Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can AI replace Teachers? Comparing AI and Human Teachers' ratings and feedback on students' essays

Herford Rei B. Guibangguibang

herfordguibang-guibang@jrmsu.edu.ph

Jose Rizal Memorial State University https://orcid.org/0000-0001-9950-1689

Kian C. Hesula

Jose Rizal Memorial State University https://orcid.org/0009-0001-5509-2390

Research Article

Keywords: ChatGPT, Bing, CiCi, Human - AI Ratings, Human - AI Feedback

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10135718/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Can AI replace Teachers? Comparing AI and Human Teachers' ratings and feedback on students' essays

Abstract

This study aimed to evaluate the consistency between human and AI-generated ratings on ESL student essays and to compare the manner and focus of their feedback. Using a standardized rubric, three human raters and three AI tools assessed essays based on content, organization, language, and mechanics. Findings revealed that while the average ratings from both groups were generally comparable, the inter-rater consistency between humans and AI was low, indicating that the two sets of raters did not closely align in how they assigned scores. Despite this, statistical analysis using the Wilcoxon signed-rank test showed no significant difference in overall ratings, suggesting that both groups arrived at similar general evaluations despite differing in scoring patterns. In terms of feedback, human raters typically provided direct and corrective remarks focused on grammar and structure. In contrast, AI tools offered more varied and structured comments that addressed broader aspects of writing, such as clarity, organization, and idea development, often balancing praise with critique. The study concludes that while AI tools like ChatGPT, CiCi, and Bing show potential in supporting writing assessment, they should be used to complement—not replace—human evaluation. Human oversight remains essential to ensure accuracy, contextual understanding, and pedagogical appropriateness in student assessment. Continued training for teachers on how to effectively use and interpret AI-assisted feedback is recommended to maximize its educational value.

Keywords: ChatGPT; Bing; CiCi; Human - AI Ratings; Human - AI Feedback

1. Introduction

It has been frequently asserted in recent years that artificial intelligence (AI) is the "new oil" or, as the UNESCO Director General stated in her keynote address for Mobile Learning Week in 2019, the greatest invention since the Paleolithic era (Holmes & Tuomi, 2022).

Recent years have seen a major development in natural language processing (NLP) thanks to large language models like the Generative Pre-trained Transformer (GPT-3) (Floridi & Chiriatti, 2020, in Kasneci et al., 2023). These models can produce text that is human-like, respond to inquiries, and perform other language-related tasks with great accuracy since they have been trained on vast volumes of text data.

In the context of education, AI has become a debatable topic regarding its integration in teaching and learning. While academic institutions worldwide engage in discussions about AI's role in assessment, the focus often centers on concerns about academic integrity, rather than its ability to improve the reliability of the assessment process (Awidi, 2024).

AI offers faster, more objective essay assessment, reducing human bias and saving time. However, it may struggle to grasp the nuances of language, tone, and context that human teachers intuitively understand and address in their feedback.

This brings to the question: Is it time for AI to replace teachers in assessing essays? While AI can certainly be a useful tool in the grading process, it should not completely replace human judgement and expertise. Teachers play a crucial role in nurturing students' growth and learning, a role that cannot entirely fulfilled by AI. The optimal solution may lie in a balanced approach, where AI assists teachers in their grading tasks, but the final authority remains with the human educators. Therefore, the core issue is not AI itself, but rather how to effectively integrate and maximize its use while ensuring its reliability and addressing potential biases. AI is here – there is no turning back.

The aim of this research is to assess the reliability of AIs in evaluating student essays by comparing its performance or ratings to traditional teacher's ratings and understand how AIs and human teachers construct its remarks. Through this comparison, the study aimed to shed light on the strengths and limitations of AIs in essay assessment, offering valuable insights for educational stakeholders and policymakers regarding the integration of AI tools in academic settings.

Artificial Intelligence

AI research began with John McCarthy, who coined the term at a 1965 Dartmouth conference (Klein & Frana, 2021). While definitions vary, the Organization for Economic Co-operation and Development (2019) as cited in

Huang et al. (2023) defines AI as machine-driven systems that make predictions, suggestions, or decisions based on human-defined goals.

IBM's Data and AI Team (2023) classifies AI into three types based on capability. The first, Artificial Narrow Intelligence (ANI), or Weak AI, is the only type that currently exists. It performs specific tasks efficiently—like Siri, Alexa, Watson, and ChatGPT—but lacks broader understanding. The second, Artificial General Intelligence (AGI), remains theoretical and refers to AI with human-like adaptability across various tasks. The third, Artificial Superintelligence (ASI), is also theoretical and describes AI that would surpass human intelligence in all areas, including reasoning, emotion, and autonomy.

AI in the Context of Education

Artificial Intelligence (AI) is gaining prominence in language education, though its trends, challenges, and applications remain underexplored. Huang et al. (2023) noted a rise in AI-related studies in this field. The 2018 Horizon Report identified AI and adaptive learning as emerging technologies, forecasting 43% growth between 2018 and 2022, with further expansion anticipated in 2019 (Educause, 2018; 2019).

In education context, AI has potentially interesting applications in educational administration, students and teachers' support. Educators are only beginning to explore the pedagogical potential of AI applications in supporting learners throughout the student life cycle. However, alongside these promising opportunities, the rise of AI in higher education also brings new ethical concerns and risks (Zawacki-Richter et al., 2019; Yang et al., 2025).

Since its initial public launch on November 30, 2022, ChatGPT attracted over one million users within just a week. This generative AI tool surprised the world with its advanced ability to handle highly complex tasks. In the realm of education, ChatGPT's impressive capabilities have sparked a range of reactions among educators, as this technological breakthrough appears to transform traditional educational practices (Baidoo-anu & Owusu, 2023). Despite some criticism, AIs present notable benefits for education.

AI has the potential to enhance the learning outcomes, boost student productivity, and increase engagement in educational tasks (Adiguzel et al., 2023; Iqbal et al., 2022). It can serve as a tool for delivering personalized feedback and educational tutorials. Additionally, teachers can utilize ChatGPT to interact with students, provide tailored feedback, facilitate interactive discussions, design lessons and assessments, and explore innovative methods for complex concepts (Iqbal et al., 2022; Rahman & Watanobe, 2023).

Generative AI tools can help address challenges educators face, such as time constraints and limited resources. By assisting with tasks like grading and feedback, tools like ChatGPT may reduce teacher workload

and stress, allowing more focus on instruction and classroom management (Kopp & Thomsen, 2023). They also offer immediate feedback on student work and can generate customized teaching materials, potentially reducing the need for expensive textbooks (Mondal et al., 2023).

The Free Generative AI and its features: ChatGPT 3.5, CiCi, and Bing AI

Generative AI tools like ChatGPT 3.5, CiCi, and Bing are widely used today as free, open-access platforms that support education. ChatGPT, developed by OpenAI, enhances learning by offering personalized explanations and interactive support. It also aids teachers by generating lesson plans and materials, reducing administrative workload and allowing more focus on instruction (Adiguzel et al., 2023; Iqbal et al., 2023).

