Jurnal Profesi Keguruan
Vol. 10 Issue 1 (2025) 66-79
DOI: https://doi.org/10.15294/jpk.v10i1.25849
Available online since June 30, 2025

Jurnal
Profesi
Keguruan
https://journal.unnes.ac.id/journals/jpk

Can We Trust AI to Assess Writing? An Analysis of Scoring Reliability and Feedback Consistency

Fitriani Fitriani✉
STAIN Mandailing Natal, Indonesia
Puput Zuli Eko Rini
Universitas PGRI Mpu Sindok, Indonesia
✉ fitriemje@gmail.com

Abstract

This study analyzes AI-generated writing assessments' scoring reliability and feedback consistency using ChatGPT. Adopting a mixed-methods approach, 23 student descriptive texts were evaluated across three assessment rounds. Quantitative findings showed high scoring reliability, with an Intraclass Correlation Coefficient (ICC) of 0.93, indicating excellent consistency across repeated evaluations. Qualitative analysis revealed that ChatGPT consistently addressed five core writing criteria—content, organization, vocabulary, language use, and mechanics. However, the feedback varied in focus and detail across rounds, and the absence of reference to prior feedback limited its support for revision as a recursive process. The findings suggest that although ChatGPT demonstrates reliable scoring and generally stable feedback themes, it lacks the continuity to facilitate sustained writing development. To enhance its pedagogical value, AI-based feedback systems should be designed to build upon previous responses, thereby enabling more effective support for students' progressive improvement in writing.

Keywords

AI-generated writing assessment; scoring reliability; feedback consistency

I. Introduction

The rapid advancement of Artificial Intelligence (AI) has profoundly transformed many sectors, including education. One of the most notable changes is in writing assessment practices. AI systems now utilize machine learning and natural language processing (NLP) techniques to automatically evaluate written texts, assign scores, and generate immediate feedback for learners (Beiting-parrish, 2024). These technologies significantly reduce the grading workload for educators, allowing more time for instructional activities instead of repetitive evaluation tasks. As such, AI has become an integral component of modern educational assessment frameworks (Roll & Ford, 2016).

Accurate and consistent assessment is fundamental to effective learning and evaluation. In writing instruction, reliable feedback is essential for helping students recognize their strengths and weaknesses and guiding their revision process. AI tools such as Grammarly, Turnitin Revision Assistant, and ChatGPT have gained widespread popularity among educators for providing automated, timely feedback on grammar, style, and content (Neill & Russell, 2019). Surveys indicate a growing reliance on AI tools in writing assessments, with some studies reporting that over 60% of teachers incorporate AI support regularly in their grading practices (Barshay, 2004).

Despite these benefits, important concerns regarding the reliability and consistency of AI-generated scores and feedback persist. When AI systems assess the same text multiple times, inconsistencies in scoring and feedback can occur, confusing students and educators alike. This variability undermines trust in AI tools and raises questions about their suitability as standalone evaluators in academic settings (J. Li et al., 2024). Inaccurate or inconsistent feedback could lead students to receive mixed messages about their writing quality, potentially hindering their progress. Given the increased reliance on AI for grading, these concerns highlight the necessity for rigorous investigation into AI reliability (Kim et al., 2024).

This study addresses this concern by investigating whether AI tools provide consistent scores and feedback when evaluating identical student texts multiple times. Specifically, this research focuses on ChatGPT (free version) developed by OpenAI, which has become widely used among educators and students due to its accessibility and ability to generate immediate writing feedback. Understanding this consistency is crucial, as stable evaluations foster more precise guidance for learners and more confident use by educators. While previous research has examined AI's accuracy compared to human raters (Kim et al., 2024), fewer studies have focused on the repeatability of AI-generated feedback and scores, particularly using ChatGPT in its free, publicly available form. This gap leaves educators without sufficient evidence to gauge the dependability of these tools over time (Barshay, 2004).

The current study employs a mixed-methods design that combines quantitative and qualitative approaches to fill this gap. The quantitative aspect uses the Intraclass Correlation Coefficient (ICC) to measure the reliability of AI-generated scores across repeated assessments of the same texts (Hove et al., 2022). Meanwhile, the qualitative aspect applies thematic analysis to evaluate the consistency and pedagogical relevance of the feedback comments produced by ChatGPT (free version) (Attali & Burstein, 2006). This dual approach allows for a comprehensive evaluation of both numerical scoring consistency and the meaningfulness of written feedback.

