AI Versus Human Graders: Assessing the Role of Large Language Models in Higher Education

Abstract

While AI grading is seeing an increase in use and adoption, traditional educational practices are also forced to adapt and function together with AI, especially in assessment grading. In retrospect, human grading, on the other hand, has long been the cornerstone of educational assessment. Traditionally, educators have assessed student work based on established criteria, providing feedback intended to support learning and development. While human grading offers nuanced understanding and personalized feedback, it is also subject to limitations such as grading inconsistencies, biases, and significant time demands. This paper explores the role of large language models (LLMs), such as ChatGPT-3.5 and ChatGPT-4, in grading processes in higher education and compares their effectiveness with that of traditional human grading methods. The study uses both qualitative and quantitative methodologies, and the research extends across multiple academic programs and modules, providing a comprehensive assessment of how AI can complement or replace human graders. In study 1, we focused on (*n=195*) scripts in (*n=3*) modules and compared GPT 3.5, GPT 4, and human graders. Manually marked scripts exhibited an average of 24% mark difference. Subsequently, (*n=20*) scripts were assessed using GPT-4, which yielded a more precise evaluation. Total average of 4% difference in results. There were individual instances where marks were higher, but this could not naturally be a marker judgment. In Study 2, the results from the first study highlighted the need for a comprehensive memorandum; thus, we identified (*n=4341*), among which (*n=3508*) scripts were used. The study found that AI remains efficient when the memorandum is well-structured. It was also found that while AI excels in scalability, human graders excel in interpreting complex answers, evaluating creativity, and picking up plagiarism. In Study 3, we evaluated formative assessments in GPT 4 (statistics *n=602*, Business Statistics *n=859* and Logistics Management *n=522*). The third study demonstrated that AI marking tools can effectively manage the demands of formative assessments, particularly in modules where the questions are objective and structured, such as Statistics and Logistics Management. The initial error in Statistics 102 highlighted the

importance of a well-designed memorandum. The study concludes that AI tools can effectively reduce the burden on educators but should be integrated into a hybrid model in which human markers and AI systems work in tandem to achieve fairness, accuracy, and quality in assessments. This paper contributes to ongoing debates about the future of AI in education by emphasizing the importance of a well-structured memorandum and human discretion in achieving balanced and effective grading solutions.

Keywords: Artificial Intelligence, LLMs, ChatGPT, Higher Education, Assessment, AI Grading

1. INTRODUCTION

The broad consensus is that technology permeates every facet of our lives, and the classroom is no exception. There are scenarios that have guided the present exploration of this study. *First*, imagine a future where assessments are graded not by human hands but by sophisticated algorithms capable of analyzing structure, coherence, and even creativity. Large language models (LLMs) like ChatGPT evolve, they present both opportunities and challenges in the realm of higher education. Are these AI systems mere tools, or can they redefine assessment standards? It can be seen that the nature of education is continually changing because of various innovations and developments. As a result, the need to adapt and transform has become increasingly important. In other words, as new technological advancements shape the landscape of higher education, integrating artificial intelligence (AI) into educational practices has emerged as a transformative force. Among the various AI applications, automated grading systems have gained significant attention. These systems promise to improve the efficiency and consistency of grading, potentially transforming traditional educational assessment practices (Patel and Ragolane, 2024).*Second,* imagine being a student who submits an assessment and promptly receives detailed feedback. You can then have meaningful discussions with your educators, who have thoroughly commented on areas for improvement. Meanwhile, your educators can use AI tools to efficiently create and manage assessments, giving them more time to support you before and after your assessments. The first scenario presents a future in which assessments are no longer graded by humans but by AI, and the second scenario emphasizes that in this process, AI and educators collaborate to improve the quality of education, potentially revolutionizing our approach to teaching and learning. In both scenarios, AI grading aims to make this vision a

reality in education. As educators, policymakers, and students, we must decide whether to embrace these advancements or risk falling behind in an evolving educational landscape. The essence of future education should be to assist students in developing a reliable compass and tools, transformative competency, and navigate an increasingly complex, volatile, and uncertain world. To achieve this primary objective, there is strong advocacy for a new set of curriculum design principles, changes in the school system, a revitalized teacher culture and an alternative assessment program (Schleicher, 2018).

The emergence of AI technologies has presented new opportunities and challenges across various industries, including education (Patel *et al.* 2024). First, we aim to explain the concepts of Large Language Models (LLMs), natural language processing (NLP), natural language processing (GPT), ChatGPT, and OpenAI. LLMs are large-scale, pre-trained, statistical language models based on neural networks (Minae *et al.*, 2024). NLP is a field of AI and Linguistics dedicated to enabling computers to understand human language statements or words (Khurana, Koli, Khatter, & Singh, 2023). OpenAI is the organization responsible for developing ChatGPT, an AI model used to train other AI models. ChatGPT is an AI chatbot developed by OpenAI that uses GPT to perform various natural language processing tasks, such as writing, generating code, and composing sonnets (Kharbach, 2024). There has been an increase in the use of (LLMs), these models include Claude, Bard, Gemini, and Microsoft Copilot. In particular, advanced AI systems, such as OpenAI's ChatGPT models, have been widely used in the public domain by students, the general public, and educators to support their day-to-day activities in school or at work. As a result, LLMs such as ChatGPT have shown remarkable progress in NLP. This groundbreaking AI tool has revolutionized educational paradigms by offering a level of personalization in learning that was previously unattainable. The potential of ChatGPT to serve as an intelligent tutoring system, on the one hand, and as a tool for academic dishonesty, on the other hand, has ignited intense debate within the education sector. Secondary and tertiary educators have expressed concern about the potential for students to abuse ChatGPT and have called for its restriction (Kamalov, Santandreu and Gurrib, 2023). In light of this trend, ChatGPT, with its sophisticated language processing capabilities, is rapidly transforming classrooms to provide personalized educational experiences tailored to each student's unique needs, strengths and weaknesses (Walter, 2024). These models are designed to generate human-like text based on the input received, which makes them potentially useful for tasks such as automated grading.