CiCi is an AI platform designed for education, offering interactive learning and real-time feedback. It creates adaptive learning paths based on student performance and customizes content to fit individual needs (Magday Jr et al., 2025). CiCi also promotes collaboration through group discussions and peer feedback, fostering a more engaging classroom environment.

Bing AI, though primarily a search engine, offers educational benefits through its integration with Microsoft tools. Its features—such as contextual search and knowledge graphs—help students and educators access accurate information quickly, supporting research and learning. As noted by Crawford et al. (2023), these AI-driven functions enhance efficiency and can enrich teaching by providing deeper, more comprehensive content.

Generative AI tools like ChatGPT, CiCi, and Bing AI are transforming education by making learning more personalized, efficient, and accessible. As they evolve, they offer promising potential to enhance teaching through innovation and inclusivity.

On Assessment: Teachers' Ratings and Feedback

Teaching was once considered a low-stress profession with flexible schedules and academic freedom. However, due to evolving political and socio-economic changes, it has become increasingly demanding. Over the past two decades, many of the profession's earlier advantages have diminished (Khan et al., 2014).

Teaching requires balancing classroom management, effective instruction, and continuous learning. Beyond maintaining discipline and designing inclusive lessons, a key responsibility is assessment—not just grading, but evaluating student progress through various methods to guide instruction and improve outcomes (Moss & Brookhart, 2019).

In essay assessments, teachers devote substantial time and effort to each stage—from designing prompts and clear instructions to setting criteria, scoring, and offering personalized feedback. The way teachers structure

writing activities and deliver feedback significantly influences how students approach writing, perceive feedback, and revise their work (Lockhart & Ng, 1995, as cited in Engeness, 2018).

Remarks are a vital part of assessment, offering feedback that highlights students' strengths and areas for improvement (Guibangguibang, 2020). This process is especially important for academically struggling students, as it helps them identify and address their weaknesses to support their progress.

Writing is a key skill for English language learners, and corrective feedback is widely recognized as essential in supporting second language (L2) development (Lee, 2017). It helps address common writing issues such as composition, style, and organization. Written corrective feedback (WCF), as defined by Al Shahrani and Storch (2013), specifically targets language errors.

Feedback is information provided by a teacher about a student's performance (Hattie & Timperley, 2007, as cited in Wisniewski et al., 2019), and its tone greatly influences its effectiveness. Shin et al. (2022) classify feedback as positive, negative, constructive, or destructive. Positive feedback highlights strengths, while negative identifies areas for improvement. Constructive feedback combines both, offering guidance for growth. In contrast, destructive feedback criticizes the individual without suggesting improvements, potentially harming motivation. As Bejtic (2024) emphasizes, tone plays a critical role in fostering student development.

According to Black and Wiliam (2009, as cited in Engeness, 2018), formative assessment feedback should clarify learning goals, elicit understanding, promote improvement, and encourage peer and self-assessment. Hattie and Timperley (2007, also cited in Engeness) expanded this by outlining three feedback questions—"Where am I going?", "How am I going?", and "Where to next?"—aligned with feed up, feedback, and feed forward. They also identified four feedback levels: task, process, self-regulation, and personal. Feedback focusing on tasks and processes was most effective, while purely personal or evaluative feedback was least helpful.

Previous Studies

Recent studies highlight the potential of AI models like ChatGPT to advance language research and assessment. Mizumoto and Eguchi (2023) found that GPT-3 reliably scored 12,100 TOEFL11 essays, with improved accuracy when combined with linguistic features. Their findings suggest that AI can effectively support human raters and enhance writing assessment practices.

Dai et al. (2024) explored the use of ChatGPT as a tool for providing student feedback. Recognizing the role of effective feedback in learning, the study assessed ChatGPT's ability to produce natural, comprehensive responses. The researchers compared the readability and quality of AI-generated feedback with that of human

instructors and evaluated it using a theoretical framework. Results showed that ChatGPT offered more detailed and coherent feedback, aligned well with instructor evaluations, and supported students' skill development by focusing on task performance.

In the study of Bouziane and Bouziane (2024) indicate that ChatGPT excels in addressing specific language-related issues such as grammar, spelling, and sentence structure, showcasing its proficiency in rule-based corrections. However, human evaluators maintain an edge in areas requiring nuanced understanding, particularly in maintaining thematic consistency.

Escalante et al. (2023) examined the ethical and effective use of generative AI tools, such as ChatGPT (GPT-4), in supporting student writing. Two longitudinal studies involving English as a New Language (ENL) learners compared AI-generated feedback with traditional human feedback. Results showed no significant difference in learning outcomes, and students expressed balanced preferences for both. The findings suggest that AI feedback can complement writing instruction when paired with human input, enhancing the strengths of each.

Guo (2024) introduced EvaluMate, an AI-assisted peer review platform using ChatGPT to help students give more effective feedback. The tool offers scaffolding to improve comment quality, showing promise for writing instruction. While beneficial, the study also notes limitations and suggests further research on student engagement and learning impact.

Seßler et al. (2025) compared five large language models (LLMs)—including GPT-3.5, GPT-4, o1-preview, LLaMA 3-70B, and Mixtral 8x7B—to human teacher ratings on 20 German student essays. Among them, closed-source models like GPT-4 and o1 showed stronger alignment with teacher evaluations, especially on language aspects. While o1 had the highest correlation, it tended to overrate. The study suggests that LLMs can aid in writing assessment but still require refinement for evaluating content quality.

While AI tools like ChatGPT have been widely studied for essay scoring and feedback generation, limited research has explored how human and AI evaluators construct their comments, particularly in terms of language and pragmatics. Moreover, few comparative studies focus specifically on ESL learners' essays in diverse contexts such as the Philippines. This study addresses these gaps and poses the following research questions:

1. How do human raters versus AI tools (ChatGPT 3.5, Bing AI, and CiCi) rate ESL students' essays?
2. What is the reliability of human and AI tools' (ChatGPT 3.5, Bing AI, and CiCi) rating of ESL students' essays?
3. What is the inter-rater reliability of human and AI tools' (ChatGPT 3.5, Bing AI, and CiCi) rating of ESL students' essays?

4. Is there a significant difference between human and AI tools' (ChatGPT 3.5, Bing AI, and CiCi) rating of ESL students' essays?
5. How do human versus AI tools (ChatGPT 3.5, Bing AI, and CiCi) provide feedback on ESL students' essays as to:
   5.1 focus, and
   5.2 manner?

2. Theoretical background

This study is grounded in the TPACK framework (Mishra & Koehler, 2006), which emphasizes the integration of content, pedagogy, and technology for effective teaching. It underscores that meaningful use of tools like generative AI requires understanding how subject matter, teaching methods, and digital technologies work together to enhance learning.

This study uses the TPACK framework to evaluate whether generative AI tools like ChatGPT can deliver feedback and assessment comparable to human teachers. By comparing AI and teacher ratings, it examines how well AI aligns with pedagogical goals and content accuracy, assessing its potential to support writing instruction without compromising quality.

3. Materials and Methods

3.1 Research Design

This study used a quantitative-qualitative comparative design, focusing on describing the sample without manipulating variables (Siedlecki, 2020). Quantitative methods compared AI and human ratings, while qualitative content analysis examined differences in how feedback was constructed.

