Trendy v podnikání - Business Trends (2024), 14(2), 25-34.
https://doi.org/10.24132/jbt.2024.14.2.25_34

COMPARING HUMAN AND AI-BASED ESSAY EVALUATION IN THE CZECH HIGHER EDUCATION: CHALLENGES AND LIMITATIONS

Tomáš Kincl[1], Daria Gunina[2], Michal Novák[3], Jan Pospíšil[4]

[1] Tomáš Kincl, Faculty of Management, Prague University of Economics and Business, email: kincl@vse.cz, ORCID: 0000-0002-9738-3348
[2] Daria Gunina, Faculty of Management, Prague University of Economics and Business, email: daria.gunina@vse.cz, ORCID: 0000-0002-4149-4962
[4] Jan Pospíšil, Faculty of Management, Prague University of Economics and Business, email: jan.zavodny@vse.cz, ORCID: 0000-0003-2054-311X

Abstract: Generative artificial intelligence (GenAI) tools offer innovative capabilities for addressing a wide array of tasks involving extensive datasets, both textual and non-textual. These tools have shown remarkable potential in the field of education, where their functionalities are increasingly leveraged not only by students but also by educators. This study investigates the extent to which human evaluator assessments align with automated evaluations conducted by large language models, with a focus on a) the complexity of the evaluated texts (academic essays that encompass literature reviews, critical assessments of sources, and reflective insights within the context of societal or economic practices) and b) the unique challenges posed by the Czech language, in which the evaluated works are submitted. The research adopts a quantitative (cross-sectional) approach, analysing 30 essays submitted as an assignment for a foundational theoretical course at the master's level. These essays were evaluated by a human evaluator and subsequently by virtual assistants utilizing large language models, specifically ChatGPT (paid version 4.0) and Claude (paid version Sonet 3.5). Statistical analysis revealed that there is a significant statistical difference between human evaluator and both automated systems. Moreover, the evaluations were not consistent when distinguishing between good and less good essays. We also discussed challenges and limitations of using GenAI tools for evaluating submitted text assignments in the context of tertiary education.

Keywords: Automated essay evaluation, generative AI, ChatGPT, tertiary education

JEL Classification: I23

INTRODUCTION

In recent years, advancements in machine learning, natural language processing, and image recognition (often confused with developments in artificial intelligence) have increasingly impacted various human activities, with the potential to fundamentally transform or even render certain tasks obsolete. This trend has not bypassed the domain of (tertiary) education (Lodge et al., 2023; Lye & Lim, 2024). Tools based on large language models (LLMs) are employed (with varying outcomes and ethical implications) not only by students in completing their assignments (Nugroho et al., 2024; Sweeney, 2023; Tossell et al., 2024) but also by educators. For teachers, these tools offer significant assistance in course design, enhancing learning experiences, predicting performance or satisfaction, recommending resources, and more (An et al., 2023; Ouyang et al., 2022). Moreover, these tools play a crucial role in automated scoring (Barrot, 2024; Xu et al., 2024), even in the context of international large-scale assessments involving multilingual student responses (Jung et al., 2024). Automated evaluation can then be a response to the differing results of evaluators with varying levels of experience (Powers et al., 2015). However, the efficacy and reliability of automated