This research aims to examine the consistency of AI-generated scores on identical student texts, analyze the relevance and consistency of the feedback provided by ChatGPT, and identify factors that may contribute to variations in scoring or feedback. These objectives align with the overall aim of understanding free-access AI systems' reliability and potential limitations in evaluating student writing.

AI in Writing Assessment

Artificial intelligence (AI) is increasingly integrated into educational contexts, particularly in writing assessment. AI-powered tools analyze and evaluate written text using machine learning algorithms and NLP techniques. These tools vary in sophistication, from basic grammar and style checkers like Grammarly and ProWritingAid to advanced systems such as ETS's Criterion and Turnitin's Revision Assistant, which provide holistic feedback and automated scoring. Large language models (LLMs) like ChatGPT have emerged, offering the potential for assessment and personalized writing instruction (M. Li, 2024).

The adoption of AI for writing assessments offers numerous benefits. These tools deliver instant feedback, promoting learner autonomy and encouraging iterative revisions (Roisah et al., 2024). They also ensure consistent application of scoring criteria, reducing subjective bias commonly found in human grading (Balfour, 2013). Moreover, AI systems can efficiently process large volumes of writing, making them ideal for large-scale assessments.

Nevertheless, the use of AI in assessment also presents challenges. A significant concern is the lack of transparency in how AI systems produce their evaluations, often called the "black box" issue (Kim et al., 2024). Furthermore, AI tools may struggle with nuanced writing elements such as creativity, tone, and rhetorical effectiveness (Polakova & Ivenz, 2024). Their reliance on standardized models may promote formulaic writing styles, potentially limiting students' expressive capabilities. Ethical concerns, including data privacy and algorithmic bias, remain significant (Lim et al., 2023).

Reliability in Writing Assessment

Reliability is a key criterion in evaluating any assessment tool, including AI-based systems. In traditional writing assessment, reliability includes inter-rater reliability (consistency between different raters) and intra-rater reliability (consistency by the same rater over time).

These forms of reliability are commonly assessed using statistical methods like Cohen's Kappa and ICC to quantify agreement levels (Hove et al., 2022).

In AI systems, reliability primarily refers to the consistency of scores and feedback generated by the algorithm. This includes test-retest reliability (whether the same input yields the same result across different instances) and concurrent validity (comparison with scores from expert human raters) (Attali & Burstein, 2006). Studies in this area often use benchmark datasets and controlled environments to evaluate whether AI tools produce stable, reproducible outcomes.

Despite efforts to ensure reliability in AI-driven assessment, challenges persist. Scoring variability may be influenced by prompt structure, genre, or non-standard language features (Evan et al., 2010). Furthermore, periodic updates or retraining of AI models can cause changes in scoring patterns, raising concerns about consistency over time (Anson et al., 2023)

Previous Studies on AI Feedback and Scoring

Numerous studies have explored the effectiveness and consistency of AI writing tools. Research on Grammarly shows high reliability in identifying grammar, punctuation, and style errors (Neill & Russell, 2019), though it is less effective in addressing deeper aspects like argument structure or coherence (Shofiah et al., 2023). Likewise, studies on Criterion demonstrate reasonable alignment with human scorers on holistic scores but limited diagnostic feedback capacity (Attali & Burstein, 2006).

More recently, generative AI tools like ChatGPT have attracted attention for their ability to simulate detailed feedback and evaluate a wide range of writing tasks. Initial findings suggest that ChatGPT can offer relevant and articulate feedback. However, concerns about its consistency have emerged—particularly when the exact text is evaluated multiple times or under slightly varied prompts (Link & Koltovskaia, 2023). Additionally, AI feedback may occasionally include biased or hallucinated content, potentially misleading learners.

Despite growing interest, research gaps remain. Few studies have systematically compared multiple AI tools using consistent datasets, making generalizations difficult. Most existing research focuses on higher education and professional writing contexts, with limited attention to EFL learners or beginner writers. Moreover, longitudinal studies that examine whether AI feedback results in actual writing improvement are scarce (Kim et al., 2024). These gaps indicate that there must be more rigorous, comparative, and context-sensitive investigations into AI's reliability and pedagogical impact in writing assessments.