Quantitative data were drawn from the essay ratings of three human raters and three AI tools (ChatGPT 3.5, CiCi, and Bing). For the qualitative analysis, researchers examined the raw feedback from both groups, coding remarks using Jakobson's (2022) model of written feedback. Content analysis was used to identify recurring patterns, themes, and key elements, providing insights into how both human and AI raters assessed writing. This method, as supported by Gómez-Galán (in Tukan & Mukarto, 2024), was effective for analyzing feedback focus and manner through Jakobson's communication framework.

3.2 Research Participants and Sampling Technique

Table 1 presents the participants of the study which involved 20 ESL students from Jose Rizal Memorial State University – Tampilisan Campus, a public university in the northwestern part of Mindanao, Philippines. These students are enrolled in a Technical Writing course that requires them to produce academic essays. Participants were selected from the initial 43 participants using a stratified sampling technique: half of the essays were randomly chosen from students whose pre-rated scores were above the average, and the other half from those below the average. This approach ensured diverse input for both human raters and AI tools, allowing for a more robust test of rating consistency.

Table 1 Participants of the Study

Initial Participants (During the Essay Writing)	Actual Participants (After Stratified Sampling Technique based on above-below the average in the pre-rated scores of their essays)		Total Participants
	Above the average score	Below the average score	
43	10	10	20

The three human raters who participated in the study were all holders of a master's degree in Language and possessed extensive experience in teaching English, each with at least five years of professional practice. Their academic background and teaching experience qualified them to evaluate student essays reliably, ensuring informed and pedagogically grounded assessments. Their involvement added credibility to the human scoring process and provided a meaningful benchmark for comparing AI-generated evaluations.

3.3 Research Instrument

To assess participants' writing skills, this study adapted a validated essay rubric from Carroll and West (1989), as cited in Tribble (1996), and previously used by Tuyen et al. (2019) and Jackaria, Hajan, and Mastul (2024). The rubric covered four equally weighted criteria: content, organization, language use, and mechanics—each rated on a 1 (poor) to 4 (excellent) scale. It was validated by three experts and tested for reliability before use.

To identify the feedback made by the human raters and AI tools on participants' writing skills, the study followed Jakobson's (2022) model for analyzing teachers' written feedback on writing. This feedback analysis model focused on two main areas: focus (what the teachers comment on) and manner (how the comments are given).

Table 2 outlines Jakobson's feedback model. For Focus, teacher comments were categorized into: Content, Organization, Language Accuracy, Reinforcement of Learning Outcomes, Reinforcement of Learning Materials, and No Focus. For Manner, feedback was classified as Information, Criticism, Suggestions, Praise, and Politeness. Information included statements or questions about content. Criticism reflected negative evaluations

of problems. Suggestions appeared as statements, questions, or imperatives. Praise was either specific or general, while Politeness referred to positive, interpersonal expressions not necessarily linked to praise.

Table 2. Feedback Analysis Model by Jackobson (2022) on Teacher's Feedback in Writing

Feedback Categories ON FOCUS	Use	Example of Teacher's Feedback/Comments
1. Content	The teacher pays attention to ideas in the writer's text.	*You also talk about what you want the character to do…*
2. Organization	Focuses on text structure, including introduction, conclusion, and paragraphing.	*It is a paragraph with a good structure…*
3. Language Accuracy		
3.1 Grammatical Structures		*Word order is difficult to learn in the beginning…*
3.2 Vocabulary		*You show variety in your vocabulary…*
3.3. Spelling	The teacher focuses about linguistic aspects.	*Work on getting the spelling right…*
3.4 Mechanics		*You need to practice writing rules such as when there should be a full stop and a capital letter…*
4. Language in General		*There is a variety in your language…*
5. Reinforcement of Learning Outcomes	The teacher focuses on reemphasizing the learning outcomes.	*The course aims for you to be able to write coherent texts that the reader understands…*
6. Reinforcement of Learning Materials	The teacher directs students to infer to a concrete material as example.	*Read ads in newspapers and online and see what they are supposed to look like.*
7. No Focus	The teacher concerns on general writing	*Write a little every day.*
Feedback Categories ON MANNER	**Use**	**Example of Teacher's Feedback/Comments**
1. Information		
1.1 Giving Information	Provides descriptive, non-evaluative statements about the current text without suggesting changes.	*The reader understands what you mean.*
1.2 Asking for Information about the Content	Questions seeking additional information about the content beyond what the current text provides.	*What style did you use?*
2. Criticism	This expresses a negative value where the teacher describes a problem.	*Your sentences are sometimes very long…*
3. Suggestions	Advises future changes or improvements, leaving it to the student to decide on making them.	*You might consider breaking your long sentences into shorter ones to improve clarity.*
3.1 Statements	These are non-evaluative neutral comments	*It is good to include the date and place as well.*
3.2 Questions	These differ from content-related questions as they focus on overall improvement rather than specific content.	*How can you vary your sentences more?*
3.3. Imperatives	These are comments that require a change in the text or action from the student.	*Rewrite so that it becomes more varied!*
4. Praise	This is a positive comment and can be interpreted as the opposite of criticism.	
4.1 Specific Praise	This is a positive comment targeting a specific aspect the teacher highlights.	*Word order is correct and varied; well done!*
4.2 General Praise	This is an unspecific positive comment.	*Good job!*
5. Politeness	This is a positive expression reflecting interpersonal communication, distinct from praise.	*Thank you for your text!*

3.4 Data Gathering

The essays used in this study were written by students in a technical writing course handled by one of the researchers. Prior to the activity, ethical protocols were strictly followed. Informed consent was obtained, participation was voluntary, and students were informed of their right to withdraw at any time without consequence. To ensure confidentiality, identifying information was removed and coded. Participants were briefed on the study's purpose and assured that their grades would remain unaffected.

During the essay-writing session, students were timed and supervised to encourage authentic, independent writing. They were asked to complete a descriptive essay on the topic of AI in education, focusing on its benefits and drawbacks. There was no word limit, allowing participants to freely develop their ideas. All handwritten essays were later transcribed verbatim into Microsoft Word documents for analysis. A total of 43 essays were collected and scored. Of these, 23 were used to establish the reliability of the scoring rubric, while the remaining 20 were included in the main data analysis. To ensure diverse writing levels, the essays were selected to represent both higher- and lower-performing students. Human raters independently evaluated the essays and provided feedback.

Student essays were submitted to ChatGPT 3.5, CiCi, and Bing AI using refined prompts (see Figs. 1–3) developed through trial runs. These prompts included task instructions, criteria, the rubric, essay count, and a request for summary tables. The process followed the same approach used by Jackaria, Hajan, and Mastul (2024). The AI-generated evaluations were then analyzed both qualitatively and statistically.