evaluation, especially in the assessment of complex works such as academic essays — where even human evaluators often disagree — remain subjects of ongoing debate (Bui & Barrot, 2024; Vo et al., 2023).
Automated scoring systems, despite significant recent advancements, still face considerable challenges. These systems tend to focus more on formal attributes of the assessed text such as grammar, vocabulary, and other linguistic dimensions, while often overlooking truly critical aspects such as cohesion, creativity, imagination, reasoning, and or idea development and structure (Ramesh & Sanampudi, 2022). Unlike automated systems, human evaluators can identify subtle nuances that machines might miss, leading to a more refined understanding of content. Additional limitations stem from the nature of the tools themselves, often operated as cloud-based applications without direct control by the institution or evaluator. Variations between different versions of these tools, which can yield dramatically different results, further impact the reliability of evaluations. Notably, there are significant differences between paid and free versions of these tools (Song et al., 2024). The outputs of these tools are also highly dependent on the prompts (with even highly experienced evaluators potentially being poor at prompting the evaluation system), the formulation and detail of the assessment rubrics, and potential biases in the training data or in previously evaluated works (Bui & Barrot, 2024; Xu et al., 2024).
A further significant challenge in automated scoring of student works is the national or linguistic context. Although large language models are typically trained on multilingual corpora, English often dominates the source data. The performance of LLMs (not just in automated scoring) can thus vary depending on the language in which the assignment is conducted. While some studies suggest that language might not significantly impact scoring model outputs, this presumption is contingent on a sufficiently large training dataset (Okubo et al., 2023). Other studies indicate that "automated systems could be used to consistently and accurately score essays written in multiple languages" (Firoozi et al., 2024), though agreement with human evaluators typically ranges between 0.7 and 0.8. Further approaches involving automated translation achieve good results (Jung et al., 2024), but questions remain about the potential loss of information or changes in the nature of the text during machine translation prior to evaluation.
To address the question of to what extent automated scoring tools can be used in the Czech tertiary education environment, this study investigates the extent to which human evaluator assessments align with automated evaluations conducted by large language models.
The context of the study is Czech economic and managerial tertiary education, which, despite long-standing efforts at convergence, is characterized by several specific features compared to Anglo-Saxon business education. One of these specifics is the Czech language. This can be overcome by machine translation, but with the limitations mentioned above. While many large language models are trained on multilingual data, the size and significance of the Czech language in these well-known LLMs are very limited relative to the entire training corpus. Another limitation is the funding of public universities in the Czech Republic, which, due to budget constraints, makes it difficult for institutions to acquire specialized automated scoring systems. This often leads to the use of general AI tools such as ChatGPT or Copilot. Additionally, Czech tertiary education is not deeply accustomed to a systematic approach to verifying learning outcomes (based on comprehensive schemes that from formulating general learning outcomes with regard to graduate profiles to specific assessment rubrics within individual course assessments), as is common in the Anglo-Saxon environment (MEYS, 2016). Evidence of this is the relatively limited number of institutions that have successfully passed international evaluations such as ACBSP or AACSB.

1. DATA AND METHODS

The aim of this study was to investigate the extent to which human evaluator assessments align with automated evaluations conducted by large language models. The research sample comprised (academic) essays submitted by students enrolled in the master's course Strategic Marketing at the Faculty of Management, Prague University of Economics and Business. This is a fundamental theoretical course

that forms part of the final state examination and significantly contributes to fulfilling the faculty's graduate profile. The purpose of the assigned essay was to assess the student's ability to understand advanced marketing concepts, conduct a review of primary academic literature regarding the chosen concept, and link theoretical knowledge with practical realities to allow students to reflect on the practical implications of theoretical insights.

The essay assignment involved selecting an area in which the student would develop their work (e.g., disruptive technology, market commoditisation, customer satisfaction). The student then conducted a review of primary academic sources (a pre-selected list of prestigious academic journals in the field of marketing, either ABS 4*, 4-star rating, or Web of Science Q1) and chose a construct to work with in the essay. Examples of such constructs include technology readiness, environmental awareness, fear of missing out, post-purchase regret, among others, depending on the chosen topic.

The academic essay itself had to comprise four parts:

1. Introduction of the Construct. The first part is dedicated to explaining the choice of construct and its relation to the chosen thematic area in which the student is working.
2. Definition. The second part focuses on the definition(s) of the selected construct based on a literature review. The aim of this part is to explain what the construct means, how its conception has evolved in the literature, and (if applicable) compare how it is variously operationalised and measured in different studies.
3. Relationships to Other Constructs. The third part then focuses on the relationship of the selected construct to other constructs, phenomena, events, etc. The aim of this part is to provide an overview of what is influenced by the selected construct, or what the selected construct is influenced by, the types of research in which it appears, the respondent groups for which it has been found significant, the industries in which it is utilised, etc.
4. Validity. The fourth part is devoted to the student's own reflection on the validity of the construct for marketing practice, whether it serves or could serve to measure marketing within a company, etc. The argumentation of the student's own opinion must be supported by sources: i.e., case studies, statistics, overviews, industry and sector reports, etc.