Figure 1 illustrates the widespread use of LLMS or GPT and presents an overview of the evolution of GPT, as well as its reception in investment and use by the public.

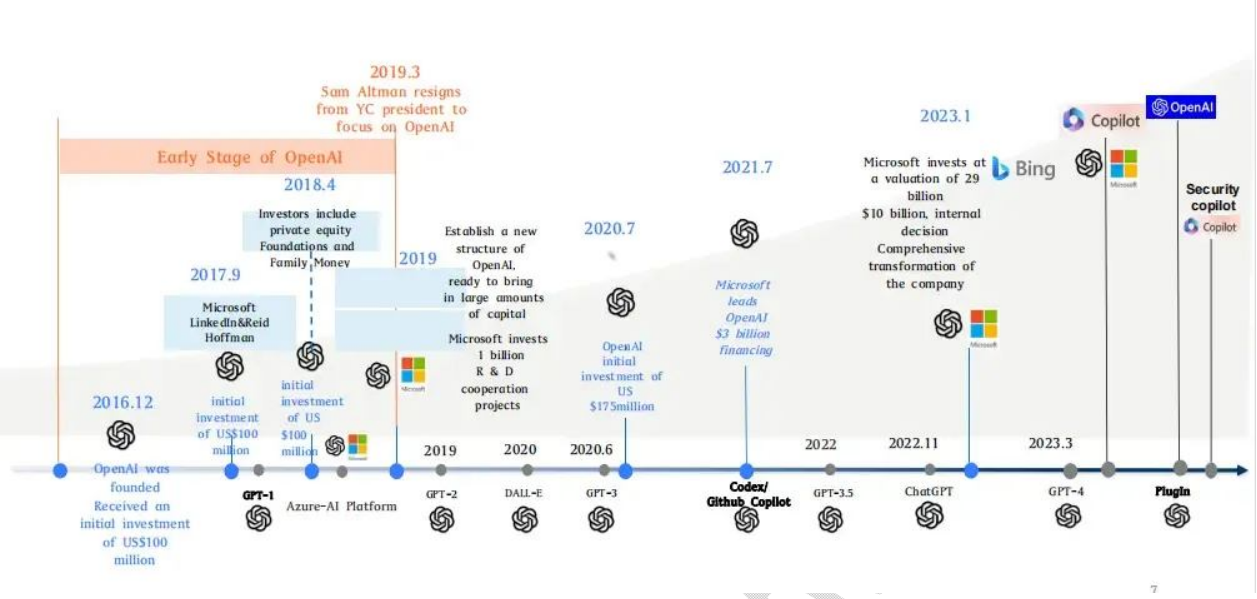


Figure 1: Evolution of AI Writing Models

Source: GameDevNews (2023)

While AI grading is seeing an increase in use and adoption, traditional educational practices are also forced to adapt and function together with AI, especially in assessment grading. In retrospect, human grading, on the other hand, has long been the cornerstone of educational assessment. Traditionally, educators have assessed student work based on established criteria, providing feedback intended to support learning and development. While human grading offers nuanced understanding and personalized feedback, it is also subject to limitations such as grading inconsistencies, biases, and significant time demands. The increasing workloads of educators and the need for fair, unbiased assessments have prompted interest in exploring AI-based grading systems as potential solutions (Patel *et al.* 2024). Despite the promising capabilities of AI, the application of LLM GPT models, such as ChatGPT-3 (GPT 3/3.5) and ChatGPT-4 (GPT 4) in grading academic work raises several questions. *How do these models compare to human graders in terms of accuracy, feedback quality, and fairness?* While AI systems can process large volumes of data quickly and consistently, their ability to replicate the nuanced understanding and judgment of human educators remains a matter of debate (Stoica, 2022; Kurzhals, 2022; Huriye, 2023).

Walvoord and Anderson (1998) found that “teachers, through their personal experiences in the classroom and from listening to faculty from various institutions at workshops around the country, have spent nearly every day of their teaching lives wrestling with the problems, the power and the paradoxes of the grading system”. This paper comprehensively explores the use of LLM GPT models compared to human graders in higher education. This study explores traditional grading methods, challenges, opportunities, and AI grading methods that shape education in South Africa. Whereas a corpus of research has already hinted at the importance of AI and its challenges in education (Patel *et al.* 2024; Opesemowo and Adekomaya, 2024; Funda and Piderit, 2024), this paper assumes that AI and human grading can co-exist and change the future of education. Therefore, this paper is timely as it seeks to address these issues by conducting a comparative analysis of AI models and traditional or human grading practices.

2. RESEARCH AIM

The primary aim of this study is to conduct a comparative analysis of large language models (LLMs), such as ChatGPT-3.5 and ChatGPT-4, against human grading practices in the context of higher education. The goal of this study is to evaluate how AI grading systems perform in terms of accuracy, consistency, and quality of feedback compared to human graders. In addition, we explore the role of memorandum in shaping AI performance and determine the potential benefits and limitations of integrating AI into the grading process. The overarching goal is to assess whether AI models can complement or replace traditional grading systems while maintaining fairness, reliability, and overall efficiency of the assessment process in higher education institutions.

3. RESEARCH METHODOLOGY AND APPROACH

This study employed a qualitative and quantitative research approach to compare the effectiveness of the AI grading systems, specifically, ChatGPT-3.5 and ChatGPT-4, with that of human graders. The research was conducted between June 2023 and November 2023, allowing for a comprehensive evaluation of the grading processes across various academic programs and modules. The study was conducted at a higher education institution in South Africa. This study uses action research design due to its nature to evaluate the instructional design of assessment

marking in higher education. According to Zuber-Skerritt (1991), in educational contexts, action research can enhance teaching practices and improve student learning outcomes. Teachers may investigate their instructional methods by implementing changes in their classrooms and assessing their impact on student engagement and performance. Action research was chosen for its iterative and reflective approach, which is ideal for examining and improving grading practices. The explorative nature of the research questions is reflected in the choice of research design, which is referred to as action research (Defrijin et al. 2007).