II. Method

This study employs a sequential explanatory mixed methods design, combining quantitative and qualitative approaches in two clear phases. As Creswell (2023) outlined, the process begins with collecting and analyzing quantitative data, followed by qualitative data collection and analysis to clarify and expand on the initial findings. The quantitative phase assesses the reliability of ChatGPT's scoring using the Intraclass Correlation Coefficient (ICC), while the qualitative phase examines the consistency of its feedback via thematic analysis. By integrating these methods, the study is expected to provide a comprehensive understanding that captures measurable outcomes and the nuanced insights behind ChatGPT's feedback in writing assessment.

This study's participants were 23 fourth-semester students from the English Education Program at STAIN Mandailing Natal. They were enrolled in an intermediate writing course and chosen because they were actively involved in academic writing activities during the semester. The writing they submitted was used as the primary data for this study.

To test how consistent ChatGPT is in scoring and giving feedback, each student's writing was submitted three times to the same AI tool, ChatGPT (free version) by OpenAI, under the same conditions each time. ChatGPT was chosen because of its strong ability to understand and produce human-like text and because it is widely used in education to help assess writing. It is known for providing detailed and helpful comments, which makes it suitable for this research (Beiting-parrish, 2024).

Each submission was done in a different session to avoid any influence from earlier interactions. For every submission, the score and the feedback ChatGPT gave were saved for further analysis. This process produced a rich data set to evaluate how consistently the AI works.

To measure how consistent ChatGPT's scores were across the three submissions of the same writing, the Intraclass Correlation Coefficient (ICC) was used. Specifically, this study used the ICC(2,1) model, a type of analysis that checks if the same tool gives similar results under the same conditions. This model is often used in education and psychology to test how reliable a measurement is when done multiple times. A higher ICC score means better consistency. Using this method, the study aimed to provide solid evidence of whether ChatGPT gives stable and trustworthy scores in academic writing situations.

For the qualitative part, the focus was on analyzing the feedback given by ChatGPT. Since each of the 23 student texts was submitted three times, the study gathered 69 feedback responses. These feedback comments usually addressed content, organization, vocabulary, language use, and mechanics. Like the scoring process, each submission for feedback was done in a different session to ensure that the AI gave fresh evaluations each time. This setup helped the researchers see whether the feedback was consistent and helpful across the three responses for each text.

The study used thematic analysis to analyze these feedback responses, as Nowell et al. (2017) explained. This method helps to identify patterns in the data. A total of 69 feedback entries were collected—three for each of the 23 students' descriptive writing assignments. All feedback was organized systematically in a Google Sheets spreadsheet, with each row representing one feedback instance labeled by the student's ID and the feedback sequence (first, second, or third). The researcher carefully read through all feedback to identify common themes and developed six main categories to classify the comments: grammar, vocabulary, organization, content, mechanics, and overall writing quality. Each piece of feedback was manually coded by marking relevant parts under these categories. Using Google Sheets' functions, the researcher counted and compared how frequently each category appeared across each student's three rounds of feedback. This process helped reveal patterns of similarity or differences, making evaluating the consistency of ChatGPT's feedback easier over time. Finally, the coded data and patterns were interpreted qualitatively, supported by selected examples from the feedback, to assess the reliability and focus of ChatGPT's evaluative comments.

In short, this methodology was carefully designed to test how reliable and consistent ChatGPT (free version) is when used to evaluate student writing. By involving real students and combining statistical tools with detailed feedback analysis, this study provides helpful information regarding the strengths and limits of using AI tools like ChatGPT in academic writing assessment.

III. Result and Discussion

This section presents the main results of the study, which aimed to analyze how consistent and proper AI-generated scores and feedback are when using ChatGPT for writing assessments. The findings are divided into two parts, which align with the study's mixed-methods approach. The first part explains the quantitative results, focusing on how consistent ChatGPT's scoring is using a statistical measure called the Intraclass Correlation Coefficient (ICC). The second part presents the qualitative results based on a thematic analysis of the feedback ChatGPT gave on three different occasions for the same student texts. This structure helps provide a complete picture of both the reliability of AI scoring and the quality of the feedback while also highlighting the strengths and limitations of AI in academic writing evaluation.