The text is expected to be an academic essay of 7–10 pages in length (excluding the bibliography), with the bibliography containing at least ten sources (primarily high-quality academic journals). The essay is assessed according to the criteria in Tab. 1.

Tab. 1: Essay Evaluation Criteria

Criterion	Description	Points
Construct choice	• The choice of construct is explained in the essay. • The chosen construct relates to the selected topic in InSIS.	5
Definition	• Definitions of the selected construct by various authors are presented. • It is explained what the given construct means, how its conception has evolved in the literature. • A comparison is made, if applicable, of how it is operationalized differently, how it is measured in various studies, etc. • This part is based on a literature review. Cross-referencing of sources and critical evaluation of sources are conducted.	8
Relationships of the selected construct to other constructs	• It is presented what the selected construct influences. • It is presented what influences the selected construct. • Discussion of the types of research in which it appears, the respondent groups for which it has been found significant, the industries in which it is utilised, etc. • This part is based on a literature review. Cross-referencing of sources and critical evaluation of sources are conducted.	12
Validity	• It is explained how the construct is used or could be used in real marketing practice, whether it serves/could serve to measure marketing in the company, etc. • The argumentation of the student's own opinion is supported by sources: i.e., case studies, statistics, overviews, industry and sector reports, etc.	10
Sources	• At least 6 high-quality sources (see list of high-quality academic journals in marketing). • At least 10 sources in total, primarily academic articles. For part four, other sources may include industry reports, studies published by professional or analytical institutions, overviews, annual reports, statistics. • All sources are cited in the text. • All sources are listed in the bibliography.	8
Formal requirements	• Compliance with the provided template. • Logical structure and coherence of individual parts of the document. • Ability to express ideas clearly and describe examined phenomena or events comprehensibly.	7
Total		50

Source: course syllabus

Students submit the essay at the end of the semester. The analysis included essays prepared by students in the part-time (combined) form of study. The submitted works were in the Czech language. A total of 15 randomly selected essays from the group of above-average essays and 15 randomly selected essays from the group of the lowest-scoring essays in the given semester (Fall semester 2023–2024) were included in the analysis. Scoring of selected essays ranged from 0 to 50 points.

The human evaluator's assessments were obtained from the faculty's information system. The automated evaluations were conducted as follows: the LLM systems used were ChatGPT in the paid version 4.0 and Claude 3.5 Sonet, also in the paid version. For the purposes of evaluation in ChatGPT, a specific ChatGPT Assistant was created; while in Claude, a standard chat was used. The prompt was identical for both platforms and was based on recommendations published on 9 May 2024 by Harvard Business Publishing (Mollick & Mollick, 2024) and a document by Adam Peruta (Peruta, 2024).

The evaluation and comparison of all assessments involved comparing the point scores obtained by both methods, specifically assessing whether the human and machine evaluations agreed on the categorisation of essays into pass/fail groups.

2. RESULTS

Descriptive characteristics of the dataset are summarised in Table 1. The evaluations of individual essays are summarised in Figure 1. Figure 2 visually displays the similarity between the evaluations by the human evaluator, ChatGPT, and Claude. Outliers at the zero level in the case of the human evaluator reflect the assessment of detected plagiarism. In these instances, LLMs were unable to detect plagiarism and therefore evaluated the essays as original works.

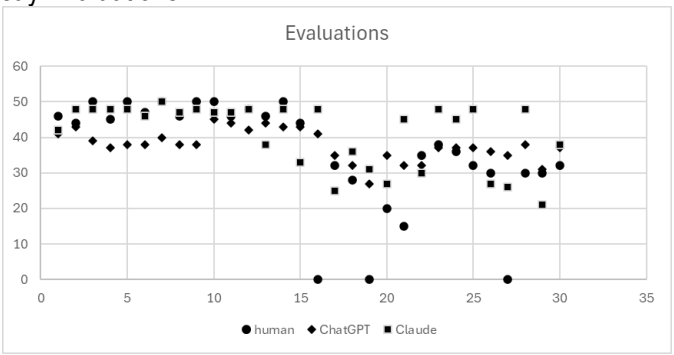
Tab. 2: Descriptive Statistics

	Human	ChatGPT	Claude
Mean	36,67	37,83	40,97
Median	41	38,00	46,50
Variance	236,44	18,90	82,31
Minimum	0	27,00	21
Maximum	50	45,00	50
Interquartile Range	17,25	6,25	15,50
Percentile 25	30,00	35,00	32,50
Percentile 50	41,00	38,00	46,50
Percentile 75	47,25	41,25	48,00