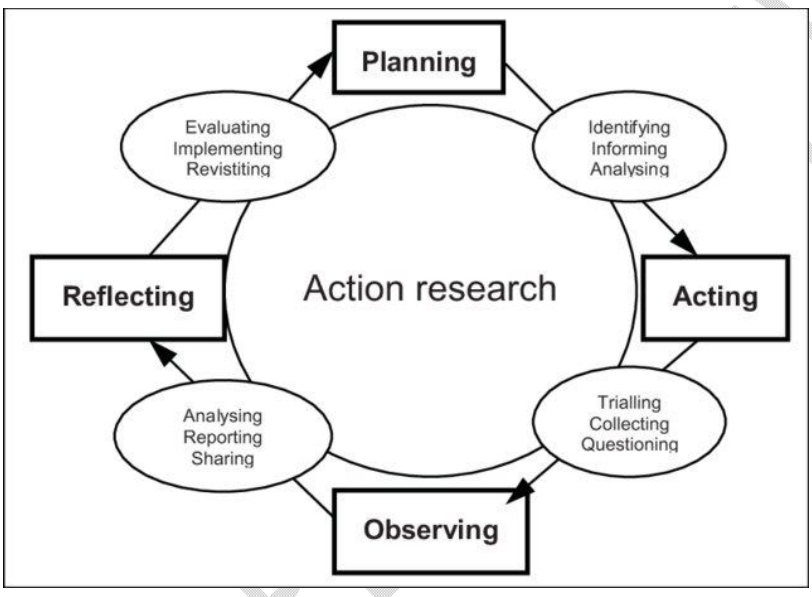


Figure 2: Action research cycle.

Source: Defrijin et al. (2007)

Therefore, the researchers used this method to plan, act, observe, and reflect on the results of the comparison between AI models and human markers. The first phase of this study was aimed at observing the first test of ChatGPT 3.5, ChatGPT 4, and human markers and then in the later cycles, continuously refining methods, data, and interpretation in light of the understanding developed in the earlier cycles (Defrijin et al. 2007). The study used data from existing formative and summative assessments within the school. The AI systems used for comparison were ChatGPT-3.5 and ChatGPT-4, which were applied to the same set of assessments to evaluate their grading accuracy and reliability against human graders.

3.1. Inclusion Criteria

To ensure the validity and comprehensiveness of the comparison, the following inclusion criteria were applied **(see Figure 3):**

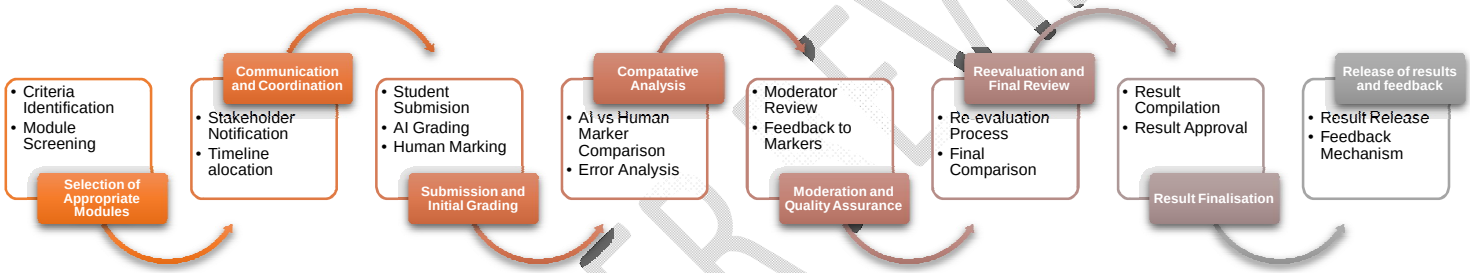
- This study included assessments from the National Qualifications Framework (NQF) Level 6, 7, and 8 programs.
- Only selected modules from the selected programs were used.

Programs	**Modules**
• *Bachelor of Commerce in Supply Chain Management* • *Bachelor of Commerce in Law,* • *Bachelor of Commerce in Retail Management,* • *Diploma in Financial Management* • *Bachelor of Business Administration* • *Bachelor of Human Resource Management*	• Business Management • Business Communication • Statistics • Business Statistics • Logistics Management • Organizational Behavior • Employee Relationship Management • Information and Communication Technology • Economics

Figure 3: Inclusion criteria for programs and modules

Furthermore, only assessments that were formally administered and graded were included. This ensured that the data were representative of typical evaluation practices in the programs. Both

formative and summative assessments were considered to provide holistic views of grading effectiveness. Data were included only from assessments in which student participation was above 80%. This threshold was established to ensure that the sample size was sufficiently large and representative of the student population. The assessments selected for AI and human grading comparison covered a range of question types and difficulty levels, including multiple-choice



questions, short-answer questions, and essay-type questions. This variety ensured that the comparison was thorough and covered different aspects of grading. Only complete and fully documented assessments were included in the analysis. Incomplete or partially recorded assessments were excluded to maintain the integrity of the comparative analysis. **Figure 4** outlines the approach that the researchers took to conduct the research and test the grading practices. This process enhances the education system by involving teachers, school directors, and other stakeholders. The goal is to improve schools and professional areas through thorough study, data gathering, critical analysis, quality planning, effective implementation, and evaluation, with regular reflection (VSO, 2019).

Figure 4: AI vs. human markers process.

