https://doi.org/10.37244/ela.2023.18.2.11
English Language Assessment, Vol. 18(2), December 2023

Applicability of Fine-Tuned ChatGPT to Automated Essay Scoring*

Yongkook Won**

Won, Yongkook. (2023). Applicability of fine-tuned ChatGPT to automated essay scoring. *English Language Assessment, 18*(2), 11-35.

ChatGPT, released in 2022, has garnered attention due to its adaptability through prompt engineering, enabling users to guide its responses. It is important to note that the extent to which users can modify ChatGPT remains limited, as its core embeddings stay unaltered through prompt engineering. Thus, this study aims to evaluate the effectiveness of ChatGPT and its fine-tuned model in essay evaluation compared to human raters. A total of 904 essays from the YELC 2011, on the subject of physical punishment, were selected for this study. Among these, 723 essays were used for fine-tuning ChatGPT, and the remaining 181 were reserved for testing the language model. Additionally, an extra set of 200 essays on different topics, such as driving and medical issues, was included to evaluate the language model's performance across various themes. Inter-rater reliability indices, including measures like correlation, agreement, Cohen's kappa, and Krippendorff's alpha, along with many-facet Rasch measurement analysis, collectively indicated that the current version of ChatGPT (gpt-3.5-turbo-0613) is not yet poised to fully supplant human raters in essay scoring. Nevertheless, through the fine-tuning process, the model demonstrated a significant level of agreement with human raters and exhibited a marked degree of consistency.

Key words: ChatGPT, automated essay scoring (AES), artificial intelligence (AI), fine-tuning, large language model

I. INTRODUCTION

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5B5A17050971).

** Visiting Researcher, Center for Educational Research, Seoul National University

Given recent strides in artificial intelligence (AI), substantial endeavors have been invested in developing AI-driven educational programs to provide adaptive learning experiences (Kabudi et al., 2021). In the realm of language education, there is an escalating demand for personalized learning based on performance assessments, leading to the introduction of various services aimed at alleviating the grading burden on educators and enabling personalized adaptive learning (Kabudi et al., 2021; Page, 1966; Shermis & Burstein, 2013). In ESL writing education, notable advancements have been made. For example, the Educational Testing Service (ETS) introduced an automated essay evaluation (AES) system called E-rater®, which employs natural language processing (NLP) to assess essays for the Test of English as a Foreign Language (TOEFL) (Burstein, 2012). Pearson offers the Intelligent Essay Assessor (IEA) system, WriteToLearn™, utilizing Latent Semantic Analysis (LSA) to gauge writing quality and furnish immediate feedback. Additionally, Cambridge English provides the Write&Improve service, using NLP technology to offer instant feedback on English writing. However, from the perspective of school teachers, current AES services have substantial limitations. Despite the fact that the underlying algorithms, such as Generative Pre-trained Transformer (GPT) (Radford et al., 2018), Bidirectional Transformers for Language Understanding (BERT) (Devlin et al., 2018), are publicly known, the scoring processes of the current commercial AES systems remain concealed in black boxes and their functions are fixed, making it impossible for teachers or researchers to customize the algorithms to suit their specific educational needs.

Since its initial release in 2022, ChatGPT (https://chat.openai.com/) has garnered the attention of numerous educators and academic researchers, with specific relevance to its application in classroom settings (Hwang et al., 2023; Kim, 2023). This is primarily attributed to ChatGPT's unique feature, which enables users to adapt its functionality to their specific requirements through a process known as prompt engineering. Prompt engineering entails the design and formulation of prompts or input queries in a manner that guides a language model toward generating desired or specific outputs. Mizumoto and Eguchi (2023) investigated the potential of ChatGPT for AES in the context of foreign language education. Their findings indicated that GPT-based AES exhibited a reasonable level of accuracy and reliability, thereby complementing human assessment. Additionally, Yancey et al. (2023) delved into the evaluation of GPT-3.5 and GPT-4 in assessing short essays created by English as a Second Language (ESL) learners. Their research concluded that GPT-4 was capable of achieving results closely aligned with modern AES techniques. Despite the positive findings from previous studies, it remains noteworthy that users' capacity to modify Large Language Models (LLMs) remains relatively constrained, as the embeddings within GPT do not readily adapt through prompt engineering. In light of this, the current study is designed to assess the efficacy of fine-tuned LLMs in the evaluation of English as a Foreign Language (EFL) students' essays, in comparison with human raters. Specifically, the research examines ChatGPT's ability to assess persuasive writing by Korean college students and assesses how

fine-tuning enhances the quality of assessment. The research questions (RQ) guiding this study are as follows:

RQ1. To what extent can ChatGPT be relied upon for essay scoring in comparison to traditional human scoring methods?
RQ2. To what extent does the fine-tuning of ChatGPT enhance its accuracy and reliability in automated essay scoring?
RQ3. How reliably can a fine-tuned ChatGPT evaluate essays on topics that were not part of the fine-tuning process?

II. LITERATURE REVIEW

1. Automated Essay Scoring

Automated essay scoring (AES), also referred to as automated writing evaluation (AWE) or automated essay evaluation (AEE), is a technology-driven approach employed for the assessment and scoring of written essays (Shermis et al., 2013). These systems are the product of interdisciplinary research endeavors and technological progress encompassing fields such as natural language processing, computer sciences, cognitive psychology, linguistics, and writing research (Hockly, 2018; Shermis et al., 2013). This approach confers substantial advantages by mitigating the constraints and potential inaccuracies associated with human scoring, notably those stemming from time limitations and fatigues (Weigle, 2013). The concept of AES has maintained enduring relevance within the academic and professional spheres for over half a century, underscoring its persistent significance in the domain (Burstein, 2012).

The inception of AES research has its roots in the 1960s, particularly exemplified by the Project Essay Grade (PEG) system (Page, 1966). The early versions of AES systems, including PEG, have been subjected to criticism for their primary emphasis on superficial structural elements while neglecting content-related features (Mizumoto & Eguchi, 2023). The 1990s saw significant advancements in AES technology, thanks to the evolution of computer capabilities and NLP (Hussein et al., 2019). ETS introduced the e-rater® system in 1999, representing a new era of AES (Burstein et al., 2013). The early versions of e-rater® utilized statistical and rule-based methods to analyze sentence structure, word structure, and meaning (Burstein et al., 2013; Deane, 2013). Pearson has also introduced the IEA system, which relies on LSA to assess the quality of written compositions and offers a feedback service named WriteToLearn™ (Foltz et al., 2013). Furthermore, services like Grammarly

and Turnitin, specializing in the provision of automated English correction services, employ NLP technology to deliver instantaneous feedback on English writing.

