Received: 23 October 2023 | Accepted: 24 May 2024

DOI: 10.1111/bjet.13494

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments

Fatih Yavuz[1] | **Özgür Çelik**[2] | **Gamze Yavaş Çelik**[2]

[1]Preparatory Department, Mudanya University, Mudanya, Turkey

[2]School of Foreign Languages, Balıkesir University, Balıkesir, Turkey

Correspondence
Fatih Yavuz, Preparatory Department, Mudanya University, Mudanya, Turkey.
Email: yavuzf@hotmail.com

Abstract

This study investigates the validity and reliability of generative large language models (LLMs), specifically ChatGPT and Google's Bard, in grading student essays in higher education based on an analytical grading rubric. A total of 15 experienced English as a foreign language (EFL) instructors and two LLMs were asked to evaluate three student essays of varying quality. The grading scale comprised five domains: grammar, content, organization, style & expression and mechanics. The results revealed that fine-tuned ChatGPT model demonstrated a very high level of reliability with an intraclass correlation (ICC) score of 0.972, Default ChatGPT model exhibited an ICC score of 0.947 and Bard showed a substantial level of reliability with an ICC score of 0.919. Additionally, a significant overlap was observed in certain domains when comparing the grades assigned by LLMs and human raters. In conclusion, the findings suggest that while LLMs demonstrated a notable consistency and potential for grading competency, further fine-tuning and adjustment are needed for a more nuanced understanding of non-objective essay criteria. The study not only offers insights into the potential use of LLMs in grading student essays but also highlights the need for continued development and research.

KEYWORDS

AI-based grading, automated essay scoring, generative AI, large language models, reliability, rubric-based grading, validity

Practitioner notes

What is already known about this topic

- Large language models (LLMs), such as OpenAI's ChatGPT and Google's Bard, are known for their ability to generate text that mimics human-like conversation and writing.
- LLMs can perform various tasks, including essay grading.
- Intraclass correlation (ICC) is a statistical measure used to assess the reliability of ratings given by different raters (in this case, EFL instructors and LLMs).

What this paper adds

- The study makes a unique contribution by directly comparing the grading performance of expert EFL instructors with two LLMs—ChatGPT and Bard—using an analytical grading scale.
- It provides robust empirical evidence showing high reliability of LLMs in grading essays, supported by high ICC scores.
- It specifically highlights that the overall efficacy of LLMs extends to certain domains of essay grading.

Implications for practice and/or policy

- The findings open up potential new avenues for utilizing LLMs in academic settings, particularly for grading student essays, thereby possibly alleviating workload of educators.
- The paper's insistence on the need for further fine-tuning of LLMs underlines the continual interplay between technological advancement and its practical applications.
- The results lay down a footprint for future research in advancing the use of AI in essay grading.

INTRODUCTION

Assessing essays is a critical aspect of language learning evaluation, particularly in English as a foreign language (EFL) contexts, where it serves as a vital measure of student proficiency in writing skills. Despite the importance of essay writing in language assessment, the traditional methods of essay grading present significant challenges, including the need for extensive time and resources, and the potential for subjective bias in human evaluation. Essay writing, which is in the category of open-ended questions that measure whether test takers have a high level of understanding, such as more complex thinking, reasoning and adapting their existing knowledge and skills to new situations (Hussein et al., 2019), is a form of assessment and evaluation that attracts attention because it can measure high-level abilities such as creative thinking, logical thinking and critical reasoning (Uto, 2021). Especially in foreign or second language learning, writing is an active and productive skill that contributes to the learning process, and it is quite challenging for learners as it requires rhetorical organization, appropriate language use and a certain level of vocabulary knowledge (Taskiran & Goksel, 2022). However, although it is a widely recognized assessment practice, measuring writing skills is not preferred, especially in exams with large numbers of participants, due to the difficulty of valid and reliable scoring in the scoring process, being time-consuming and expensive and requiring a high number of well-trained human evaluators (Meyer et al., 2023).

Traditionally, educators have used various methods to grade student essays, including holistic and rubric-based scoring. Holistic scoring provides an overall judgement, often

represented as a letter grade, percentage or a number on an established scale (Bacha, 2001). On the other hand, rubric-based scoring evaluation is done based on specific, clearly outlined rubrics or characteristics. Each rubric has a defined scale, with scores explained comprehensively. The final grading for the essay is a cumulative sum of all the individual rubric scores (Fazal et al., 2013). Each of these methods has its advantages and disadvantages. Holistic scoring is quick and easy, but it can be subjective and difficult to ensure reliability. Rubric-based scoring is more objective, but it can be time-consuming and difficult to grade all of the components of an essay. Although it is more reliable than holistic scoring, it is also challenging to ensure validity (accurately assessing what it is intended to assess) and reliability (consistency of an assessment tool) in rubric-based scoring.

During the essay grading process, especially in exams with a large number of participants, such as massive open online courses, researchers are working on automated essay scoring systems to reduce the workload on teachers (Escalante et al., 2023), obtain valid and reliable results and reduce intrarater and interrater disagreement (Wu et al., 2023). Ifenthaler (2023) states that automated essay scoring (AES)—likewise termed automated essay grading, automated writing evaluation or automated essay evaluation—is defined as a computer-based process for applying standardized measurements to open-ended or structured response test items and that developments in the fields of computer technology, data analytics and AI have enhanced the utility, objectivity, reliability and validity of the evaluation of written texts in AES systems. Ifenthaler (2023) also states that AES systems can provide instant feedback during evaluation. AES systems imitate the human evaluation of written texts using various scoring methods such as statistics, machine learning and natural language processing (NLP) techniques (Ifenthaler, 2023). Therefore, educators have struggled to overcome the difficulties in essay grading by integrating neural language processing and machine learning processes into AES systems (Hussein et al., 2019). Studies that started with Project Essay Grader, the first widely known automatic article evaluation system in the 1960s, continued with systems such as e-rater, IEA and IntelliMetric, and open code systems such as AKOVIA, which can be used in research (Ifenthaler, 2023).

AES systems are grouped in the literature as handcrafted and neural based (Hussein et al., 2019; Uto, 2021). In their study, Shin and Gierl (2021) compared two different AES frameworks in terms of effectiveness and performance and found that recent deep neural approaches using convolutional neural networks (CNN) have significantly better AES performance compared to traditional systems using support vector machines with Coh-Metrix features. They stated that it provided satisfactory results and that the compatibility with human raters was high. However, Wu et al. (2023) found that these neural-based approaches produce predicted holistic scores that do not provide sufficient pedagogical information. In addition, Hussein et al. (2019) state that although deep learning-based systems make better results than their predecessors, they cannot directly evaluate the internal qualities of the paper as much as human raters and they may not be very good at using the complex linguistic and cognitive features that are important for articles. However, they suggest that handcraft AES systems may be better in rubric-based evaluation than AES systems that use neural network algorithms. From this perspective, it can be argued that feature-based AES systems can be evaluated as fair and more objective for students in the assessment process (Hussein et al., 2019). On the other hand, instead of thinking as if there are two mutually exclusive methods in AES, such as featured based and deep learning based, there are also studies that see the two systems as complementary to each other and approach the subject from a pedagogical perspective in terms of relevancy (Kumar & Boulanger, 2020).

