliMetric, Intelligent Essay Assessor, and MY Access (Barrot, 2023). These systems adopt different methods to assess writing quality. For example, ETS's e-rater evaluates essays by examining linguistic features and applying statistical models to judge aspects like organization, development, grammar, and mechanics. IntelliMetric, on the other hand, relies on AI-based algorithms designed to replicate the evaluative reasoning of expert human raters. In addition to these, there is a study in which the ChatGPT application is used for automatic scoring (Mizumoto & Eguchi, 2023). However, it revealed that although a certain level of accuracy was achieved using ChatGPT, the application was insufficient in reaching an agreement with human raters. Despite their popularity, the accuracy and consistency of these tools have been rigorously examined for many years (Almusharraf & Alotaibi, 2023; Attali, 2013; Bui & Barrot, 2024; Dikli & Bleyle, 2014; Powers et al., 2015; Shermis, 2014). Research indicates that there remains room for improvement in their validity, particularly in systems such as Criterion (Dikli & Bleyle, 2014) and Grammarly (Almusharraf & Alotaibi, 2023).

There are several studies focusing on the validity and reliability of evaluation carried out with AES tools (Almusharraf & Alotaibi, 2023; Dikli & Bleyle, 2014; Powers *et al.*, 2002; Shermis, 2014). For example, Shermis (2014) revealed that different AES tools yielded similar scores to a human rater and regarded these tools as reliable. On the contrary, Dikli and Bleyle

(2014) found out that the AES tool they used in their study (Criterion) was unable to identify students' mistakes or misidentified the mistakes. Similarly, Almusharraf and Alotaibi (2023) more recently stated that there was a significant difference between the evaluation results made with the AES tool and the human rater in teaching English as a foreign language (EFL). More specifically, the study revealed that the AES tool tends to give higher scores than the human rater. The importance of ensuring the reliability of scoring with AES is also emphasized in different literature review studies (Barrot, 2023; Huawei & Aryadoust, 2023; Hussein *et al.*, 2019; Ramesh & Sanampudi, 2022; Susanti *et al.*, 2023).

Empirical Data on ChatGPT as an AES Tool

In addition to other AES tools, there are several studies have been carried out to find ways for ChatGPT to be used for assessment purposes (e.g., Bui & Barrot, 2024; Chen *et al*., 2025; Latif & Zhai, 2024; Manning et al., 2025; Mizumoto & Eguchi, 2023; Shin & Lee, 2024; Tsai *et al*., 2024; Uyar & Büyükahıska, 2025; Yamashita, 2024; Yavuz *et al.*, 2024). All conducted in recent years, these studies used different ways to assess to what extent ChatGPT can be used to assess students' writing skills by using various criteria such as making comparisons with different AI and AES tools as well as human raters. Some are elaborated below.

In their 2024 study, Bui and Barrot investigated how well ChatGPT's essay scores matched those of a human rater (Bui & Barrot, 2024). They looked at 50 argumentative essays of different proficiency levels, each rated at several points in time. The results showed only a weak to moderate match, and ChatGPT almost always gave lower scores than the human rater. The researchers suggested several possible reasons: its scoring system isn't yet as sophisticated as it could be, its training data may be incomplete, periodic model updates can change how it behaves, and there's a certain built-in randomness to its responses.

A similar pattern showed up in research by Chen *et al.* in 2025, which also tested ChatGPT 3.5 (Chen *et al*., 2025). This time, the focus was on IELTS writing tasks, with scores from ChatGPT compared against those from official IELTS examiners on criteria like coherence, vocabulary, and grammar. Again, ChatGPT tended to mark essays more harshly and wasn't consistent enough to rely on. The authors concluded that, at least for now, ChatGPT shouldn't be used for official IELTS scoring.

Another version of ChatGPT was used by Uyar and Büyükahıska who tested ChatGPT 4o mini by comparing its essay scores with those given by two human raters (Uyar & Büyükahıska, 2025). They evaluated 50 essays of various types, namely, descriptive, narrative, and compare-and-contrast essays written by 10 B2-level students, using the IELTS writing criteria as the benchmark. The results echoed earlier research. ChatGPT consistently awarded lower scores than both human evaluators. Even with the upgraded model, the researchers concluded that it is still premature to depend on ChatGPT for AES. They emphasized the need for further refinement to capture the subtle and complex elements of essay assessment.

Yamashita (2024) evaluated the performance of a more advanced model, ChatGPT 4, by comparing its assessments of 136 argumentative essays with ratings from 80 human evaluators. While the AI successfully sorted the essays into three CEFR proficiency bands, its scoring matched human judgments only partially, namely, showing moderate to strong correlations, with exact agreement in just about half of the cases. Even so, these results were more encouraging than those reported in many earlier studies, suggesting that ChatGPT 4 may hold greater potential as an automated essay scoring tool.