Even with the benefits of using AES, AES also confronts notable challenges, such as disparities that arise from the misalignment between the criteria of AES and the heterogeneous nature of student's writing, apprehensions pertaining to fairness and bias in sentiment analysis, and the imperative for flexibility across a spectrum of academic disciplines (Latif & Zhai, 2023). In addition, it is noteworthy that AES may exhibit limitations in fully encapsulating the multifaceted nature of writing, primarily due to their inability to encompass the collaborative and interactive aspects of the writing construct and their incapacity to accommodate intricate features of writing (Huawei & Aryadoust, 2022). The constraints inherent in AES have been partially mitigated through the introduction of embeddings, which transform natural languages into purposeful vectors, demonstrated to uphold semantic information (Andersen et al., 2023; Delgado et al., 2020; Mikolov et al., 2013). Given that LLMs are built upon these embeddings, it is anticipated that advanced AES systems will overcome the limitations of conventional AES systems.

2. Large Language Models

In recent years, LLMs, exemplified by GPT-3 (Floridi & Chiriatti, 2020), have achieved notable progress in the domain of NLP. LLMs represent a category of AI models engineered for the comprehension and generation of human language. These models are constructed employing *transformer* architectures, incorporating the foundational *attention* mechanism (Vaswani et al., 2017). Their training is constructed on extensive datasets comprising textual content sourced from the internet, literary works, and diverse textual origins. The salient attribute characterizing LLMs is their expansive scale and intricacy, typified by the encompassment of millions or billions of parameters (e.g., 175 billion parameters for GPT-3 and 1.7 trillion parameters for GPT-4). This substantial parameterization empowers these models with the capacity to comprehend and generate natural language text with remarkable fluency and cohesiveness.

A notable advancement within the domain of LLMs pertains to the incorporation of pre-training, which encompasses the initial phase of model development (Radford et al., 2018). During pre-training, the LLMs undergo training on an extensive and diverse dataset, which includes text sourced from a variety of origins. This process entails the model learning to anticipate the subsequent word in a sentence by considering the context provided by the preceding words (Vaswani et al., 2017). Consequently, pre-training serves as the foundational stage through which the LLMs acquire an understanding of fundamental linguistic constructs, encompassing grammar, syntax, semantics, and general world knowledge. Following the pre-training phase, LLMs undergo further refinement through fine-tuning, a process wherein the model is adapted to address specific NLP tasks or datasets.

Fine-tuning facilitates the model's capacity to excel in particular applications, such as text generation, translation, sentiment analysis, and an array of related tasks (Rothman, 2022).

Recent advancements within the realm of NLP encompass the emergence of ChatGPT, an intricately trained model, drawing from an extensive dataset extracted from an expansive web corpus (OpenAI, 2022). ChatGPT has exhibited pioneering excellence across an array of natural language tasks, including translation, question answering, essay composition, and program generation (Kasneci et al., 2023). Additionally, substantial research efforts have been directed towards the refinement of these models through fine-tuning on comparatively smaller datasets and the strategic utilization of transfer learning methods to address novel challenges (Kasneci et al., 2023; Rothman, 2022). In an educational context, these approaches hold the promise of substantially diminishing scoring discrepancies between human and artificial raters, guaranteeing equitable and uniform assessment of student answers (Latif & Zhai, 2023).

3. Previous Studies on ChatGPT for Automated Essay Scoring

Since the launch of ChatGPT in November 2022, there has been ongoing discussion about the potential utilization of ChatGPT for AES (Christodoulou, 2023). However, there is a lack of comprehensive research on the application of ChatGPT for AES. Yancey et al. (2023) investigated the effectiveness of GPT-3.5 and GPT-4 in assessing short essays written by ESL learners. They reported that, with calibration examples, GPT-4 could achieve performance levels close to contemporary AES methods. However, the alignment with human ratings was subject to variation based on the examinee's first language (L1). Yeşilyurt and Sezgin (2023) presented preliminary evidence suggesting ChatGPT's potential as an AI-powered tool for automated writing evaluation. The intraclass correlation coefficient analysis indicated moderate to high agreement between the scores assigned by a human rater and ChatGPT across various rubric criteria. This implies that ChatGPT can provide reliable and valid ratings comparable to human raters. Altamimi (2023) examined the efficacy and reliability of ChatGPT, particularly GPT-4, compared to conventional grading approaches. The study asserted that GPT-4 demonstrated notable effectiveness in automating the grading of the ASAP essay dataset (Hamner et al., 2012). However, the study has limitations, including a small sample size of 50 essays and a failure to validate the model's consistency. On the other hand, Khademi (2023) investigated the reliability of ChatGPT and Google Bard in assessing the complexity of writing prompts, a task analogous to essay scoring. The study compared the performance of ChatGPT and Bard to experienced human raters. The findings revealed that both OpenAI ChatGPT and Google Bard exhibited significantly low inter-rater reliability when compared to the benchmark set by human ratings.

In some cases, ChatGPT was employed indirectly via modifications rather than being used directly. Mizumoto and Eguchi (2023) explored the capability of ChatGPT in the application

of AES within the domain of foreign language education. The authors used ChatGPT (text-davinci-003 model) as their baseline GPT for regression models and determined that the integration of linguistic measures could enhance the precision of AES. They found that the benchmark levels of essays were best predicted when GPT scores were combined with a range of linguistic features, including lexical, syntactic, and cohesion features. Latif and Zhai (2023) investigated the potential of fine-tuned ChatGPT (GPT-3.5) in automating the scoring of student-written responses in science education assessment tasks, not in AES. The results showed that GPT-3.5, after fine-tuning, achieved a significant increase in automatic scoring accuracy compared to BERT across various tasks. The study affirmed the effectiveness of fine-tuned GPT-3.5 for accurately scoring student responses in science education.

In AES, ChatGPT is often used without modification, and so far, the only case where fine-tuning has been used to score writing is for scoring limited production answers in science education. Therefore, to assess the performance of ChatGPT in AES, it is necessary to evaluate the performance of fine-tuned ChatGPT in addition to ChatGPT in its original form.

III. RESEARCH METHOD

1. Dataset

The essays used in this study were extracted from the Yonsei English Learner Corpus (YELC) 2011. The entire YELC 2011 is a compilation of essays from Korean English language learners, gathered by collecting essays written by a total of 3,286 prospective students (1,958 males and 1,328 females) from Yonsei University in 2011. They participated in the English writing test between January and February 2011, producing essays of 100 characters or fewer (Part 1) and 300 characters or fewer (Part 2) (Rhee & Jung, 2014). The essay scores, evaluated by human raters, were provided with the corpus. Essay grading was reported to be based on the Common European Framework of Reference for Languages (CEFR), featuring six grades subdivided into nine levels (Rhee & Jung, 2014). However, a detailed grading rubric was not provided. For this study, only Part 2 essays were used in order to reduce the influence of topic variation on training the LLM. A manual check of all 3,286 essays confirmed that each essay was written on at least one of the six topics listed in Table 1. A total of 904 essays focused on the topic of physical punishment were selected for the current study, with 723 allocated for training and 181 for testing the LLM. Essays on physical punishment were chosen for LLM training and testing due to their abundance in the available dataset. Additionally, 200 essays from two additional topics, namely driving and medical issues, were included to assess the model's performance across diverse topics. The use of a single topic for LLM training aimed to ensure that the trained LLM could effectively evaluate essays from various topics.