Reliability of AI-Generated Scores

The statistical analysis of 23 descriptive texts scored in three separate rounds revealed some significant patterns. The average score was 11.52 in the first round, 11.00 in the second, and 10.87 in the third. These small changes show that ChatGPT's scores were fairly stable over time. The overall average score across all sessions was 11.13.

The lowest score given was 5, and the highest was 19, showing that ChatGPT could recognize differences in writing quality. The standard deviation was 3.22, indicating moderate variation among the students' scores.

The average variance in scores for each text was 0.71, which is low. This means that most of the time, the AI gave similar scores for the same text. Some texts—like Texts 3, 10, 11, 12, and 23—even received the same score in all three rounds. A few others—like Texts 6, 9, and 15 to 18—showed greater score changes (up to 3 points), but the differences were still within a reasonable range.

Table 1. Intraclass Correlation Coefficient (ICC) Results

Parameter	Value
Model	ICC(2,1)
Number of Subjects (n)	23
Number of Raters (k)	3
Mean Square (Subjects)	31.05
Mean Square (Raters)	2.74
Mean Square Error	0.62
ICC(2,1)	0.93

To test how consistent the scores were, the study used ICC(2,1)—a statistical model that measures agreement in repeated ratings. The ICC result was 0.93, which is considered excellent reliability (Koo & Li, 2016). This means ChatGPT gave consistent scores across the three evaluations.

The results also show that most score differences were due to actual differences in the texts, not inconsistent scoring. So, the scoring process can be trusted.

Consistency of ChatGPT Feedback

Thematic analysis of the feedback focused on five areas: content, organization, vocabulary, language use, and mechanics. Across all three rounds, ChatGPT consistently focused on these key writing components. This shows that AI can reliably guide students in improving their writing. In every round, the AI highlighted the same main areas needing improvement:

Table 2. Consistent Themes in ChatGPT's Feedback Across Three Rounds

Criterion	Consistent Focus
Content	Asked for more vivid, detailed descriptions.
Organization	A clear structure is recommended: introduction, body, and conclusion.
Vocabulary	Suggested using richer, more expressive words.
Language Use	Emphasized complete sentences and correct grammar.
Mechanics	Highlighted mistakes in punctuation and capitalization.

For example, in terms of content, ChatGPT repeatedly asked for more sensory descriptions of the beach—what it looks like, sounds like, and feels like. It also stressed the importance of organizing the text well, with a clear beginning, middle, and end.

Variations in Feedback Detail and Style

The table below presents three sets of feedback across five writing criteria—content, organization, vocabulary, language use, and mechanics—to illustrate how ChatGPT's feedback varies based on different writing drafts. Each column shows a different way the AI provides suggestions to improve the same piece of student writing.

The analysis of feedback generated by ChatGPT on students' writing reveals that while the AI consistently focuses on five key writing criteria—content, organization, vocabulary, language use, and mechanics—the specific points of feedback vary across drafts. These variations occur without clear indicators of whether previously identified issues have been resolved. This inconsistency can make it difficult for students to track their progress or understand which revisions were effective. Although the feedback generally aims to enhance overall writing quality, the lack of continuity and follow-up on earlier comments limits its effectiveness in supporting revision as a recursive process.

Table 3. Sample of Variation of ChatGPT Feedback Across Three Identical Drafts Based on the Writing Criteria

Criteria	Feedback 1	Feedback 2	Feedback 3
Content	Add details like what the water looks like or how the wind feels on your face.	Add vivid details, like the sound of waves or the smell of ocean breeze.	Add sensory details—what do you see, hear, feel, or smell? Describe the sound of waves, warm sand, etc.
Organization	Use transitions like "Next," "Also," and "One more thing I like..."	Use linking words like "Next," "Also," or "After that."	Start with a hook ("The beach is my happy place"), and end with a conclusion ("That is why I love going there").
Vocabulary	Use expressive words like "peaceful," "relaxed," and "joyful."	Try "massive," "peaceful," and "relaxing" for more colorful writing.	Use more expressive words like "relaxing," "peaceful," and "joyful"; swap "big" with "vast" or "endless."
Language Use	Fix subject-verb agreement and plurals—minor adjustments.	Watch subject-verb agreement: "It makes," "families come," "sandcastles."	Focus on subject-verb agreement and plurals. Read out loud to catch issues.
Mechanics	Double-check for grammar bits; reading aloud might help.	Double-check sentences: fix "gives" vs. "give."	Double-check grammar and punctuation. Use a spell checker or ask a friend to proofread.