As illustrated above, a large number of studies have been conducted to examine the applicability of ChatGPT as an AES tool. The primary concern is ensuring reliability between ChatGPT and other AI or AES tools as well as human raters. Therefore, there has been a call for further research to explore to what extent it is effective across diverse settings and types of writing (Bui & Barrot, 2024). The current study builds upon these studies and investigates the reliability of the human rater and ChatGPT in terms of evaluating critical writing pieces in EFL context.

Method

Research Design

This study employed a quantitative, comparative correlational research design (Lee, 2008). It compared scores assigned by ChatGPT 3.5 and a human rater on critical writing tasks and examined the degree of agreement between the two raters. No experimental intervention was conducted; instead, the design focused on identifying the relationship and differences between two independent scoring sources.

Participants

The participants of the current study were second-grade undergraduate students majoring in teaching English as a foreign language at the Faculty of Education at a state university. A total of 59 students participated in the study, with 42 of them being female and 17 male. All participants took a compulsory course titled "Critical Writing" in the fall term. It was a two-hour course, and the medium of instruction was English. Aiming to enhance students' argumentation and critical thinking skills, this course encompassed both theoretical and practical aspects of critical writing. It also sought to equip students with the ability to compose well-constructed and compelling pieces of writing including well-argued debates and the exchange of viewpoints. The course's focus established the assessment rubric for the students' essays in their final exams. The criteria in this rubric formed the basis for assessing the essays by both the human rater and ChatGPT.

Research Procedure

The steps followed in the research process are presented in Figure 1. The critical writing course in the research context was a two-hour compulsory course that spanned 14 weeks, with the objective of equipping students with essential skills for critically approaching, analyzing, and evaluating various real-world and literary texts across different genres, styles, and registers. The course focused on analyzing texts using a range of critical thinking and writing skills, including skimming, scanning, analyzing, summarizing, inferring, inducing, deducing, and reasoning. Additionally, it aimed to enhance students' argumentation skills in writing through practical activities.

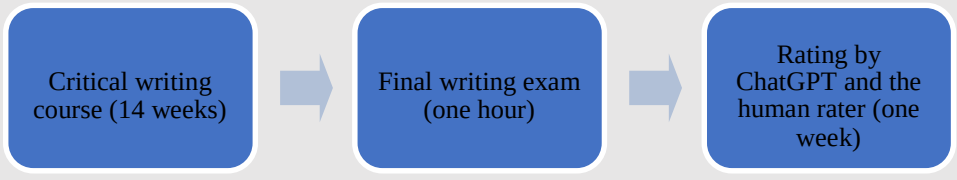

Figure 1. Research procedure

At the end of the course, a one-hour final exam was administered, requiring students to compose a critical essay on a specific topic—school uniforms. They were provided with ample information, including arguments from various sources, empirical study results, statistics, and

detailed information about references. In addition to crafting an essay following the principles of critical writing, students were expected to include in-text citations and end-text references following APA 7 guidelines.

Both the human rater and ChatGPT were provided with evaluation criteria and were requested to evaluate in line with these. It is generally known that AI tools produce more accurate results as learning data are generated and models are updated. Therefore, it is important to carry out the evaluations in the same time period to compare the results obtained from these tools and ensure the accuracy of evaluations. Considering this, both evaluations of ChatGPT in the research were conducted within the same week. Two different chats were started as ChatGPT1 and ChatGPT2 during the evaluation of participants' essays, and evaluations were carried out through these chats.

Data Collection Tools

In order to evaluate the critical essays written by the participants, a detailed literature review was conducted (Barnet *et al.*, 2017; Bean & Melzer, 2021; Canagarajah, 2012; Graff & Birkenstein, 2018; Huang, 2012; Nosich, 2022), and the following evaluation criteria emerged. Following the determination of these criteria, they were checked and confirmed by an English instructor with extensive teaching experience in critical writing skills, and the final version of the evaluation criteria was created. The evaluation criteria include eight categories: "grammar & spelling," "citation & references," "thesis statement," "topic sentence," "recap/summary in the conclusion," "support & evidence," "organization," and "coherence."