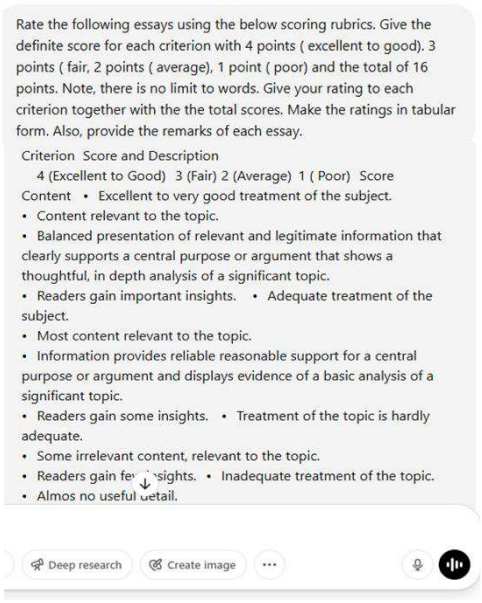


Fig. 1 Prompt inputted to ChatGPT 3.5 to rate and provide remarks on students' essays

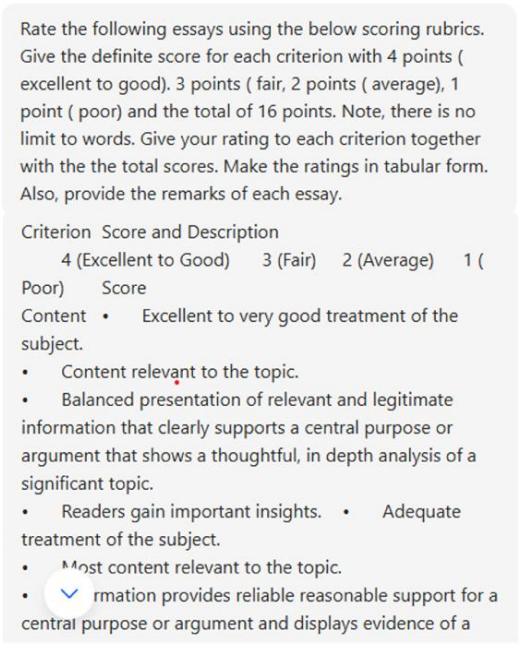


Fig. 2 Prompt inputted to CiCi to rate and provide remarks on students' essays

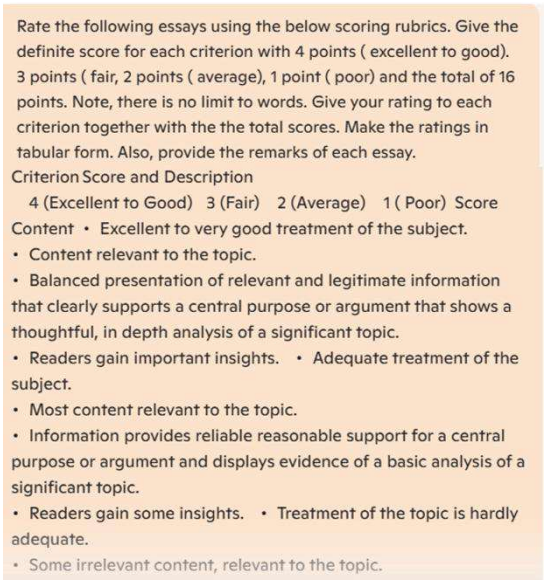


Fig. 3 Prompt inputted to Bing AI to rate and provide remarks on students' essays

3.5 Data Analysis

The written responses from ChatGPT 3.5, Bing AI, and CiCi were collected, with careful documentation of the ratings assigned for each criterion. To assess the level of agreement among the three human raters as well as among ChatGPT 3.5, Bing AI, and CiCi, the intraclass correlation coefficient (ICC) was employed. Specifically, the study used ICC estimates with 95% confidence intervals, based on a mean-rating (k = 3), absolute-agreement,

two-way mixed-effects model. Additionally, ICC was applied to examine the consistency of ratings between the human raters and ChatGPT across each criterion. ICC was chosen as it is widely recognized as a meaningful indicator of inter-rater agreement (Laschinger, 1992). The interpretation of the ICC values followed the scale proposed by Koo and Li (2016), classifying results as poor (below 0.5), moderate (0.5–0.75), good (0.75–0.9), and excellent (0.90 and above). This method was also utilized by Jackaria, Hajan, and Mastul (2024).

It is important to note that ICC reflects only the linear consistency between two datasets; therefore, even when datasets differ, a high reliability coefficient may still result if the differences are consistent. To address this limitation, the Wilcoxon signed-rank test was used to further investigate differences between paired samples. This non-parametric test helped determine whether the differences in ratings between AI and human raters were statistically significant. Finally, all results were systematically tabulated, analyzed, and discussed to draw conclusions regarding the consistency and reliability of AI-generated ratings in comparison to human ratings.

On the other hand, the written feedback of AIs and human teachers, content analysis was used. The researchers utilized a table to organize and systematically analyze the collected data. Feedback from both human raters and AI tools on students' writing exemplars were first identified and classified into predefined categories based on Jakobson's outlined model, focusing on aspects such as content, organization, language accuracy, reinforcement of learning outcomes, reinforcement of learning materials, and no focus. The manner of feedback delivery was also coded into categories such as information, criticism, suggestions, praise, and politeness, along with their subtypes. Once classification was completed, the researchers analyzed each feedback type to determine patterns and trends in both the focus and manner of feedback. This analysis was then used to address the study's research questions, providing insights into the similarities and differences between AI-generated and human-generated feedback.

4. Results

4.1 On Human versus AI tools ratings on ESL students' essay

Table 3 presents the ratings given by human raters and AI tools on 20 student essays, assessed using the rubric by Caroll and West (1986). The human raters scored the essays with weighted means of 3.07 (SD = 0.73) for content, 2.83 (SD = 0.92) for organization, 2.73 (SD = 0.66) for language, and 2.72 (SD = 0.80) for mechanics — all falling under the interpretation of Fair. The overall weighted mean for human raters was 2.84 (SD = 0.79), likewise interpreted as Fair.

Similarly, the AI tools produced ratings with weighted means of 2.78 (SD = 0.80) for content, 2.77 (SD = 0.76) for organization, 2.67 (SD = 0.72) for language, and 2.62 (SD = 0.71) for mechanics, with an overall

weighted mean of 2.71 (SD = 0.75) — also under the Fair category. These results show that both human and AI raters provided relatively comparable evaluations, although human raters generally gave slightly higher scores than the AI tools.

This finding slightly contrasts with the results reported by Jackaria, Hajan, and Mastul (2024), who observed that ChatGPT tended to give marginally higher ratings than human evaluators.

Notably, among the AI tools assessed, CiCi demonstrated remarkable consistency, assigning the same mean score of 2.10 across all four criteria: content, organization, language, and mechanics. This uniformity may suggest a rigid scoring pattern or a limitation in CiCi's capacity to differentiate performance levels across diverse writing components.