3.2. Data Collection and Analysis

Data collection involved retrieving assessment records from the school's database. The records were anonymised to ensure confidentiality. Each assessment was first graded by human evaluators according to established grading rubrics. Subsequently, the same assessments were processed and graded by both ChatGPT-3.5 and ChatGPT-4. The grading accuracy comparison involved statistical analysis to assess the consistency and correlation between the grades awarded by AI systems and human graders. Descriptive statistics, such as mean scores and standard deviations, were calculated for each grade level. Qualitative evaluation was performed by reviewing discrepancies in grading and analyzing feedback from human graders about the AI grading process. This involved thematic analysis of comments and observations to understand the strengths and limitations of each grading method. The performance of ChatGPT-3.5 and ChatGPT-4 was assessed based on their grading accuracy, consistency, and alignment with human grading standards. The differences in grading patterns between the two AI versions were also examined to determine any variations in performance. The study adhered to ethical standards by ensuring the confidentiality and anonymity of student and assessment data. Informed consent was obtained from all relevant stakeholders, including the school administration and faculty. Data were securely stored and access was restricted to authorized personnel. The study was conducted in accordance with institutional guidelines and data protection regulations to ensure ethical integrity throughout the research process.

4. RESULTS AND FINDINGS

4.1. STUDY 1: GPT 3.5 AND GPT 4 vs. human markers

The first study to compare GPT 3.5, GPT 4 and human markers. **Figure 5** presents a summary of the programs, modules, and total number of scripts provided to prospective service providers for assessment using the AI grading tool. The question paper and memorandum were also provided.

Program	**Module**	**No scripts GPT 3.5**	**No scripts GPT 4**	**Human Markers**
Bachelor of Commerce in Human Resource Management Honors ((BCOMHHRM)	Organizational Behavior	50	10	50

Bachelor of Commerce in Human Resource Management Honors (BCOMHHRM)	Employee Relationship Management	45	10	45
Bachelor of Public Administration (BPA)	Information and Communication Technology (ICT)	100	20	50

Figure 5: Programs and modules

4.1.1. Accuracy and Consistency

Table 1: AI Marking of BPA ICT–GPT 3.5

Name	**Overall Score % GPT3.5**	**Overall Score % Human Marker**	**Difference**
Student	48	63	-15
Student	56	64	-8
Student	50	62	-12
Student	30	40	-10
Student	61,5	76	-14,5
Student	33	66	-33
Student	49	86	-37
Student	73	94	-21
Student	57	81	-24
Student	68	89	-21
Student	63	92	-29
Student	62	70	-8
Student	61	84	-23
Student	41	66	-25
Student	56,5	86	-29,5
Student	22,5	39	-16,5

Student	43,5	79	-35,5
Student	30	55	-25
Student	38	59	-21
Student	36	41	-5
Student	57,5	83	-25,5
Student	56	84	-28
Student	37	79	-42
Student	30	56	-26
Student	58	85	-27
Student	38,5	85	-46,5
Student	66,5	76	-9,5
Student	71	83	-12
Student	45,5	81	-35,5
Student	67,5	97	-29,5
Student	46,5	93	-46,5
Student	66,5	86	-19,5
Student	20	32	-12
Student	49	56	-7
Student	54,5	81	-26,5
Student	64,5	89	-24,5
Student	75	95	-20
Student	63,5	98	-34,5
Student	73	86	-13
Student	45,5	78	-32,5
Student	9	16	-7
Student	50,5	75	-24,5
Student	48	82	-34
Student	37,5	67	-29,5
Student	50,5	72	-21,5
Student	33,5	61	-27,5

Student	46	84	-38
Student	24,5	64	-39,5
Student	38,5	79	-40,5
Student	51,5	88	-36,5
Student	61	88	-27
Student	56	92	-36
Student	66	86	-20
Student	46,5	62	-15,5
Student	77,5	93	-15,5
Student	53,5	93	-39,5
Student	81,5	96	-14,5
Student	45,5	95	-49,5
Student	20,5	32	-11,5
Student	43,5	66	-22,5
Student	31,5	71	-39,5
Student	33,5	67	-33,5
Student	7,5	22	-14,5
Student	30,5	32	-1,5
Student	68,5	92	-23,5
Student	38,5	65	-26,5
Student	66,5	95	-28,5
Student	55,5	79	-23,5
Student	67	91	-24
Student	65	81	-16
Student	44	70	-26
Student	35,5	64	-28,5
Student	38,5	66	-27,5
Student	51,5	49	2,5
Student	72	93	-21
Student	46	80	-34

Student	48	72	-24
Student	62	74	-12
Student	12,5	27	-14,5
Student	66	88	-22
Student	59,5	89	-29,5
Student	42,5	76	-33,5
Student	43	76	-33
Student	55,5	82	-26,5
Student	34	44	-10
Student	46,5	78	-31,5
Student	60,5	82	-21,5
Student	77	95	-18
Student	55,5	84	-28,5
Student	63,5	89	-25,5
Student	49,5	72	-22,5
Student	46,5	73	-26,5
Student	54	76	-22
Student	36	70	-34
Student	40	58	-18
Student	43,5	66	-22,5
Student	63	78	-15
Student	51,5	89	-37,5
Student	37	50	-13

The accuracy of the AI models was a key area of focus, particularly in terms of how closely their scores aligned with those of human markers. The analysis revealed that GPT 3.5 showed significant discrepancies in its grading. For example, in the BPA: Information and Communication Technology (ICT) module, GPT 3.5 produced an average score difference of 24% ***(see Table 1)*** compared with human markers, with individual cases displaying even larger discrepancies. Some students' marks differed by as much as 145%, highlighting GPT 3.5's

difficulty in accurately grading application-based or case study-style questions, where answers vary significantly depending on the student's approach. The lack of nuance in GPT 3.5's understanding of these complex responses resulted in inconsistent grading, making it less reliable for such assessments.