Source: own elaboration

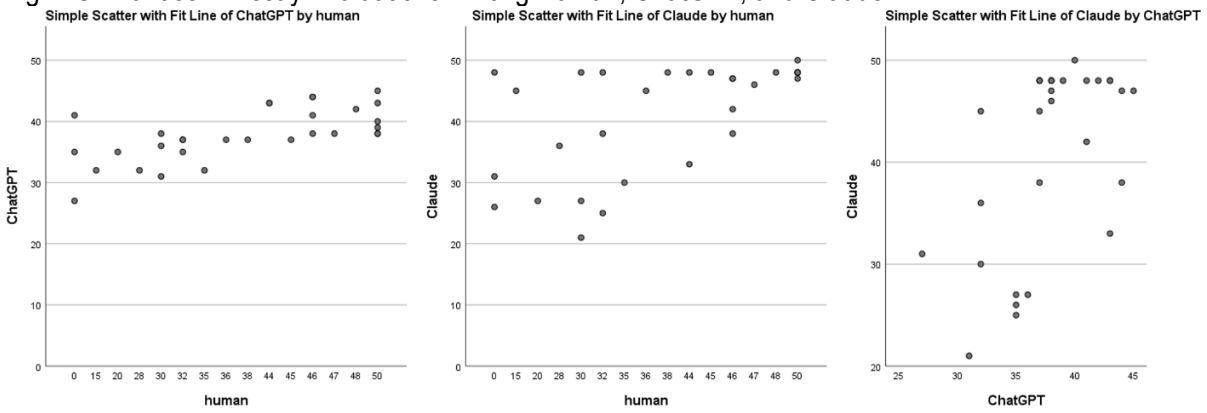
Figure 1 and Figure 2 show that the human evaluator utilised a much broader range of the scoring scale (0–50). Additionally, several instances of plagiarism were detected and were assigned a score of zero. Despite both LLMs being initiated with the same prompt, their results differed. Besides numerical scores, these models also provided verbal feedback on various aspects of the evaluated essays.

Fig. 1 Comparison of Essay Evaluations



Source: own elaboration

Fig. 2 Similarities in Essay Evaluations Among Human, ChatGPT, and Claude



Source: own elaboration

Generally, ChatGPT offered much more positive verbal feedback. This feedback was also more general and, although it essentially reflected the quality of the work, it did not provide specific guidance for further improvement. For example:

> **ChatGPT (Essay 12), total score 42/50 points:**
> *"The definition of the construct is carefully elaborated and supported by academic sources. Various definitions and methods of operationalisation have been presented, indicating a good literature review. Critical analysis could be expanded with a deeper comparison of how different authors approach the operationalisation of this construct."*

> **ChatGPT (Essay 28), total score 31/50 points:**
> *"The essay provides a quality overview of definitions and the historical development of the concept of international segmentation. The author cites key academic sources and shows various approaches to the operationalisation of the construct. However, it would be appropriate to further develop a critical comparison of different segmentation methods and analyse their practical use in specific industries."*

In contrast, Claude was better at identifying shortcomings in individual aspects and reflected them both in the unstructured commentary and in the point evaluation.

> **Claude (Essay 12), total score 48/50 points**:
> *"The essay provides a comprehensive overview of the development of the construct's definition. The author presents various definitions and perspectives on the construct from different authors. The evolution of the construct's conception in the literature is very well described, including changes in its perception. The author critically evaluates different approaches to defining the construct."*

> **Claude (Essay 28), total score 21/50 points:**
> *"The essay provides a general overview of distribution channels but does not meet the main requirement of the assignment, which is to focus on a specific, well-defined construct in the field of marketing. This significantly affects the evaluation across all main criteria. The work contains useful information on various aspects of distribution channels but lacks the depth and academic rigour expected at this level of study. The insufficient focus on a specific construct means that the essay could not provide a detailed analysis of definitions, relationships to other constructs, and validity for marketing practice."*