Table 3. Means and Interpretation of the human and AI tools essay ratings

Rubric Criterion	Human Raters						Weighted Mean	Standard Deviation	Interpretation	AI Raters						Weighted Mean	Standard Deviation	Interpretation
	Teacher 1		Teacher 2		Teacher 3					ChatGPT 3.0		Bing AI		CiCi				
	x̄	SD	x̄	SD	x̄	SD				x̄	SD	x̄	SD	x̄	SD			
Content	3.10	0.64	2.90	0.85	3.20	0.69	**3.07**	0.73	**F**	3.25	0.63	3.00	0.64	2.10	0.64	**2.78**	0.80	**F**
Organization	3.40	0.68	2.20	0.69	2.90	0.96	**2.83**	0.92	**F**	3.15	0.58	3.05	0.60	2.10	0.64	**2.77**	0.76	**F**
Language	3.10	0.44	2.40	0.68	2.70	0.65	**2.73**	0.66	**F**	3.10	0.55	2.80	0.61	2.10	0.64	**2.67**	0.72	**F**
Mechanics	2.75	0.63	2.30	0.73	3.10	0.85	**2.72**	0.80	**F**	3.15	0.58	2.60	0.50	2.10	0.64	**2.62**	0.71	**F**
General Weighted Mean							**2.84**	0.79	**F**	**General Weighted Mean**						**2.71**	0.75	**F**

** Note: Poor "**P**" (1.00-1.75); Average "**A**"(1.76-2.50); Fair "**F**"(2.51-3.25); Excellent-Good "**EG**"(3.26-4.00)*

4.2 On Reliability of Human and AI Tools Raters in Rating Student Essays

Table 4 displays the level of inter-rater consistency among the two groups of raters — human raters and AI raters — based on the criteria outlined in the essay writing rubric adapted from Carrol and West (1989).

For the human raters, the results indicate excellent consistency in the Content criterion (ICC = 0.911), Organization (ICC = 0.865), and Language (ICC = 0.765), while Mechanics showed good consistency (ICC = 0.715). The overall ICC score for the human group is excellent, at 0.815. This suggests that human raters were highly reliable in their evaluations.

Meanwhile, the AI raters demonstrated a higher level of agreement, with an overall ICC of 0.879, which is interpreted as excellent consistency. Individually, the consistency across content (0.897), organization (0.888), language (0.876), and mechanics (0.855) all fall within the excellent range. This suggests that the AI tools applied a more uniform evaluation standard across the four criteria, possibly reflecting algorithmic precision or less subjective interpretation compared to human evaluators.

In support to this finding, Latif and Zhai (2023) found that ChatGPT has the potential to be effectively adapted for evaluating students' written outputs, such as essays. Their study highlighted that with appropriate

training or fine-tuning, the AI model can provide consistent, reliable, and human-like scoring, making it a promising tool for automated writing assessment tasks.

Table 4 Intraclass Correlation results indicating reliability of the Human and AI tools essay ratings

Rubric Criterion	**Reliability Coefficient**			
	Human raters	Interpretation	AI Raters	Interpretation
Content	0.911	Good	0.897	Good
Organization	0.865	Good	0.888	Good
Language	0.765	Good	0.876	Good
Mechanics	0.715	Good	0.855	Good
Overall Score	**0.815**	**Good**	**0.879**	**Good**

When the consistency between human and AI ratings was examined, the results revealed poor agreement, as shown in table 5. All criteria — content, organization, language, and mechanics — demonstrated similarly low levels of consistency, culminating in an overall consistency score of .353. This indicates that the ratings provided by human raters and AI tools were not closely aligned or linearly correlated.

These findings suggest that, in the context of assessing students' academic work such as essays, AI tools (e.g., ChatGPT, CiCi, and Bing) still require significant improvement in order to match the consistency levels of human evaluators. The results support the conclusion that current AI models may not yet be suitable replacements for human raters in high-stakes academic assessment.

This conclusion is consistent with the findings of Jackaria, Hajan, and Mastul (2024), who also reported poor consistency between ChatGPT and human raters. Similarly, Paz et al. (2023) highlighted that ChatGPT's outputs may contain inaccuracies and errors, underscoring the continued need for human verification in AI-assisted evaluation systems.

Table 5. Intraclass Correlation results indicating inter-rater reliability between Human and AI tools essay ratings

Human Raters	**AI Raters**				**Total Score**	**Interpretation**
	Content	Organization	Language	Mechanics		
Content	0.440	---	---	---	---	**Fair**
Organization	---	0.339	---	---	---	**Poor**
Language	---	---	0.379	---	---	**Poor**
Mechanics	---	---	---	0.185	---	**Poor**
Total Score	---	---	---	---	**0.353**	**Poor**

Although the Intraclass Correlation Coefficient (ICC) measures the degree of linear association between two sets of ratings, it primarily indicates how closely the average scores of raters align. In this study, while human and AI ratings showed differences, these were consistent and patterned, rather than random. This warranted further analysis to determine whether the differences were statistically significant.

Table 6 presents the results of the Wilcoxon signed-rank test, comparing the mean scores of human and AI raters across four key criteria: Content, Organization, Language, and Mechanics. The findings show that human

raters assigned slightly higher scores than AI raters across all dimensions. No significant differences were found for Content ($z = 2.78$, $p = 0.804$), Organization ($z = 2.77$, $p = 0.767$), Language ($z = 2.67$, $p = 0.729$), or Mechanics ($z = 2.62$, $p = 0.715$). Furthermore, the overall comparison between human and AI ratings revealed no significant difference ($z = 2.71$, $p = 0.753$), suggesting a generally comparable pattern of assessment between the two groups.

These results underscore the emerging potential of generative AI tools—such as ChatGPT, CiCi, and Bing—as valuable supports in the assessment of students' written work. With proper instructional guidance and carefully constructed prompts, AI tools can assist in reducing the workload associated with essay evaluation, thereby allowing educators to devote more time to instructional planning and student engagement. This study demonstrates that AI, when used responsibly, can meaningfully contribute to academic assessment. However, it is essential to apply these tools with discernment. In alignment with prior findings (Jackaria, Hajan, & Mastul, 2024; Bitzenbauer, 2023; Lo, 2023), the results reaffirm that human oversight remains crucial, particularly to validate AI-generated scores and ensure fairness, contextual accuracy, and pedagogical relevance in student evaluation.

Table 6.

Wilcoxon signed-rank test results on significant differences between the human and AI essay ratings

	Human Raters		**AI Raters**		**Wilcoxon**		
Criterion	X	SD	X	SD	z-score	*p*-value	**Interpretation**
Content	2.78	0.804	2.78	0.804	2.78	0.804	Not Significant
Organization	2.83	0.924	2.77	0.767	2.77	0.767	Not Significant
Language	2.73	0.660	2.67	0.729	2.67	0.729	Not Significant
Mechanics	2.72	0.804	2.62	0.715	2.62	0.715	Not Significant
Total Score	**2.84**	**0.809**	**2.71**	**0.753**	**2.71**	**0.753**	**Not Significant**

** Significant at $\alpha \leq 0.05$*

4.3 On Human versus AI tools' Remarks on students' essays

4.3.1Feedback by focus

The figures below illustrate the algorithmic patterns by which both human raters and AI tools provide feedback on students' essays, particularly in terms of focus. As shown in figure 4, the remarks given by human teachers were generally categorized into three types: language accuracy-focused, organization-focused, and no remarks. Prompted by the Caroll and West (1986) rubric, teachers primarily concentrated on language accuracy, offering explicit corrections related to grammar, spelling, vocabulary, and usage. Examples include: *"Error in spelling 'addition'"* (T1), *"Faulty parallelism – you, we, their (not parallel)"* (T2), and *"Wrong preposition!"* (T3). These

comments reflect the teachers' detailed attention to the formal aspects of language, particularly grammatical structure and mechanics.