Table 2: BPA: ICT RESULTS–GPT 4

Name	**Overall Score % GPT4**	**Human Marker**	**Difference**
student	60,5	62	1,5
student	54	66	12
student	73,5	81	7,5
student	70	70	0
student	63	66	3
student	36,5	39	2,5
student	66,5	56	-10,5
student	92	95	3
student	70,5	78	7,5
student	65,5	75	9,5
student	86	88	2
student	51	62	11
student	22	22	0
student	52	64	12
student	73	82	9
student	89	89	0
student	52	58	6
student	56	66	10
student	78,5	78	-0,5

student	52	50	-2

However, AI models have shown accuracy in recent tests. For example, Gobrecht *et al.* (2024) found that AI systems have demonstrated high accuracy in grading tasks, often outperforming human raters. Their findings revealed that a novel automatic short answer grading system showed a median absolute error that was 44% smaller than that of human graders. Kortemeyer*et al.* (2024) cautioned that AI grading is precise in identifying passing examinations but lacks accuracy and reliability for failing grades, requiring human validation due to Optical Character Recognition (OCR) failures and nuances missed in grading criteria. This study found that GPT 4 performed significantly better in terms of both accuracy and consistency. For the same ICT module, GPT 4's average difference from human grading was only 4%, with only a few outlier cases showing significant variation. This demonstrated GPT 4's enhanced ability to interpret diverse responses and evaluate them more in line with human expectations. This is similar to the findings of Rutner and Scott (2022) that while not extensively used in academia like in other fields, such as the medical community, AI offers potential for students and faculty members, and recent versions of AI will be more updated to handle the expectations of the academic community. For example, the researchers found that GPT 4 was able to reduce the extreme outliers observed in GPT 3.5's performance, showing improved consistency across different types of questions and student responses. The results demonstrate that GPT 4 is a more reliable and accurate tool for grading, particularly in complex, application-driven assessments.

4.1.2. Quality of Feedback

Feedback is essential for meta-cognition because it enables students to comprehend and improve their mistakes (Stoica, 2022). The researchers discovered that the feedback provided by the AI models differed in their ability to provide constructive guidance to students. For example, GPT 3.5 excelled in offering detailed feedback, often breaking down student responses into sub-questions and providing clear explanations of where marks were awarded or deducted. This detailed approach enabled students to understand specific areas for improvement. However, despite its high level of detail, the feedback often lacked nuance in more complex cases. GPT 3.5 could not fully grasp alternative ways to answer questions that fell outside the narrow scope of

the memorandum, resulting in feedback that occasionally misrepresented the quality of the student's work. Stoica (2022) warns that one may question how fair an AI-supported tool can be when trained with the help of past examinations. For example, if the training dataset considers previously graded assignments, there may be cases in which an assessor made a mistake and incorrectly graded a student. GPT 4, on the other hand, offered a notable improvement in this area. Its feedback was more personalized and closer to the style of human markers, with greater depth and clarity. GPT 4 demonstrated a better understanding of students' varied responses, offering tailored feedback that guided students in a more human-like manner. Feedback was not only more accurate but also more reflective of the student's overall performance, balancing technical evaluation with constructive suggestions for improvement. The researchers noticed that in assessments where subjective judgment was required, GPT 4 was able to provide comments that better aligned with human expectations, enabling students to learn more effectively from their mistakes.

4.1.3. Role of Memorandum in AI Performance Shaping

Table 3: BCOMHHRM: Organizational Behavior

Name	Q1 Total mark counts	Q1 Marks Awarded GPT3.5	Human Markers	Difference
Student	25	25	15	10
Student	25	19	14	5
Student	25	23	12	11
Student	25	8	18	-10
Student	25	18	15	3
Student	25	17	15	2
Student	25	17	12	5
Student	25	21	16	5
Student	25	21	15	6
Student	25	16	12	4

Student	25	25	13	12
Student	25	16	14	2
Student	25	19	14	5
Student	25	19	6	13
Student	25	17	14	3
Student	25	19	14	5
Student	25	17	16	1
Student	25	14	12	2
Student	25	19	15	4
Student	25	17	15	2
Student	25	9	14	-5
Student	25	17	12	5
Student	25	16	12	4
Student	25	19	12	7
Student	25	13	14	-1
Student	25	15	15	0
Student	25	13	14	-1
Student	25	13	15	-2
Student	25	13	14	-1
Student	25	13	14	-1
Student	25	13	13	0
Student	25	13	12	1
Student	25	11	11	0
Student	25	17	14	3
Student	25	13	14	-1
Student	25	13	13	0
Student	25	13	14	-1
Student	25	15	13	2
Student	25	13	13	0
Student	25	13	14	-1

Student	25	17	15	2
Student	25	16	14	2
Student	25	17	15	2
Student	25	15	14	1
Student	25	20	16	4
Student	25	21	16	5
Student	25	19	15	4
Student	25	17	14	3
Student	25	19	16	3
Student	25	15	14	1
Student	25	9	6	3

In education, memoranda are used to assess student performance and offer guidance on the expectations from the teacher to the students. Therefore, the researchers aimed to assess the importance of memorandum in shaping the performance of AI models. In assessments in which the memorandum was comprehensive and detailed, both GPT 3.5 and GPT 4 performed relatively well. For example, in the BCOMHHRM: Organizational Behavior module, where the memorandum was designed to cover all expected answers, the average difference between GPT 3.5 scores and human markers was 8.4%. In this context, GPT 3.5 could adhere to the structured criteria and deliver consistent results, albeit with minor deviations.

However, in assessments like the BPA: ICT module, where the questions required a broader range of interpretations and the memorandum may not have captured all possible correct responses, GPT 3.5 struggled significantly. The variability in responses posed a challenge for the model, which relied heavily on the preset answers in the memorandum. This limitation led to various mark differences and highlighted the need for memoranda that are sufficiently flexible to accommodate diverse student answers.

GPT 4, with its more advanced capabilities, demonstrated greater flexibility in interpreting responses, even when the memorandum was not as detailed. It was better able to evaluate responses that fell outside the expected framework, thereby reducing the reliance on rigid marking criteria. Despite this, the findings suggest that the design of the memorandum remains

crucial for maximizing the accuracy and effectiveness of AI grading tools. A well-constructed memorandum allows AI models to function at their best, ensuring that both structured and unstructured responses are graded fairly and consistently.