TABLE 1

Topic Distribution of 3,286 Essays in the YELC Corpus

Topics	Physical Punishment	Driving	Medical Issues	Internet	Smoking	Military Service
Number of Essays	904	458	485	504	486	449

2. Fine-Tuning ChatGPT

The ChatGPT model employed in this study is the gpt-3.5-turbo-0613, a variant of OpenAI's GPT-3, released on November 30, 2022 (OpenAI, 2022). The gpt-3.5-turbo-0613 (hereafter referred to as "original ChatGPT") was used for fine-tuning because it was the most recent model available at the time of the fine-tuning process. GPT-3 is a robust language model with a staggering 175 billion parameters, enabling it to handle a wide range of language-related tasks, such as translation, summarization, question answering, and text generation. Consequently, ChatGPT can be utilized for essay grading without the need for any additional model adjustments. Nevertheless, in this study, supplementary fine-tuning was conducted to explore whether it could enhance the grading accuracy of ChatGPT. A dataset comprising 723 essays, encompassing 1,748,883 tokens (including both the prompt featured in the subsequent chapter and the students' essays), was used for this fine-tuning process, which was performed over three epochs. For the fine-tuning process, the data set was prepared with a JSON file format as in Appendix 1. This dataset was used for fine-tuning the model and scoring the essays using the original ChatGPT model. The scoring rubric, adapted from the CEFR written communication band descriptors, was included in the prompt. This choice was made because the YELC was originally assessed according to the CEFR standard. The prompt encompasses various descriptors from (a) overall written production, (b) vocabulary, (c) grammatical accuracy, (d) creative writing, and (e) reports and essay sections in the CEFR descriptors. An example of the actual prompt utilized in the study can be found in Appendix 2. The fine-tuning process was executed automatically via the original ChatGPT chat completions API provided by OpenAI (https://platform.openai.com/docs/guides/fine-tuning). The Python code and prompt used for fine-tuning are available on the author's GitHub repository (https://github.com/linguistry/Fine-Tuned_ChatGPT_Essay_Scoring).

3. Procedures

This study followed the procedures outlined below. First, a dataset comprising 904 essays on the topic of physical punishment was partitioned into two sets, one for training or fine-tuning of the model and one for testing the model, in an 80:20 ratio. The training set consisted of 723 essays, while the testing set comprised 181 essays. This division allowed for the training of the language model using the training data and subsequent evaluation of its

performance on the test data, ensuring a robust assessment of the model's generalization. Secondly, a set of 181 essays focusing on the subject of physical punishment was scored using both the original ChatGPT provided by OpenAI and the fine-tuned model. To obtain the scoring results for these essays, both models processed prompts and students' essays in accordance with the data format specified in Appendix 1. It is important to note that manual input of score information was omitted, in contrast to the fine-tuning process, as the models predicted this data. In the subsequent phase, an examination was conducted to measure the inter-rater reliability between human raters and the original ChatGPT model within the context of the 181 essays on the subject of physical punishment. Similarly, inter-rater reliability was also explored between human raters and the fine-tuned model, with the aim of assessing potential enhancements in essay scoring quality attributed to the fine-tuned model. Additionally, the rating consistency of the fine-tuned model was validated through multiple assessments of the physical punishment essays using the fine-tuned model on two separate occasions. Lastly, a comparison was made between the 100 scores assigned by human raters for essays related to driving and another 100 scores for essays discussing medical issues, and the scores assigned by the fine-tuned model. The purpose of this comparative analysis is to determine whether the fine-tuned model can be extended to grading essays on topics not included in the fine-tuning process.

4. Statistical Analysis

In the current study, several inter-rater reliability measurements were used to explore the applicability of ChatGPT in automated essay scoring, as a potential replacement for human raters. These measurements included Spearman's rank-order correlation, percentage agreement, Cohen's kappa (Cohen, 1960), Krippendorff's alpha (Krippendorff, 1970), and many-facet Rasch measurement (MFRM) analysis (Linacre, 1989). Spearman's correlation was employed to examine the correlation between the two sets of ratings, which could be among human ratings, ChatGPT ratings, or fine-tuned model ratings. Spearman's correlation, rather than Pearson's correlation, was chosen because it provides a valuable metric for gauging the level of agreement between two sets of ordinal or ranked data. Percentage agreement was also used to quantify the extent to which ChatGPT and human raters concurred in their assessments. Cohen's kappa is a statistical measure designed to quantify the level of agreement between two raters, surpassing what might occur purely by chance. It takes into account the potential for agreement due to random chance, offering a more resilient indicator of consensus. The Kappa coefficient spans from -1 to 1, where a value of 1 signifies perfect agreement, 0 denotes agreement at chance levels, and -1 indicates perfect disagreement. Krippendorff's alpha is an alternative statistical metric used to assess inter-rater reliability and is recognized for its higher robustness compared to Cohen's kappa. The Krippendorff's alpha reliability calculation is strategically employed in projects with a

substantial number of raters, diverse measurement scales, and instances of missing data, making it a highly versatile tool. A Krippendorff's alpha value of 1 indicates perfect reliability, while a value of 0 indicates the absence of reliability. MFRM analysis is an extension of the Rasch analysis method, designed to handle multiple factors (e.g., examinees, raters, and evaluation criteria) simultaneously. MFRM unites these factors onto a common scale for comparative analysis. It offers several advantages, including the capacity to provide comprehensive insights into rater severity, rater self-consistency, and rater bias related to various facets (Eckes, 2015; Pratitis & Purwono, 2018). Spearman's correlation, percentage agreement, Cohen's kappa, and Krippendorff's alpha were measured with the R package *irr* (Gamer et al., 2012), while MFRM was analyzed using the computer program FACETS version 3.80. FACETS is a statistical software package specifically designed for conducting MFRM. FACETS can analyze data with multiple facets or sources of variation, such as raters or items. It enables researchers to model and assess the influence of these facets on the measurement.

IV. RESULTS

1. Outline of the Findings

Table 2 displays the data analysis outcomes that address the research questions, drawing upon the quantitative analysis findings presented in the preceding section. This section commences by examining the correlation between human-assigned scores and the original ChatGPT scores (RQ1). Subsequently, an examination is conducted to assess the comparative performance of the fine-tuned ChatGPT in relation to human raters (RQ2) and its efficacy in grading essays on topics not incorporated in the fine-tuning process (RQ3). MFRM techniques were employed to scrutinize the grading patterns across four rounds of evaluation: human grading, original ChatGPT grading, and two rounds of grading using fine-tuned ChatGPT. In summary, the findings revealed a significant divergence between the original ChatGPT scores and the remaining scores, whereas the fine-tuned ChatGPT scores and the human scores exhibited a closer alignment.