In addition to language-related feedback, teachers also focused on the organizational structure of the essays, offering remarks related to text development and formatting. Comments such as *"Margin and indentions!"* (T1), *"It does not have a strong thesis statement."* (T2), and *"It did not follow the proper way of writing an essay—intro, body, and conclusion."* (T3) indicate concern for coherence, logical flow, and adherence to academic essay conventions.

Interestingly, several essays received no remarks from the human raters. This could suggest several possibilities: that the essays were considered satisfactory and did not warrant feedback; that time constraints or cognitive fatigue may have limited the depth of commentary; or that teachers selectively prioritized certain aspects, such as grammar, over others. This variation in feedback aligns with the findings of Tukan and Mukarto (2024), who reported that teacher feedback on student writing tends to focus primarily on form and content, often at the expense of other elements.

By contrast, AI tools exhibited a wider range of focus areas in their feedback, including language accuracy, organization, content, and general language use as reflected in figure 5. While they addressed structural and linguistic concerns, AI tools also offered content-focused feedback, engaging with the ideas presented in the text, as well as more generalized language evaluations (*e.g., "clear, insightful, and strong content with minor language issues" or "There is a variety in your language"*). Notably, every student essay received some form of feedback from the AI tools—there were no instances of missing remarks, in contrast to the human raters.

This suggests that AI tools, such as ChatGPT, CiCi, and Bing, provide broader and more consistent coverage across multiple aspects of writing. In comparison, human teachers tended to focus narrowly, often prioritizing grammar or structure. This difference implies that while human feedback may reflect pedagogical judgment or time-bound constraints, AI-generated feedback can offer more comprehensive, automated input—though still requiring critical oversight from educators for contextual accuracy and appropriateness.

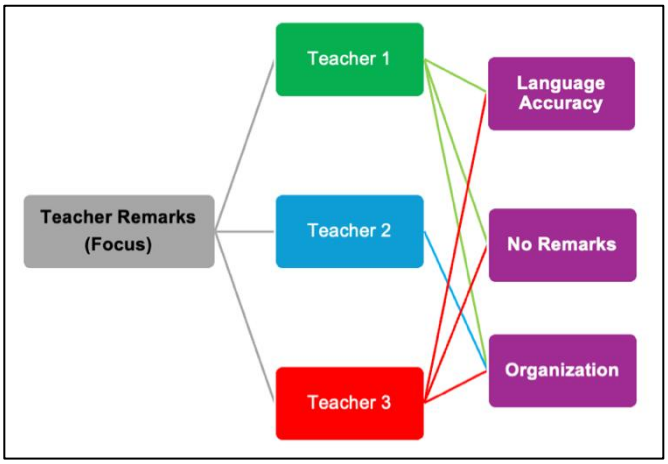


Fig. 4 Human Teachers' Remarks on Students' Essays in terms of Focus

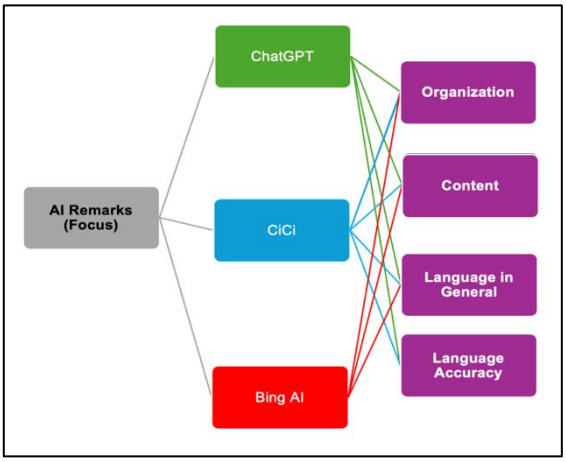


Fig. 5 AI Tools' Remarks on Students' Essays in terms of Focus

4.3.2 Feedback by manner

Table 7 shows recurring pattern of the feedback of human teachers and AI tools on students' essays. The manner of feedback provided by human raters was generally direct, corrective, and assertive. Teachers used concise, declarative statements that clearly identified issues in students' writing. Comments such as "Lacks concluding statement," "Subject agreement is erratic," and "Faulty parallelism" demonstrate a straightforward style that emphasizes clarity and precision. This manner reflects an instructional tone focused on identifying specific areas for improvement, assuming that students will understand and apply the corrections independently.

In contrast, AI tools delivered feedback in a more balanced and measured tone, often combining critique with encouraging language. Remarks such as "Clear and strong points but there are minor language issues" or "The content is informative… however, some grammatical issues detract from clarity" illustrate a neutral and supportive manner. This style tends to soften critical observations, making the feedback feel more constructive and accessible. Overall, AI-generated feedback maintains a consistent and polite tone, which may help foster student receptiveness to revision.

The differing manners in which human raters and AI tools provide feedback carry important implications for writing instruction and assessment. The direct and assertive tone used by human teachers may be effective for learners who are accustomed to explicit correction and who benefit from clear, targeted guidance. However, this approach may be perceived as overly critical by some students, potentially affecting motivation or confidence, especially in more formative learning environments. On the other hand, the balanced and supportive tone used by AI tools—characterized by a mix of encouragement and gentle critique—may help reduce anxiety and promote a more positive reception of feedback. This suggests that integrating AI tools into the feedback process can complement traditional teacher evaluation by offering a consistent, neutral, and approachable source of formative

input. This result is aligned with Abdullah (2025), who pointed out the high potential of AI to provide more frequent, personalized, and objective feedback. Educators may consider using AI-generated feedback as a preliminary layer to guide student revisions, while reserving human feedback for deeper pedagogical engagement and clarification, especially on complex writing concerns.

Table 7.

Manner of Remarks of Human Teachers and AI Tools on student essays

Subject	Remarks	Manner
Teacher 1	"Lacks concluding statement."	Criticism
	"Subject agreement is erratic."	
Teacher 2	"Error in spelling ang capitalization."	
	"Your punctuations are not appropriate."	
Teacher 3	"It does not have a strong thesis statement."	
	"Faulty parallelism."	
Subject	**Remarks**	**Manner**
Teacher 1	*"To improve their teaching [insert comma]…"*	Suggestive (Specific)
	"students table… [add apostrophe]."	
Teacher 2	*"It make…[makes instead of make.]"*	
	"ChatGPT are one of the… [verb tense-'is' to be used]	
Teacher 3	*"Your going to…[you're instead of your]…"*	
	"spacing in the 2nd and 3rd paragraphs…"	
Subject	**Remarks**	**Manner**
ChatGPT	"*Clear and strong points but there is a minor language issue.*"	Praise (specific) + Criticism
	"Good insights but minor grammar issues."	
CiCi	"*The essay provides relevant points about AI in education but lacks depth in analysis."*	
	"The content is informative and detailed, with good organization. However, some grammatical and mechanical issues slightly detract from clarity."	
Subject	**Remarks**	**Manner**
ChatGPT	*"Insufficient Content; Vague arguments."*	Criticism
	"Off-topic; lacks educational focus'	
Bing	*"Needs more depth; lacks specific examples"*	
	"Very weak; one-sided argument; lacks details"	

5. Conclusions and Recommendations

This study investigated the consistency between AI-generated and human-generated ratings of ESL student essays, as well as the differences in the focus and manner of their feedback. Both groups evaluated the essays using the same rubric and generally arrived at similar categorical assessments. Human raters tended to assign slightly higher scores than AI tools, but statistical testing revealed no significant difference in the overall ratings between the two groups. Although the inter-rater consistency between AI and human evaluators was low, the overall scoring outcomes were relatively aligned. This suggests that while human and AI raters may apply different approaches and judgment styles, they often reach comparable conclusions when assessing student writing.

