Harnessing AI for CEFR-Aligned Writing Assessment: A Study of a Customised GPT

Chua Wei Chuan[1*], Melor Md Yunus[2], Harwati Hashim[2]

[1]SBP Integrasi Sabak Bernam, Selangor

[2]Faculty of Education, Universiti Kebangsaan Malaysia

[*]Corresponding Author

DOI: https://doi.org/10.47772/IJRISS.2026.100400326

Received: 08 April 2026; Accepted: 14 April 2026; Published: 08 May 2026

ABSTRACT

Artificial intelligence (AI) tools, specifically ChatGPT, has garnered considerable interest in the field of English as a Second Language (ESL) as demand grows in recent times. The introduction of ChatGPT has transformed the assessment landscape, thus marking the beginning of a new era in AI-assisted assessment. While existing research has largely addressed the comparison of ChatGPT and human raters with regards to assessment, limited study has focused on customised GPTs and the context of Common European Framework of References for Languages (CEFR). To address this gap, this paper investigates the effectiveness of a customised GPT as a formative and summative assessment tool in a CEFR-aligned written task pitched at B2 CEFR level. Adopting a quantitative research design, the respondents' attitudes towards using the GPT for formative assessment were examined through a questionnaire administered to 31 English teachers. In parallel, the scores assigned by the GPT and the teachers for the same writing tasks were compared via inter-rater reliability analysis. Findings revealed that the respondents hold a generally positive view regarding the effectiveness of GPT in providing formative feedback. However, the results also indicated that the GPT demonstrates a moderate level of agreement with the teacher scores in most assessment constructs. The data further emphasized the need of prompt engineering in developing the GPT to be an effective formative and summative assessment assistant for teachers. This paper concludes by discussing the practical implications of employing the customised GPT in assessment, thereby contributing to the discourse on AI-assisted assessment.

Keywords: Assessment, Writing, AI, Customised GPT, CEFR

INTRODUCTION

In recent years, ChatGPT, one of the latest breakthroughs in Artificial Intelligence (AI) technology, has rapidly surpassed many other digital tools in prominence (Kohnke et al., 2023). Its rapid global adoption can be attributed to its versatility and its ability to deliver instant, informative and personalised support (Alm & Ohashi, 2022; Susnjak, 2022). Currently the AI chatbot continues to undergo extensive training on vast datasets as it moves towards functioning not only as an information retrieval tool but also as a system capable of engaging users in increasingly human-like conversational interactions (Chung & Jeong, 2024; Kim & Su, 2024). In the field of assessment, ChatGPT is deemed an effective tool that provides feedback and assistance to aid learners in acquiring language (Abduljawad, 2024). Upon receiving feedback generated by the chatbot, learners are able to develop their own understanding in the learning area as they take charge of their learning, hence cultivating an active learning environment (Dahri et al., 2024).

Writing is arguably one of the effective tools that can improve a student's academic achievement and help develop one's critical thinking, communication and deep learning, thus highlighting the importance of emphasising writing skills to support multifaceted student development (Elkatmış, 2024). Despite that, writing

is perceived as one of the most significant challenges for English as a Second Language (ESL) due to its intricacies in grammar and syntax, hence underscoring the significance of formative feedback from teachers that impact writing (Almashy et al., 2024; Anderson & Ayaawan, 2023; Hyland & Annan, 2006). Although formative feedback provided by teachers offers support to students in the process of writing, it is often difficult to address individual learners' needs in the form of feedback due to time constraints and large classroom sizes (Golzar et al., 2022; Mahapatra, 2024; Zhu et al., 2020). Additionally, assessing writing involves the process of editing, providing quality feedback and evaluating the written work by assigning a numerical grade or score (Geckin et al., 2023). The aforementioned process is inevitably time-consuming, physically taxing and inconsistent among and within the raters (Hua & Wind, 2019; Uto & Ueno, 2018). In other words, the human-related factors are the limitations to providing effective feedback to students in writing.

Computer-mediated feedback, on the other hand, as exemplified by ChatGPT can complement the teacher's role by offering advice, providing writing samples that match the learner's proficiency level and facilitating guided writing (Barrot, 2023; Bonner et al., 2023; Imran & Almusharraf, 2023; Kohnke et al., 2023). Yet, in reality present studies have primarily focused on how ChatGPT brings impact to writing in a general sense without addressing the context of the English language curriculum. Currently the Common European Framework of Reference for Languages (CEFR) model is a framework that is globally borrowed inside and outside Europe via policy borrowing (Rozana Sani, 2016). Despite being initially developed specifically for Europe, the framework has been adapted and adopted by many non-European countries due to its dynamic nature and adaptability (Mohamad Uri et al., 2023). A survey even claimed that the CEFR is the most influential model that is being used worldwide pertaining to innovation in curriculum, pedagogy and assessment (Council of Europe, 2026). However, the effectiveness of ChatGPT in relation to the CEFR model that is widely adopted has yet to be explored by future researchers.

Besides, studies in the present literature have placed great emphasis on exploring ChatGPT as an assessment tool with regards to formative and summative assessment in general. In the context of summative assessment, researchers have explored how reliable and consistent the AI chatbot scores student work when its evaluations are compared to those of human raters (Geckin et al., 2023; Mizumoto & Eguchi, 2023). For formative assessment, learners who received feedback from ChatGPT reported improvements in their writing, noting that the AI-generated feedback not only identified and corrected errors more effectively than teachers but also offered personalised guidance to support their learning (Bhutoria, 2022; Jamshed et al., 2024; Mahapatra, 2024). Nevertheless, the customised version of ChatGPT that can be developed by individual users, referred to as GPTs, has not been addressed in the existing literature. A GPT can be programmed to perform specific tasks through the input of task-specific instructions and domain-specific knowledge provided by the user, thereby demonstrating significant potential as an assessment tool for particular writing genres (OpenAI, 2025).

Malaysia's educational landscape and assessment paradigms have likewise evolved when the CEFR model was adopted by the Ministry of Education Malaysia (MoE) with the aim of benchmarking the country's ESL education system, aligning it to international standards and calibrating the components of the English language programmes in Malaysia (English Language Standards and Quality Council [ELSQC], 2015). With regards to English language, students need to attempt an extended writing task which is focused on the B2 CEFR level in a national public examination (Cambridge Assessment English, 2020; Examinations Syndicate, 2021; Mohd Salleh et al., 2023). Therefore, this study aims to examine the effectiveness of a customised GPT as an assessment tool for writing within the B2 CEFR-aligned educational context, guided by the following research questions.

RQ1: How effective is the customised GPT as a formative assessment tool for writing as perceived by ESL teachers?

RQ2: To what extent do the scores generated by the customised GPT align with the scores awarded by ESL teachers in writing?

LITERATURE REVIEW

In this section, researchers firstly discuss the role of ChatGPT in writing assessment in the current literature followed by a background review of the CEFR model and the contextualised writing assessment focus. Furthermore, similar studies are also reviewed at the end of the section.