1. Grammar & Spelling: Referring to the correctness and accuracy of language use in writing, "grammar & spelling" is regarded as an important component of critical writing (Barnet *et al.*, 2017; Graff & Birkenstein, 2018; Nosich, 2022).
2. Citation & References: This is another crucial criterion concerning the use of external sources both within and at the end of the essay in a particular format, which is APA 7 guidelines in this case (Barnet *et al.*, 2017; Nosich, 2022).
3. Thesis Statement: It is the central argument that encapsulates the main idea of the critical writing process. Shaping the trajectory of the essay, it should be clear and specific as well as have the role of a guide for the reader (Barnet *et al.*, 2017; Graff & Birkenstein, 2018; Nosich, 2022).
4. Topic Sentences: "Topic sentences" are related to paragraph-level coherence, indicating the initial sentences in the paragraphs (Nosich, 2022).
5. Recap & Summary in the Conclusion: "Recap & Summary in the Conclusion" indicates the level of a concise revisit of key points in the essay and provides a summary in the conclusion paragraph (Barnet *et al.*, 2017; Huang, 2012; Nosich, 2022).
6. Support & Evidence: It is related to the elaboration of details to support the arguments, providing evidence and examples to enhance persuasiveness and relevance (Barnet *et al.*, 2017; Bean & Melzer, 2021; Graff & Birkenstein, 2018; Nosich, 2022).
7. Organization: As a crucial part of critical writing, "organization" refers to the logical flow and sequence of ideas, the order of connected points, and the structure of the essay (Barnet *et al.*, 2017; Bean & Melzer, 2021; Graff & Birkenstein, 2018; Nosich, 2022).
8. Coherence: The last criterion included in the evaluation is "coherence", which deals with the smooth transition between paragraphs and ideas to ensure a comprehensible

and well-integrated set of arguments (Bean & Melzer, 2021; Canagarajah, 2012; Nosich, 2022).

Characteristics of Evaluators

In the study, the data obtained from the students' essays were evaluated by a human rater and ChatGPT. Detailed information about the human rater and ChatGPT is provided below.

1. Human: The human rater was an expert in academic English with study-abroad experience in an English-speaking country. He held a PhD in Academic English from a UK university and possessed extensive experience in teaching English as a foreign language to various age groups. In addition, he had substantial expertise in teaching academic writing and conducting research in the field, with many years of professional experience. Given this background, the rater's evaluations were considered reliable, independent of ChatGPT's scoring. Considering these qualifications, he was invited to assess exam papers on critical writing based on the provided rubric. The rater reliability of the rubric scores was examined using Cronbach's alpha and McDonald's omega. The analyses indicated good internal consistency across the eight rubric criteria ($\alpha = .81$; $\omega = .80$), suggesting that the rater's scoring was reliable.
2. ChatGPT1: ChatGPT1 refers to the evaluation tool that is carried out without providing any contextual information about the participants. During the ChatGPT1 evaluation process, several commands were used, including the role of ChatGPT, evaluation criteria, and exam instructions.
3. ChatGPT2: ChatGPT2 refers to the evaluations that were carried out by providing contextual information about students. In addition to the role of ChatGPT, evaluation criteria, and exam instructions, ChatGPT2 included information about students' grades, the country they lived in, and the qualifications participants would have in the future.

ChatGPT 1 prompts can be found in Appendix A.

Data Analysis

Intraclass correlation coefficient (ICC) statistics were used to calculate the reliability level between the human evaluator and ChatGPT. ICC is one of the statistical methods used to determine the level of consistency and agreement among multiple raters (Shrout & Fleiss, 1979). Although ICC scores can be used to examine the reliability level between more than two raters, raters were compared pairwise for the purpose of the study. An ICC score of less than 0.5 is interpreted as poor reliability, values between 0.5 and 0.75 as moderate reliability, values between 0.75 and 0.9 as good reliability, and values greater than 0.90 as excellent reliability (Koo & Li, 2016). Differences between mean scores of raters were examined using Repeated-Measures ANOVA.

Ethics

Ethics was paramount during all stages of the study. First, ethics approval was obtained from the related university's ethics committee with a confirmation number 2023.11.269. Consent was granted from the participants who were informed of the study both orally and in a written way. No identifying personal information was put into ChatGPT or given to the examiner. In summary, as the British Education Research Association (BERA) emphasizes, it was ensured that participants of this research were "treated fairly, sensitively, with dignity…" (BERA, 2024 p. 11).

Results

The reliability level between the scores of the human rater and ChatGPT

Inter-rater reliability obtained from ChatGPT and the human rater was calculated in terms of "grammar & spelling", "citation & references", "thesis statement", "topic sentence", "recap/summary in the conclusion", "support & evidence", "organization", and "coherence", and their total scores were compared and examined. Findings regarding the level of reliability between the two raters are presented in Table 1 with different parameters.