One of the latest developments that brings promising innovations to language teaching, learning and research is generative AI. Currently, users can easily access AI text generation tools through publicly available interfaces (large language models—LLMs) such as ChatGPT, Bing or Bard. ChatGPT, developed by OpenAI and Bard, developed by Google, has been

researched in many aspects, but there are very few studies on their validity and reliability in automatic essay grading. Generative pre-trained transformer (GPT) and bidirectional encoder representations from transformers (BERT) are transformer-based LLMs. While BERT is generally used for natural language understanding tasks, GPT is used for natural language generation tasks (Mizumoto & Eguchi, 2023). Transformers, a deep learning neural network architecture designed to learn context and meaning from sequential data, are an evolution or extension of previous neural network architectures that combine the benefits of CNNs and recurrent neural networks (RNNs) (Mizumoto & Eguchi, 2023). In their study, Mizumoto and Eguchi (2023) found that AES using GPT has a certain level of correctness and reliability and that LLMs such as ChatGPT can be used effectively as AES tools, showing potential in both research and practice, writing evaluation and feedback methods. On the other hand, a study investigating the reliability of ChatGPT and Bard against human raters was conducted by Khademi (2023). In this study, Khademi (2023) used intraclass correlation (ICC) to compare human raters with ChatGPT and Bard, finding that reliability is lower for both ChatGPT and Bard compared to human raters.

The current literature on using AI tools as grading assistants for student essays presents several gaps, most probably due to the rapidly evolving nature of the technology. Many of these studies have taken into account previous versions of NLP models, such as GPT2, GPT3 and RNN (Fazal et al., 2013; Madala et al., 2018; Ramalingam et al., 2018; Shehab et al., 2016). The few studies that utilize upgraded AI models often fail to provide a comparative analysis against human-generated scores (Kumar & Boulanger, 2020; Mizumoto & Eguchi, 2023) or use very few (ie, one or two) human raters (Kumar & Boulanger, 2021; Ramalingam et al., 2018; Shermis, 2014; Zhao et al., 2023). Another gap in the literature is the rubric-based reliability and validity analysis of AI tools. Lastly, the temperature of LLM models was not addressed in many studies. 'Temperature' is a critical parameter influencing the model's output variability and creativity. A lower temperature value (closer to 0) makes the model's responses more deterministic and less varied. Conversely, a higher temperature (closer to 1) increases randomness, encouraging more diverse and creative outputs. The temperature setting plays a pivotal role in applications where the balance between predictability and creativity is essential, such as in automated essay grading. Given the complex and multidimensional nature of essay grading, testing the performance of AI tools based on a well-constructed rubric with clear and detailed performance indicators is critical for the discussions on AI-assisted essay grading. With this in mind, in this study, we set out to explore the validity and reliability of two LLMs, namely ChatGPT and Bard,[1] in grading university-level EFL students' essays of diverse quality based on a given rubric. This study addressed the research questions below:

RQ1: To what degree do LLMs provide reliable grading for student essays in accordance with a provided rubric?
RQ2: To what extent can LLMs reliably grade student essays across specific domains based on a rubric?
RQ3: How accurately do LLMs grade student essays based on a given rubric compared to human raters?
RQ4: In what domains do the scores assigned by LLMs align with those given by human raters?

METHODS

Materials

The materials for this study comprised three distinct essays written by three university students learning EFL. The participant students are third-year pre-service students at the

English language teaching department of a university in Türkiye. Their proficiency levels are B2/C1. In order to explore the performance of LLMs in assessing essays of varying quality, these essays were carefully chosen, to represent a poor, an average and a good essay, from a larger essay collection of student works previously assigned to the students of the second author of this paper. The essays are argumentative essays written for the same class but on different topics. They were pre-classified as poor (<2 points), average (3 points) and good (>3 points) quality, as per the grading rubric used to grade the student essays by the second author in his classes. Students' permissions were taken to use the essays for this study. Each one of the essays can be found in Appendix S1.

A revised version (Şahan, 2019) of Han's (2013) analytic writing scoring scale served as the grading rubric for student essays in this research. This refined edition of the scale encapsulates five critical elements: grammar, content, organization, style & expression and mechanics. Each element further contains five tailored performance indicators that provide clear guidelines for scoring. Contrarily to Han's original rubric, where scoring weights were attached to each component (for instance, grammar had a weight of 1.5 points, 3 points for content, 2.5 points for organization and 1 point for mechanics), the scoring weight for this study was rationalized to a common scale. We used a 5-point scale ranging from 1 (poor) to 5 (excellent) for evaluating performance on each of the five components. This modification aided in providing a uniform grading structure for overall essay evaluation, making comparison across different aspects more straightforward. The rubric can be found in Appendix S2.

Participants and data collection

The participants in this research study constituted 15 experienced ($\overline{X} = 12.6$ years) EFL instructors in higher education. Among these educators, 11 individuals held PhD degrees in English language teaching. Every participant had a background in teaching academic writing, ensuring their proficiency in utilizing the chosen grading rubric. Initially, the participants were introduced to the analytical grading rubric, enabling them to familiarize themselves with the varying performance indicators. The goal behind this activity was to ensure consistency and minimize discrepancies during the grading process. The participating raters were then acquainted with the designated essays; however, we did not share any prior information regarding the anticipated quality of the essays (labelled as Essay 1, Essay 2 and Essay 3). This strategy aimed to prevent any preconceived biases that might influence the evaluation process. We created a Google Form incorporating links to the essays and the scoring rubric (illustrated in Appendix S3) created for the grading of the essays. Then, we provided the participants with this Google Form link and asked them to assess the essays using the rubric.

Alongside the human grading, automated rating by OpenAI's ChatGPT and Google's Bard was undertaken. LLM scoring reliability data were collected by assessing each essay 10 times at separate intervals using both ChatGPT (Model GPT4) and Bard (Version: 2023.07.13). Two separate approaches, Default and FineTuned, were implemented on ChatGPT to explore the best results. The default mode engaged a simple prompt and a standard temperature level (0.7). Conversely, the fine-tuned mode utilized a more detailed prompt and a lowered temperature of 0.2. Google's Bard was used in its default setting as it does not support fine-tuning. The prompts used in this process can be referred to in Appendix S4.