TABLE 2

Outline of the Data Analysis Results

Sections in the Results	Findings	Research Questions
Comparison of Human Rating and AES Rating	▪ Moderate to very low agreement between human raters and ChatGPT	▪ RQ1. Original ChatGPT vs. Human raters

Effect of Fine-tuning	▪ Moderate to high agreement between human raters and fine-tuned ChatGPT ▪ Hight consistency of scoring with fine-tuned ChatGPT	▪ RQ2. Fine-tuned ChatGPT vs. Human raters
Transfer Learning of Fine-tuned ChatGPT	▪ Moderate correlation between scoring of different topics with the fine-tuned ChatGPT	▪ RQ3. Fine-tuned ChatGPT for topics that were not part of the fine-tuning process
Analysis of Rater Variations using MFRM	▪ Original ChatGPT was significantly severer than human raters and fine-tuned ChatGPT ▪ The scoring of fine-tuned ChatGPT was too predictable	▪ RQ1 & RQ2

Table 3 displays the descriptive statistics of essay scores on the topic of physical punishment assigned by human raters, the original ChatGPT, and the first and second fine-tuned ChatGPT models. Human raters assigned an average score of 4.27 to essays on physical punishment, with a moderate level of variation (*SD* = 1.27) and a wide range of scores, spanning from 1.00 to 8.00. The original ChatGPT model, without fine-tuning, provided lower average scores (3.48) with a standard deviation of 1.10. The scores ranged from 1.00 to 6.00, indicating slightly less variability compared to human raters. In contrast, the first scoring with the fine-tuned ChatGPT model exhibited a higher mean score (4.28) and lower variability (*SD* = 0.85), with scores ranging from 2.00 to 6.00. The second scoring with the fine-tuned ChatGPT model had an average score of 4.34, similar to the first scoring, and a standard deviation of 0.90. Its scores ranged from 2.00 to 7.00. Both scorings with the fine-tuned ChatGPT model showcased a narrower spectrum of variability.

TABLE 3

Descriptive Statistics of Grading of Essays on the Topic of Physical Punishment (*n* = 181)

Raters	Mean	Median	SD	IQR	Minimum	Maximum
Human Rater	4.27	4.00	1.27	2.00	1.00	8.00
Original ChatGPT	3.48	3.00	1.10	1.00	1.00	6.00
Fine-tuned ChatGPT 1st	4.28	4.00	0.85	1.00	2.00	6.00
Fine-tuned ChatGPT 2nd	4.34	4.00	0.90	1.00	2.00	7.00

Note. IQR: Interquartile ranges, SD: Standard deviation

2. Comparison of Human Rating and AES Rating

Table 4 shows different types of inter-rater reliability indices comparing human raters' assessments to those by the original ChatGPT. Table 4 presents a comprehensive set of statistical measures for assessing agreement and correlation within the dataset. Notably, Spearman's correlation coefficient revealed a statistically significant moderate positive

relationship between variables, with a value of .27. The percentage agreement varies with and without a tolerance of 1 point, showing agreements of 27% and 64%, respectively, while a high agreement of 88% is observed with a tolerance of 2 points. In contrast, Cohen's kappa indicates very low agreement (.05), considering chance agreement, and Krippendorff's alpha stands at .13, suggesting relatively low agreement between human raters and the original ChatGPT. Overall, the results depict diverse degrees of agreement and correlation in the dataset, ranging from moderate to very low. This analysis suggests that it is not feasible to conclude that the original ChatGPT can replace human raters for essay scoring.

TABLE 4

Inter-rater Reliability Between Human Raters and the Original ChatGPT (n = 181)

Spearman's ρ	Percentage Agreement	Percentage Agreement (Tolerance = 1)	Percentage Agreement (Tolerance = 2)	Cohen's Kappa	Krippendorff's Alpha
.27 ($p < .001$)	.27	.64	.88	.05 ($p = .13$)	.13

3. Effect of Fine-Tuning

Table 5 presents the inter-rater reliability between human raters and two iterations of scoring with the fine-tuned ChatGPT model based on 181 essays. Spearman's correlation coefficients were notably high in both phases, registering values of .67 and .68 (both $p < .001$) for the first and second scoring with the fine-tuned model, respectively. These outcomes reinforced the robust agreement witnessed between the model's rankings and the ordinal preferences of the human raters, emphasizing the model's adeptness in replicating the ordinal judgments made by the human raters. The assessment of percentage agreement without tolerance revealed a 45% concurrence in the first scoring with the fine-tuned model, while a slightly lower percentage of 43% was observed in the second scoring. This metric illuminated the extent to which the model and human raters consistently assigned identical scores to the evaluated assessments. When a tolerance of 1 point was allowed, the agreement substantially improved to 91% for both the first and second scoring with the fine-tuned model. Notably, with a tolerance of 2 points, the agreement remained at 100% in both scorings with the fine-tuned model, underscoring the model and human raters' shared consensus, even when relatively larger score disparities were deemed acceptable, further substantiating the model's capability to capture the essence of human rater assessments. Considering that the rating scale ranged from 1 to 9, the agreement between the scorings with the fine-tuned ChatGPT model and human raters could have improved if the rating scale was smaller.

TABLE 5

Inter-rater Reliability Between Human Raters and the Fine-Tuned ChatGPT (n = 181)

Fine-tuned ChatGPT	Spearman's ρ	Percentage Agreement	Percentage Agreement (Tolerance = 1)	Percentage Agreement (Tolerance = 2)	Cohen's Kappa	Krippendorff's Alpha
1st scoring	.67 (p < .001)	.45	.91	1.00	.24 (p < .001)	.64
2nd scoring	.68 (p < .001)	.43	.91	1.00	.22 (p < .001)	.65

With respect to Cohen's kappa values, human raters' scoring with the first and second scoring with the fine-tuned model yielded coefficients of .24 and .22, respectively. These findings indicated a level of agreement beyond chance, although the reliability appeared moderate, which could be attributed to the inherent complexity of the assessment task. Concerning Krippendorff's alpha, values of .64 and .65 were obtained between the human scoring and the first and second scoring with the fine-tuned models, respectively. These values denoted a degree of agreement surpassing chance expectations when assessing the reliability between human raters and the fine-tuned ChatGPT model. However, they also implied the potential for enhancing the model's consistency compared to human raters' evaluations. It is worth noting that Krippendorff's alpha values nearing 1.00 indicate a heightened degree of reliability. Hence, the interpretation of these metrics requires a nuanced consideration of contextual factors and agreement standards. In summary, both the first and second scoring with the fine-tuned model demonstrated substantial alignment with human raters, particularly in terms of Spearman's correlation and percentage agreement, particularly when minor score discrepancies were accounted for. Nonetheless, there are discernible opportunities for improving the model's consistency, as evidenced by the Krippendorff's alpha values.