Table 4: BCOMHHRM–Employee Relationship Management GPT 3.5 vs. Human Graders

Name	Total Score Percentage (GPT) 3.5	Human Marker	Difference
Student	35,5	62	26,5
Student	57,5	75	17,5
Student	55,5	75	19,5
Student	42	38	-4
Student	67,5	61	-6,5
Student	32	62	30
Student	71	62	-9
Student	61	60	-1
Student	57,5	70	12,5
Student	42	59	17
Student	77	62	-15
Student	35	56	21
Student	63	66	3
Student	79	52	-27
Student	52	60	8
Student	55	71	16
Student	47	33	-14
Student	50	70	20
Student	76	51	-25
Student	55,5	61	5,5
Student	39,5	54	14,5

Student	73	60	-13
Student	52,5	49	-3,5
Student	36	63	27
Student	55	51	-4
Student	46,5	60	13,5
Student	63	55	-8
Student	49	25	-24
Student	46	56	10
Student	46	56	10
Student	46	70	24
Student	84,5	70	-14,5
Student	46,5	43	-3,5
Student	42,5	62	19,5
Student	50	65	15
Student	50,5	66	15,5
Student	45	64	19
Student	58,5	58	-0,5
Student	37	35	-2
Student	51	52	1
Student	61	48	-13
Student	51	34	-17
Student	56	40	-16
Student	54	43	-11
Student	56,5	70	13,5
Student	69	77	8
Student			

4.1.4. Efficiency and Time Savings

Although the focus of this study was accuracy and consistency, it is important to note the significant efficiency gains provided by the AI models. Both GPT 3.5 and GPT 4 were able to

process large volumes of student scripts much faster than human markers, providing immediate feedback to students. This was especially valuable in modules with high enrollment numbers, such as the BPA: ICT module, in which hundreds of scripts could be evaluated simultaneously. In contrast, human markers though more flexible and nuanced in their assessments, were susceptible to fatigue, bias, and inconsistencies over time. The immediate feedback offered by the GPT models also played an important role in enhancing the learning process because students were able to understand and correct their mistakes much faster than they would if they were waiting for human grading.

Table 4: BCOMHHRM–Employee Relationship Management GPT 4 vs. Human Graders

Name	**Total Score Percentage GPT 4**	**Human Marker**	**Difference**
Student	61,5	62	-0,5
Student	63	61	2
Student	61	60	1
Student	60	56	4
Student	75,5	70	5,5
Student	59	54	5
Student	50	49	1
Student	71	70	1
Student	46	40	6
Student	75	77	-2

4.2. STUDY 2: GPT 4 vs. human markers

This study compared the performance of GPT 4 AI marking tools against human markers across several Bachelor of Commerce modules during the November 2023 examination period. The key focus was on assessing the accuracy, consistency, and appropriateness of the marks awarded by AI tools versus human markers, as well as understanding the role that memorandum plays in

ensuring reliable outcomes. Based on the analysis and findings from the first study, the researchers sought to determine the impact of a memorandum when it is presented in detail. The study identified n=*4341* scripts in Business Communication 101, Business Management 1, and Economics 1 (see table 5). However, the researchers managed to include n=*3 508* scripts to mark using GPT 4 (see Table 10).

Table 5: Pilot 2 AI grading and the memorandum

Bachelor of Commerce,	**Business Management 1**	**627**
Bachelor of Commerce in Supply Chain Management,	Business Management 1	711
Bachelor of Commerce in Law,	Business Management 1	297
Bachelor of Commerce in Retail Management,	Business Management 1	19
Bachelor of Commerce,	Economics 1	673
Bachelor of Commerce in Supply Chain Management,	Economics 1	741
Bachelor of Commerce in Law,	Economics 1	309
Bachelor of Commerce in Retail Management,	Economics 1	19
Bachelor of Commerce in Accounting	Economics 1	313
Bachelor of Commerce,	Business Communication	240
Bachelor of Commerce in Supply Chain Management,	Business Communication	272
Bachelor of Commerce in Law,	Business Communication	93
Bachelor of Commerce in Retail Management,	Business Communication	27
		4341

Table 6: Business communication

Bachelor of Commerce,	**Business Communication 101**
Bachelor of Commerce in Retail	Business Communication

Management,	
Bachelor of Commerce in Law,	Business Communication
Bachelor of Commerce in Supply Chain Management,	Business Communication 101

4.2.1. Accuracy and Consistency

One of the key themes emerging from the analysis was the significant discrepancies between the marks awarded by AI tools and those given by human markers. These differences were most pronounced in Economics 1, where questions tended to require high levels of interpretation and subjectivity. For example, of the n=108 scripts reviewed in this module, a considerable number exhibited large mark differences between AI and human markers:

Parameter	**Number of Scripts**
1–5 marks	9
6–10 marks	16
11–15 marks	19
16–20 marks	32
21–25 marks	24
26-30 marks	4

Figure 6: Difference Between the Final Marks Awarded for Human and AI Markers

The variance, as shown in **Figure 6**, was primarily driven by AI's inability to interpret complex responses, particularly when students presented alternative or nuanced answers. This was especially evident in economics questions, where the application of diagrams and paraphrasing played a significant role in determining the final mark. Human markers were better equipped to assess these elements, leading to greater discrepancies in the awarding of marks. The majority of

variances fell between n=15 and n=25 marks, highlighting AI's challenges in subjective assessments.

In contrast, AI marking showed smaller discrepancies in modules such as Business Management 1 and Business Communication 101. For these modules, AI performed more consistently because of the structured nature of the assessments and the relatively straightforward application of the memoranda. For example, in Business Communication 101, most scripts had a mark difference of only n=1 to n=12 marks between AI and human markers, indicating that AI's leniency was not as drastic. Similarly, in Business Management 1, where human markers tended to be more lenient, the discrepancy between AI and human markers was in the range of n = 3 to n = 15. However, even in these modules, AI tools generally awarded higher marks compared to human markers, reflecting a trend toward leniency in AI grading.