To conclude, the final dataset (Appendix S5) for this study comprised scores given by 15 raters across five domains: grammar, content, organization, style and expression and mechanics. These scores were paralleled with 10 rating scores obtained from AI models (ChatGPT Default, ChatGPT FineTuned and Bard) across the same five domains.

Data analysis

In order to evaluate the reliability of the scores produced by the LLMs in our study, we initially used the intraclass correlation coefficient (ICC), which can be used for repeated measures over time, on the Statistical Package for the Social Sciences (SPSS). Through the employment of a two-way mixed model, we strived to investigate consistency and calculated average measures. By using a two-way mixed model, we acknowledge the individual differences between our specific raters, while ensuring that our focus remains on the consistency of their assessments using the chosen rubric. This approach is crucial for understanding the reliability of human grading in an EFL context, especially when comparing it to the performance of AI models (Koo & Li, 2016). Following the establishment of ICC scores, the mean rating scores for each of the five domains were calculated.

Following the analysis of the LLMs grading, we proceeded to calculate the ICC of the human raters participating in our study. This was undertaken to investigate the level of consistency across human raters' assessments. Subsequent to the computation of ICC scores, mean scores for each domain, as assessed by the human raters, were obtained. This twofold data analysis empowered us to compare and contrast the consistency and reliability of LLMs and human grading techniques based on the same evaluation criteria.

Limitations

It is important to note that this study has certain limitations. First, our sample size was relatively small, consisting of only 3 student essays with varying levels of quality and 15 EFL expert academic writing instructors. Due to the expertise required of instructors and the workload involved in grading essays, it can be difficult to find volunteers to participate in these types of studies.

For this study, two LLMs, namely ChatGPT and Bard, were used as they were the most accessible tools available at the time of the research in Türkiye. It was a deliberate decision not to include BingChat, which is Microsoft's LLM, as it uses the same algorithm as OpenAI's ChatGPT. Another LLM called Claude was not accessible from Türkiye during the study, hence it was not included.

This study is limited to the specific domains of grammar, content, organization, style & expression and mechanics, which were measured in the rubric.

Lastly, to evaluate the reliability and accuracy of FineTuned LLMs, we employed the FineTuned version of ChatGPT. However, it is worth mentioning that the fine-tuning was done only through prompt engineering and temperature adjustment. No algorithmic fine-tuning or data training was carried out.

Ethical considerations

In this study, we prioritized protecting participant rights and adhering to the principles of research integrity during the study. The participants of the study consisted of 3 university students and 15 raters (university instructors). The participants were informed in a detailed verbal interview before they were included in the study. The rater participants were given clear information about the purpose of the study, the data collection processes, the use of the data and the way the results would be published. It was stated that the scores obtained from the raters would be used anonymously in the study and that the results would be published anonymously as Supplementary Material in the journal in which the results would be published. Under these conditions, 15 instructors voluntarily participated in the study. The

student participants were verbally informed about the purpose for which their essays would be used and that these essays would be anonymized, stored and uploaded to AI tools and their essays would be added as Supplementary Material in the journal in which the study was published after removing any identifying information. Initially, their verbal consent was obtained, and they voluntarily participated in the study. In addition, written consent was also obtained from the students before starting the study. All personal identifying information about the participants was kept confidential and the data were anonymized. The results to be published are organized in such a way that they do not contain any personal identifying information.

We investigated whether the AI tools used in the research, OpenAI's ChatGPT and Google's Bard, have policies for academic research. When OpenAI's Sharing & Publication Policy was examined, it was seen that no special permission was required for the use of ChatGPT for research purposes (OpenAI, 2022, para. 11). However, no such policy was found in Google's DeepMind policies. For both tools, care was taken to protect the confidentiality of participant data and not to enter any personal identifying information into the systems.

RESULTS

RQ1: To what degree do LLMs provide reliable grading for student essays in accordance with a provided rubric?

To address this research question, we began by calculating the ICC to determine the reliability of LLMs in evaluating student essays. Using a two-way mixed model, we analysed consistency and calculated average measures. Our analysis revealed that Default ChatGPT had an ICC score of 0.947 (0.905–995) $F(4,116) = 30.43$, $p < 0.001$ over 10 measurements. In contrast, ChatGPT FineTuned had a score of 0.972 (0.920–997) $F(4,116) = 36.25$, $p < 0.001$, while Google's Bard had a score of 0.919 (0.765–990) $F(4,116) = 12.33$, $p < 0.001$.

We also calculated and compared the descriptive data of 10 measurements (Figure 1). Starting with Essay 1, which was of good quality, the grades given by LLMs show that both versions of the ChatGPT models (Default and FineTuned) were largely consistent in their scoring, with most grades lying between 4.2 and 4.6. On the other hand, Bard seemed to be more conservative in its scoring, assigning grades ranging from 3.2 to 4. It can be seen that, despite grading the same high-quality essay, there was a variation between the different LLMs, suggesting there is a level of inconsistency. Furthermore, based on the relatively low standard deviation in the scores between the 10 instances of measurements, the FineTuned ChatGPT ($SD = 0.00$) provides a higher degree of reliability in grading the essay in alignment with the provided rubric, followed by Default ChatGPT ($SD = 0.21$) and Bard ($SD = 0.28$).

Moving on to Essay 2, the average quality essay, all models display a narrower spread of scores between 3 and 3.6. The ChatGPT (FineTuned) showed a tendency to assign a stable grade of 3.6 ($SD = 0.00$) across all 10 measurements, demonstrating considerably reliable grading. Both the ChatGPT (Default) and Bard, unlike their assessment of Essay 1, maintain relatively uniform scoring patterns with a small standard deviation ($SD^{\text{ChatGPT\_Default}} = 0.20$; $SD^{\text{Bard}} = 0.21$), suggesting an improved agreement with the grading rubric.

The third essay (Essay 3), considered to be of poor quality, again demonstrates contrasting grading behaviours of the different LLMs. While ChatGPT (FineTuned) consistently scored the essays ($SD = 0.00$), the other two models show a wider spread between 2 and 3.4. Despite assigning higher scores than the FineTuned model, the Default and Bard models produced less reliable results in Essay 3 ($SD^{\text{ChatGPT\_Default}} = 0.28$; $SD^{\text{Bard}} = 0.18$).