In terms of content, the AI initially recommends adding specific sensory details, such as what the water looks like or how the wind feels. Subsequent feedback shifts to suggestions like describing the sound of waves or the smell of the ocean breeze. By the third round, the AI advises including multiple sensory descriptions—sight, sound, touch, and smell—without referencing whether earlier content suggestions had been addressed. This pattern also appears in feedback on organization. At the same time, the first and second rounds encourage transitions like "Next," "Also," or "After that." The third revision suggests adding a hook at the beginning and a firm conclusion, again without evaluating the implementation of earlier suggestions.

For vocabulary, the AI suggests using expressive words such as "peaceful," "relaxed," or "joyful" in the first round. Later, it recommends more varied adjectives like "massive" or "endless." Although this approach enriches students' word choices, the AI provides no feedback on whether previously recommended vocabulary had been integrated. Similarly, language use feedback consistently points out subject-verb agreement issues, yet there is no indication of whether students have improved or repeated the same mistakes. Regarding mechanics, the AI repeatedly advises proofreading strategies—reading aloud, using spell check, or peer review—but does not mention whether prior mechanical errors have been corrected.

The findings suggest that although ChatGPT delivers relevant and constructive feedback aligned with standard writing traits, its shifting focus and lack of follow-up make it difficult for students to monitor their writing improvement. More effective support would involve continuity in feedback, including explicit references to past revisions and whether issues have been resolved or persist.

This study clearly shows that AI-generated feedback can offer consistency and instructional flexibility, which are essential in helping students develop their writing skills. By analyzing 69 feedback instances from 23 student texts over three rounds, the study presents a detailed picture of how ChatGPT behaves during repeated assessments.

The ICC analysis and other statistics prove that ChatGPT's scores are stable. The scores did not change much from one round to the next, and the average variance per text was minimal. Several texts even received the same score in all three rounds. The result matches findings from Attali & Burstein (2006) and Hackl et al. (2023), who found that AI scoring systems like e-rater and Criterion consistently produce reliable results. Such stability is important because it builds trust among students and teachers in the fairness and objectivity of AI scoring.

Beyond scores, the feedback itself also showed thematic consistency. Each time, ChatGPT emphasized the same five writing elements: content, structure, vocabulary, grammar, and mechanics. Such consistency shows that the AI has a stable instructional approach, which is helpful for teachers and students trying to strengthen their writing skills.

For example, under the content category, all feedback rounds suggested making the writing more descriptive using sensory language. Structural advice was also consistent—suggesting a clear beginning, middle, and end. These areas are commonly highlighted in writing feedback literature, such as by Bitchener et al. (2017), who said that the best feedback focuses on both big-picture elements (like content and structure) and more minor details (like grammar and punctuation).

Other studies, like those by Anson et al. (2023), also show that AI tools like Grammarly and Write & Improve often focus on these key areas, helping students identify repeated weaknesses, especially in grammar and word choice.

However, this strength also highlights a subtle weakness in terms of consistency. While ChatGPT focuses on broad writing categories, the specific points it addresses within each category can shift from round to round. For example, initial feedback might focus on general vocabulary improvement, while later feedback could emphasize stylistic choices or sentence variety. Although enriching the feedback, this variation may unintentionally create challenges for learners.

One potential issue is the lack of clear progress markers. Students often benefit from repeated reminders and explicit follow-ups on earlier feedback to understand which issues remain unresolved. If ChatGPT moves on to new suggestions without clearly revisiting previous points, students may feel uncertain about their progress or miss out on reinforcing important writing skills. This shifting focus could also confuse or overwhelm, making it difficult for learners to prioritize which feedback to address first.