Table 1. Correlation and reliability levels between human rater and ChatGPT

Variable	Rater	ChatGPT1			ChatGPT2		
		r	ICC	p	r	ICC	p
Grammar & spelling	Human	.225	.054	.163	.291	.084	.073
	ChatGPT1				.329	.489	.006*
Citation & references	Human	.076	.070	.297	.088	.097	.255
	ChatGPT1				.874	.919	.000*
Thesis statement	Human	.197	.106	.252	.142	.095	.306
	ChatGPT1				.301	.390	.010*
Topic sentence	Human	.048	.049	.416	.250	.239	.151
	ChatGPT1				.295	.362	.011*
Recap/summary in the conclusion	Human	.138	.087	.332	.214	.139	.251
	ChatGPT1				.214	.337	.050*
Support & evidence	Human	.400	.259	.027*	.329	.242	.047*
	ChatGPT1				.640	.777	.000*
Organization	Human	.279	.139	.129	.132	.077	.274
	ChatGPT1				.523	.685	.000*
Coherence	Human	.152	.059	.275	.158	.077	.227
	ChatGPT1				.440	.603	.000*
Total Score	Human	.425	.175	.018*	.452	.237	.005*
	ChatGPT1				.783	.846	.000*

*p < .05

The detailed examination of inter-rater reliability between human rater and ChatGPT1 reveals that the ratings are statistically significant in "support & evidence" (ICC=.259, p<.05) and total scores (ICC=.175, p<.05). However, ICC values are less than 0.5 in both variables, and it indicates that there is poor reliability between human and ChatGPT1 raters. Examination of the inter-rater reliability between the human rater and ChatGPT2 yielded similar findings to those obtained from the human rater and ChatGPT1 comparison. In other words, "support & evidence" (ICC=.242, p<.05) and total scores (ICC=.237, p<.05) appear to be statistically significant. Similarly, since ICC values in both variables are less than 0.5, it can be argued that there is poor reliability between the human rater and ChatGPT2 scoring. This finding shows that the scoring reliability between ChatGPT and the human rater is low in terms of critical writing skills, regardless of the information about the users' contextual information.

The examination of inter-rater reliability between ChatGPT1 and ChatGPT2 shows weak inter-rater reliability in the variables "grammar & spelling", "thesis statement", "topic sentence", and "recap/summary in the conclusion" (ICC < .50, p < .05). Moreover, the findings show that there is moderate reliability in the "organization" and "coherence" variables (.50 < ICC <.75, p<.05), good reliability in the "support & evidence" and "total score" variables (.75<ICC<.90, p<.05), and excellent reliability in the "citation & references" variable (ICC>.90, p<.05). It could

be interpreted that there may be differences in scoring when participants' contextual information is included in the variables "grammar & spelling", "thesis statement", "topic sentence" and "recap/summary in the conclusion". On the other hand, contextual information in the variables of "organization", "coherence", "support & evidence", "total score", and "citation & references" is largely ignored by ChatGPT.

To what extent is there a significant difference between the average scores obtained from the human rater and ChatGPT?

In line with the second RQ, this section focuses on the findings regarding the comparison of the mean scores given by human and ChatGPT raters in the context of the evaluation criteria of critical writing skills. It elaborates on descriptive statistical findings as well as post-hoc analysis results for significant differences between raters.

Table 2. Paired-sample t-test findings of human and ChatGPT ratings

	Human		ChatGPT1		ChatGPT2			
Variables	M	SD	M	SD	M	SD	F	η^2
Grammar & spelling	4.54	2.03	8.63	0.64	8.51	0.75	228.02*	.80
Citation & references	13.36	4.37	18.44	2.88	18.15	3.62	49.82*	.46
Thesis statement	6.39	3.42	9.07	0.81	8.47	0.86	29.40*	.34
Topic sentence	7.78	2.67	8.80	0.85	8.12	0.79	6.21*	.10
Recap/summary in the conclusion	6.78	3.33	8.56	0.73	8.29	0.72	14.58*	.20
Support & evidence	10.90	3.11	13.64	1.09	13.47	1.18	47.38*	.45
Organization	10.44	3.22	13.80	0.92	13.83	1.07	63.44*	.52
Coherence	6.59	1.94	9.14	0.54	9.08	0.68	93.83*	.62
Total Score	66.78	16.12	90.07	5.81	87.93	6.93	133.43*	.70

*p<.05

Findings about the scores made between the human rater, ChatGPT1, and ChatGPT2 based on critical writing variables are presented in Table 2. The results show that there is a significant difference between the raters in terms of "grammar & spelling" F(1.26, 73.15)=228.02, MSE=2.21, p=.000, η2=.80. Based on the results of post-hoc pairwise comparisons with a Bonferroni adjustment, there is a significant difference between the human rater (M=4.54, SD=2.03) and ChatGPT1 (M=8.63, SD=0.64) (p=.000) for "grammar & spelling". Similarly, there is a significant difference between the mean scores of the human rater (M=4.54, SD=2.03) and ChatGPT2 (M=8.51, SD=0.75) (p=.000). On the other hand, the score averages given by ChatGPT1 and ChatGPT2 do not show a significant difference in the "grammar & spelling" variable (p=.798). This finding indicates that ChatGPT produces similar results when contextual information is put in the "grammar & spelling". Additionally, the examination of score differences reveals that ChatGPT gives higher scores for "grammar & spelling" than the human rater.

