TEFL Praxis

Automated Grammar Scoring with AI Chatbots

Paul Daniels
Kochi University of Technology, Japan

Abstract: *Artificial intelligence (AI) continues to revolutionize the way that we work and learn and its impact on education is rapidly unfolding. Recent research is shedding light on how AI can best support teachers and learners. Studies on generative AI and large language models (LLMs) indicate that educators are optimistic AI will play a productive role in education (Polak et al., 2022; Caines et al., 2023). Research in the language learning field has demonstrated that AI is as good or even better at assessing writing than humans (Bridgeman et al., 2012). To continue the investigation of how AI can facilitate automated essay scoring (AES) and language learning in general, this study examines several of the latest AI Chatbots to better understand how accurately and reliably they rate the grammar of writing samples of EFL engineering students in Japan.*

Keywords: automated essay scoring, chatbots, AI, writing, EFL

Introduction

Writing skills are a vital component of an EFL curriculum and grammar skills are instrumental in composing one's ideas and thoughts. Japan's Course of Study guidelines, established by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), highlights the importance of teaching speaking and writing (MEXT, 2018). However, writing is often a neglected subject in EFL settings, particularly in Japan where learners have some of the lowest TOEFL iBT writing scores in Asia. Reasons for the lack of writing instruction may be due to the large number of students enrolled in language courses, lack of necessity of English writing skills and lack of teacher training for writing instruction (Harwood & Hirose, 2019). Writing is often introduced using 'translation tasks' where learners are asked to translate a sentence in the L1 into English (Kobayakawa, 2011). In these types of translation tasks, the writing process lacks a communicative purpose, and the outcome is controlled, as only a limited set of answers are correct. This approach to teaching writing eases the assessment by the instructor but does not allow for learners to develop creative or communicative writing skills. In addition, because of the large number of students typically enrolled in EFL classes, learners often receive limited personalized feedback on their writing. Consequently, Japanese students often enter university with limited writing experience, and have difficulty expressing ideas in writing in their L2.

ISSN 2758-5700
TEFL Praxis Journal, 2023, Vol. 2, pp.59–68
DOI:10.5281/zenodo.10402530

Recent advances in AI may hold promise for both writing instructors and language learners. With the surge in AI and improvements in translation technology, there has been considerable debate on the role of this new technology in the language learning classroom, particularly with writing and translation tasks. This paper focuses on how AI Chatbots can assist in the L2 writing process by alleviating the teacher bias, fatigue and score inconsistencies when evaluating student writing, and by providing efficient personalized learner feedback in large classroom settings.

Automated Essay Scoring (AES)

Automated essay scoring (AES) systems are used to automatically assess writing quality by replicating the scoring procedures that human raters employ. AES systems attempt to quantify similar writing structures gathered from past writing samples which have been scored by experienced writing instructors. AES calculates the relationships between these quantifiable structures and the student writing levels. Some of the writing structures that AES identifies and quantifies include:

- *grammar and mechanics:* spelling, punctuation, capitalization, and grammar rules.
- *word usage and sentence structure:* appropriate use of vocabulary and sentence structures, as well as sentence variety and complexity.
- *organization and coherence:* essay organization, flow of ideas, and use of transition words and phrases that signal the relationship between ideas.
- *content and development:* relevance of the topic, the depth of development, and use of evidence to support claims.
- *style:* use of figurative language and the overall tone and voice of the writing.

Research on initial AES systems, such as Project Essay Grade (PEG), focused primarily on essay punctuation, spelling, and grammar scores. Researchers have indicated that PEG is able to calculate essay scores that highly correlate with human raters (Valenti et al., 2003; Page, 2003). Early research also explored the capabilities of Pearson Educational Technologies Intelligent Essay Assessor (IEA), a popular AES system that several standardized exam companies have adopted. Landauer (2003) observed that the scores generated by Pearson's IEA strongly correlated with human graders.