Moreover, ChatGPT's feedback lacks the human tutor's ability to hold students accountable by tracking whether earlier suggestions have been implemented. This absence of explicit continuity may reduce the coherence of the revision process and make it harder for students to internalize recurring strengths or weaknesses in their writing.

IV. Conclusion

This study shows that ChatGPT is a reliable and consistent tool for assessing student writing. The scoring results were highly consistent across three rounds, with an excellent Intraclass Correlation Coefficient (ICC) of 0.93. This finding indicates that ChatGPT can objectively evaluate differences in students' writing quality. In addition to stable scoring, its feedback remains focused on key aspects such as content, organization, grammar, vocabulary, and mechanics. One strength of using ChatGPT in this research is its ability to offer progressively detailed and helpful advice rather than simply repeating the same comments, which enhances its instructional value across multiple revisions.

However, a notable limitation is the lack of follow-up on previously mentioned issues. While the AI consistently addresses general writing criteria, it tends to vary specific comments without indicating whether earlier problems have been resolved. Such an oversight can make it harder for students to track their progress or see which areas still need attention. Unlike human tutors, ChatGPT does not yet provide continuity or accountability in its feedback over time.

ChatGPT offers potential as a writing assistant, primarily providing quick, consistent evaluations and broad instructional guidance. Future versions should include features that can track prior feedback, revisit unresolved issues, and support students through a more structured and goal-oriented revision process to further improve its educational usefulness.

V. References

Anson, O. K., Rix, C., McKay, C., Brisk, K., Clark, E., Doherty, A., & Shibani, A. (2023). *Digital writing technologies in higher education*. https://doi.org/10.1007/978-3-031-36033-6

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and Calibrated Peer Review™. *Research & Practice in Assessment*, *8*, 40–48.

Barshay, J. (2004). The potential of AI feedback to improve student writing. *FutureEd*. https://www.future-ed.org/the-potential-of-ai-feedback-to-improve-student-writing/

Beiting-Parrish, M. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. https://doi.org/10.59863/MIQL7785

Bitchener, J., Young, S., & Cameron, D. (2017). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*. https://doi.org/10.1016/j.jslw.2005.08.001

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications.

Evans, N. W., Harthorn, K. J., & Tuioti, E. A. (2010). Written corrective feedback: Practitioners' perspectives. *International Journal of English Studies*, *1*(801), 47–77.

Hackl, V., Müller, A. E., Granitzer, M., Sailer, M., & Aug, C. L. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *International Journal of Artificial Intelligence in Education*, 1–14.

Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022). Updated guidelines on selecting an ICC for interrater reliability: With applications to incomplete observational designs. *Psychological Methods*. https://doi.org/10.1037/met0000516

Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., & Beck, J. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. https://doi.org/10.31274/isudp.2024.154.06

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications*, 1–9. https://doi.org/10.1057/s41599-024-03755-2

Li, M. (2024). Leveraging ChatGPT for second language writing feedback and assessment. *International Journal of Computer-Assisted Language Learning and Teaching*, *14*(1), 1–11. https://doi.org/10.4018/IJCALLT.360382

Lim, T., Gottipati, S., & Cheong, M. L. F. (2023). Ethical considerations for artificial intelligence in education assessments. In *Artificial Intelligence Applications in Education* (pp. xx–xx). https://doi.org/10.4018/979-8-3693-0205-7.ch003

Link, S., & Koltovskaia, S. (2023). *Automated scoring of writing*. Springer International Publishing. https://doi.org/10.1007/978-3-031-36033-6

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*, 1–13. https://doi.org/10.1177/1609406917733847

O'Neill, R., & Russell, A. M. T. (2019). Grammarly: Help or hindrance? Academic learning advisors' perceptions of an online grammar checker. *Journal of Academic Language and Learning*, *13*(1), 88–107.

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2410101

Roisah, S., Widyaningsih, T. L., & Sedya, Y. (2024). The impact of ChatGPT use on EFL students' writing ability. *Journal of Language Education and Research*, *3*(3).

Roll, I., & Ford, H. (2016). Evolution and revolution in artificial intelligence in education. In

Artificial Intelligence in Education (pp. 582–599). https://doi.org/10.1007/s40593-016-0110-3