Regarding the "citation & references" variable, there is a significant difference between the average scores between raters F(1.17, 67.74)= 49.82, MSE=16.54, p=.000, η2=.46. Post-hoc pairwise comparison results of "citation & references" show a significant difference (p=.000) between the mean scores of the human (M=13.36, SD=4.37) and ChatGPT1 (M=18.44, SD=2.88). Similarly, there is a significant difference (p=.000) between the mean scores of the human rater (M=13.36, SD=4.37) and ChatGPT2 (M=18.15, SD=3.62). However, there is no significant difference in the same variable between the mean scores given by ChatGPT1 and ChatGPT2 (p=.657). It can be interpreted that the ChatGPT application gives higher scores than the scores

obtained from the human rater. In the "citation & references" variable, there is no difference between the average scores in two different evaluations made by ChatGPT.

Regarding the variable of "thesis statement", there is a significant difference between the average scores in the evaluations F(1.12, 65.49)=29.40, MSE=7.03, p=.000, η2=.34. The results of post-hoc pairwise comparisons show that there is a significant difference (p=.000) between the mean scores of the human rater (M=6.39, SD=3.42), ChatGPT1 (M=9.07, SD=0.81), and ChatGPT2 (M=8.47, SD=0.86). This finding can be interpreted as the human rater's scores being lower than those given by ChatGPT. Additionally, it is revealed that higher scores are given for the "thesis statement" variable by ChatGPT when no contextual information is provided.

When the mean score results for the "topic sentence" were examined, a significant difference was found between the raters F(1.20, 69.28)=6.21, MSE=4.27, p=.011, η2=.10. Post-hoc pairwise comparison results show that there is a significant difference (p=.019) between the mean scores of the human rater (M=7.78, SD=2.67) and ChatGPT1 (M=8.80, SD=0.85). The findings show that no significant difference was found in "topic sentence" between the average scores of the human rater (M=7.78, SD=2.67) and ChatGPT2 (M=8.12, SD=0.79) (p=.957). On the other hand, there is a significant difference (p=.000) between the mean scores of ChatGPT1 (M=8.80, SD=0.85) and ChatGPT2 (M=8.12, SD=0.79).

In terms of the variable of "recap/summary in the conclusion", there is a significant difference between the raters F(1.12, 64.73)=14.58, MSE=6.67, p=.000, η2=.20. For this variable, post-hoc pairwise comparisons with a Bonferroni adjustment show that there is a significant difference (p= .000) between the human rater (M=6.78, SD=3.33) and ChatGPT1 (M=8.56, SD=0.73) score averages. Similarly, a significant difference was found in the mean scores of the human rater (M=6.78, SD=3.33) and ChatGPT2 (M=8.29, SD=0.72) in terms of this variable (p=.003). However, the mean scores given by ChatGPT1 (M=8.56, SD=0.73) and ChatGPT2 (M=8.29, SD=0.72) do not show a significant difference in the "recap/summary in the conclusion" variable (p= .076).

The findings also show that there is a significant difference between the average scores between the raters in the "support & evidence" variable. F(1.17, 67.83)=47.38, MSE=5.04, p=.000, η2=.45. Post-hoc pairwise comparisons with a Bonferroni adjustment reveal that there is a significant difference (p=.000) between the mean scores for the human rater (M=10.90, SD=3.11) and ChatGPT1 (M=13.64, SD=1.09) and ChatGPT2 (M=13.47, SD=1.18). However, the mean scores given by ChatGPT1 (M=13.64, SD=1.09) and ChatGPT2 (M=13.47, SD=1.18) do not show a significant difference in the "support & evidence" variable (p= .551).

There is also a significant difference between the average scores of the raters in the "organization" variable. F(1.14, 66.19)=63.44, MSE=6.18, p=.000, η2=.52. Based on the post-hoc results regarding which raters differ in the mean scores, there is a significant difference between the mean scores (p=.000) of the human rater (M=10.44, SD=3.22) and ChatGPT1 (M=13.80, SD=0.92) as well as the human rater (M=10.44, SD=3.22) and ChatGPT2 (M=13.83, SD=1.07). However, there is no significant difference between the average scores of ChatGPT1 (M=13.80, SD=0.92) and ChatGPT2 (M=13.83, SD=1.07) (p=1.000).

In terms of the "coherence" variable, there is a significant difference between the average scores of the raters F(1.17, 68.08)=93.83, MSE=2.26, p=.000, η2=.62. The post-hoc results regarding which raters differ in the mean scores reveal that there is a significant difference (p=.000) between the mean scores of the human rater (M=6.59, SD=1.94) and ChatGPT1 (M=9.14,

SD=0.54) as well as the human rater (M=6.59, SD=1.94) and ChatGPT2 (M=9.08, SD=0.68). However, there is no significant difference between the average scores of ChatGPT1 (M=9.14, SD=0.54) and ChatGPT2 (M=9.08, SD=0.68) (p=1.000).