Within the last decade, researchers have also extensively examined the Educational Testing Service (ETS) e-rater, a robust automated essay scoring program that was introduced commercially in 1999. ETS's e-rater engine is used to both generate writing scores and provide targeted writing feedback. The e-rater engine is used in language practice tests such as TOEFL Practice Online. It is also used as a second rater for double-rater evaluation of high-stakes tests such as SAT and GRE examinations. (Chen et al., 2017; Attali & Burstein 2006). E-rater has scored close to a million Graduate Management Admission Test (GMAT) writing tasks at a 97% or better agreement level between human and computer scores (Valenti et al., 2003). Studies also indicated that e-rater was able to assess similar features within essays written by native and non-native English speakers (Burstein & Chodorow, 1999), indicating that it could fairly judge essays written by L2 writers that may contain more syntax errors.

Although studies indicate that AES systems can churn out reliable and valid writing scores, most of these systems remain a black box. There is little information available on how they score essays, the code is proprietary, and the usage fees are high. The online GMAT exam runs $250

USD and the TOFEL Practice Online costs $45 USD. Due to these obstacles, smaller educational institutions and individual language instructors are pursuing alternative AES options that make use of AI Chatbots such as ChatGPT, Google Bard and Microsoft Bing.

AI Chatbots

Automated essay scoring (AES), guided by AI Chatbots, is a disruptive innovation that has been gaining attention as researchers try to interpret the complexities of AI and AES (Kumar & Boulanger, 2020; Hussein et al., 2019). AI chatbots are computer programs that can simulate human conversation with a user. In recent years, AI powered chatbots have become accessible to the public. At the time of this writing, OpenOffice's ChatGPT, Microsoft's Bing Chat and Google's Bard are the three leading AI chat contenders. These AI chatbots are trained on large datasets of text, which helps them to understand and respond to almost any question or request in a natural way. Natural language processing (NLP), machine learning, and deep learning work in conjunction to generate possible replies to user requests. NLP helps AI chatbots understand the meaning of human language, and machine learning together with deep learning can help AI Chatbots learn from past experiences and make improvements (IBM, 2023). In education, AI chatbots can be used to provide personalized instruction to students, answer their questions, and help them learn at their own pace.

Large Language Models

AI Chatbots, such as ChatGPT, Bing Chat and Bard, produce and understand language using LLMs. The LLMs are trained with authentic text databases, so that they can identify the patterns and connections between words and phrases. Once they are able to identify the patterns in human language, LLMs can generate text, for example, poems, computer code, or email using these different patterns. LLM can also translate text from one language to another and can informatively and comprehensively answer open-ended questions. LLMs formulate answers by first understanding the meaning of the question, and then by searching for relevant information within its database of text gathered from websites, books, news articles, and scientific journals. LLMs identify patterns in language using artificial neural networks that are very similar to networks in the human brain. Each neural network consists of interconnected nodes that perform calculations and pass the results to other nodes. LLMs appear to be the next wave of innovation as they are already impacting the way teachers and learners interact with computers (Gokul, 2023). They are currently being used to create engaging computer interfaces, improve educational instruction, and provide immediate and personalized user feedback (Rahman & Watanobe, 2023; Milano et al., 2023).

AI Chatbots and AES

While AI Chatbots are still very recent and educators may feel uneasy about using this technology, early studies suggest that AI Chatbots can be effective at evaluating writing samples. More recent studies have been conducted to evaluate how effective Chat AIs can be at scoring student essays (Mizumoto & Eguchi, 2023; Dao, 2023). AI chatbots can be used to assess writing in a variety of ways, for example, they attempt to check:

- plagiarism by comparing writing samples to a database of existing sources.

- grammar and spelling errors and suggest corrective feedback.
- writing style, including sentence structure, vocabulary, and tone.
- content of essays, including essay organization, thesis development, and use of supportive evidence.
- clarity and conciseness of the writing, and areas where the writer could be more specific or concise.
- organization of the essay and suggest how to improve the flow of arguments.
- overall mechanics of the writing, such as the use of formatting, citations, and references.

Research questions

With recent web-based AI advancements, small-scale AES systems can be deployed easily and inexpensively by institutions to assist in scoring essays and providing learner feedback. This particular study investigates a small-scale AES project for scoring grammar keeping the following research questions in mind:

Research Q1: How similar are AI-generated grammar scores with human-generated grammar scores using the same set of writing samples?

Research Q2: Is the inter-rater reliability of human raters and AI raters similar when scoring grammar of L2 learners?