The first hypothesis posits that there is no significant difference between the evaluations of the individual evaluators. To test this hypothesis, Kendall's W metric was utilised. Kendall's W (or Kendall's coefficient of concordance) is not sensitive to violations of the normality assumption and allows for assessing inter-rater reliability when more than two raters are involved. The coefficient ranges from 0 (no agreement) to 1 (complete agreement) and serves as an effect size measure for the Friedman test (Privitera, 2023). The Friedman test did not reveal statistically significant differences between the evaluations of the three evaluators, $\chi^2(2)=4.949$, $p=0.084$, with Kendall's W = 0.082, suggesting no significant agreement among evaluations. Human evaluation, ChatGPT, and Claude LLMs do not show a statistically significant level of concordance.

Tab. 2: Pairwise Comparison of Evaluations

	ChatGPT	**Claude**	**Human**
ChatGPT		0.391** (p=0.005)	0.532** (p=0.000)
Claude	0.391** (p=0.005)		0.448** (p=0.001)
Human	0.532** (p=0.000)	0.448** (p=0.001)	

*Note: **. Correlation is significant at the 0.01 level (2-tailed).* *Source: own elaboration*

Further analysis focused on pairwise comparison of evaluations. For this purpose, correlation analysis using Kendall's tau was employed. Kendall's tau is more suitable than Spearman's correlation coefficient because it is less sensitive to small sample sizes and outliers. As shown in Table 2, there is a statistically significant moderate positive correlation between the human evaluation and ChatGPT ($\tau b = 0.532$, $p = 0.000$), and similarly between the human evaluation and Claude ($\tau b = 0.448$, $p = 0.001$). Therefore, while substantial differences existed among the evaluations, pairwise comparisons with the human evaluator indicated a moderately strong, statistically significant relationship with the automated scoring using LLMs ChatGPT and Claude. The differences may have been primarily due to the fact that the automated scoring systems were unable to detect plagiarism in the essays and therefore evaluated the manuscripts as original texts. The human evaluator then attempted to assess the plagiarised essays similarly to original authored texts.

Tab. 3: Pairwise Comparison of Evaluations Without Considering Plagiarism

	ChatGPT	**Claude**	**Human**
ChatGPT		0.391** (p=0.005)	0.560** (p=0.000)
Claude	0.391** (p=0.005)		0.518** (p=0.000)
Human	0.560** (p=0.000)	0.518** (p=0.000)	

*Note: **. Correlation is significant at the 0.01 level (2-tailed).* *Source: own elaboration*

As shown in Table 3, when unethical behaviour by students in writing the essays was not considered, there was a statistically significant moderate positive correlation between the human evaluation and ChatGPT ($\tau b = 0.560$, $p = 0.000$), and similarly between the human evaluation and Claude ($\tau b = 0.518$, $p = 0.000$). Although the correlations between the evaluations are significant, fundamental differences still exist between the assessments by the human evaluator and the automated scoring systems utilising the most commonly used LLMs. In many cases, however, when evaluating complex texts, the focus is not so much on assigning a specific point value as on categorising the works into good and poor (i.e., passed/failed, or a more detailed classification into multiple groups).

The final part of the analysis focused on comparing whether the different approaches could at least similarly categorise the essays as good/poor (according to the point evaluation, without considering plagiarism). See Tables 4 and 5 for more details. For the comparison between the human evaluator and Claude, both sensitivity and specificity are 0.67. For the human evaluator and ChatGPT, sensitivity and specificity are both 0.76. Thus, both models can correctly categorise (assuming the human evaluator's assessment is "correct") approximately two-thirds to three-quarters of the evaluated cases.