TABLE 6

Inter-rater Reliability Between the First Rating and the Second Rating by the Fine-Tuned ChatGPT (n = 181)

Spearman's ρ	Percentage Agreement	Percentage Agreement (Tolerance = 1)	Percentage Agreement (Tolerance = 2)	Cohen's Kappa	Krippendorff's Alpha
.92 (p < .001)	.90	.99	1.00	.84 (p < .001)	.92

Table 6 displays inter-rater reliability between the first and second ratings conducted by the fine-tuned ChatGPT on a dataset of 181 essays on physical punishment. The results indicated very strong agreement between these two sets of ratings. Specifically, Spearman's correlation coefficient also demonstrated a robust positive rank correlation (ρ = .92, p < .001). The percentage agreement between the first and second ratings was high, with a raw

agreement rate of 90%, and even more agreement is observed when allowing for a tolerance of 1 point (99%) or 2 points (100%). Furthermore, Cohen's kappa generated a high value of .84 ($p < .001$), indicating substantial agreement beyond chance, and Krippendorff's alpha produced a value of .92, signifying the degree of agreement beyond what would be expected by chance. These results collectively suggest a strong consistency between the first and second rounds of ratings conducted by the fine-tuned ChatGPT, highlighting the model's reliability in assessing the given dataset.

4. Transfer Learning of Fine-Tuned ChatGPT

Table 7 presents the descriptive statistics for the grading of essays on two different topics that were not used for the fine-tuning of the ChatGPT model: driving and medical issues, with a sample size of 100 essays for each topic. As depicted in Table 7, for essays on the topic of driving, the human rater assigned an average score of 4.38, with a median score of 4.00, and exhibited considerable variation as indicated by a standard deviation of 1.84. The interquartile range (IQR) was 3.00, with scores ranging from a minimum of 1.00 to a maximum of 8.00. In comparison, the fine-tuned ChatGPT assigned slightly higher scores, with a mean of 4.50, a median of 4.00, and lower variability, reflected in the lower standard deviation of 1.12. The IQR was 1.00, and scores ranged from 2.00 to 7.00. For essays on the topic of medical issues, the human rater provided an average score of 4.20, with a median of 4.00, and a standard deviation of 1.83. The IQR was 3.00, and the scores ranged from 1.00 to 8.00. In this case, the fine-tuned ChatGPT assigned an average score of 4.34, with a median of 4.00, and a standard deviation of 1.17. The IQR was 1.00, and scores ranged from 2.00 to 7.00. These statistics reflect the scoring patterns of both human raters and the fine-tuned ChatGPT for essays on the topics of driving and medical issues, providing insights into the consistency and distribution of scores in both cases.

Table 8 displays inter-rater reliability metrics between human raters and the fine-tuned ChatGPT for essays on two distinct topics: driving and medical issues, based on 100 ratings. For the topic of driving, the analysis revealed a robust agreement in rankings as indicated by a Spearman's correlation coefficient of .82 ($p < .001$). The percentage agreement between human raters and ChatGPT for this topic, without tolerance, was relatively low at 29%, suggesting a substantial difference in their assigned scores. However, when allowing a tolerance of 1 point, the agreement substantially increases to 81%, and with a tolerance of 2 points, it reaches 100%, indicating near-identical scores between human raters and ChatGPT. The Cohen's kappa value was .15 ($p < .001$), signifying some degree of agreement beyond chance, although the reliability remains relatively low. Krippendorff's alpha, which assesses agreement beyond chance, was estimated at .76. For the topic of medical issues, the analysis similarly demonstrated a robust rank correlation with a Spearman's coefficient of .74 ($p < .001$). The percentage

agreement without tolerance was 31%, indicating differences in score assignments between human raters and the fine-tuned ChatGPT. Yet, with a tolerance of 1 point, the agreement increased to 75%, and with a tolerance of 2 points, it reached 98%, indicating a high level of agreement when slight score variations are considered. The Cohen's kappa value was .17 ($p < .001$), suggesting some agreement beyond chance but relatively low reliability. Krippendorff's alpha, signifying agreement beyond chance, was estimated at .68. These reliability metrics provide insights into the alignment between human raters and the fine-tuned ChatGPT on the assessment of essays on the topics of driving and medical issues.

TABLE 7

Descriptive Statistics of Grading of Essays on the Topic of Driving and Medical Issues

Topic	Rater	Mean	Median	SD	IQR	Minimum	Maximum
Driving (n = 100)	Human Rater	4.38	4.00	1.84	3.00	1.00	8.00
	Fine-tuned ChatGPT	4.50	4.00	1.12	1.00	2.00	7.00
Medical Issues (n = 100)	Human Rater	4.20	4.00	1.83	3.00	1.00	8.00
	Fine-tuned ChatGPT	4.34	4.00	1.17	1.00	2.00	7.00

Note. IQR: Interquartile ranges, SD: Standard deviation

TABLE 8

Inter-rater Reliability Between Human Raters and the Fine-Tuned ChatGPT on the Topic of Driving and Medical Issues

Topic	Spearman's ρ	Percentage Agreement	Percentage Agreement (Tolerance = 1)	Percentage Agreement (Tolerance = 2)	Cohen's Kappa	Krippendorff's Alpha
Driving (n = 100)	.82 ($p < .001$)	.29	.81	1.00	.15 ($p < .001$)	.76
Medical Issues (n = 100)	.74 ($p < .001$)	.31	.75	.98	.17 ($p < .001$)	.68

5. Analysis of Rater Variations using MFRM

To explore variations in rater severity, MFRM analysis was conducted on the topic of physical punishment with examinees and raters (i.e., human raters, original ChatGPT, fine-tuned ChatGPT first grading, and fine-tuned ChatGPT second grading) as facets. This MFRM analysis compares the severity of human raters, original ChatGPT, and fine-tuned ChatGPT, thus determining whether ChatGPT scores align with human raters. In addition, infit mean-square and outfit mean-square values of raters in the MFRM analysis reveal each rater's rating patterns, such as outlying or inlying rating pattern.

Table 9 presents summary statistics for essay scores using Rasch measurement. The mean essay score for examinees was 4.09 (*SE* = 0.81), while rater severity was centered at zero. The root mean square error (RMSE) for examinees was 0.79, indicating the average residual error in the model, whereas for raters, it was lower at 0.12, suggesting less error. The adjusted (true) standard deviation for examinees was 1.83, signifying the variability in examinee scores, and for raters, it was 0.89, indicating lower variability. Infit and outfit mean-squares provide insights into the fit statistics, with examinees having an infit mean-square mean of 0.95 and an outfit mean-square mean of 0.98. Raters exhibit similar fit statistics with an infit mean-square mean of 0.99 and an outfit mean-square mean of 0.98. All these fit statistics fell within the acceptable range (0.5-1.5) (Linacre, 2023). Separation (G) and Strata (H) values suggest the model's ability to distinguish between different levels of proficiency. Examinees have a separation of 2.32 and a strata of 3.42, indicating moderate reliability (.84). In contrast, raters show a higher separation of 7.63 and a strata of 10.51, with strong reliability (.98), indicating that raters were distinctly separated into ten different severity levels.