Research Q3: How consistent are AI Chatbots amongst themselves at scoring the grammar of student writing?

Method

Participants and procedure

This research evaluates how similar AI Chatbots and humans rate the grammar of writing samples from 152 Japanese university engineering students enrolled in a 4-skills general first-year English course. For the writing task, students were asked to draft a 200-word report as part of the course requirement. Because the students are engineering majors and have had little experience writing longer creative-type essays, they were asked to write a ‘how to’ report, i.e., ‘giving instructions’. Students were encouraged to choose a topic that they were interested in and something that they wanted to teach their classmates. In addition to the written report, the students were asked to give an oral presentation about their topic. A sample ‘instructions’ report was introduced to students, and later they were asked to write their own instructions using imperative verbs and were requested to provide details that included prepositional phrases as well as details such as sizes, shapes, and materials. Finally, students were asked to check their spelling and grammar of their writing.

Data collection and data analysis

Each report was human scored based on several requirements including report length, details, and grammatical accuracy. With this type of multi-judgement rating, descriptions of the several levels of achievement that were expected were provided to the learners. For this study, the focus was only on the grammar judgment, which was scored by both human raters and the Chatbots. To better determine how effectively the grammar of each report could be automatically scored using AI tools, each writing sample was individually rated on a scale of 1-10 using both AI models and human raters. The writing samples were rated by a total of four qualified TEFL-certified English teachers.

For the AI grammar rating of student writing samples, OpenAI's ChatGPT, Google Bard and Microsoft Bing Chat were used. For the human raters, four TEFL-certified instructors were asked to read and rate the grammar of each of the 152 writing samples based on a common grammar scoring rubric.

The writing scores of both the human raters and the AI chatbots were analyzed using basic descriptive statistics and the Pearson correlation coefficient was used to determine the strength of the linear relationship between the human and AI generated grammar scores. In addition, the intraclass correlation coefficient (ICC) was used to measure the inter-rater reliability of the four human raters and the three AI Chatbots.

Results

Descriptive statistics were used to evaluate the first research question 'How similar are AI-generated grammar scores with human scores using the same set of writing samples?'. Figures 1-3 show the mean grammar scores of the human raters and the AI Chatbots.

The data for both the human scores and the AI scores is presumed to be normal as the skewness is between -2 and +2 and kurtosis is between +5 and –5 for both datasets (Hair et al., 2010; Bryne, 2010). The Pearson correlation coefficient was employed to examine the relationship between the human scores and the AI scores, as this is the most common coefficient used to evaluate the accuracy of automated scoring (Liu et al., 2016; Chen et al., 2022).