Tab. 4: Comparison of Categorisation Between Human Evaluator and Claude

		Human	
		Good	Bad
Claude	Good	12	3
	Bad	3	12

Source: own elaboration

Tab. 5: Comparison of Categorisation Between Human Evaluator and ChatGPT

		Human	
		Good	Bad
ChatGPT	Good	13	2
	Bad	2	13

Source: own elaboration

3. DISCUSSION

This study explored the extent to which automated scoring tools can be utilised within the context of Czech tertiary education, specifically by examining the alignment between human evaluator assessments and automated evaluations conducted by LLMs such as ChatGPT and Claude. The results indicate that (despite considerable enthusiasm and high expectations generated by the advent of LLMs among students and, to a lesser extent, educators) commonly available LLMs are only marginally capable of providing reliable and balanced automated scoring, especially when evaluating complex tasks such as essays. Our empirical evidence demonstrates a significant discrepancy between the scoring provided by AI tools (Claude, paid version 3.5 Sonet, and ChatGPT, paid version 4.0) and that of an experienced human evaluator, coupled with a lack of internal consistency in the AI's scoring methodology. These findings corroborate the results of other studies (e.g. Bui & Barrot, 2024; Guo & Wang, 2024; Klyshbekova & Abbott, 2024; Schmidt-Fajlik, 2023; Shin & Lee, 2024). However, contrary to some prior research (Almusharraf & Alotaibi, 2023), our models did not exhibit a tendency to assign lower scores compared to the human evaluator: instead, their scores were generally higher, and the models were less critical in their evaluations.

Several factors may have contributed to these outcomes. One significant consideration is the quality of the input prompt used for automatic evaluation. Prompt engineering is emerging as a demanding discipline, and educators typically lack the necessary competencies and experience in this area. In addition to other aspects (such as access to paid versions of LLMs), the setting of parameters, over which regular users have limited control through standard web interfaces (e.g., temperature settings), also plays a crucial role (Tang et al., 2024).

Another important factor is the quality and specificity of the assignment instructions and evaluation criteria. Human evaluators may perceive and interpret these differently, potentially incorporating contextual knowledge or long-term experience that is not explicitly part of the assignment. In the Czech context, universities are often not accustomed to the explicit formulation of learning outcomes, including detailed decomposition down to individual subjects, components of evaluation, and the formulation of rubrics that distinguish between levels of achievement of learning outcomes.

The training datasets on which LLMs are based also influence their performance. While these models are often trained on extensive corpora obtained through web scraping, the majority of this data is in English or other widely used languages. Consequently, less commonly used languages like Czech are underrepresented in these corpora, potentially affecting the models' ability to accurately process and evaluate texts in these languages. Although some studies suggest that LLMs may possess capabilities for more exhaustive error detection and can simultaneously evaluate multiple facets of written composition (Bui & Barrot, 2024), it often occurs that these systems focus more on formal attributes of the assessed text (such

as grammar, vocabulary, and other linguistic dimensions) and may overlook critical aspects like cohesion, creativity, imagination, reasoning, or idea development and structure (Ramesh & Sanampudi, 2022).
Another factor contributing to the observed discrepancies is the continuous development and evolution of LLMs. Ongoing changes in algorithms, data storage, representation methods, and the addition of training data and retraining of models can lead to variations in performance. Furthermore, the complex nature of LLMs can introduce instability and randomness, with identical prompts entered at similar times but in different instances of the model potentially yielding significantly different results (Ray, 2023).
Despite the expectations that may arise from limited experience with LLMs on straightforward tasks requiring mainly encyclopaedic knowledge, our study highlights shortcomings when these models are applied to truly complex tasks that demand consistency and reliability in evaluating individual inputs or even in repeated evaluations of the same input.

CONCLUSION

Our study confirms that the use of automated scoring systems based on widely available LLMs—even in their paid versions—still has significant limitations and that these systems do not provide evaluations sufficiently consistent with those of human evaluators. Additionally, inconsistencies emerge between individual models, as well as between different instances or runs within the same model. Randomness and inconsistency in LLM outputs can cause significant problems in the acceptance of such results by students and raise questions regarding the accountability of automated evaluations. This does not imply that automated scoring systems are without utility in essay evaluation. They can assist in automating or expediting a range of supplementary activities. However, to become truly valid assistive technologies, further development and research are necessary. This includes the creation and training of models tailored to specific contexts, as well as the development of educators' competencies to fully exploit the opportunities presented by LLMs. Importantly, it should also involve educating students on the ethical and responsible use of LLMs.

Acknowledgemnt
This study was conducted as a part of the International Visegrad Fund no 22410207: Innovation of the education process towards implementing AI tools.