Table 7: Business management

Bachelor of Commerce,	Business Management 1
Bachelor of Commerce in Retail Management,	Business Management 1
Bachelor of Commerce in Law,	Business Management 1
Bachelor of Commerce in Supply Chain Management,	Business Management 1

4.2.2. Feedback Quality

Though the study did not focus heavily on the feedback aspect, it became evident that AI's ability to offer meaningful, constructive feedback remained limited compared to human markers. AI tools, especially in subjective areas like Economics 1, struggled to recognize nuances in student responses. For instance, in some cases, AI failed to account for copied and pasted content, which was a factor that human markers were more adept at capturing. Furthermore, AI marking lacked the capacity to appreciate the creative or interpretive elements of student answers, which often resulted in higher marks being awarded without sufficient consideration of errors or incomplete application of concepts.

Table 8: Economic 1

Bachelor of Commerce,	Economics 1
Bachelor of Commerce in Retail Management,	Economics 1
Bachelor of Commerce in Law,	Economics 1
Bachelor of Commerce in Supply Chain Management,	Economics 1
Bachelor of Commerce in Accounting	Economics 1

4.2.3. Role of the Memorandum

The design and comprehensiveness of the memorandum played a significant role in the accuracy and reliability of AI marking. As observed in the Business Communication 101 module, issues with the memorandum directly affected the consistency of AI grading. The absence of clear mark allocations for key questions led to discrepancies between the marks awarded by AI tools and human markers. For example, Q1 in Business Communication 101 was allocated n=25 marks, but no detailed mark breakdown was provided, which created the potential for wide variations in how AI interpreted and marked student responses. Similarly, Q2.3, which carried n=10 marks, also lacked clear mark allocation, which led to further inconsistencies.

In contrast, Economics 1 contained a more comprehensive memorandum, yet the discrepancies between AI and human markers remained significant. This suggests that even with a well-structured memorandum, AI struggles to manage questions that require subjective interpretation and creative responses. The findings from Business Management 1 reinforced this point, where memoranda with clearer structures led to smaller differences in the marks awarded by AI and human markers.

The study highlights the fact that while the design of the memorandum is crucial, it is not sufficient to ensure accurate AI grading in modules with high levels of subjectivity. Human markers still excel at interpreting a range of acceptable responses and applying discretion when necessary, particularly in areas such as application and creativity.

4.2.4. Efficiency and Practical Application

One of the undeniable strengths of AI-based marking tools is their efficiency. In this study, AI tools were able to process large volumes of scripts much more quickly than human markers, making them particularly valuable for modules with high student enrollment. For example, the Business Management 1 module saw over n=600 scripts from the Bachelor of Commerce in Supply Chain Management and another n=499 scripts from the standard Bachelor of Commerce program. The shear volume of scripts would be a challenge for human markers to grade in a timely manner; however, AI tools provide nearly instant results.

Similarly, in Economics 1, AI marked over n=600 scripts, demonstrating its scalability and ability to handle large datasets efficiently. This advantage, however, comes with a trade-off in accuracy, particularly for subjective questions. Although AI tools are highly efficient, their inability to interpret nuanced or creative responses limits their applicability in modules where subjective assessments are central to student evaluation.

Table 9: summative assessments Scripts Marked by AI (GPT 4)

BCOM__Jan 23-Business Management 1_07 Nov 23	
BCOMSC_BM100	602
BCOMRM_BM100	13
BCOMLW_BM100	225
BCOM_BM100	499
BCOM__Jan 23-Economics 1_09 Nov 23	
BCOMSC_ECO100	627
BCOMRM_ECO100	13
BCOMLW_ECO100	237
BCOMAC_ECO100	252
BCOM_ECO100	542
BCOM__BC101-Business Communication 101_10 (November 23)	
BCOMLW_BC101-AD-10100	221
BCOMRM_BC101-AD-10135	17
BCOMLW_BC101-AD-10100	65

BCOM-BC-101	195
Total	3508

4.3. STUDY 3: AI MARKING FOR FORMATIVE ASSESSMENTS, JANUARY 2024

This third study explored the application of AI marking tools in formative assessments across several key modules: Statistics 102, Business Statistics 102, and Logistics Management 2. The focus of this study was to assess the ability of AI tools to handle large volumes of formative assessments efficiently and provide accurate and timely feedback to students in low-stake settings. Unlike summative assessments, formative assessments prioritize feedback and learning, making this an important pilot for AI's potential in ongoing student evaluation.

Table 10: formative assessments marked on the GPT 4

Programs	**Module**	**Number of Actual Scripts**
Bachelor of Commerce,	Statistics 102	602
Bachelor of Commerce in Supply Chain Management	Business Statistics, 102	859
Bachelor of Commerce in Supply Chain Management	Logistics Management 2	522

The AI tool was used to mark *n=602* scripts from the Bachelor of Commerce program. To ensure accuracy, the first *n=100* scripts were sent to a moderator for comparison with human marking. During this review, a 5-mark discrepancy was discovered in one question due to an error in the memorandum's mark allocation. This mistake led to *n=49* student queries because their marks were lower than expected. Once the error was identified, the memorandum was corrected, and the students' marks were adjusted accordingly. After this correction, the AI-marked scripts showed no further issues, demonstrating the tool's ability to accurately mark large volumes of

assessments once the memorandum was properly calibrated. This experience highlighted the importance of ensuring the memorandum's precision to prevent such discrepancies.

In the Business Statistics module, the AI tool was applied to *n=859* scripts from the Bachelor of Commerce in Supply Chain program. Unlike the Statistics 102 pilot, no significant discrepancies or student queries were noted. The module's structure and nature of the questions lent themselves well to AI marking because the assessments were largely objective. The smooth performance of the AI tool in this module underscores its effectiveness in grading objective, structured assessments with high accuracy and consistency, even when using several scripts.