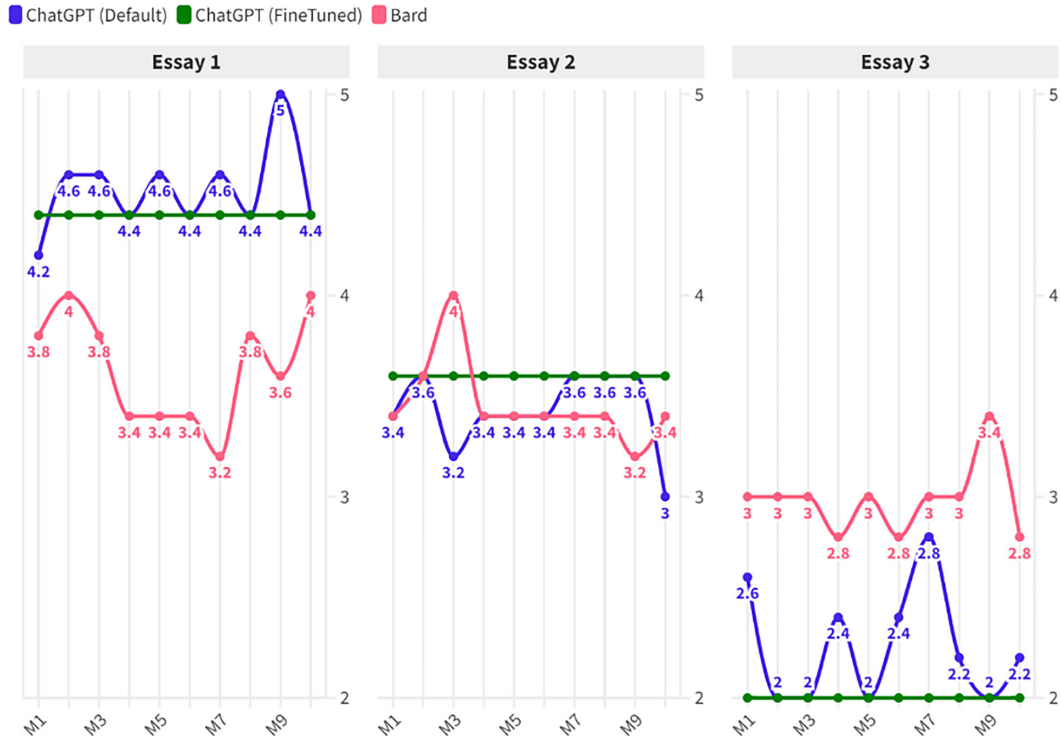

FIGURE 1 Grades assigned by LLMs to student essays.

To summarize, in response to the first research question, these findings suggest that, as shown in Figure 1, the ChatGPT models—both Default and FineTuned—establish proficiency in grading the essays and distinguishing their varying levels of sophistication. On the other hand, Bard, while being a reliable grading tool, showed a less advantageous ability to distinguish the essays based on their quality level. Furthermore, the ChatGPT FineTuned model, displaying remarkable grading consistency, appears to provide the most reliable grading among these LLMs, followed by ChatGPT Default and Bard.

RQ2: In which specific domains do LLMs demonstrate reliability for grading student essays based on a rubric?

To address this research question, we assessed the grading reliability of LLMs in five different domains of rubric: grammar, content, organization, style and mechanics, based on their mean values (*M*) and standard deviations (*SD*). We excluded ChatGPT FineTuned data from our analysis as all measurements had an *SD* of 0.00. Therefore, we compared ChatGPT Default and Bard data instead. We assumed that, in this context, a lower standard deviation indicates higher grading reliability, as it signifies a more consistent scattering of grading scores around the mean. The results are summarized in Table 1.

In the grammar domain, ChatGPT assigned a mean score of 4.3, which indicates a high grammatical quality for Essay 1, with an *SD* of 0.48, suggesting reasonable reliability in the grading. In contrast, Bard assigned a lower mean score ($M = 3.6$), although it showed a similar standard deviation ($SD = 0.52$), suggesting slight inconsistency in its grading. For Essay 2, ChatGPT marked a mean score of 2.9 with a *SD* of 0.32. In contrast, Bard's grading was perfect in its consistency ($SD = 0$) despite providing a higher mean score ($M = 3.0$). In Essay

TABLE 1 Domain-based mean scores assigned by LLMs.

	Essay 1				Essay 2				Essay 3			
	ChatGPT		Bard		ChatGPT		Bard		ChatGPT		Bard	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Grammar	4.30	0.48	3.60	0.52	2.90	**0.32**	3.00	**0.00**	1.70	0.48	2.70	0.48
Content	4.90	0.32	4.00	**0.00**	3.90	**0.32**	4.10	0.32	3.00	0.47	3.10	0.32
Organization	5.00	**0.00**	3.60	0.52	3.70	0.48	3.60	0.52	2.40	0.52	3.10	0.32
Style	4.10	0.32	3.50	0.53	3.00	0.47	3.10	0.32	2.20	0.42	3.00	**0.00**
Mechanics	4.30	0.48	3.50	0.53	3.60	0.52	3.50	0.53	2.00	**0.00**	3.00	**0.00**
Mean	4.52	0.32	3.64	0.42	3.42	0.42	3.46	0.34	2.26	0.38	2.98	0.22

Bold indicates the lowest SD.

3, both models demonstrated parity in their standard deviation ($SD = 0.48$), yet ChatGPT had a lower mean score ($M = 1.7$) than Bard ($M = 2.7$).

In the context domain, ChatGPT assigned a higher mean score ($M = 4.90$) than Bard ($M = 4.00$) for Essay 1. With respect to the standard deviation, ChatGPT had a higher standard deviation ($SD = 0.32$), indicating slight inconsistency, while Bard marked a perfect *SD* of 0.00. In Essay 2, ChatGPT again scored a lower mean value ($M = 3.90$) as compared to Bard ($M = 4.10$). Yet, both models showed the same degree of consistency with an *SD* of 0.32. For Essay 3, ChatGPT and Bard presented almost similar mean scores ($M = 3.00, 3.10$) and standard deviations ($SD = 0.47, 0.32$), showcasing equivalent reliability.

In the organization domain, Essay 1's perfect organization as per ChatGPT ($M = 5.00$, $SD = 0.00$) contrasted Bard's considerably lower mean score of 3.60 ($SD = 0.52$). In the case of Essay 2, ChatGPT offered a slightly higher mean score ($M = 3.70$) as compared to Bard (M = 3.60), although Bard maintained lower consistency ($SD = 0.52$) against ChatGPT ($SD = 0.48$). For Essay 3, Bard gained the lead with a higher mean score ($M = 3.10$) and a lower standard deviation ($SD = 0.32$), therefore coming off as more reliable than ChatGPT ($M = 2.40$, $SD = 0.52$).

In the style and expression domain, ChatGPT assessed the style of Essay 1 ($M = 4.10$, $SD = 0.32$) as being slightly higher than Bard ($M = 3.50$, $SD = 0.53$), hence appearing more reliable. For Essay 2, Bard delivered a marginally superior performance ($M = 3.10$, $SD = 0.32$) compared to ChatGPT's grading ($M = 3.00$, $SD = 0.47$). Lastly, in Essay 3, Bard was definitively more reliable in grading style and expression, given its perfect standard deviation ($SD = 0.00$) despite marginally higher mean scores ($M = 3.00$) than ChatGPT ($M = 2.20$, $SD = 0.42$).