ChatGPT in Writing Assessment

According to Alsaweed and Aljebreen (2024), the AI chatbot can locate and correct both minor and major writing errors in the user's writing. However, it fails to recognise slang words or informal language expressions and the chatbot often replaces casual language with more formal ones upon correction (Punar Özçelik & Yangın Ekşi, 2024). Additionally, earlier versions of ChatGPT have been shown to struggle with addressing issues related to meaning, cohesion, grammatical accuracy, and unnatural expressions (Alsaweed & Aljebreen, 2024). In contrast, the more advanced ChatGPT-4 demonstrates an improved ability to recognise various word categories and word-formation processes, enabling it to correct linguistic errors more effectively (Ronan & Schneider, 2023).

ChatGPT is able to generate model texts and writing templates across a range of registers, thus supporting students in understanding the conventions of different writing styles (Alm & Ohashi, 2024; Hadizadeh, 2024). By creating texts that are comprehensible, coherent and of the right length, ChatGPT facilitates cognitive offloading as it supports users in managing cognition (Barrett & Pack, 2023). The texts generated sometimes do not contain a clear story and the content is somehow similar and predictable for users (Kotmungkun et al., 2024; Zindela, 2023). Despite that, ChatGPT exhibits an exemplary performance in creating texts discussing academic topics in the form of complex ideas and less common lexis, thereby making it suitable for advanced learners with good reading skills (Kotmungkun et al., 2024; Nkhobo & Chaka, 2023; Zindela, 2023).

In terms of feedback provided, it is generated immediately after being given input by the user and the feedback is generally personalised and caters to needs of the individual users (Naz & Robertson, 2024). Devoid of societal pressure, learners can learn the subject more deeply and flexibly via personalised responses that are not usually found in conventional classrooms (Grassini, 2023; Maghsudi et al., 2021; Pang, 2022). Through iterative cycles of responses and feedback, this aids the learning process and improves the learner's language proficiency over time (Shaikh et al., 2023). Furthermore, AI-generated feedback is more beneficial than feedback from teachers as it leads to less language errors in second drafts based on the existing literature (Jamshed et al., 2024). By engaging in life-like conversational interactions with ChatGPT, students remain more motivated and are better able to grasp grammatical concepts through guided practice and corrective feedback (Almashy et al., 2024). Also, the feedback provided by ChatGPT is more specific and detailed and this helps students in acquiring language and mastering writing (Al-khresheh, 2024; Bhutoria, 2022; Mahapatra, 2024).

The CEFR

The CEFR provides a common basis for countries practising different educational systems with regards to the development of the curriculum, pedagogy and assessment by defining measurable levels of proficiency and objectives for language learners in a comprehensive and explicit way regardless of context (Council of Europe, 2001). Thus, this facilitates the mutual recognition of language qualifications gained in various contexts and increases the mobility of language learners accordingly (Council of Europe, 2001). Based on the CEFR framework, six levels are used to describe a learner's language proficiency (A1, A2, B1, B2, C1 and C2) in which each includes "can do" descriptors covering oral production, written production, listening, reading, and interaction for every reference level (English language programmes in Malaysia (English Language Standards and Quality Council [ELSQC], 2015).

For those who are exposed to the language until primary school level, they are regarded as A1 or A2 learners as they are able to use the language for basic needs that are of immediate relevance to them (ELSQC, 2015). When they enter secondary school and college or university, they become independent users with a B1 or B2 proficiency level because they are able to use the language independently in familiar or foreign contexts

(ELSQC, 2015). For users who are proficient in the language with a C1 or C2 CEFR level, they can use the language with ease and fluidity regardless of whether the language is their first, second or third language (ELSQC, 2015).

Assessing Writing

The writing assessment focus of the Malaysian national public examination for English can be divided into task completion (Content) and linguistic competence (Communicative Achievement, Organisation and Language) (Examinations Syndicate, 2021). When assessing the content, primary focus is given to whether the candidate fulfills the task requirements. Misinterpretation of task often results in low scores for content depending on whether it is a global misinterpretation or local misinterpretation (Examinations Syndicate, 2021). However, it is not to be assumed that there is a one-to-one relationship between the task requirements and the content scale as the content assessment is primarily based on the cumulative effect of the response towards the target reader (Examinations Syndicate, 2021).

As for communicative achievement, the focus is on whether the conventions of the communicative task from the aspect of tone, format and register are followed appropriately (Examinations Syndicate, 2021). Besides, the text is assessed to what extent it can hold the target reader's attention by allowing the reader to derive meaning without being distracted or forced to read closely as the text is difficult to make sense of (Examinations Syndicate, 2021). In addition, how simple, straightforward or complex ideas are being communicated throughout the task based on the CEFR level is given attention with regard to the construct (Examinations Syndicate, 2021). When assessing the organisation sub-scale, the text is evaluated on the coherence and cohesion of the writing by employing linking words and cohesive devices as well as paragraphing and punctuation (Examinations Syndicate, 2021). For the language subcomponent, the type of vocabulary used and complexity of sentence structures are evaluated according to the degree of errors made (Examinations Syndicate, 2021).

Previous Studies

In the existing literature, numerous studies have examined the extent to which AI-generated scores align with those awarded by human raters, with the findings yielding mixed results. Mizumoto and Eguchi (2023) used ChatGPT to grade essays (n = 12000) written by ESL learners after introducing it to a simplified writing rubric that had been developed in alignment with the criteria used by human raters. As a result, the scores assigned by ChatGPT were found to be consistent with the scores of human raters. Another study conducted by Atasoy and Moslemi Nezhad Arani (2025) revealed similar outcomes but human raters were found to exhibit a superior performance over AI in rating. Similarly, when ChatGPT was provided with example-rich rubric in the form of prompts, it demonstrated a consistent and fair grading compared to the unprompted version of ChatGPT (García-Varela et al., 2025).

In contrast, ChatGPT showed a slight to fair level of agreement with the human raters with regards to scores of a small sample size (n = 43) based on the study conducted by Geckin et al. (2023). Despite that, human raters only exhibited some level of agreement among themselves in the same study, thus suggesting that the scores assigned by a human rater could be more reliable when a human rater is paired up with an application in assessing written work. In the same vein, Shabara et al. (2024) and Manning et al. (2025) reported that ChatGPT displayed a consistent leniency in its analytical scoring across all criteria as compared to teachers. Additionally, results suggested that agreement of ChatGPT-4 with human ratings varies depending on the learner's first language in a previous study that involved rating short essay responses written by ESL learners (Yancey et al., 2023).

In terms of error correction, the chatbot can detect global writing errors which impede meaning and local writing errors which do not affect the comprehension of sentences well. According to Alsaweed and Aljebreen (2024), ChatGPT corrected global writing errors related to verb forms, word order, conditional sentences and passive voice with a high degree of accuracy. As for local writing errors, errors involving spelling, subject-verb agreement, singular/ plural nouns, word forms and articles are detected and corrected well. Similarly, the AI chatbot exhibited superior performance especially in correcting spelling, verb form and subject-verb agreement-related errors in a study conducted by Almashy et al. (2024) which compared the effectiveness of different

Computer-Assisted Language Learning (CALL) tools. However, a study conducted by Saricaoglu and Bilki (2025) showed that ChatGPT might make mistakes by misclassifying errors pertaining to the grammar and lexis, thus providing mixed results regarding the chatbot's ability in identifying and correcting errors effectively.