Finally, there is a significant difference between the total score averages given by the raters F(1.14, 65.86)=133.43, MSE=129.07, p=.000, η2=.70. Post-hoc pairwise comparisons with a Bonferroni adjustment for total show that there is a significant difference (p=.000) between the human rater (M=66.78, SD=16.12) and ChatGPT1 (M=90.07, SD=5.81). Similarly, a significant difference (p=.000) was found between the mean scores of the human rater (M=66.78, SD=16.12) and ChatGPT2 (M=87.93, SD=6.93). Finally, a significant difference (p=.001) was found between the mean scores of ChatGPT1 (M=90.07, SD=5.81) and ChatGPT2 (M=87.93, SD=6.93).

As demonstrated by effect size analyses, the differences between human and ChatGPT scores are not only statistically significant across all criteria, but also practically significant. The data analysis revealed considerable effect sizes in the Grammar & Spelling (.80), Coherence (.62) and Total Score (.70) variables. This finding suggests that ChatGPT tends to achieve significantly higher scores than human evaluators in these domains. Large effect sizes were identified in the Citation & references (.46), Support & evidence (.45), and Organization (.52) criteria, indicating that ChatGPT produced significantly higher values than human scores in these criteria. The Thesis statement (.34) and Recap/summary in the conclusion (.20) variables also demonstrated large effect sizes. Conversely, the topic sentence (.10) demonstrated an effect size ranging from medium to large, signifying a comparatively diminished level of discrepancy relative to the other criteria. The results indicate that ChatGPT consistently outperformed human raters on nearly all criteria, with the majority of these differences proving to be statistically significant. The interpretation of effect sizes is based on the thresholds proposed by Cohen (1988); therefore, $\eta^2 \approx .20$ indicates a small effect, .50 indicates a moderate effect, and $\geq .80$ indicates a large effect.

Discussion and Conclusion

This study aimed to use ChatGPT as an automatic rater for the evaluation of critical writing skills. Moreover, it was aimed to investigate the inter-reliability between ChatGPT and the human rater as well as the average score differences between these two raters in terms of several key components of critical writing skills, including “grammar & spelling”, “citation & references”, “thesis statement”, “topic sentence”, “recap/summary in the conclusion”, “support & evidence”, “organization”, and “coherence”.

Whether or not contextual information was given, ChatGPT's dependability was shown to be low when compared to the human rater's scores in the study. Although the level of inter-rater correlation in the evaluation of ChatGPT1 and ChatGPT2 in total scores is at a medium level (r(human, ChatGPT1)=.425, r(human, ChatGPT2)=.452), the poor inter-rater reliability between ChatGPT and the human rater means that ChatGPT is not sufficient to be used alone for scoring purposes. This is largely consistent with the findings of Dikli and Bleyle's (2014) study, which showed that feedback from an AES tool (Criterion) and a human rater differed significantly. Some errors in the study were missed by the AES tool, which is consistent with the results of the current study's use of ChatGPT. In a similar vein, the results are consistent with the research of Mizumoto and Eguchi (2023), which showed that there was little agreement in terms of overall score between ChatGPT and the human rater.

In the study, comparing the scores of the human rater, ChatGPT1, and ChatGPT2 regarding critical writing skills revealed that there was no significant difference between the human rater and

ChatGPT2, only in the "topic sentence" criterion. Although there is no significant difference between raters in the “topic sentence” variable, the lack of significance in the inter-rater reliability level shows that ChatGPT cannot be used alone for automatic scoring purposes. Figure 2 presents the statistical significance of the pairwise comparisons between the raters, as well as the average scores in the groups based on the evaluation criteria of critical writing skills.