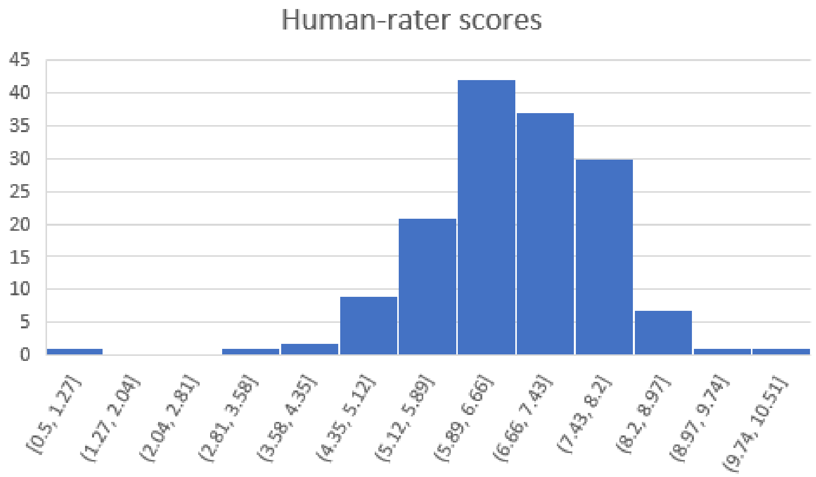

Figure 1. *Histogram of human-rater scores*

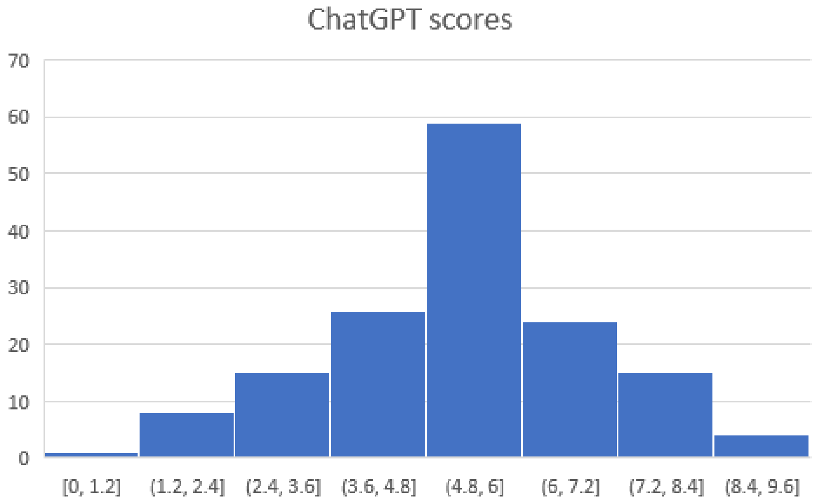

Figure 2. *Histogram of ChatGPT scores*

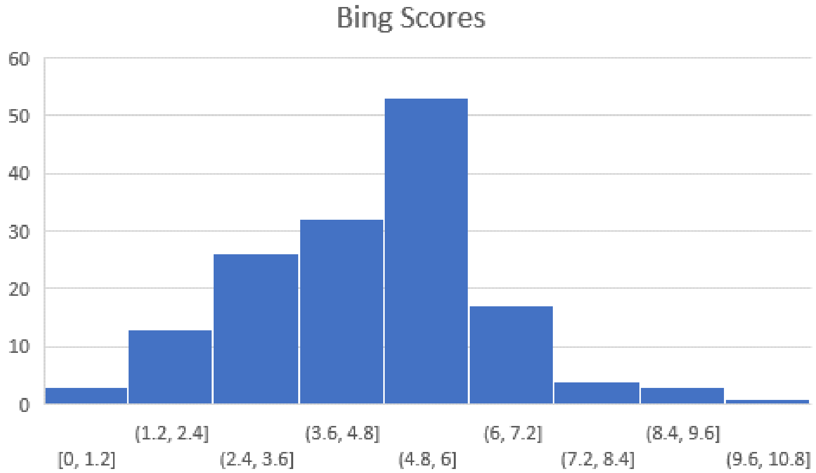

Figure 3. *Histogram of Bing scores*

Figure 4 represents the mean grammar scores of the four human raters and the ChatGPT generated grammar scores which were found to be positively correlated, r(152) = .55, p < .001. Likewise, as shown in figure 5, a positive correlation was observed between the human and the Bing Chatbot scores r(152) = .55, p < .001. These relational results indicate that both the ChatGPT and the Bing Chatbots assigned comparatively similar grammar scores as the human raters to each writing sample. On the other hand, a weak relationship was observed between the human scores and the Bard Chatbot scores, r(152) = .26.

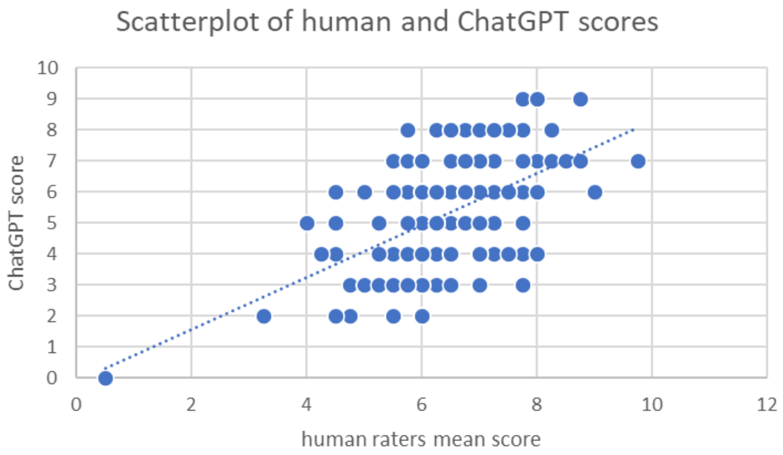

Figure 4. *Scatter plot of the relationship between human scores and ChatGPT scores*