Current literature further identifies ChatGPT as a valuable writing tool, noting its ability to deliver meaningful and insightful formative feedback that supports learners' writing development. Based on the study carried out by Jamshed et al. (2024), feedback provided by ChatGPT is more effective than human feedback which results in improvement of writing proficiency and reduction in common language errors among secondary school ESL learners. This is due to the fact that AI-generated feedback is tailored to the individual needs of language learners which helps improve one's writing skills (Bhutoria, 2022; Mahapatra, 2024). Studies conducted by Mahapatra (2024), Abduljawad (2024) and Shi et al. (2025) also reported similar findings stating that tertiary level ESL students' writing skills have improved significantly after utilising ChatGPT as a formative feedback tool which possesses the dialogic feature. ChatGPT provides feedback in the form of interactive dialogues and this approach engages learners and promotes deeper linguistic comprehension (Almashy et al., 2024). Besides, Fitria (2023) discussed the idea that the structured responses offered by ChatGPT facilitate language learning.

Research Gaps

Although previous studies have examined the role of ChatGPT as an assessment tool in writing extensively, experimental research on customised versions of ChatGPT or GPTs remains scarce, as most existing work continues to focus on ChatGPT-3 and ChatGPT-4. The role of a GPT which is a custom version of ChatGPT that is created or configured by the user to cater to assessing writing remains underexplored, thereby leaving a gap in the present literature. Apart from that, none of the existing studies examine writing within the CEFR framework despite the framework being widely adopted by countries around the world. Therefore, this study seeks to address the aforementioned gaps by studying the effectiveness of a customised GPT as an assessment tool in the CEFR context with regards to writing.

METHODOLOGY

The Methodology section is presented by the discussion of the research approach as well as research design followed by the elaboration of the research context and research instruments to provide a nuanced understanding of the study. This section ends with the explanation of how the researchers collected and analysed the data to answer the research questions in the Introduction section.

Research Approach and Design

The researchers employed a quantitative research approach in this study to collect numerical data from respondents in order to study the effectiveness of a customised GPT chatbot as an assessment tool for writing in the Malaysian CEFR-aligned English curriculum (Creswell, 2012). For this research purpose, the effectiveness of the chatbot was examined through the formative and summative lenses. Initially the respondents and the customised GPT were asked to grade the same essays using the analytical scales aligned to the CEFR. To study the chatbot's role in terms of formative assessment, the respondents were asked to give their opinion regarding the AI-generated formative feedback with regards to the essays written via survey design to identify their attitude towards using the customised GPT as a formative assessment tool (Creswell, 2012). As for the summative component, the scores awarded by the respondents and the chatbot for the same essays were analysed through inter-rater reliability study as it considers the differences between the real scores assigned and the correlation between raters in measuring the relationship between different measures in the same construct (Cole et al., 2013; Larsen-Hall, 2010).

Research Context

The English teachers who are teaching in secondary schools in a certain district located in Malaysia were selected via multistage cluster sampling as the research samples based on the criteria shown in Table 1.

Table 1Selection criteria for the samples

No.	Selection Criteria
1	Currently teaches upper secondary English
2	Has taught upper secondary English for a minimum of five years
3	Holds at least a bachelor's degree in education with a specialisation in teaching ESL
4	Is an English optionist
5	Has made professional contributions related to English at least once a year at the district level or higher
6	Participates in CEFR-related professional development courses at least once a year at the district level
7	Participation is voluntary

A sample size of 31 respondents was selected out of approximately 80 upper secondary ESL teachers serving in the district to support the attainment of a normally distributed sampling distribution (Chua, 2013; Jamaludin, personal communication, September 22, 2025). All participants claimed owning a laptop, although a significant proportion did not own a desktop computer. Interestingly, every teacher possessed at least one mobile device and reported using gadgets on a regular basis for the purpose of teaching and learning. In short, these technological profiles indicate that the English teachers in this study demonstrate a comparatively high level of digital literacy.

The district under study can be classified as a rural area with 17 government schools for secondary students (Sabak Bernam District Council, 2025). Among the secondary schools in the district which comprise national schools, religious schools and boarding schools, 9 schools are considered as town schools while the others are located in the countryside (Ministry of Education Malaysia, 2025). Multistage cluster sampling was conducted to choose the research respondents as the population of English teachers in Malaysia cannot be easily identified, hence making it difficult to obtain a complete list of the members in the population per se (Creswell, 2012). Thus, the researchers first selected one district for research in order to obtain a list of upper secondary English teachers. Subsequently, participants were randomly drawn from each type of secondary school within the district to ensure that the sample accurately represented the population.

Research Instruments

The customised GPT was developed by the researchers following the steps outlined by Hassanien et al. (2025) as part of the Analysis, Design, Development, Implementation and Evaluation (ADDIE) model, henceforth referred to as EssayGPT. Prior to developing the AI chatbot, the researchers subscribed to ChatGPT Plus followed by configuring the chatbot. After providing a name for the customised GPT and including description for the chatbot, the following prompt was provided as input in the Instructions section.

As an English teacher, I want to build a GPT for marking English Paper 2 Part 3. Prepare prompts to build a GPT for EssayGPT where this GPT will help teachers to mark the uploaded papers and give comments according to the SPM Part 3 marking scheme. The marking must follow the marking scheme and references given. The language used must be clear, understandable for a teacher and can improve the students' skills. The comments can help teachers and can improve the students' writing skills. Besides, the comments must comply with the requirements of the marking scheme given. Prepare this prompt in English.

Next, conversation starters were prescribed to give users the choice regarding which type of writing (article/ story/ review/ report) they want the AI chatbot to assess. Concurrently, to elicit a desirable first response by the chatbot after the user clicks at the conversation starter, a sample first response as follows was provided by the researchers to the chatbot.

State the full question and the content points for the question requirement.

If the essay question is:

"You see this notice on the board outside the school library.

Articles wanted!

Ways to socialise

Video games and board games help us socialise. Which do you think is better? Support your choice by giving your reasons. Suggest ways to encourage people to socialise. Write us an article answering these questions. The best article will be displayed in the school magazine.

Write your article."

The user should state:

Question: "You see this notice on the board outside the school library.

Articles wanted!

Ways to socialise

Video games and board games help us socialise. Which do you think is better? Support your choice by giving your reasons. Suggest ways to encourage people to socialise. Write us an article answering these questions. The best article will be displayed in the school magazine.

Write your article."

C1 - Whether video games or board games help us socialise (Choose one only)

C2 - Reasons why video games or board games help us socialise

C3 - Ways to encourage people to socialise

Besides, CEFR-aligned resources namely the marking rubric, sample scripts for all genres and sample essays from the English textbooks were uploaded as well. Lastly, the customised GPT was published in the form of a link that can be shared to other users or it can also be published in the GPT store to be downloaded by the general users. During the Design and Development phases, EssayGPT has received expert reviews to ensure that the AI chatbot is a reliable and valid tool for formative and summative assessment (Branch, 2009).