For the mechanic domain, ChatGPT outperformed Bard with higher mean scores ($M = 4.30$, $M = 3.50$) and similar standard deviations ($SD = 0.48$, $SD = 0.53$), indicating higher reliability in Essay 1. For Essay 2, ChatGPT and Bard exhibited similar mean scores ($M = 3.60$, $M = 3.50$) and standard deviations ($SD = 0.52$, $SD = 0.53$), falling in line with reliability. For Essay 3, ChatGPT and Bard both exhibited absolute reliability with an *SD* of 0.00, with Bard providing a higher mean score ($M = 3.00$) than ChatGPT ($M = 2.00$).

In conclusion, our analysis reveals that LLMs, ChatGPT and Bard have shown good reliability in grading essays across various domains. While both produced notable results, distinctive strengths were observed. Bard exhibited absolute consistency in grading grammar for average essays, content for good and poor essays and organization and mechanics for poor essays. On the other hand, ChatGPT outperformed in evaluating style and grammar for good essays, exhibited more nuanced sensitivity for content across the performance range and showed outstanding reliability in grading organization for good essays.

RQ3: How accurately do LLMs grade student essays based on a given rubric compared to human raters?

After examining the reliability of LLMs in grading student essays based on a provided rubric, our goal was to determine the accuracy of these grades by comparing them with human grading. To achieve this, we first calculated the ICC scores of human raters to determine the level of agreement between them. This was done to investigate the consistency across the assessments of human raters. The ICC score for the human raters was 0.773 (0.405–0.962) $F(5,230) = 4.40$, $p < 0.001$, indicating good consistency among the human raters (Koo & Li, 2016).

In general, Figure 2 shows that both LLMs and human raters appear successful in that they reflected different scores for essays of varying qualities. This shows that both humans and LLMs could distinguish between different levels of essay quality. For Essay 1, the one in good quality, the mean score given by ChatGPT ($M = 4.40$) was higher than that of humans ($M = 3.89$). This might indicate that ChatGPT tends to be more lenient or perhaps perceives more merit in good essays than the human raters do. Bard's evaluation ($M = 3.64$), on the other hand, was slightly lower than the human mean. With respect to the average quality essay (Essay 2), both LLMs rated the essay higher than human raters did. This may suggest that both ChatGPT and Bard tend to be more generous or less critical when grading essays of medium quality. Bard, albeit slightly more generous, had a mean score ($M = 3.46$) closer to the human mean score than did the mean score from ChatGPT ($M = 3.60$), indicating its grading style was more aligned with the human raters for this essay quality. Interestingly, Essay 3, which was of a poorer quality, elicited different grading patterns from the two LLMs. While the mean score given by ChatGPT ($M = 2.00$) aligned perfectly with the mean score of

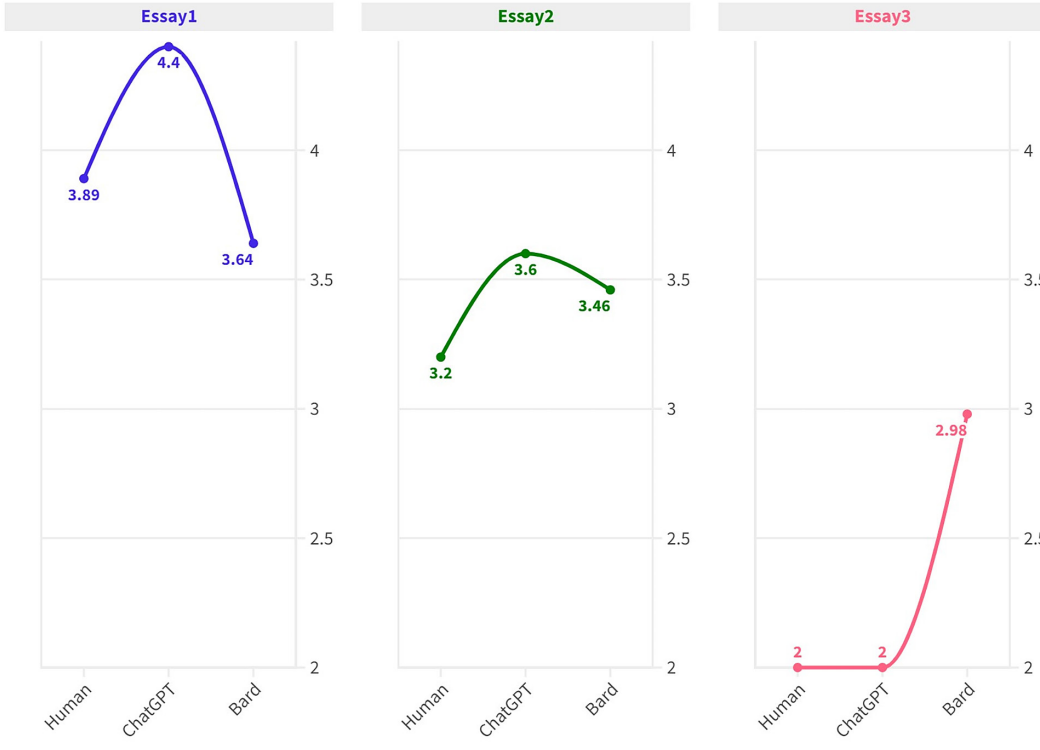

FIGURE 2 Mean scores of grades by humans and LLMs.

the human raters ($M = 2.00$), Bard's mean score was higher ($M = 2.98$), implying a problem in grading low-quality essays.

In sum, the LLMs exhibited varying degrees of alignment with the human raters throughout the essay quality spectrum. With ChatGPT showing a tendency to be lenient with high-quality and average-quality essays but aligning well with human raters' scores on poorer-quality essays and Bard demonstrating a general tendency to award higher scores across the essay quality range, we can infer that while LLMs show potential as grading assistants, their patterns do not uniformly align with human grading patterns across all essay quality levels.

RQ4: In what domains do the scores assigned by LLMs align closely with those given by human raters?

After assessing the overall grading accuracy of LLMs, we aimed to investigate the alignment between human raters and LLMs in grading essays across different domains. To achieve this, we calculated the average scores given by human raters and LLMs for three essays in each domain. We then used a radar chart (Figure 3) to illustrate the degree of similarity in grading scores between human raters and LLMs in each domain. As shown in the radar chart in Figure 3, there is a particular overlap in the grading scores of human raters and LLMs.