TABLE 9

Rasch Measurement Summary Statistics of Essay Scores

Statistic	Examinees	Raters
Measures		
Mean (SE)	4.09 (0.81)	4.03 (0.31)
RMSE	0.79	0.12
Adjusted (True) SD	1.83	0.89
Infit MS		
Mean	0.95	0.99
Outfit MS		
Mean	0.98	0.98
Separation (G)	2.32	7.63
Strata (H)	3.42	10.51
Separation reliability	.84	.98

Note. Rater variable was centered at zero.

Figure 1 illustrates the Wright map derived from the rating scale analysis of 181 examinees by four raters: human raters, original ChatGPT, and two iterations of fine-tuned ChatGPT grading. The first column of the Wright map displays a logit scale, while the second column represents the examinees' abilities. The third column depicts the severity of the raters across the four grading sessions, highlighting the divergence of the original ChatGPT from the other three ratings. The last column provides score categories.

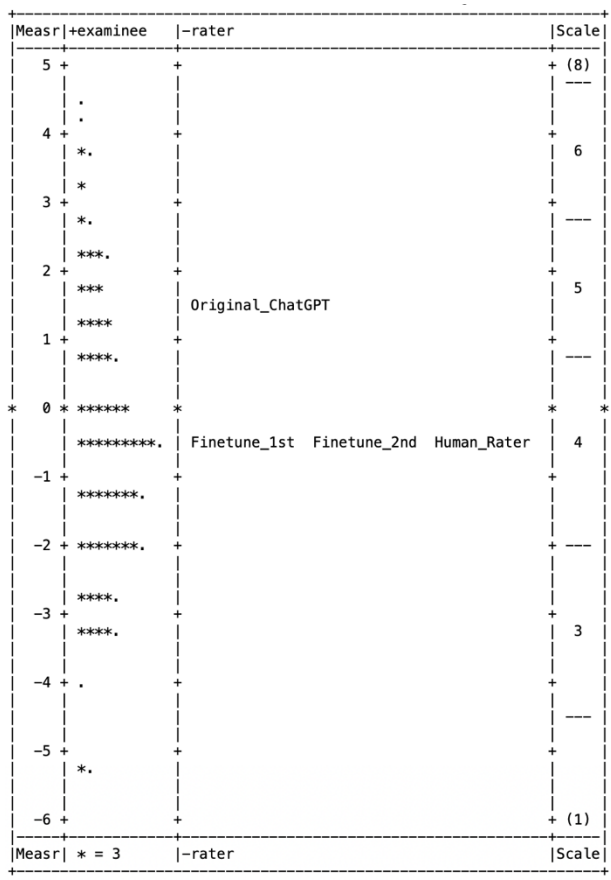

FIGURE 1

Wright Map for Examinees and Raters

Table 10 presents rater severity, infit mean-square, and outfit mean-square for different raters. Human raters exhibited a rater severity measure of -0.45 logits (*SE* = 0.12), with a 95 % confidence interval ranging from -0.68 to -0.22. Their infit mean-square and outfit mean-square values were 1.21 and 1.19, respectively, falling within the acceptable range (0.5-1.5) (Linacre, 2023). In contrast, the original ChatGPT displayed a higher rater severity of 1.55 logits (*SE* = 0.12), with a 95% confidence interval between 1.32 and 1.78. Both the infit mean-square and outfit mean-square values for the original ChatGPT were approximately 1.94, indicating the presence of unexpected outliers and disturbed rating patterns (Linacre, 2023). The first scoring with the fine-tuned ChatGPT had a rater severity of -0.48 logits (*SE* = 0.12), with a confidence interval from -0.71 to -0.25. Its infit mean-square was 0.40, and the outfit mean-square was 0.38. Lastly, the second scoring with the fine-tuned ChatGPT exhibited a rater severity of -0.61 logits (*SE* = 0.12), with a confidence interval from -0.84

to -0.38. The infit mean-square and outfit mean-square for the second scoring with the fine-tuned ChatGPT were 0.42 and 0.41, respectively. The infit and outfit mean-square values of the fine-tuned model scorings were smaller than the acceptable level, indicating that the ratings were overly predictable (Linacre, 2023). These low mean-square values may suggest that the fine-tuned ChatGPT might have applied scores mechanically without considering the actual content of the responses. Considering the confidence intervals of rater severity, only the original ChatGPT exhibited a statistically higher value compared to the other three raters. Given that the true measure of rater severity should typically fall within ± 1.96 times the standard error about 95% of the time (Eckes, 2015), this result indicates that the rater severity of the original ChatGPT was not equivalent to human raters, while that of the fine-tuned ChatGPT was almost equivalent to human raters.

TABLE 10

Rater Severity and Infit and Outfit Mean-Squares on the Topic of Physical Punishment

Raters	Rater Severity			Infit Mean-Square	Outfit Mean-Square
	Measures (logits)	Standard Error (SE)	Confidence Interval		
Human Rater	-0.45	.12	[-0.68, -0.22]	1.21	1.19
Original ChatGPT	1.55	.12	[1.32, 1.78]	1.94	1.94
Fine-tuned ChatGPT 1st	-0.48	.12	[-0.71, -0.25]	0.40	0.38
Fine-tuned ChatGPT 2nd	-0.61	.12	[-0.84, -0.38]	0.42	0.41

V. DISCUSSION

Automated essay evaluation has been recognized as a captivating tool in the field of writing education. When ChatGPT was first introduced, it received much attention, but it also raised concerns as *hallucinations* soon emerged (Emsley, 2023; Meyer et al., 2023). Various types of AES algorithms, such as LSA and linear regression (Ramalingam et al., 2018), have been introduced to improve the accuracy of automated scoring. However, it was not until LLMs were introduced in 2018 (Devlin et al., 2018) that practitioners, such as teachers and researchers in the field of language education who lacked engineering expertise, gained the ability to adapt AES models to their educational contexts. For this reason, when ChatGPT was introduced, stakeholders in the educational sector welcomed the service because they could simply engage in a chat with ChatGPT and obtain the answers they needed.

In this context, the first research question aimed to determine how LLMs, which were not specifically trained for essay scoring in EFL writing, could score essays written by Korean college students compared to human raters. The results did not support the idea that AES

using ChatGPT could be performed with a reasonable degree of accuracy. Inter-rater reliability between human raters and the original ChatGPT in AES indicated that the un-fine-tuned ChatGPT did not perform in line with human raters. This observation could also be related to the phenomenon of hallucinations as GPT models tend to "make things up", especially when lacking prior information (O'Brien, 2023). Considering that ChatGPT was trained on publicly available sources, such as Wikipedia, Reddit, books, and articles, with only minimal EFL essays and proficiency scores publicly available, it is reasonable to infer that ChatGPT lacks prior information about the proficiency levels of academic essays, especially those written by EFL students. Consequently, the model is less likely to accurately assess EFL essays. This finding differs from Mizumoto and Eguchi's (2023) research, which suggested that ChatGPT could be utilized for AES especially when combined with linguistic features. The discrepancy may be due to differences in the ChatGPT version used (text-davinci-003) and a greater focus on comparing GPT scores with linguistic features in their study.