Upon receiving the student's essays as input, EssayGPT grades the essays according to the CEFR-aligned analytical scale to provide an overall score for the essays per se. In addition, comments and suggestions on how to improve the quality of the written work are provided in the form of formative feedback. The essays comprise candidates' written responses to the B2 CEFR writing task according to the format of the Malaysian national examination for English. The task prompts them to write a story, review, report or article of word length 200 to 250 words based on text stimuli designed for the B2 CEFR level although it can elicit responses ranging from B1 to C1 CEFR level (Examinations Syndicate, 2021). The items in the specified task are designed by English language teachers selected at state and national level by adhering to the prescribed Table of Specifications, thus ensuring that they are items of high validity and reliability (Mehrens & Lehmann, 1991). The B2 CEFR writing task was chosen as the focus of this study because the task requires candidates to produce clear and detailed texts that address topical issues from multiple perspectives (ELSQC, 2015). However, since students are expected to reach only a B1 level by the end of secondary school, a proficiency gap remains and this highlights the need for appropriate scaffolding to support learners in meeting the demands of this task (Curriculum Development Division, 2015).

A questionnaire was employed in the survey design as it allows the participants to complete and return the form to the researchers after providing basic demographic information and choosing answers to questions (Creswell, 2012). To determine the effectiveness of EssayGPT as a formative assessment tool, a questionnaire (as presented in Table 2) was administered to the 31 respondents to gather their perceptions of the formative feedback generated by the chatbot.

Table 2 Research constructs and items for the questionnaire

Section	Construct	Item Number	Source
A	Demographic information	1 - 6	Adapted from Kutluk & Gülmez (2014)
B	Teacher's Attitude towards Using EssayGPT	7 - 18	Adapted from Jamshed et al. (2024)

The questionnaire was adapted from existing questionnaires designed by Jamshed et al. (2024) and Kutluk & Gülmez (2014) for previous studies to ensure the reliability and validity of the items. Section A requires participants to provide basic information related to their background and experience of using gadgets while section B contains 12 items that measure the respondent's opinion upon using EssayGPT as a formative assessment based on a 5-point Likert scale. The Cronbach's Alpha value for section B was 0.908 based on the survey administered, hence establishing the reliability of the items in the construct.

Data Collection and Analysis

In the early phase of data collection, both the respondents and EssayGPT were required to grade the same set of essays using the CEFR-aligned analytical scoring scales (Content, Communicative Achievement, Organisation, Language). Next, the respondents interacted with the GPT to elicit formative feedback regarding the essays written before giving their opinion on its role as a formative assessment tool via the survey administered by the researchers. The items in the questionnaire were analysed statistically using the Statistical Package for the Social Sciences (SPSS) software to determine their level of agreement with respect to the chatbot's effectiveness in conducting formative assessment (Creswell, 2012).

In addition, to determine its effectiveness as a summative assessment tool, the intraclass correlation coefficients (ICCs) for the four criteria: Content, Communicative Achievement, Organisation, Language in the analytical scale between the teacher and EssayGPT were calculated using the SPSS program in order to compare the reliability of the scores between human raters and AI raters (Shabara et al., 2024).

FINDINGS

This section presents results of the study structured into two subsections. The first subsection discusses the ESL teachers' attitudes towards using EssayGPT to measure its effectiveness as a formative assessment tool for writing as perceived by them. This is followed by the comparison of EssayGPT-generated scores and the teacher-rated scores for the same writing tasks in the subsequent subsection.

Teachers' Attitudes towards Using EssayGPT

Table 3 provides an overview of the descriptive statistics with respect to the 31 respondents' levels of agreement with the questionnaire items. The interpretation of the mean scores was further guided by the classification scale proposed by Allehyani and Algamdi (2023).

Table 3 Respondents' attitude towards using EssayGPT

Item No.	Item	SD	D	N	A	SA	*M*	*SD*
1	The EssayGPT interface is easy to navigate.	0	2	5	13	11	4.06	.892
2	I can easily read and understand the text provided by EssayGPT.	0	0	3	15	13	4.32	.653
3	The grammar and writing suggestions provided by EssayGPT are clear and understandable.	0	0	4	17	10	4.19	.654
4	The examples used by EssayGPT are relevant and helpful for improving the student's essay.	0	0	5	17	9	4.13	.670
5	Interacting with EssayGPT makes assessing English essays enjoyable.	0	0	4	17	10	4.19	.654

6	I feel motivated to assess English essays regularly using EssayGPT.	0	0	3	22	6	4.10	.539
7	My understanding of my students' essays has improved since using EssayGPT.	0	0	9	12	10	4.03	.795
8	I feel more confident in assessing English essays after practising with EssayGPT.	0	0	4	17	10	4.19	.654
9	EssayGPT addresses my specific questions towards my students' essays effectively.	0	1	3	18	9	4.13	.718
10	The feedback from EssayGPT is tailored to my level of English proficiency.	0	1	2	17	11	4.23	.717
11	I am satisfied with my overall experience with EssayGPT.	0	0	3	13	15	4.39	.667
12	I recommend using EssayGPT to others who are assessing English essays.	0	0	5	10	16	4.35	.755

Notes: SD = Strong Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree

$N = 31$

First and foremost, the teachers agreed that the chatbot's interface is easy to navigate ($M = 4.06$, $SD = .892$) and EssayGPT provides clear and comprehensible feedback ($M = 4.32$, $SD = .653$). Besides, the grammar and writing suggestions provided by EssayGPT are easy to understand ($M = 4.19$, $SD = .654$). Similarly, the examples offered by the AI chatbot are relevant and helpful in improving the quality of students' essays ($M = 4.13$, $SD = .670$). In addition, teachers indicated that EssayGPT addresses the specific queries regarding students' writing effectively ($M = 4.13$, $SD = .718$) and the feedback generated is aligned with their own level of English proficiency ($M = 4.23$, $SD = .717$). Apart from the quality of feedback and responses generated, respondents reported that interacting with EssayGPT has made the assessment process more enjoyable ($M = 4.19$, $SD = .654$).

Many teachers also noted an increased level of motivation to use EssayGPT regularly when grading essays ($M = 4.10$, $SD = .539$). Furthermore, respondents believed that they understood their students' written assignments better after using EssayGPT ($M = 4.03$, $SD = .795$), and they are more confident with regards to written assessment after a series of interactions with EssayGPT ($M = 4.19$, $SD = .654$). Interestingly, the item measuring the overall user satisfaction recorded the highest mean score ($M = 4.39$, $SD = .667$), thus reaffirming the effectiveness of EssayGPT as a formative assessment tool. Additionally, most teachers expressed that they would recommend the use of EssayGPT to their colleagues for assessing written essays in the future ($M = 4.35$, $SD = .755$). Overall, the findings indicated that the respondents perceive EssayGPT positively as a formative assessment tool. They view the customised GPT as an effective assessment assistant that delivers formative feedback to facilitate the teaching and learning process.

AI-Generated Scores Versus Teacher-Rated Scores

Concurrently, a total of 36 scripts were graded by the same English teachers and EssayGPT which resulted in their scores being analysed in terms of descriptive statistics and ICC denoted by *r* as presented in Table 4.