In the grammar domain, there appears to be a close alignment between the LLMs and human raters. The average grade score of human raters was 3.0. Similarly, ChatGPT reported an equal mean score ($M = 3.0$) which is an exact match with the human raters. On the other hand, Bard's mean score ($M = 3.1$) was slightly higher than human raters and ChatGPT. The close agreement in assessing grammar between human raters and LLMs suggests that the LLMs closely matched the performance of the human raters in assessing grammar structures. It can be argued that when it comes to grading grammar, LLMs provide valid grades based on a rubric.

Moving onto the content domain, a divergence between humans and LLMs can be observed. The mean score of human raters was 3.1, whereas ChatGPT and Bard assigned a higher mean score ($M = 3.7$). It is notable that the divergence between humans and LLMs is highest in the content domain when compared to other domains. Although the difference is not major ($MD = 0.6$), this divergence signifies a potential area of fine-tuning where LLMs'

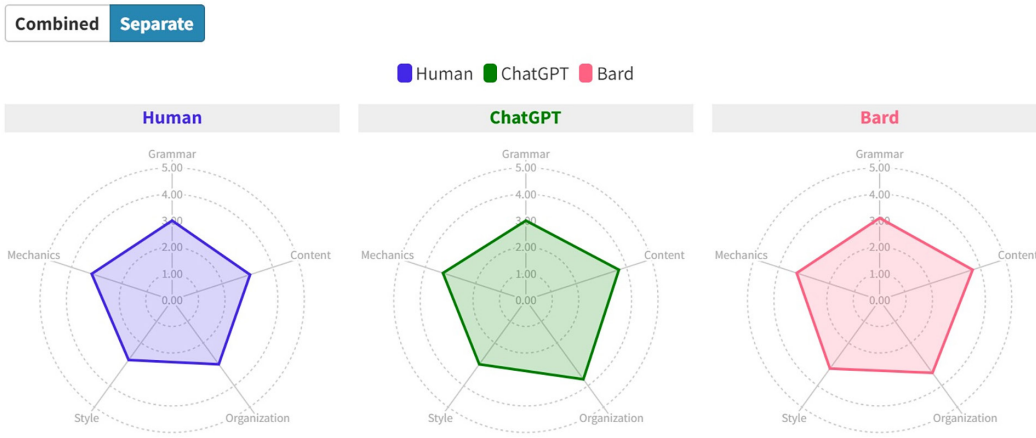

FIGURE 3 Radar chart showing the comparison of grade mean scores according to rubric domains.

grading could better align with human grading in the content domain. A similar observation can be made in the organization domain. Human raters reported a mean score of 3.0, while ChatGPT ($M = 3.7$) and Bard ($M = 3.4$) assigned higher mean scores. Again, organization domain deserves a closer investigation.

A very slight divergence can be observed in the style & expression domain. Human raters yielded a mean score of 2.8. On the other hand, LLMs offered slightly higher grades, with ChatGPT at $M = 3.0$ and Bard at $M = 3.2$. Although LLMs continued to provide higher scores, the slight difference (0.2 for ChatGPT and 0.4 for Bard) implies a closer alignment in grading the stylistic aspects of student essays. Similarly, another instance of close alignment can be observed in the mechanics domain, where the human raters reported a mean score of 3.2. Both ChatGPT and Bard recorded a marginally higher score ($M = 3.3$) in the mechanics domain, which indicates that human raters and LLMs are largely in close agreement with grading mechanics-related aspects of student essays, also suggesting that LLMs are proficient in detecting technical aspects of essay writing, such as punctuation, capitalization and spelling, in a manner consistent with human raters.

Overall, in terms of domain-based alignment between human raters and LLMs, while grammar and mechanics showcased closer alignments, divergencies were observed in other domains like content and organization. Style and expression served as an interesting middle ground, with LLMs demonstrating a slightly higher but converging grading compared to human raters.

DISCUSSION

The analysis of the first research question provides valuable insights into the reliability of LLMs in grading student essays based on a given rubric. The results clearly show that both ChatGPT models, FineTuned and Default, demonstrated a high level of reliability, as indicated by their high ICC scores of 0.972 and 0.947 respectively. Remarkably, the FineTuned model showed an impressive performance with $SD = 0.00$ in grading essays. This is largely due to the reduced temperature (0.2) of the model, which produces a more focused and deterministic output (Lo, 2023). When the temperature of the model is closer to 1.00, it produces more creative responses. Setting the temperature low makes the model more deterministic about the grading, and it records the same grades in each measurement, which dramatically increases the reliability of the grading. On the other hand, while Bard recorded a lower ICC score of 0.919 when compared to ChatGPT, it still demonstrates substantial reliability in grading essays. Similarly, in their study, where they aimed to explore the reliability of GPT-3 model in an automated essay scoring system, Mizumoto and Eguchi (2023) found that GPT-3 provides reliable scores to an extent that exhibits a close but not perfect alignment with human raters.

The results of the second research question contribute to a deeper understanding of the varying performance of LLMs in grading student essays across the domains defined in the rubric. At first glance, the findings show that LLMs not only offer reliability in overall grading but they are also capable of grading particular domains like grammar, content, organization, style and mechanics. It can be argued that although slight variations were observed in the domain-based grades, LLMs can provide consistent grades across different measurements. Particularly, Bard showed notable consistency in several domains, recording absolute reliability ($SD = 0.00$) in domains such as grammar for average essays, content for good and poor essays and organization and mechanics for poor essays. Similarly, the Default version of ChatGPT showed a good performance in the grading of style and grammar for good-quality essays and a perfectly consistent performance ($SD = 0.00$) in assessing organization for high-quality essays. Notably, ChatGPT consistently excelled at grading the content

domain in all essays. In alignment with the study of Fazal et al. (2013), which discusses the development of an automated essay grading (AEG) system focusing on spelling, our study found similar reliability in AI tools when grading mechanics. Also, similar to our findings, the study of Madala et al. (2018) found that the machine learning model they use effectively grades grammar and mechanics domain. However, it should be noted that these two studies use different language models from our study. These findings together establish that although the grades given by LLMs tend to have small differences across different essays, which can be seen as a disadvantage, these minor variations actually prove the grading mechanism of these tools to be reliable. The small differences in grading are similar to the subjectivity in human scoring and do not affect the overall reliability of the grades.