In terms of feedback style, human raters commonly gave direct, corrective, and concise comments that emphasized grammar, sentence structure, and organization. Their manner was often critical or suggestive, reflecting their instructional focus and professional judgment. AI tools, on the other hand, demonstrated a more formal and structured tone in their responses. Their feedback extended beyond surface-level errors, addressing broader aspects of writing such as clarity of ideas, cohesion, and language use, and often included both praise and critique in a balanced manner. This distinction indicates that human feedback tends to prioritize form and structure, while AI-generated feedback provides a more expansive commentary across multiple dimensions of writing.

Based on the findings, it is recommended that generative AI tools like ChatGPT, CiCi, and Bing AI be used to complement—not replace—human evaluation in writing assessment. Their consistency and broad feedback can aid formative assessment and reduce workload, but the nuanced judgment of human raters remains essential. Teachers should guide and validate AI-generated feedback to ensure its accuracy and educational value. Professional development is also advised to support responsible AI integration in classrooms.

Appendix 1

Table 8 Scoring Rubrics in Assessing Essays adapted from Caroll and West (1989, as cited by Tribble, 1986)

Writing Components	1 Poor	2 Average	3 Fair	4 Excellent to Good
Content	Inadequate treatment of the topic • almost no useful detail •Reader is confused or may be misinformed •Central purpose or argument is not clearly identified •Analysis is vague or no evident.	Treatment of the topic is hardly adequate •Some irrelevant content, relevant to the Topic •Analysis is basic or general •Reader gains few insights.	Adequate treatment of the topic • Most content relevant to the topic • Information provides reasonable support for a central purpose or argument and displays evidence of a basic analysis of a significant topic •Readers gain some insights.	Excellent to very good treatment of the subject• Content relevant to the topic •Balanced presentation of relevant and legitimate information that clearly supports a central purpose or argument and shows a thoughtful, in-depth analysis of a significant topic •Readers gain important insights.
Organization	Lack fluent expression, ideas very difficult to follow • Little sense of paraphrasing /	Very uneven expression • Ideas difficult to follow/stand out • Paraphrasing or section	Uneven expression • but main ideas stand out •Paraphrasing or section	Fluent expression • ideas clearly stated/supported • Appropriately organized

	organization• No sense of logical sequence (coherence) •Connectives not used (cohesion).	organization does not help the reader, logical sequence difficult to follow (coherence) •Connectives largely absent (cohesion).	organization evident •Logical sequenced (coherence) • Connectives appropriately used (cohesion).	paragraphs or sections •Logical sequenced (coherence) •connectives appropriately used (cohesion).
Language Use (Vocabulary+ Grammar)	Uncomfortably frequent mistakes in word/ idiom choice and usage • Register, no appropriate sense of register • Major problems with structures even simple ones • frequent errors of negation agreement, tense, number, word order/function, articles, pronouns, prepositions, meaning often obscured	A noticeable number of mistakes in word choice and usage •Register not always Appropriate • insufficient range of structures with control only shown in simple constructions • Mostly appropriate structures •Frequent appropriate Structures • Some errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, meaning some times obscured	Occasional mistakes in word choice and Usage • Register not always appropriate • acceptable grammar- but problems with more complex structures •Mostly appropriate structures • some errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, meaning some times obscured	Accurate word and usage •Appropriate selection to match register • Confident handing of appropriate structures • Hardly any errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, meaning never obscured.
Mechanics	There are so many errors on punctuation, spelling and APA style that meaning is obscured.	The writing has many errors on punctuation, spelling and APA style, and the reader is distracted.	There are occasional errors on punctuation, selling and APA style, but they don't represent a major distraction or obscure meaning.	The writing is free or almost free of errors on punctuation, spelling and APA style.

Acknowledgement

Not applicable.

Author Contributions

HRG conceptualized the research, conducted the analysis, and led the drafting and finalization of the manuscript. KH contributed by overseeing the data collection process.

Availability of data and materials

To safeguard participants' privacy, the authors have not made the dataset publicly available. This study was reviewed and approved by the Institutional In-House Review Board under protocol number 2025-29, ensuring adherence to ethical standards for participant protection.

Declarations

Competing interests

The authors declare that they have no competing interests.

References

Abdullah, A.H. S. (2025). Using AI-powered tools to provide constructive feedback during field training for English teachers. *Educational Sciences Journal, 1(3),* 21-49. 10.21608/SSJ.2025.421746

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence* 6 (2024): 100226. https://doi.org/10.1016/j.caeai.2024.100226

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. Contemporary Educational Technology, 15(3), Article 429. https://doi.org/10.30935/cedtech/13152

Al Shahrani, A. A., & Storch, N. (2013). Investigation of written corrective feedback in an EFL context: beliefs of teachers, their real practices and students' preferences. *Unpublished MA thesis, University of Melbourne, Australia*. https://shorturl.at/llvTC

Bouziane, K., & Bouziane, A. (2024). RETRACTED: Exploring the Role of AI in Essay Evaluation: A Comparative Analysis of ChatGPT and Human Corrections. https://www.researchsquare.com/article/rs-4139088/v1.pdf

Baidoo-anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI, 7(1), 52-62. https://doi.org/10.61969/jai.1337500

Bejtic, Z. (2024). Nurturing Feedback and its Impact on Self-Efficacy, Empowerment, and Professional Growth in Educational and Corporate Environments. *International Journal of Advanced Corporate Learning*, *17*(2). https://openurl.ebsco.com/EPDB%3Agcd%3A2%3A29819186/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A177144365&crl=c&link_origin=scholar.google.com

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. Contemporary Educational Technology, 15(3), ep430. https://doi.org/10.30935/cedtech/13176

Crawford, J., Cowling, M., Ashton-Hay, S., Kelder, J. A., & Middleton, R. (2023). Artificial intelligence and authorship editor policy: ChatGPT, Bard Bing AI, and beyond. *Journal of university teaching and learning practice*, *20*(5), 1-11. https://search.informit.org/doi/abs/10.3316/informit.T202411270002040124448774l

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., ... & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, 100299. https://doi.org/10.1016/j.caeai.2024.100299

EDUCAUSE. (2018). Horizon report: 2018 higher education edition. Retrieved from EDUCAUSE Learning Initiative and The New Media Consortium website: https://library.educause.edu/~/media/files/library/2018/8/2018horizonreport.pdf

EDUCAUSE. (2019). Horizon report: 2019 higher education edition. Retrieved from EDUCAUSE Learning Initiative and The New Media Consortium website: https://library.educause.edu/-/media/files/library/2019/4/2019horizonreport.pdf