Table 4 Quantitative analysis of the scores between the teachers and EssayGPT

Criteria	C		CA		O		L		Total	
Rater	1	2	1	2	1	2	1	2	1	2
M	3.47	4.14	2.17	2.81	2.17	2.53	1.72	2.72	9.53	12.19
SD	1.000	.867	.811	.786	1.028	.810	1.059	.741	3.194	2.724
r	.470		.589		.765		.495		.661	

Notes: 1 = Teacher; 2 = EssayGPT

$N = 36$

Based on Table 4, across the four assessment criteria, EssayGPT consistently recorded relatively higher mean scores than those assigned by the teachers, thus indicating more leniency by the AI rater. However, the teachers' scores exhibited variations and this reflects their stronger ability to differentiate between variation in essay quality.

To examine the level of agreement between AI-generated scores and teacher-assigned scores, inter-rater reliability was analysed using the ICC interpreted according to the guidelines by Koo and Li (2016). The results reveal poor agreement between EssayGPT and teacher scores for the Content and Language criteria (r_C = .470; r_L = .495). Furthermore, a moderate level of reliability is observed for Communicative Achievement construct and the overall score (r_{CA} = .589; r_{Total} = .661). Despite that, the Organisation criterion exhibits a strong level of agreement between EssayGPT and the teachers (r_O = .765).

Generally these findings suggest concerns regarding the reliability and accuracy of AI raters due to weak to moderate correlations observed across most criteria and the non-correspondence between the EssayGPT's scores and the teachers' judgement. Moreover, the consistently higher scores awarded by EssayGPT compared to the teachers' further highlight the doubts in its effectiveness as a summative assessment tool as this corroborates the leniency in AI-based scoring.

DISCUSSION

The Discussion section provides a discussion of the findings according to the two research questions in the Introduction section. The first subsection discusses the effectiveness of EssayGPT as a formative assessment tool from the theoretical lenses while the second subsection presents an elaborated explanation to justify the alignment between EssayGPT and teachers in terms of summative scoring. Both subsections are also supported with the mechanism of ChatGPT in writing considering that EssayGPT is a customised version of ChatGPT.

EssayGPT As A Formative Assessment Tool

From a theoretical viewpoint, the use of EssayGPT to support formative feedback can be grounded in two theoretical perspectives. Firstly, Winstone and Carless (2020) define feedback as a dialogic process in which learners clarify expectations, seek knowledge and make progress. AI-generated feedback, being interactive in nature, encourages meaningful dialogues between the chatbot and the learner which supports learners' deeper understanding of writing as a process that promotes critical thinking (Abduljawad, 2024). As for EssayGPT, it operationalises the dialogic feedback process by responding to queries related to writing and offering actionable suggestions to improve the quality of written work. As a result, the feedback cycle being an interactive learning process enhances language acquisition (Winstone & Carless, 2020). In line with previous research, ChatGPT-generated feedback is reported to provide more specific and detailed guidance to improve a learner's writing skills (Mahapatra, 2024; Marzuki et al., 2022; Xhao, 2022). In other words, as a formative assessment tool, EssayGPT acts as an agent that fosters learner engagement by responding to questions and offering feedback through its feedback mechanism (Abduljawad, 2024).

Secondly, Barrot (2023) views AI as effective writing tools that can provide immediate and personalised feedback tailored to individual learners' needs throughout different stages of writing. Previous studies suggest that personalised feedback can significantly increase learner engagement and confidence (Maghsudi et al., 2021). Moreover, existing literature asserts the notion that timely and personalised feedback allows learners to address errors more efficiently, thus leading to higher-quality revisions (Raheem et al., 2023). Because of the responsive and interactive nature demonstrated by EssayGPT, users can receive instant yet accurate feedback which can improve motivation in learning and build self-confidence (Jamshed et al., 2024; Washington, 2023). According to Naz and Robertson (2024), ChatGPT supports personalised learning by offering tailored examples, relevant evidence and individualised guidance which creates a continuous feedback cycle. Such sustained interaction allows learners to develop a more nuanced understanding of the content, thus resulting in improved learning outcomes.

ChatGPT demonstrates potential in playing multiple roles as a full participant that can be positioned in different stages in learning writing (Sharples, 2023). First and foremost, it can be employed to brainstorm ideas to generate

alternate perspectives and provide explicit corrective feedback (Mizumoto & Eguchi, 2023; Poole & Coss, 2024; Sharples, 2023). Additionally, by providing feedback, the relationship between the feedback provider (ChatGPT) and receiver (student) can be established which shapes the learning experience (Yang & Carless, 2013). By incorporating discipline-specific guidance into the content of the feedback process, ChatGPT aids the student's development of disciplinary knowledge in writing (Zhang & Liu, 2025). Through iterative cycles of feedback from ChatGPT, students can learn to assess their strengths and weaknesses according to the task criteria via self evaluation, thus encouraging the users to make constructive use of the feedback and become self-regulative learners (Yang & Carless, 2013). Grounded in the notion of dialogic feedback that is immediate and personalised in nature, teachers can leverage the AI chatbot throughout the writing process to address traditional classroom challenges, namely large classroom sizes and heavy assessment workloads (Beaumont et al., 2011; Zhang & Liu, 2025).

EssayGPT As A Summative Assessment Tool

The weak correlations between EssayGPT's scores and human raters' are consistent with prior studies that have reported leniency in AWE scoring (Jackaria et al., 2024; Manning et al., 2025; Shabara et al., 2024). With respect to analytical scoring, the findings align with previous research demonstrating moderate levels of agreement between ChatGPT and human raters for certain criteria (Geckin et al., 2023; Shabara et al., 2024). The observed disparities between AI-generated and human-assigned scores may be attributed to how Automated Writing Evaluation (AWE) systems interpret writing rubrics compared to human raters. AWE systems are highly data-dependent and require a large corpus of human-graded essays to serve as guidelines (Zhang et al., 2020). The input for EssayGPT may be deemed inadequate for the chatbot to grade essays accurately. In addition, AWE systems typically focus on text length, syntax and lexis while overlooking the sociocultural dynamics in writing (Wilson & Roscoe, 2020). Moreover, such systems are more inclined to prioritise grammatical accuracy and word count over meaning which possibly result in inflated scores in EssayGPT's assessments (Shi & Aryadoust, 2022).

A peculiar outcome is that EssayGPT exhibited good reliability in assessing the Organisation construct. It is a finding that does not resonate much with the prior literature which has frequently highlighted AWE systems' limitations in assessing organisation (Barkaoui & Woodworth, 2023; Gardner et al., 2020). In other words, this substantiates the fact that a customised GPT exhibits potential to function as a reliable and accurate assessment tool when supported by carefully engineered prompts. As an AWE system, ChatGPT typically finds it hard to operationlise the understanding of the depth of ideas, context, logic, meaning and purpose, hence explaining the poor correlation between EssayGPT and human raters in the Content, Communicative Achievement and Language criteria (Fu et al., 2022). On the other hand, identification of explicit linguistic devices, namely cohesive devices and connectors, is comparatively easier for ChatGPT, thus suggesting a high level of reliability in the Organisation construct (Crossley et al., 2013). In short, comprehension beyond the surface-level text features and evaluation of coherence are deemed challenging for ChatGPT (Saricaoglu & Bilki, 2025). Assessing CEFR-aligned essays is complex in nature as examiners must not only consider the marking rubric but also the detailed guidelines provided in the writing assessment manual. ChatGPT may either overemphasise or overlook certain aspects in the writing criteria, thus suggesting the use of carefully designed and task-tailored prompts when employing AI chatbots for assessment purposes.