The second research question aimed to determine whether fine-tuning the original ChatGPT could improve the model's accuracy in AES. The results, comparing the accuracy of human raters with that of the fine-tuned ChatGPT in essay scoring, provide valuable insights into the potential of using a fine-tuned ChatGPT for AES. Inter-reliability indices significantly improved compared to those of human raters when the original ChatGPT was used. Pre-trained language models, like GPT, can be effectively adapted to specific domains using optimization techniques such as fine-tuning or other transfer learning methods (Quinn, 2022). The results of this study support the hypothesis that fine-tuning ChatGPT with Korean student essays and their human-scored grades can enhance accuracy in AES. Thanks to this domain-specific adaptation, the fine-tuned scoring model can capture the unique complexities and intricacies found in essays within EFL education. Moreover, both the first and second scoring with the same fine-tuned ChatGPT exhibited the highest inter-rater reliability, indicating that the model is exceptionally consistent, surpassing even the consistency of human raters. This result further supports the idea that fine-tuned ChatGPT could potentially replace human raters in terms of consistency, as suggested in previous studies (Latif & Zhai, 2023; Mizumoto & Eguchi, 2023; Yancey et al., 2023).

This study also explored whether a fine-tuned ChatGPT can effectively assess students' essays in the same format as the essays used for fine-tuning but on different topics. Given the nature of English placement tests at the university, where essay topics are typically not reused for test security reasons, students are assigned different topics each year. In this academic environment, there is a need for an AES system capable of handling new topics that were not used during fine-tuning. Fortunately, the results of the inter-rater reliability analysis between human raters and the fine-tuned ChatGPT, using topics that were not part of the fine-tuning process, also showed a high level of agreement. This result indicates that once ChatGPT is trained for a specific task, it can be applied to that task or AES for different

topics. This observation suggests that AES with LLMs may not only learn the content itself but also comprehend the characteristics of the language and proficiency categories. This result aligns with the findings of Yancey et al. (2023), who discovered that the newly introduced ChatGPT (GPT-4) can perform well in AES when calibration examples are provided. In their study, however, agreement with human ratings varied depending on the examinees' L1. They suggested that this variation could be due to the fact that essays in some L1s are harder to distinguish, leading to less reliable human ratings, or it might be attributed to inconsistencies in human raters' behaviors in their study.

These concerns raise potential limitations in this study. Firstly, LLM-based AES studies, including this one, often compare the model's performance to that of human raters, which ideally should be validated in advance. Due to limitations regarding the availability of EFL essay samples with proficiency levels, however, the essays used for fine-tuning were not pre-validated. The YELC corpus (Rhee & Jung, 2014) did not provide intra-rater reliability information, so it was not considered in this study. Secondly, even when human raters' rating performances are reported in their original corpus-building studies, human rating is not always considered the gold standard (Ethayarajh & Jurafsky, 2022). Therefore, independent standards should be prepared to assess the quality of AES, rather than solely relying on raw scoring data by human raters.

Despite these limitations, the study offers some practical insights. It suggests that advanced LLMs are likely capable of understanding human writings, albeit possibly in a different manner than human readers, and can score EFL essays. This implies the potential for replacing human raters with LLM-based AES, especially when the stakes are not high. Such a shift could result in cost reduction and improved time efficiency in essay scoring. However, it is important to acknowledge the limitations of using AES with ChatGPT. While fine-tuning may enhance its performance, it should be used as a complementary tool alongside human evaluation. Given the relatively high threshold for fine-tuning, local school teachers may face challenges in conducting this process, so they should exercise caution when using the original ChatGPT and relying on its AES results.

VI. CONCLUSION

The study examined the reliability of LLMs in automated essay scoring compared to traditional human scoring methods, the impact of fine-tuning on scoring accuracy, and the model's ability to evaluate essays on different topics. The results uncovered that it is not currently feasible to conclude that the current ChatGPT (gpt-3.5-turbo-0613 as of October 31, 2023) can fully replace human raters for essay scoring. However, when ChatGPT was fine-tuned, the model demonstrated substantial alignment with human raters, and its

consistency was high. Furthermore, this study revealed the alignment between human raters and the fine-tuned ChatGPT in the assessment of essays on topics that were not used for model fine-tuning.

Future research possibilities include exploring the use of more advanced GPT models, such as GPT-4, to enhance the efficiency of evaluating student essays. Additionally, it would be valuable to develop a fine-tuned model capable of evaluating EFL essays not only based on holistic scoring scales, as seen in prior studies, but also on analytic scoring scales. This approach would facilitate the provision of specific writing feedback and contribute to improving student motivation. These studies will address the challenge of developing explainable AI for AES as well.

REFERENCES

Altamimi, A. B. (2023, August 14-16). Effectiveness of ChatGPT in essay autograding. *Proceedings of the 2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 102-106). Swansea, United Kingdom. https://doi.org/10.1109/iCCECE59400.2023.10238541

Andersen, M. R., Kabel, K., Bremholm, J., Bundsgaard, J., & Hansen, L. K. (2023). Automatic proficiency scoring for early-stage writing. *Computers and Education: Artificial Intelligence*, *5*, 100168. https://doi.org/https://doi.org/10.1016/j.caeai.2023.100168

Burstein, J. (2012). Automated essay evaluation and scoring. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 309-315). Blackwell Publishing Ltd.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55-67). Routledge.

Christodoulou, D. (2023, January 10). *Could ChatGPT mark your students' essays?* TES. https://www.tes.com/magazine/teaching-learning/general/ai-marking-teachers-could-chatgpt-mark-your-students-essays

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7-24. https://doi.org/https://doi.org/10.1016/j.asw.2012.10.002

Delgado, H. O. K., de Azevedo Fay, A., Sebastiany, M. J., & Silva, A. D. C. (2020). Artificial intelligence adaptive learning tools: The teaching of English in focus. *BELT-*

Brazilian English Language Teaching Journal, *11*(2), e38749-e38749. https://doi.org/10.15448/2178-3640.2020.2.38749

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv:1810.04805. https://doi.org/10.18653/v1/N19-1423

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang GmbH.