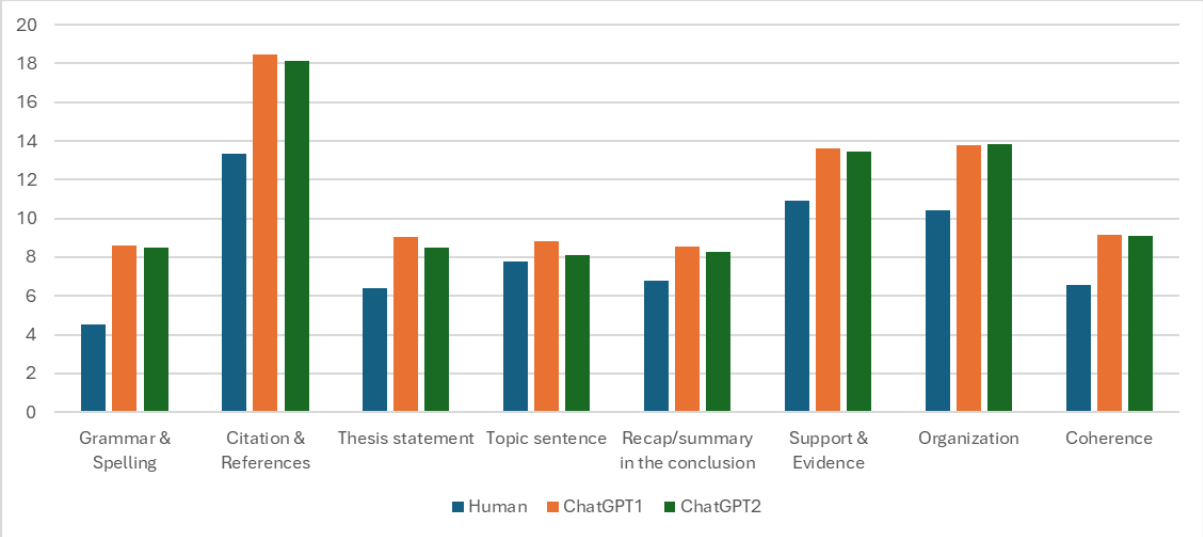

Figure 2. Mean scores of raters on critical writing skills

This study shows that ChatGPT tends to give high scores to students, which is similar to what Almusharraf and Alotaibi (2023) obtained from the Grammarly application. The findings of this study support previous research that favors the use of AES tools as an aid or complementary to human raters rather than standalone use (Almusharraf & Alotaibi, 2023; Attali *et al.*, 2013; Mizumoto & Eguchi, 2023).

The results of this study also show that ChatGPT 3.5, as an AES tool, has a low level of reliability for the assessment of critical writing skills. Therefore, in its current form, AES tools can only be used as a supplement; they cannot take the role of human raters (Mizumoto & Eguchi, 2023). AES users should integrate the tools into their learning environments, taking into account the potential limitations (Li, 2021), and should not assume that the results regarding the assessment tools used are completely accurate.

The research results show that ChatGPT tends to award higher scores than human raters across all dimensions of critical writing skills. This finding is consistent with previous studies (Manning *et al.*, 2025; Shin & Lee, 2024). Therefore, ChatGPT models can be recommended for use in an educational context as a support tool; however, current findings indicate that these models cannot be used directly as evaluators. The tendency to assign higher scores could pose a significant risk, particularly when distinguishing between low- and high-performing students within a group.

In the study, the use of ChatGPT 3.5 was tested, and the findings are limited to this tool. Future research can focus on carrying out evaluations with the help of different versions of ChatGPT and other widely used AES tools (e.g., E-rater and IntelliMetric) in the context of critical writing skills and examining the reliability between the tools as well as the reliability with the human rater.

This research focused on the use of ChatGPT as a rater. The findings show that the reliability level of the results obtained from ChatGPT is low; however, ChatGPT can provide detailed feedback, taking into consideration each evaluation criterion. Future research is invited to develop a ChatGPT-integrated online platform in which it is used as a support tool, providing feedback to its users.

Future research should also address the ethical and technical issues that could affect the adoption of AI-based scoring tools. Data privacy is a critical concern, especially when processing student writing samples, which may contain personally identifiable information. To protect user data, it will be essential to implement secure storage, anonymization protocols and compliance with relevant regulations (e.g. General Data Protection Regulation (GDPR) and The Family Educational Rights and Privacy Act (FERPA)). Another important consideration is ensuring consistent responses across multiple instances of the same artificial intelligence model, as well as across different model updates. Longitudinal evaluations, test-retest reliability assessments and calibration mechanisms can help to maintain consistent scoring behavior over time. Addressing these issues will increase the reliability of AI-based assessment systems and facilitate their sustainable integration into educational assessment practices.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

Conflicts of Interest

The authors declare that there is no conflict of interest

Ethics

The research was carried out with the approval of Nevşehir Hacı Bektaş Veli University Ethics Commission dated 27.09.2023 and numbered 2023.11.269.

References

Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and Automated Essay Scoring Approaches. *Technology, Knowledge and Learning, 28*(3), 1015-1031. https://doi.org/10.1007/s10758-022-09592-z

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). Routledge.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141. https://doi.org/10.1177/0265532212452396

Barnet, S., Bedau, H., & O'Hara, J. (2017). *Critical thinking, reading, and writing: A brief guide to argument.* Macmillan.

Barrot, J. S. (2023). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12083-y

Bean, J. C., & Melzer, D. (2021). *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom.* Jossey-Bass.

British Educational Research Association (BERA) (2024). *Ethical Guidelines for Educational Research.* Available at: http://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2024

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies, 30*(2), 2041-2058. https://doi.org/10.1007/s10639-024-12891-w

Canagarajah, A. S. (2012). Understanding critical writing. In Luria, H., Seymour, D. M., & Smoke, T. (Eds.) *Language and Linguistics in context* (pp. 307-314). Lawrence Erlbaum Associates.