Evidently ChatGPT exhibited a variety of evaluation perspectives autonomously without explicit guidance in the form of prompts (Li & Ng, 2024; Yancey et al, 2023). When there is a lack of shared reference framework, it enables ChatGPT to overemphasise or underemphasise certain dimensions in the evaluation constructs (Garcia-Varela et al., 2025). By prescribing a comprehensive rubric design, rigorous prompt control and good examples, its reliability and accuracy as an assessment tool increases significantly but variability still exists, highlighting the strengths and limitations of rubric-based AI assessment as well as the need of precise prompt instructions (Garcia-Varela et al., 2025; Wu et al., 2024). In other words, perfecting ChatGPT to be an effective assessment tool requires not only pedagogical clarity, but also prompt engineering (Lo, 2023). Despite eliminating the need for large datasets or retraining, ChatGPT still fails to produce perfect stability due to minimal variations caused by student responses (Garcia-Varela et al., 2025). Such responses can be considered

as borderline responses where interpretation is subjective due to the vagueness and complexity of the responses themselves (Garcia-Varela et al., 2025). Therefore, this study suggests the design of a human-AI hybrid system as human judgement is still superior when it comes to evaluation of intricate textual contents (Garcia-Varela et al., 2025; Wang & Demszky, 2023; Yang et al., 2023).

CONCLUSION

In light of the information above, the present study examines the effectiveness of EssayGPT, a customised GPT, as a formative and summative assessment tool in the context of CEFR-aligned English language written curriculum. Findings reveal that the teachers hold a generally positive attitude towards the use of EssayGPT as a formative assessment tool. The customised GPT was perceived by respondents as user-friendly and capable of providing useful feedback for assessing essays, which in turn increased their motivation to grade using the tool and to recommend the AI chatbot to others. However, the study also concludes that EssayGPT is not sufficiently accurate in analytically scoring written assignments based on the weak to moderate correlations between the AI-generated and human-assigned scores in most criteria, thus not yet viable to function as a summative assessment tool.

Despite that, the findings of this study have several implications for educational policymakers and English language practitioners who employ ChatGPT in the teaching and learning process. First and foremost, it is essential to equip teachers with the knowledge of AI, particularly ChatGPT, in order for them to leverage the AI chatbot as an assessment tool. The knowledge of ChatGPT should not be merely restricted to utilising ChatGPT in general, but to explore the possible GPTs that are available in the GPT store for assessment purposes. Additionally, teachers should be made aware that AI tools should not be used as a substitute for teachers, but to be used alongside with teachers as it complements the teacher's role in the classroom. Besides, policymakers should also embrace the idea of ChatGPT by integrating it in professional development programmes to enable teachers to develop GPTs via collaboration to suit the localised contexts.

While the study provides valuable insights into the effectiveness of GPTs, it also acknowledges some limitations that point to opportunities for future research. Firstly, EssayGPT is still at its infancy stage as a product developed adhering to the ADDIE model which has not received sufficient training data and well-engineered prompts. Future research should continue examining the effectiveness of EssayGPT after a series of further revisions. Furthermore, the aim of the research is to study its effectiveness with regard to writing tasks that are pitched at B2 CEFR level only. Future research should expand its focus by considering utilising customised GPTs to evaluate writing tasks pitched at other CEFR levels to determine its capability as a comprehensive assessment tool. Apart from that, the results obtained are only limited to teachers teaching in a Malaysian rural area. In order for the findings to be more generalisable, a larger quantitative sample from teachers in both urban and rural areas that covers a wider geographical distribution is needed to provide richer insights.

In conclusion, the multifaceted insights from the study has contributed to a thorough understanding of a GPT's effectiveness from the perspectives of formative and summative assessment. As this timely study provides an early exploration of a customised GPT, it establishes a baseline for potential research by addressing the aforementioned research directions. By improving the prompt design via effective prompt engineering, the full potential of AI can be more effectively harnessed by future researchers and practitioners.

ACKNOWLEDGEMENTS

The researchers would like to extend appreciation to the 31 English teachers who participated in this study. Their insights and assistance have greatly contributed to the successful completion of this research.

Ethics Approval and Consent to Participate

This study was approved by the Faculty of Education, Universiti Kebangsaan Malaysia. Permission has been granted by the Ministry of Education Malaysia (MoE) to collect empirical data for the research questions related

to the study. All necessary approvals for data collection at district and state level have been obtained. Besides, informed consent was obtained from all participants in the study.

Conflicts of Interest

The authors declare that there is no conflict of interest associated with this publication.

Data Availability

The data will be made available upon requests to the corresponding author.

REFERENCES

1. Abduljawad, S. A. (2024). Investigating the impact of ChatGPT as an AI tool on ESL writing: Prospects and challenges in Saudi Arabian higher education. International Journal of Computer-Assisted Language Learning and Teaching, 14(1), 1-19. https://doi.org/10.4018/IJCALLT.367276
2. Al-khresheh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. Computers and Education: Artificial Intelligence, 6, Article 100218. https://doi.org/10.1016/j.caeai.2024.100218
3. Allehyani, S. H., & Algamdi, M. A. (2023). Digital competences: Early childhood teachers' beliefs and perceptions of ChatGPT application in teaching English as a Second Language (ESL). International Journal of Learning, Teaching and Educational Research, 22(11), 343-363. https://doi.org/10.26803/ijlter.22.11.18
4. Alm, A., & Ohashi, L. (2024). A worldwide study on language educators' initial response to ChatGPT. Technology in Language Teaching & Learning, 6(1), 1141. https://doi.org/10.29140/tltl.v6n1.1141
5. Almashy, A., Ahmed, A., Jamshed, M., Ansari, M., Banu, S., & Warda, W. (2024). Analyzing the impact of CALL tools on English learners' writing skills: A comparative study of errors correction. World Journal of English Language, 14(6), p657. doi:http://dx.doi.org/10.5430/wjel.v14n6p657
6. Alsaweed, W., & Aljebreen, S. (2024). Investigating the accuracy of ChatGPT as a writing error correction tool. International Journal of Computer-Assisted Language Learning and Teaching, 14(1), 1-18. https://doi.org/10.4018/IJCALLT.364847
7. Anderson, J. A., & Ayaawan, A. E. (2023). Formative feedback in a writing programme at the University of Ghana. In A. Esimaje, B. van Rooy, D. Jolayemi, D. Nkemleke, & E. Klu (Eds.), African perspectives on the teaching and learning of English in higher education (pp. 197–213). Routledge.
8. Atasoy, A., Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students' written texts?. Educ Inf Technol 30, 20385–20415. https://doi.org/10.1007/s10639-025-13553-1
9. Barkaoui, K., & Woodworth, J. (2023). An exploratory study of the construct measured by automated writing scores across task types and test occasions. Studies in Language Assessment, 12(1), 1–38. https://doi.org/10.58379/QCFS2805
10. Barrett, A., & Pack, A. (2023). Not quite eye to AI: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. International Journal of Educational Technology in Higher Education, 20(1), Article 59. https://doi.org/10.1186/s41239-023-00427-0
11. Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745
12. Beaumont, C., O'Doherty, M. & Shannon, L. (2011). Reconceptualising assessment feedback: A key to improving student learning? Studies in Higher Education, 36(6), 671–687. https://doi.org/10.1080/03075071003731135
13. Bhutoria, A. (2022). Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model. Computers and Education: Artificial Intelligence, 3, 100068. https://doi.org/https://doi.org/10.1016/j.caeai.2022.100068

14. Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. Teaching English with Technology, 23(1), 23–41. https://doi.org/10.56297/BKAM1691/WIEO1749
15. Branch, R. M. (2009). Instructional design: The ADDIE approach. Springer.
16. Cambridge Assessment English. (2020). SPM – Writing. Video. Putrajaya, Ministry of Education Malaysia.
17. Chua, Y. P. (2013). Mastering research statistics. McGraw-Hill Education.
18. Chung, J. Y., & Jeong, S.-H. (2024). Exploring the perceptions of Chinese pre-service teachers on the integration of generative AI in English language teaching: Benefits, challenges, and educational implications. Online Journal of Communication and Media Technologies, 14(4), e202457. https://doi.org/10.30935/ojcmt/15266
19. Cole, W. R., Arrieux, J. P., Schwab, K., Ivins, B. J., Qashu, F. M., & Lewis, S. C. (2013). Test–retest reliability of four computerized neurocognitive assessment tools in an active duty military population. Archives of Clinical Neuropsychology, 28, 732–742. https://doi.org/10.1093/arclin/act040
20. Council of Europe. (2001). Common European Framework of Reference for Languages: Language, teaching, assessment. Cambridge University Press.
21. Council of Europe. (2026). Historical overview of the development of the CEFR. https://www.coe.int/en/web/common-european-framework-reference-languages/history
22. Creswell, J. W. (2012). Educational research: Planning, conducting and evaluating quantitative and qualitative research (4th ed.). Pearson.
23. Crossley, A. S., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science (Vol. 7926, pp. 134–143). Springer. https://doi.org/10.1007/978-3-642-39112-5_28
24. Curriculum Development Division. (2015). English Language Curriculum Framework Secondary. Ministry of Education Malaysia.
25. Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., Shutaleva, A., & Soomro, R. B. (2024). Extended TAM based acceptance of AI-Powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. Heliyon, 10(8), e29317. https://doi.org/10.1016/j.heliyon.2024.e29317
26. Elkatmış, M. (2024). Chat GPT and Creative Writing: Experiences of Master's Students in Enhancing. International Journal of Contemporary Educational Research, 11(3), 321-336. https://doi.org/10.52380/ijcer.2024.11.3.597
27. English Language Standards and Quality Council (ELSQC). (2015). English language education reform in Malaysia: The roadmap 2015-2025. Ministry of Education Malaysia.
28. Examinations Syndicate. (2020). Format Pentaksiran Bahasa Inggeris Yang Dijajarkan Kepada CEFR. Examinations Syndicate.
29. Examinations Syndicate. (2021). Sijil Pelajaran Malaysia English Language: Instructions For Writing Examiners (To Be Used With Revised Examination). Examinations Syndicate.
30. Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. ELT Forum: Journal of English Language Teaching, 12(1), 44-58. https://doi.org/10.15294/elt.v12i1.64069
31. Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. Computer Assisted Language Learning, 37(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787
32. García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. Education Sciences, 15(8), 946. https://doi.org/10.3390/educsci15080946
33. Gardner, J., O'Leary, M., & Yuan, L. (2020). Artifcial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' Journal of Computer Assisted Learning, 37, 1207–1216. https://doi.org/10.1111/jcal.12577

34. Geckin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. Human raters. Journal of Educational Technology and Online Learning, 6(4), 1096-1108. https://doi.org/10.31681/jetol.1336599
35. Golzar, J., Momenzadeh, S. E., & Miri, M. A. (2022). Afghan English teachers' and students' perceptions of formative assessment: A comparative analysis. Cogent Education, 9(1), 2107297. https://doi.org/10.1080/2331186X.2022.2107297
36. Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. Education Sciences. 13(7), 692. https://doi.org/10.3390/educsci13070692
37. Hadizadeh, A. (2024). ChatGPT, the end of L2 academic writing or a blessing in disguise?. Acuity: Journal of English Language Pedagogy, Literature and Culture, 9(2), 183 -201. https://doi.org/10.35974/acuity.v9i2.3128
38. Hassanien, M. A., Elsamanoudy, A. Z., Ghoneim, F. M., Hegazy, G. A., Amin, H. A., Mustafa, H. N., et al. (2025). Six-step approach for developing customized GPT in medical education. Journal of Advanced Pharmacy Education & Research, 15(2), 107–115. https://doi.org/10.51847/1GHL2Sws00
39. Hua, C., & Wind, S. A. (2019). Exploring the psychometric properties of the mind-map scoring rubric. Behaviormetrika, 46(1), 73-99.
40. Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. System, 34(4), 509-519. https://doi.org/10.1016/j.system.2006.09.001
41. Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. Contemporary Educational Technology, 15(4), ep464. https://doi.org/10.30935/cedtech/13605
42. Jackaria, P. M., Hajan, B. H., & Mastul, A.-R. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. International Journal of Learning, Teaching and Educational Research, 23(2), 478–492. https://doi.org/10.26803/ijlter.23.2.23
43. Jamshed, M., Abu Saleh Md Manjur, A., Md, S., & Wahaj Unnisa, W. (2024). The impact of ChatGPT on English language learners' writing skills: An assessment of AI feedback on mobile. International Journal of Interactive Mobile Technologies (iJIM), 18(19), pp. 18-36. https://doi.org/10.3991/ijim.v18i19.50361
44. Kim, A., & Su, Y. (2024). How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. System, 122, Article 103256. https://doi.org/10.1016/j.system.2024.103256
45. Kohnke, L., & Ulla, M. B. (2024). Embracing generative artificial intelligence: The perspectives of English instructors in Thai higher education institutions. Knowledge Management & E-Learning, 16(4), 653–670. https://doi.org/10.34105/j.kmel.2024.16.030
46. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language learning and teaching. RELC Journal. https://doi.org/10.1177/003368822311628
47. 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
48. Kotmungkun, S., Chompurach, W., & Thaksanan, P. (2024). OpenAI ChatGPT vs Google Gemini: A study of AI chatbots' writing quality evaluation and plagiarism checking. English Language Teaching Educational Journal, 7(2), 90-108. https://doi.org/10.12928/eltej.v7i2.11572
49. Kutluk, F. A., & Gülmez, M. (2014). A Research about Mobile Learning Perspectives of University Students who have Accounting Lessons. Procedia - Social and Behavioral Sciences, 116, 291-297. https://doi.org/https://doi.org/10.1016/j.sbspro.2014.01.210
50. Larsen-Hall, J. (2010). A Guide to Doing Statistics in Second Language Research Using SPSS (1st ed.). Routledge.
51. Li, S., & Ng, V. (2024). Automated essay scoring: Recent successes and future directions. Proceedings of the 33rd International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea.
52. Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), 410.