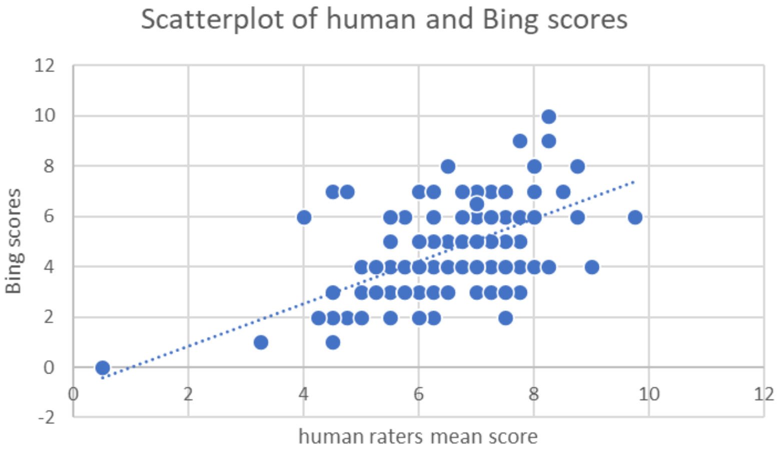

Figure 5. *Scatter plot of the relationship between human scores and Bing scores*

The second research question investigated the inter-rater reliability of both human raters and AI raters when scoring writing grammar. The intraclass correlation coefficient (ICC) was used as a statistical measure of inter-rater reliability as the writing scores were continuous quantitative variables. The ICC score for the four human raters was 0.32. ICC scores range from 0 to 1, with a score below 0.5 indicating weak inter-rater reliability (Koo & Li, 2016). When calculating the ICC score for all four human raters in addition to two of the AI raters- ChatGPT and Bing, the ICC score dropped slightly to 0.30, again indicating weak inter-rater reliability.

Finally, to evaluate the consistency of the grammar scores generated among the different AI Chatbots, again the Pearson correlation coefficient was used. The grammar scores generated by ChatGPT strongly correlated with Bing's scores r(152) = .56, p < .001, whereas the scores generated by Bard showed weak correlations with both ChatGPT r(152) = .24 and Bing r(152) = .05.

Discussion and Practical Implications

The results of this study indicated a moderate positive relationship between the human and ChatGPT generated grammar scores suggesting that AI Chatbots can play a valuable role in assessing language skills, which can be particularly advantageous when teaching productive skills with a large number of students. In addition to generating reliable grammar scores for student writing, the AI Chatbots were able to provide personal feedback on how to improve the grammar, as well as other features of the writing samples. This type of personal writing feedback can have a positive impact on how writing is taught in EFL settings. AI can be used to recognize frequent errors in speaking or writing tasks and generate tailored practice activities to help language learners improve these errors.

When choosing a Chat AI to assist with scoring writing, both ChatGPT and Bing generated similar scores, and the scores were in line with the human raters. Bard on the other hand was not as successful at scoring grammar. ChatGPT and Bing are powered by OpenAI's generative pre-trained transformer model, so the similar score results are not a surprise. Writing instructors are encouraged to use multiple AI chatbots to ensure a more accurate assessment of the writing samples, and the AI Chatbot results need to be carefully reviewed as the scores are not a definitive assessment of the writing. Writing instructors must use their own judgment when making decisions about student writing.

It is also worthy to note that the AI inter-rater reliability was on par or higher than the human raters. Even if the human raters were experienced TEFL instructors and asked to score grammar samples on a scale of 1 to 10 using the same grammar rubric, the inter-rater reliability was not strong. Strong inter-rater reliability requires extensive training and practice, which is typically employed when administering high-stake exams, but seldom in the writing classroom. Therefore, AI scores used in conjunction with human scores may lead to more reliable writing scores.