Chen, X., Zhou, Z., & Prado, M. (2025). ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams. *International Journal of Assessment Tools in Education, 12*(1), 62-77. https://doi.org/10.21449/ijate.1496193

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1-17. https://doi.org/10.1016/j.asw.2014.03.006

Graff, G., & Birkenstein, C. (2018). *They say/I say: The moves that matter in academic writing*. W. W. Norton & Company.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY access. *Language Assessment Quarterly, 13*(4), 359-376. https://doi.org/10.1080/15434303.2016.1230121

Huang, S. Y. (2012). The integration of 'critical' and 'literacy' education in the EFL curriculum: Expanding the possibilities of critical writing practices. *Language, Culture and Curriculum, 25*(3), 283-298. https://doi.org/10.1080/07908318.2012.723715

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies, 28*(1), 771-795. https://doi.org/10.1007/s10639-022-11200-7

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, 1-24. https://doi.org/10.7287/peerj.preprints.27715v1

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, 1-10. https://doi.org/10.1016/j.caeai.2024.100210

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of educational research, 78*(3), 608-644. https://doi.org/10.3102/0034654308324427

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System, 99*, 1-14. https://doi.org/10.1016/j.system.2021.102505

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. ETS Research Report Series, (1), 1-23. https://doi.org/10.1002/ets2.12009

Manning, J., Baldwin, J., & Powell, N. (2025). Human versus machine: The effectiveness of ChatGPT in automated essay scoring. *Innovations in Education and Teaching International. 62*(2), 1-14. https://doi.org/10.1080/14703297.2025.2469089

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

Moore, K. A., Rutherford, C., & Crawford, K. A. (2016). Supporting postsecondary English language learners' writing proficiency using technological tools. *Journal of International Students, 6*(4), 857-872.

Nosich, G. (2022). *Critical writing: A guide to writing a paper using the concepts and processes of critical thinking*: Rowman & Littlefield.

Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational research, 42*(3), 237-249. https://doi.org/10.1080/001318800440579

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407-425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A

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

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653-674. https://doi.org/10.1080/09588221.2018.1428994

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*, 29(18), 24735-24757. https://doi.org/10.1007/s10639-024-12817-6

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Spector, J. M., & Ma, S. (2019). Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. *Smart Learning Environments, 6*(8), 1-11. https://doi.org/10.1186/s40561-019-0088-z

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33-48. https://doi.org/10.1016/j.compedu.2018.12.005

Susanti, M. N. I., Ramadhan, A., & Warnars, H. L. H. S. (2023). Automatic essay exam scoring system: A systematic literature review. *Procedia Computer Science, 216*, 531-538. https://doi.org/10.1016/j.procs.2022.12.166

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*(2), 1-19. https://doi.org/10.1007/s10639-024-12722-y

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education, 12*(1), 20-32. https://doi.org/10.21449/ijate.1517994

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), 1-14. https://doi.org/10.1016/j.rmal.2024.100133

Yavuz, F., Çelik, Ö., & Çelik, G. Y. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150-166. https://doi.org/10.1111/bjet.13494

Appendix A: ChatGPT Prompts

ChatGPT1 Prompts

I want you to act as a teacher of critical writing. I will give you the answers of some students in the final exam and the criteria for evaluating these answers. Your task is to evaluate the students' answers against the criteria using artificial intelligence tools such as natural language processing.

First of all, I will share the evaluation criteria:

1. Grammar & Spelling: 10 points
2. Citation & References (APA style): 20 points
3. Thesis Statement: 10 points
4. Topic sentence: 10 points
5. Recap/summary in the conclusion: 10 points
6. Support & Evidence: 15 points
7. Organization: 15 points
8. Coherence: 10 points

ChatGPT2 Prompts

I want you to act as a teacher of critical writing. I will give you the answers of some students in the final exam and the criteria for evaluating these answers. Your task is to evaluate the students' answers against the criteria using artificial intelligence tools such as natural language processing. Before this, I would like to give you some background information about the students and context. These students are second year pre-service English as a foreign language teachers studying in a four-year program at a state university in Türkiye. Since they are meant to be English teachers in the future, their writing skills as well as overall English level (around intermediate/upper intermediate) should be sufficient to maintain efficient teaching environment.

First of all, I will share the evaluation criteria:

1. Grammar & Spelling: 10 points
2. Citation & References (APA style): 20 points
3. Thesis Statement: 10 points
4. Topic sentence: 10 points
5. Recap/summary in the conclusion: 10 points
6. Support & Evidence: 15 points
7. Organization: 15 points
8. Coherence: 10 points