53. Maghsudi, S., Lan, A., Xu, J., & Van der Schaar, M. (2021). Personalized education in the AI era: What to expect next? IEEE Signal Processing Magazine, 38(3), 37-50. https://doi.org/10.48550/arXiv.2101.10074
54. Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. Smart Learning Environments, 11(1), 9. https://doi.org/10.1186/s40561-024-00295-9
55. Manning, J., Baldwin, J., & Powell, N. (2025). Human versus machine: The effectiveness of ChatGPT in automated essay scoring. Innovations in Education and Teaching International, 62(5), 1500–1513. https://doi.org/10.1080/14703297.2025.2469089
56. Marzuki, S., Widiati, U., Rusdin, D., Darwin, R. & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. Cogent Education, 10(2), 1–17. https://doi.org/10.1080/2331186X.2023.2236469
57. Mehrens, W. A., & Lehmann, I. J. (1991). Measurement and evaluation in education and psychology (4th ed.). Wadsworth Publishing.
58. Ministry of Education Malaysia. (2025, October 31). Risalah maklumat asas pendidikan. Ministry of Education Malaysia. https://emisonline.moe.gov.my/risalahmap/
59. 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), 1-13.
60. Mohamad Uri, N. F., Mohmud, M. S., & Abd Aziz, M.S. (2025). Challenges in CEFR Adoption: Teachers' Understanding and Classroom Practice. International Journal of Modern Languages and Applied Linguistics, 7(1). https://doi.org/10.24191/ijmal.v7i1.7522
61. Mohd Salleh, N. F. A., Nimehchisalem, V., Jalaluddin, I., & Mukundan, J. (2023). Washback effects of School-Based Assessment (SBA) on Malaysian secondary school students' English language learning. PASAA Journal, 66, 168-201.
62. Naz, I., & Robertson, R. 2024. Exploring the Feasibility and Efficacy of ChatGPT3 for Personalized Feedback in Teaching. The Electronic Journal of e-Learning, 98-111, https://doi.org/10.34190/ejel.22.2.3345
63. Nkhobo, T., & Chaka, C. (2023). Student-written versus ChatGPT-generated discursive essays: A comparative coh-metrix analysis of lexical diversity, syntactic complexity, and referential cohesion. International Journal of Education and Development using Information and Communication Technology (IJEDICT), 19(3), 69-84.
64. OpenAI. (2025). GPTs FAQ. https://help.openai.com/en/articles/8554407-gpts-faq
65. Pang, Y. (2022). The role of web-based flipped learning in EFL learners' critical thinking and learner engagement. Frontiers in Psychology. 13. https://doi.org/10.3389/fpsyg.2022.1008257
66. Poole, F. J., & M. D. Coss. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). Journal of Technology & Chinese Language Teaching, 15(1). https://doi.org/10.35542/osf.io/3r2zb
67. Punar Özçelik, N., & Yangın Ekşi, G. (2024). Cultivating writing skills: The role of ChatGPT as a learning assistant—a case study. Smart Learning Environments, 11(1), 10. https://doi.org/10.1186/s40561-024-00296-8
68. Raheem, B. R., Anjum, F., & Ghafar, Z. N. (2023). Exploring the profound impact of artificial intelligence applications (Quillbot, Grammarly and ChatGPT) on English academic writing: A Systematic Review. International Journal of Integrative Research, 1(10), 599–622. DOI: 10.59890/ijir.v1i10.366
69. Ronan, P., & Schneider, G. (2023). Can Chat GPT solve a linguistics exam? https://arxiv.org/abs/2311.02499
70. Rozana Sani. (2016). Journey to master English. 18 July. The New Straits Times Online. https://www.nst.com.my/news/2017/03/159164/journey-master-english
71. Sabak Bernam District Council. (2025, November 24). Brief history of Sabak Bernam district. Sabak Bernam District Council. https://www.mdsb.gov.my/index.php/en/visitors/sabak-bernam-info
72. Saricaoglu, A., & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. Annual Review of Applied Linguistics, 45, 253–273. doi:10.1017/S0267190525100160

73. Shabara, R., ElEbyary, K., & Boraie, D. (2024). Teachers or ChatGPT: The issue of accuracy and consistency in L2 Assessment. Teaching English with Technology, 24(2), 71–92. https://doi.org/10.56297/vaca6841/LRDX3699/XSEZ5215
74. Shaikh, S., Yayilgan, S. Y., Klimova, B., & Pikhart, M. (2023). Assessing the usability of ChatGPT for formal English language learning. European Journal of Investigation in Health, Psychology and Education, 13(9), 1937-1960. https://doi.org/10.3390/ejihpe13090140
75. Sharples, M. (2023). Towards social Generative AI for education: Theory, practices and ethics. Learning: Research and Practice, 9(2), 159–167. https://doi.org/10.1080/23735082.2023.2261131
76. Shi, H., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. Education and Information Technologies, 28. https://doi.org/10.1007/s10639-022-11200-7
77. Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. Computer Assisted Language Learning, 1–28. https://doi.org/10.1080/09588221.2025.2454541
78. Susnjak, T. (2022). ChatGPT: The end of online exam integrity? (2212.09292). arXiv. https://doi.org/10.48550/arXiv.2212.09292
79. Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. Heliyon, Elsevier 4(5), 1-32.
80. Washington, J. (2023). The impact of generative artifcial intelligence on writer's self-efcacy: A critical literature review. Social Science Research Network. https://doi.org/10.2139/ssrn.4538043
81. Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. Journal of Educational Computing Research, 58(1), 87–125. https://doi.org/10.1177/0735633119830764
82. Winstone, N., & Carless, D. (2020). Designing effective feedback processes in higher education. Routledge.
83. Wu, X., Saraf, P. P., Lee, G., Latif, E., Liu, N., & Zhai, X. (2024). Unveiling scoring processes: Dissecting the differences between LLMs and human graders in automatic scoring. arXiv.
84. Xhao, X. (2022). Leveraging Artifcial Intelligence (AI) technology for English writing: Introducing wordtune as a digital writing assistant for EFL writers. RELC Journal, 54(3), 890–894. https://doi.org/10.1177/00336882221094089
85. Yancey, K. P., LaFlair, G. T., Verardi, A. R., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49
86. Yang, M., & D. Carless. (2013). The feedback triangle and the enhancement of dialogic feedback processes. Teaching in Higher Education, 18(3), 285–297. https://doi.org/10.1080/13562517.2012.719154
87. Zhang, M., Yao, L., Haberman, S., & Dorans, N. (2020). Assessing scoring accuracy and assessment accuracy for spoken responses. In K. Zechner & K. Evanini (Eds.), Automated Speaking Assessment: Using Technologies to Score Spontaneous Speech (pp. 32–57). https://doi.org/10.4324/9781315165103-3
88. Zhang, Y., & Liu, Y. (2025). Designing ChatGPT-mediated feedback activities in EFL writing: a design-based study of the dialogic feedback triangle. Assessment & Evaluation in Higher Education, 1–21. https://doi.org/10.1080/02602938.2025.2571846
89. Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. Computers & Education, 143, 103668. https://doi.org/10.1016/j.compedu.2019.103668
90. Zindela, N. (2023). Comparing measures of syntactic and lexical complexity in Artificial Intelligence and L2 human-generated argumentative essays. International Journal of Education and Development using Information and Communication Technology, 19(3), 50-68.