With the third research question, we aimed to explore how accurate LLMs are in grading student essays, comparing their outcomes with those of human raters. Evidently, the results associated with this research question revealed significant observations about the grade alignment between LLMs and human raters. First, it can be seen that both LLMs, ChatGPT and Bard, demonstrated the capacity to distinguish varying quality levels in essays, thereby validating their reliability findings. This finding aligns closely with those of Powers et al. (2002), who explored the validity and reliability of automated essay scoring systems, particularly remarking on their consistency across different essays and raters. However, some inconsistencies were spotted in the grades assigned by LLMs and human raters across different essay quality levels and rubric domains. As in the study of Mizumoto and Eguchi (2023), both ChatGPT and Bard have shown to be more generous in awarding grades to good- and average-quality essays compared to the human raters in our study. However, the two tools align closely with human grades when it comes to poor essays. One possible reason for this is that they are predominantly trained on high-quality examples. However, further investigation is required to confirm this hypothesis. Many other studies also found close alignment between the grades assigned by LLMs and human raters (Kumar & Boulanger, 2021; Ramalingam et al., 2018; Shehab et al., 2016; Shermis, 2014; Shin & Gierl, 2021; Suresh et al., 2023; Yamamoto et al., 2018). Although there is some deviation in the grading, the variance between the grades assigned by the LLMs and humans is low, indicating that these LLMs offer valid grading.

The fourth research question aimed to explore the domain-based alignment in grades between humans and LLMs and offers valuable perspectives on this. The findings reveal a considerable overlap in the grades assigned to the essays by human raters and LLMs, with the grammar and mechanics domains showing an almost perfect alignment. It can be inferred that these domains, performance indicators of which are fairly objective and rule bound, can be regarded as suitable for automated grading, suggesting that LLMs provide reliable and valid grades in rubric-based grading of grammar and mechanics domains (Fazal et al., 2013; Madala et al., 2018). On the other hand, a divergence was observed in content and organization domains where LLMs were assigned higher scores than human raters. This indicates that LLMs might be recognizing or valuing certain elements within those domains that human graders do not, or vice versa, revealing a potential opportunity for fine-tuning the grading models. An interesting middle ground is found in grading style and expression, where LLMs were expected to judge the language errors that interfere with meaning, weak and inappropriate vocabulary and unrelated and repetitive sentences. The LLMs demonstrated slightly higher but converging grading trends compared to human raters, indicating an evolving alignment in grading the more subjective aspects of essays, such as semantic understanding at the word and sentence level. It can be argued that this slight variation in grades should not be seen as a grading failure of LLMs, rather it can be seen as a result of different interpretations of quality in grading essays as can be seen in human raters.

CONCLUSION

With the purpose of offering pedagogical insights into the emerging use of LLMs in educational contexts, our study was designed to assess the reliability and validity of two leading LLMs—ChatGPT and Bard—in grading student essays based on a given rubric with an aim to shed light on the potential performance of LLMs in real-world classroom settings where accurate, efficient and quick responses are pivotal. Our methodology is carefully designed involving a diverse set of participants, including 15 human raters, reflecting a mix of expertise and perspectives, and two widely used LLMs, ChatGPT and Bard. Within this scope, participants (humans and LLMs) were assigned to grade three student essays that varied from good to poor quality, thereby ensuring a wide spectrum of performance for scrutiny. Our grading rubric included five distinct domains: grammar, content, organization, style & expression and mechanics. The resulting scores provided both an overall rating and detailed insights across distinct domains, thereby allowing a multidimensional analysis of the grading competency of both human raters and LLMs. This approach enabled us to explore the capabilities and potential limitations of LLMs in essay grading and the integration of such technologies in educational settings.

The analysis of data revealed significant findings regarding the validity and reliability of LLMs in grading student essays in accordance with a provided rubric. First, we observed a significant performance in terms of reliability from both ChatGPT and Bard, as can be seen in their high intraclass correlation coefficient (ICC) scores. Specifically, the FineTuned ChatGPT model demonstrated a very high level of reliability with an ICC score of 0.972 and an *SD* of 0.00 across the 10 measurements. The Default ChatGPT model also exhibited a high reliability with an ICC score of 0.947, despite a slightly higher standard deviation. Bard, despite having a lower ICC score of 0.919, still presented a substantial level of reliability. These results indicate good consistency in overall and domain-based grading by LLMs despite the varying quality of essays. Furthermore, a notable overlap was observed across various quality levels of essays when comparing the grades assigned by LLMs to those of human raters, suggesting a potential for scoring competence of the LLMs. However, this comparison also revealed some deviance in grading between human raters and LLMs. For example, both ChatGPT and Bard tended to be more generous in grading good- and average-quality essays compared to human raters, while in low-quality essays, their grades were closely aligned with those of human raters. In the most general sense, these findings suggest that LLMs demonstrate perfect reliability and acceptable (human-like) validity in grading student essays based on a given rubric.

Recommendations

The findings obtained in this study offer an understanding of the validity and reliability capacities of LLMs in grading student essays based on a given rubric and establish a ground for discussions on how LLMs can be integrated into essay grading in educational contexts. It is noteworthy that the overall high-reliability capacity exhibited by both ChatGPT and Bard is a strong indicator of their potential use as an effective grading assistant. The grading reliability of LLMs, particularly of the FineTuned ChatGPT model, offers potential advancements in grading integrity. It can clearly be articulated that instructors can increase the reliability of LLMs, particularly ChatGPT, by performing very basic fine-tuning such as lowering the temperature of the model and writing a more detailed prompt. In terms of domain-based grading, this study revealed both alignment and divergence between LLMs and human raters. The close alignment between LLMs and human raters in grading the grammar and mechanics domains suggests that LLMs are competent at evaluating aspects of writing that are rule bound and more objective. This ability could significantly reduce the grading load for educators, allowing

them to focus their attention on formative assessment practices and enriched feedback (Zhao et al., 2023). However, the divergence observed in content and organization domains signals the need for fine-tuning LLMs. These domains require a more nuanced understanding of the text and could be areas where LLMs could benefit from further fine-tuning to better align with human grading. The slight divergence observed in the style and expression domain offers additional implications. Although assessing the style and expression of writers leans more towards the subjective side, the closeness of LLMs grading to human grading could be seen as a promising domain to work on. It can be suggested that with proper fine-tuning and training, LLMs can perform human-like grading in style and expression domain.

On the other hand, incorporating LLMs into educational settings, particularly to evaluate writing performance, may present several practicality and integration challenges. First, the overreliance and irresponsible use of LLMs in writing assessments can be a significant issue. Instructors should critically evaluate the output produced by LLMs during the decision-making process. Overreliance on LLM outputs may distort the integrity of the assessment. Therefore, deploying LLMs in writing assessments requires robust frameworks or policies and instructor training before incorporating such tools into educational settings. Second, as this study suggests, the most effective and accurate use of LLMs in writing assessment is possible with the accurate fine-tuning of these models. Fine-tuning can be achieved in two ways: prompt engineering or temperature adjustment and algorithmic adjustments or data feeding. Although the latter is expected to produce better results, it requires expertise and may not be possible to incorporate by users who have no training in LLM training. Institutional support mechanisms can overcome this challenge.