Emsley, R. (2023). ChatGPT: These are not hallucinations—they're fabrications and falsifications. *Schizophrenia (Heidelb)*, *9*(1), 52. https://doi.org/10.1038/s41537-023-00379-4

Ethayarajh, K., & Jurafsky, D. (2022, December 7-11). The authenticity gap in human evaluation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 6056-6070). Abu Dhabi, United Arab Emirates. https://doi.org/10.18653/v1/2022.emnlp-main.406

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68-88). Routledge.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr: Various coefficients of interrater reliability and agreement.* https://CRAN.R-project.org/package=irr

Hamner, B., Morgan, J., lynnvandev, Shermis, M., & Ark, T. V. (2012). *The Hewlett Foundation: Automated Essay Scoring*. Kaggle. https://kaggle.com/competitions/asap-aes

Hockly, N. (2018). Automated writing evaluation. *ELT Journal*, *73*(1), 82-88. https://doi.org/10.1093/elt/ccy044

Huawei, S., & Aryadoust, V. (2022). 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*, e208. https://doi.org/10.7717/peerj-cs.208

Hwang, K.-H., Heywood, D., & Carrier, J. (2023). The implementation of ChatGPT-assisted writing instruction in ESL/EFL classrooms. *The New Korean Journal of English Language and Literature*, *65*(3), 83-106. https://doi.org/10.25151/nkje.2023.65.3.004

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017. https://doi.org/https://doi.org/10.1016/j.caeai.2021.100017

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274

Khademi, A. (2023). *Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance.* arXiv:2304.05372. https://doi.org/10.48550/arXiv.2304.05372

Kim, M. K. (2023). Towards a Critical-PBLL utilizing ChatGPT and Google Bard within college English education. *Korean Journal of English Language and Linguistics*, *23*, 741-767. https://doi.org/10.15738/kjell.23.202309.741

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Education and Psychological Measurement*, *30*, 61-70. https://doi.org/10.1177/001316447003000105

Latif, E., & Zhai, X. (2023). *Fine-tuning ChatGPT for automatic scoring.* arXiv:2310.10072. https://doi.org/10.48550/arXiv.2310.10072

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2023). *A user's guide to FACETS*. Retrieved September 15, 2023, from https://www.winsteps.com/a/Facets-ManualPDF.zip

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P. C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Min*, *16*(20), 1-11. https://doi.org/10.1186/s13040-023-00339-9

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013, December 5-10). *Distributed representations of words and phrases and their compositionality* [Conference session]. The 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA. arXiv:1310.4546. https://doi.org/10.48550/arXiv.1310.4546

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/https://doi.org/10.1016/j.rmal.2023.100050

O'Brien, M. (2023, August 2). *Chatbots sometimes make things up. Is AI's hallucination problem fixable?* The Associated Press. https://apnews.com/article/artificial-intelligence-hallucination-chatbots-chatgpt-falsehoods-ac4672c5b06e6f91050aa46ee731bcf4

OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. Retrieved October 16, 2023, from https://mkai.org/chatgpt-optimizing-language-models-for-dialogue/

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243. http://www.jstor.org/stable/20371545

Pratitis, N. T., & Purwono, U. (2018). The architectural creativity test development: A many facet Rasch model analysis to establish inter-rater reliability. *Journal of Educational, Health and Community Psychology*, *7*(3), 225-247. https://doi.org/10.12928/JEHCP.V7I3.11698

Quinn, J. (2022). *Dive into deep learning: Tools for engagement*. Cambridge University Press.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI.

Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018, January 5-6). Automated essay grading using machine learning algorithm. In A. Govindarajan, E. P. Siva, & K. Suja (Eds.), *Proceedings of the National Conference on Mathematical Techniques and its Applications (NCMTA 18)* (pp. 223-229). Kattankulathur, India. https://doi.org/10.1088/1742-6596/1000/1/012030

Rhee, S.-C., & Jung, C. K. (2014). Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *The Journal of the Korea Contents Association*, *14*(11), 1019-1029. https://doi.org/10.5392/JKCA.2014.14.11.1019

Rothman, D. (2022). *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3* (2nd ed.). Packt Publishing Ltd.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1-15). Routledge.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017, December 4-9). *Attention is all you need* [Conference session]. The 31st Annual Conference on Neural Information Processing Systems, Long Beach, California, USA. arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). Routledge.

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023, July 13). Rating short L2 essays on the CEFR scale with GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576-584). Toronto, Canada. https://doi.org/10.18653/v1/2023.bea-1.49

Yeşilyurt, Y. E., & Sezgin, S. (2023). *Automated writing evaluation with a large pre-trained language model: A preliminary study* [Unpublished manuscript]. https://ssrn.com/abstract=4545244.

APPENDICES

APPENDIX 1 Example Data Format for Fine-Tuning and Scoring

<fine-tuning>

```
{"messages": [{"role": "system", "content": " prompt."},
   {"role": "user", "content": "student essay 1"},
   {"role": "assistant", "content": "score 1"}]}
{"messages": [{"role": "system", "content": " prompt."},
   {"role": "user", "content": "student essay 2"},
   {"role": "assistant", "content": "score 2"}]}
{"messages": [{"role": "system", "content": " prompt."},
   {"role": "user", "content": "student essay 3"},
   {"role": "assistant", "content": "score 3"}]}
```

<scoring>

```
{"messages": [{"role": "system", "content": " prompt."},
   {"role": "user", "content": "student essay 1"}
{"messages": [{"role": "system", "content": " prompt."},
   {"role": "user", "content": "student essay 2"}
{"messages": [{"role": "system", "content": " prompt."},
   {"role": "user", "content": "student essay 3"}
```

APPENDIX 2 Prompt Excerpt for Fine-tuning

PROMPT:
Evaluate the quality of an essay written by English as a foreign language (EFL) learner on a scale of 1 to 9, with 9 being the highest and 1 being the lowest. Please provide only a score without providing the rationale for your rating.

The evaluation rubric for your rating, adapted from CEFR Descriptors, is as follows:
[Score 9]
<Overall written production>
Can produce clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points.
<Vocabulary>
Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.
Consistently correct and appropriate use of vocabulary.
<Grammatical Accuracy>
Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).
<Creative writing>
Can relate clear, smoothly flowing and engaging stories and descriptions of experience in

a style appropriate to the genre adopted.
Can exploit idiom and humour appropriately to enhance the impact of the text.
<Reports and essays>
Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works.
Can provide an appropriate and effective logical structure which helps the reader identify significant points.
Can set out multiple perspectives on complex academic or professional topics, clearly distinguishing their own ideas and opinions from those in the sources.

[…]

[Score 1]
<Overall written production>
Can give information about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words/signs and basic expressions.
Can produce simple isolated phrases and sentences.

[…]

ESSAY:

Note. To ensure the reproducibility of the fine-tuning process, access has been granted to the data (excluding the raw data from the YELC 2011) and the Python code (including the full prompt) used in the study on OSF (osf.io/3qxgn) and the author's GitHub repository (https://github.com/linguistry/Fine-Tuned_ChatGPT_Essay_Scoring).

Applicable levels: Secondary and Tertiary

Won, Yongkook
Visiting Researcher/Center for Educational Research, Seoul National University
linguistry@gmail.com

Received: November 03, 2023
Reviewed: November 22, 2023
Accepted: December 01, 2023