Engeness, I. (2018). What teachers do: facilitating the writing process with feedback from Essay Critic and collaborating peers. *Technology, Pedagogy and Education,* 27 (3), 297-311. https://doi.org/10.1080/1475939X.2017.1421259

Escalante, J., Pack, A. & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *Int J Educ Technol High Educ* **20**, 57. https://doi.org/10.1186/s41239-023-00425-2

Guibangguibang, H. R. B. (2020). Association between Oral Error Corrections of University Teacher and English Majors' Language Anxiety in Philippine Higher Education Context. *International Journal of Language Education*, *4*(2), 183-193. https://eric.ed.gov/?id=EJ1270891

Guo, K. (2024). EvaluMate: Using AI to support students' feedback provision in peer assessment for writing. *Assessing Writing*, *61*, 100864. https://doi.org/10.1016/j.asw.2024.100864

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European journal of education*, *57*(4), 542-570. https://doi.org/10.1111/ejed.12533

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, Research Issues and Applications of Artificial Intelligence in Language Education. *Educational Technology & Society*, *26*(1), 112–131. https://www.jstor.org/stable/48707971

Hyland, F., & Hyland, K. (2001). Sugaring the Pill: Praise and Criticism in Written Feedback. Journal of Second Language Writing, 10, 185-202. https://doi.org/10.1016/S1060-3743(01)00038-8

IBM Data and AI team (2023). *Understanding the different types of artificial intelligence.* IBM. https://www.ibm.com/about?lnk=flatitem

Iqbal, N., Ahmad, H., & Azhar, K. (2023). Exploring teachers' attitudes towards using ChatGPT. Global Journal for Management and Administrative Sciences, 3(4), 97–111. https://doi.org/10.46568/gjmas.v3i4.163

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

Jakobson, A. (2015). *Analyzing teacher feedback in Swedish L2 contexts*. Journal of Language Teaching and Research, 16(4), 356-369.

Jakobson, L. (2022). A model for analyzing teachers' written feedback on adult beginners' writing in Swedish as a second language. *Languages*, *7*(2), 74. https://doi.org/10.3390/languages7020074

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & 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/10.1016/j.lindif.2023.102274

Klein, M. J., & Frana, P. L. (2021). Encyclopedia of Artificial Intelligence. https://www.torrossa.com/en/resources/an/5526530

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

Kopp, W., & Thomsen, B. (2023). How AI can accelerate student's holistic development and make teaching more fulfilling. World Economic Forum. https://www.weforum.org/agenda/2023/05/ai-accelerate-students-holistic-development-teaching-fulfilling/

Khan, F., Yusoff, R. M., & Khan, A. (2014). Job demands, burnout and resources in teaching a conceptual review. *World Applied Sciences Journal*, *30*(1), 20-28. DOI: 10.5829/idosi.wasj.2014.30.01.21029

Laschinger, H. K. (1992). Intraclass correlations as estimates of interrater reliability in nursing research. Western Journal of Nursing Research, 14(2), 246–251. https://journals.sagepub.com/doi/pdf/10.1177/019394599201400213

Latif, E., & Zhai, X. (2023) Fine-tuning ChatGPT for automatic scoring. Computers and Education: Artificial Intelligence, 6, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, I. (2017). Classroom writing assessment and feedback in L2 school contexts. Singapore: Springer Singapore. https://link.springer.com/content/pdf/10.1007/978-981-10-3924-9.pdf

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), Article 410. https://doi.org/10.3390/educsci1304041

Magday Jr, D. W., Jerwin, M. F., Joel, A. T., Gail, G. G., Bella, C. G. A., Lhea, C. T., ... & Boie, T. B. E. (2025). Enhancing students' subject-verb agreement skill by using CiCi artificial intelligence-based app: Embedded mixed-methods study. *HO CHI MINH CITY OPEN UNIVERSITY JOURNAL OF SCIENCE-SOCIAL SCIENCES*, *15*(2). https://scholar.googleusercontent.com/scholar?q=cache:JlvKMFtQvRoJ:scholar.google.com/+Cici+AI&hl=en&as_sdt=0,5

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for integrating technology in teachers' knowledge. Teachers College Record, 108 (6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mondal, H., Marndi, G., Behera, J. K., & Mondal, S. (2023). ChatGPT for teachers: Practical examples for utilizing artificial intelligence for educational purposes. Indian Journal of Vascular and Endovascular Surgery, 10, 200–205. https://doi.org/10.4103/ijves.ijves_37_23

Moss, C. M., & Brookhart, S. M. (2019). *Advancing formative assessment in every classroom: A guide for instructional leaders*. ASCD. https://shorturl.at/9jC7g

Paz, M. A., Turner, K., & Racila, E. (2023). Evaluating the performance of ChatGPT in writing autopsy clinicopathological correlations. American Journal of Clinical Pathology, 160(1), S125. https://doi.org/10.1093/ajcp/aqad150.272

Rahman, M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. Applied Sciences, 13(9), Article 5783. https://doi.org/10.3390/app13095783

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025, March). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462-472). https://doi.org/10.1145/3706468.3706527

Shin, D. D., Kim, S. I., Lee, M. J., Jiang, Y., & Bong, M. (2022). Role of perceived competence and task interest in learning from negative feedback. *Frontiers in Psychology*, *13*, 830462. https://doi.org/10.3389/fpsyg.2022.830462

Siedlecki, S. (2020). Understanding research designs and methods. *Clinical Nurse Specialist CNS, 34, 8-12.* 10.1097/NUR.0000000000000493

Straub, R., & Lunsford, R. (2006). Key Works on Teacher Response: An Anthology. Boynton Cook. https://eric.ed.gov/?id=ED495444

Sutton, P. (2012). Conceptualizing feedback literacy: Knowing, being, and acting. *Innovations in Education and Teaching International*, *49*(1), 31-40. https://doi.org/10.1080/14703297.2012.647781

Tribble, C. (1996). Writing. Oxford: Oxford University Press.

Tukan, A. I. A. L., & Mukarto, F. X. (2024). Unpacking teachers' feedback on students' writing: A content analysis of writing exemplars. *Journal of English Language and Education*, *9*(4), 1-11. https://www.jele.or.id/index.php/jele/article/view/506

Tuyen, T., Osman, S. B., Ahmad, N. S. B., & Dan, T. C. (2019). Developing and validating scoring rubrics for the assessment of research papers writing ability of EFL/ESL undergraduate students: The effects of research papers writing intervention program using process genre model of research paper writing. https://www.semanticscholar.org/paper/Developing-and-Validating-Scoring-Rubrics-for-the-Tuyen-Osman/e86657da7ec761a7d403b4f8d85cc8bf19f99923

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10.* https://doi.org/10.3389/fpsyg.2019.03087

Yang, T., Cheon, J., Cho, M.H., Huang, M., & Cusson, N, (2025). Undergraduate students' perspectives of generative AI ethics. *International Journal of Educational Technology in Higher Education, 22 (35),* https://doi.org/10.1186/s41239-025-00533-1

Zawacki-Richter, O., Marín, V.I., Bond, M., & Gouverneur, F. *(2019).* Systematic review of research on artificial intelligence applications in higher education – where are the educators?. *Int J Educ Technol High Educ* **16**, 39. https://doi.org/10.1186/s41239-019-0171-0