Future studies may contribute to the findings of this study by using a larger volume of essays, including texts with more diverse topics and styles to further test the capability of these LLMs. Furthermore, exploring more LLMs (ie, Claude or Bing) for grading can help create a comprehensive understanding of the strengths and weaknesses of different LLMs. It may be worthwhile to consider adopting a qualitative approach along with the quantitative assessment of LLMs in essay grading, which could involve an in-depth analysis of essays graded by LLMs. Adopting a mixed-method approach could help in providing a more holistic understanding of machine grading efficiency. Lastly, future research can focus on investigating potential biases in LLMs essay grading performance in different genres (non-academic texts, creative writing, etc.) that contain cultural references, discourse features, figurative language, idioms, etc. Exploring the use of LLMs across a broad spectrum of educational contexts might propose valuable pedagogical implications.

Another recommendation for future research on AES using LLMs can address exploring their application across various academic disciplines and extend to different types of academic texts. Exploring LLMs' grading capacity in diverse subjects and their adaptability to different formats can broaden our understanding of automated assessment's potential. Moreover, understanding the impact of LLM-based grading on pedagogy and learning outcomes can offer valuable insights into its educational implications. Lastly, future research should rigorously address the ethical considerations associated with the use of LLMs in AES. Researchers can explore strategies to mitigate risks and ensure that the deployment of these technologies aligns with educational equity and fairness principles. Similarly, the lack of transparency in the assessment processes of LLMs is a particular concern. In cases where students challenge or question automated assessment decisions, the ability to explain how decisions were made, as human assessors do, may not be available in LLMs. This can jeopardize student rights and accountability in education. Therefore, the evaluation methodologies of AI systems need to be clearly defined and shared. In order to maximize the potential benefits and minimize the potential harms of using LLMs in education, it is crucial that the details of AI decision-making processes are transparently revealed and managed. Future studies should develop a deeper understanding of how LLMs make decisions

in assessment processes and provide recommendations on how to integrate human intervention and supervision into these processes.

ACKNOWLEDGEMENTS

In this paper, generative AI tools were used for literature search (Elicit), outline creation (ChatGPT) and language improvement (GrammarlyGo) purposes.

FUNDING INFORMATION

This study did not receive any funding.

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY STATEMENT

The materials and data used in this study can be found in Appendices.

ETHICS STATEMENT

In this study, we prioritized the protection of participant rights and adherence to research integrity. Participants, including three university students and fifteen raters, were fully informed about the study's purpose, data collection processes, and the anonymization of their data. Informed consent was obtained both verbally and in writing. We also ensured compliance with OpenAI's policy for using ChatGPT in research.

ORCID

Fatih Yavuz https://orcid.org/0000-0003-2645-2710
Özgür Çelik https://orcid.org/0000-0002-0300-9073
Gamze Yavaş Çelik https://orcid.org/0000-0003-1571-9686

ENDNOTE

[1] During the period our research was conducted, the tool was officially known as Google Bard. It has been rebranded to Google Gemini as of February 2024. This study refers to the tool as Google Bard throughout, as our data collection and analysis were completed prior to this name change.

REFERENCES

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Fazal, A., Hussain, F. K., & Dillon, T. S. (2013). An innovative approach for automatically grading spelling in essays using rubric-based scoring. *Journal of Computer and System Sciences*, *79*(7), 1040–1056. https://doi.org/10.1016/j.jcss.2013.01.021

Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of EFL students' classroom- based writing assessments in Turkish universities: An investigation of problems and solutions* [Unpublished Doctoral Dissertation]. Atatürk University.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Ifenthaler, D. (2023). Automated essay scoring systems. In O. Zawacki-Richter & I. Jung (Eds.), *Handbook of open, distance and digital education* (pp. 1057–1071). Springer Nature. https://doi.org/10.1007/978-981-19-2080-6_59

Khademi, A. (2023). Can ChatGPT and bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching*, *6*(1), 75–80. https://doi.org/10.37074/jalt.2023.6.1.28

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

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, *5*, 572367. https://doi.org/10.3389/feduc.2020.572367
Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, *31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5
Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720
Madala, D. S. V., Gangal, A., Krishna, S., Goyal, A., & Sureka, A. (2018). *An empirical analysis of machine learning models for automated essay grading*. https://doi.org/10.7287/peerj.preprints.3518v1
Meyer, J., Jansen, T., Fleckenstein, J., Keller, S., & Köller, O. (2023). Machine learning im Bildungskontext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch. *Zeitschrift Für Pädagogische Psychologie*, *37*(3), 203–214. https://doi.org/10.1024/1010-0652/a000296
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
OpenAI. (2022). *Sharing & publication policy*. https://openai.com/policies/sharing-publication-policy
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
Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018). Automated essay grading using machine learning algorithm. *Journal of Physics: Conference Series*, *1000*, 012030. https://doi.org/10.1088/1742-6596/1000/1/012030
Şahan, Ö. (2019). The impact of rater experience and essay quality on the variability of EFL writing scores. In S. Papageorgiou & K. M. Bailey (Eds.), *Global perspectives on language assessment* (1st ed., pp. 32–46). Routledge. https://doi.org/10.4324/9780429437922-3
Shehab, A., Elhoseny, M., & Hassanien, A. E. (2016). A hybrid scheme for automated essay grading based on LVQ and NLP techniques. *12th International Computer Engineering Conference (ICENCO)*, 65–70. https://doi.org/10.1109/ICENCO.2016.7856447
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
Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247–272. https://doi.org/10.1177/0265532220937830
Suresh, V., Agasthiya, R., Ajay, J., Gold, A. A., & Chandru, D. (2023). AI based automated essay grading system using NLP. *7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 547–552. https://doi.org/10.1109/ICICCS56967.2023.10142822
Taskiran, A., & Goksel, N. (2022). Automated feedback and teacher feedback: Writing achievement in learning English as a foreign language at a distance. *Turkish Online Journal of Distance Education*, *23*(2), 120–139. https://doi.org/10.17718/tojde.1096260
Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*(2), 459–484. https://doi.org/10.1007/s41237-021-00142-y
Wu, Y., Henriksson, A., Nouri, J., Duneld, M., & Li, X. (2023). Beyond benchmarks: Spotting key topical sentences while improving automated essay scoring performance with topic-aware BERT. *Electronics*, *12*(1), 150. https://doi.org/10.3390/electronics12010150
Yamamoto, M., Umemura, N., & Kawano, H. (2018). Automated essay scoring system based on rubric. In R. Lee (Ed.), *Applied computing & information technology* (pp. 177–190). Springer International Publishing. https://doi.org/10.1007/978-3-319-64051-8_11
Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies*, *28*(6), 7031–7063. https://doi.org/10.1007/s10639-022-11473-y

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *00*, 1–17. https://doi.org/10.1111/bjet.13494