REFERENCES

Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and Automated Essay Scoring Approaches. *Technology, Knowledge and Learning*, *28*(3), 1015–1031. https://doi.org/10.1007/s10758-022-09592-z

An, X., Chai, C. S., Li, Y., Zhou, Y., & Yang, B. (2023). Modeling students' perceptions of artificial intelligence assisted language learning. *Computer Assisted Language Learning*, 1–22. https://doi.org/10.1080/09588221.2023.2246519

Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, *29*(6), 7155–7179. https://doi.org/10.1007/s10639-023-12083-y

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, 1–18.

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2024). Using Automated Procedures to Score Educational Essays Written in Three Languages. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12406

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Jung, J., Tyack, L., & von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *LARGE-SCALE ASSESSMENTS IN EDUCATION*, *12*(1). https://doi.org/10.1186/s40536-024-00199-7

Klyshbekova, M., & Abbott, P. (2024). ChatGPT and Assessment in Higher Education: A Magic Wand or a Disruptor?. *Electronic Journal of E-Learning*, *22*(2), 30–45.
Lodge, J., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *AUSTRALASIAN JOURNAL OF EDUCATIONAL TECHNOLOGY*, *39*(1), 18–18. https://doi.org/10.14742/ajet.8695
Lye, C., & Lim, L. (2024). Generative Artificial Intelligence in Tertiary Education: Assessment Redesign Principles and Considerations. *EDUCATION SCIENCES*, *14*(6). https://doi.org/10.3390/educsci14060569
MEYS. (2016). *Využití výsledků učení na vysokých školách: Příručka pro pedagogickou praxi a vedení VŠ*. https://msmt.gov.cz/uploads/odbor_30/Jakub/Prirucka_Vyuziti_vysledku_uceni_na_vysokych_skolach_Impuls.pdf
Mollick, L., & Mollick, E. (2024). *Instructor prompts*. More Useful Things. Retrieved October 10, 2024, from https://www.moreusefulthings.com/instructor-prompts
Nugroho, A., Andriyanti, E., Widodo, P., & Mutiaraningrum, I. (2024). Students' appraisals post-ChatGPT use: Students' narrative after using ChatGPT for writing. *INNOVATIONS IN EDUCATION AND TEACHING INTERNATIONAL*. https://doi.org/10.1080/14703297.2024.2319184
Okubo, T., Houlden, W., Montuoro, P., Reinertsen, N., Tse, C. S., & Bastianic, T. (2023). *AI scoring for international large-scale assessments using a deep learning model and multilingual data*.
Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, *27*(6), 7893–7925. https://doi.org/10.1007/s10639-022-10925-9
Peruta, A. (2024). *Prompts for faculty*. Peruta. Retrieved October 10, 2024, from https://peruta.com/ai/prompts-for-faculty.html
Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education*, *28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920
Privitera, G. J. (2023). *Statistics for the behavioral sciences*. Sage publications.
Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2
Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154.
Schmidt-Fajlik, R. (2023). ChatGPT as a Grammar Checker for Japanese English Language Learners: A Comparison with Grammarly and ProWritingAid. AsiaCALL Online Journal, 14(1), 105–119. https://doi.org/10.54855/acoj.231417
Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*, 1–23.
Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES*, *17*, 1920–1930. https://doi.org/10.1109/TLT.2024.3396873
Sweeney, S. (2023). Who wrote this? Essay mills and assessment-Considerations regarding contract cheating and AI in higher education. *INTERNATIONAL JOURNAL OF MANAGEMENT EDUCATION*, *21*(2). https://doi.org/10.1016/j.ijme.2023.100818
Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, *10*(14). https://doi.org/10.1016/j.heliyon.2024.e34262
Tossell, C., Tenhundfeld, N., Momen, A., Cooley, K., & de Visser, E. (2024). Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. *IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES*, *17*, 1069–1081. https://doi.org/10.1109/TLT.2024.3355015
Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for english learners in a statewide evidence-based writing assessment. *Assessing Writing*, *56*, 100719. https://doi.org/10.1016/j.asw.2023.100719
Xu, W., Mahmud, R., & Hoo, W. (2024). A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios? *IEEE ACCESS*, *12*, 77639–77657. https://doi.org/10.1109/ACCESS.2024.3399163