Finally, a potentially more serious issue is with how the learner makes use of AI to generate his or her writing assignments. If a learner is writing in their L1 and submitting an assignment using an AI translation, then the grammar most likely does not need to be assessed, and if the learner is not actively reading and comprehending the AI-generated L2 text, then the amount of learning taking place is questionable. Students need to be guided on how to best use AI Chatbots to improve their writing. One approach may be to introduce and guide writers on how to make use of AI Chatbots with several practice writing assignments, while instructing them that their final product will be composed without the use of AI.

Limitations and Summary

It is important to note that AI chatbots are not perfect. They are unable to judge whether content is applicable or accurate, and they may misinterpret a writer's style or voice. However, AI chatbots can be a valuable and inexpensive tool for assessing L2 writing, especially when they are used in conjunction with human assessors. AI tends to make superficial corrections to student writing and is often not able to advise on overall structure, content, and organization (Hoang, 2022). There are also serious concerns regarding false information, academic dishonesty and breaches of privacy that need to be assessed when employing AI in educational settings. To summarize, AI Chatbots should not be trusted to make unsupervised decisions.

Regardless of whether teachers and learners embrace these emerging AI tools, AI will continue to evolve as a dynamic learning tool. One can gallantly argue that the printed book, the blackboard, the ball-point pen, the personal computer, and the Internet all transformed the way that we teach and learn. All these technologies were highly criticized in their times.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. *Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing* (pp. 68-75). College Park, MD. https://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Routledge.

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., ... & Buttery, P. (2023). On the application of Large Language Models for language teaching and assessment technology. *arXiv preprint.* https://doi.org/10.48550/arXiv.2307.08393

Chen, H., & Pan, J. (2022). Computer or human: A comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. *Asian Journal of Second and Foreign Language Education, 7*, 34. https://doi.org/10.1186/s40862-022-00171-4

Chen, J., Zhang, M. and Bejar, I. (2017), An Investigation of the e-rater® Automated Scoring Engine's Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model. *ETS Research Report Series, 1*, 1–14. https://doi.org/10.1002/ets2.12131

Dao, X. (2023). Which Large Language Model should you use in Vietnamese education: ChatGPT, Bing Chat, or Bard? *SSRN Electronic Journal.* http://dx.doi.org/10.2139/ssrn.4527476

Gokul, A. (2023). LLMs and AI: Understanding Its Reach and Impact. *Preprints 2023.* https://doi.org/10.20944/preprints202305.0195.v1

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson.

Harwood, C., & Hirose, K. (2019). Factors influencing English as a Foreign Language (EFL) writing instruction in Japan from a teacher education perspective. In L. Seloni and S. H. Lee (Eds.), *Second language writing instruction in global contexts* (pp. 71–90). Multilingual Matters. https://doi.org/10.21832/9781788925877-008

Hoang, G. (2022). Feedback precision and learners' responses: A study into ETS Criterion automated corrective feedback in EFL writing classrooms. *The JALT CALL Journal, 18*(3), 444–467. https://doi.org/10.29140/jaltcall.v18n3.775

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

IBM. (n.d.). What is a chatbot? Retrieved from https://www.ibm.com/topics/chatbots

Kobayakawa, M. (2011). Analyzing writing tasks in Japanese high school English textbooks: English I, II, and writing. *JALT Journal, 33,* 27-48.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education, Assessment, Testing and Applied Measurement, 5*(2020). https://doi.org/10.3389/feduc.2020.572367

Landauer, T. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295–308. https://doi.org/10.1080/0969594032000148154

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215–233. https://doi.org/10.1002/tea.21299

Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence, 5*(7), 333–334. https://doi.org/10.1038/s42256-023-00644-2

Ministry of Education, Culture, Sports, Science and Technology [MEXT]. (n.d.). *Improvement of Academic Abilities (Courses of Study).* https://www.mext.go.jp/en/policy/education/elsec/title02/detail02/1373859.htm.

Mizumoto, A., & Eguchi, M. (2023, March 31). Exploring the potential of using an AI language model for automated essay scoring. https://doi.org/10.31219/osf.io/2uahv

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum.

Polak, S., Schiavo, G., & Zancanaro, M. (2022). Teachers' perspective on artificial intelligence education: An initial investigation. *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7). New Orleans, USA. https://doi.org/10.1145/3491101.3519866

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

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research, 2*(1), 319-330.