The AI tool was also tested on *n=522* scripts in the Logistics Management 2 module. The pilot was considered highly successful, with only minor discrepancies noted. The average difference between AI-marked and human-marked scripts was only *n = 2* marks, reflecting the tool's accuracy in this module. The consistency in grading, combined with the minimal variance, suggests that AI can be a reliable tool for formative assessment, particularly in modules in which questions are straightforward and evaluation criteria are clear.

The results of this study demonstrate that AI marking tools can effectively manage the demands of formative assessments, particularly in modules where the questions are objective and structured, such as Statistics and Logistics Management. Although the initial error in Statistics 102 highlighted the importance of having a well-designed memorandum, the overall performance of the AI tool in this study was encouraging. It demonstrated that, once properly calibrated, AI can provide accurate, efficient, and scalable grading, offering students timely feedback while significantly reducing the burden on human markers. This makes AI a valuable tool for formative assessments, especially in large courses with high enrollment.

5. DISCUSSION

The findings of this study provide valuable insights into the performance of AI grading tools compared to human markers in higher education. A significant observation is the difference between AI and human grading, especially in subjective or application-based assessments. Colonna (2024) supported this claim by stating that teachers should have the authority to override decisions made by automatic assessment software and should be provided with clear instructions and criteria for interpreting the results of an automated assessment system. This

would enable students to make informed assessments about whether or not to accept the results. ChatGPT-3.5, in particular, showed a marked inconsistency in its grading accuracy, especially in cases where students provided nuanced or varied responses. AI struggled to grasp the complexity of these answers, often adhering rigidly to the predetermined memorandum, resulting in larger discrepancies between AI and human-assigned marks. Ragolane and Patel (2024) raised this concern that students fear that AI will not have the capacity to grade essay questions that require grassroots understanding as opposed to multiple-choice questions, despite the use of a memorandum. In contrast, ChatGPT-4 demonstrated noticeable improvements in both accuracy and consistency, particularly in more objective assessments where the memorandum clearly outlined the expected answers. However, even the more advanced GPT-4 model encountered challenges when dealing with subjective questions, where human markers were better at evaluating creativity, critical thinking, and unique problem-solving approaches. The results indicate that although AI is effective for structured, objective assessments, it still requires human oversight for more interpretive tasks. These findings are similar to Kurzhals (2022), who found that although some students believe that AI will be able to grade examinations independently in the future, the majority stated that AI still needs the support of teachers. This suggests that the trust and acceptance of AI among students are not yet fully established. In addition, a few students in his study expressed a preference not to take AI grade exams at all, further supporting this implication.

A key factor influencing AI grading performance was the quality of the memorandum. In modules where the memorandum provided detailed, clear instructions, AI tools—especially ChatGPT-4—performed better, delivering more accurate and consistent results. However, when the memorandum was vague or did not include specific mark breakdowns, discrepancies between AI and human grading were more pronounced. This underscores the importance of designing a well-structured memorandum to guide AI models in grading complex responses. Another critical finding was the efficiency of AI grading. Both ChatGPT-3.5 and ChatGPT-4 saved significant time by processing large volumes of assessments much faster than human markers. Previous research (see Opesemowo et al. 2024; Ragolane and Patel, 2024)) demonstrated that have the potential to save educators time and enhance efficiency in grading student work, thus giving them enough time to engage with students. In courses with many students, the use of AI enabled quicker turnaround times, allowing students to receive feedback

more promptly. This efficiency can play a pivotal role in formative assessment, where timely feedback is essential for student learning and improvement. However, while AI offers speed and scalability, there are trade-offs in terms of accuracy and feedback quality. AI-generated feedback is often more generic than that provided by human markers, particularly when evaluating subjective assessments. Although ChatGPT-4 was able to offer more personalized feedback than ChatGPT-3.5, it still lacked the depth and nuance of human commentary, which is essential for guiding students toward areas of critical thinking and creativity.

One thing is certain: the study suggests that AI grading tools, especially ChatGPT-4, can significantly reduce the grading burden on educators, particularly for large-scale and objective assessments. However, human graders are indispensable for assessments that require subjective judgment. A hybrid approach that combines the efficiency of AI with the discretion and interpretive skills of human markers is recommended. This approach allows institutions to leverage the strengths of both systems while mitigating their respective limitations. AI models like ChatGPT-4 show great promise in enhancing the grading process in higher education, particularly when used in tandem with human graders and supported by well-designed memorandum. This integration of AI into the educational system could lead to more efficient and scalable grading solutions; however, human oversight remains crucial for ensuring fairness, accuracy, and quality in assessment, especially in subjective and application-based contexts.

6. CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH

This study highlighted the potential and limitations of using large language models (LLMs) such as ChatGPT-3.5 and ChatGPT-4 for higher education grade classification. Although AI tools offer significant advantages in terms of efficiency and scalability, particularly in handling large volumes of assessments, they are not without limitations. ChatGPT-4 demonstrated improved accuracy and consistency compared to ChatGPT-3.5, especially in objective assessments. However, both models struggled with subjective, application-based questions, where human markers outperformed AI in terms of nuance, creativity, and critical thinking. The results also underscore the crucial role of well-structured memos in guiding AI grading systems, reinforcing the need for comprehensive grading criteria to realize better AI performance. Ultimately, AI can significantly reduce the grading burden; however, human oversight remains essential, especially

for complex assessments that require interpretive judgment.This study also has several limitations that must be addressed:

- The research was conducted within a short period of time and focused on specific modules in a few academic programs. This limits the generalizability of the findings to other disciplines or educational settings.
- The results revealed that AI performance was heavily dependent on the quality of the memorandum. Incomplete or poorly structured marking guides lead to greater discrepancies, limiting AI's effectiveness in such contexts.

Addressing these limitations in future research will provide a more comprehensive understanding of AI's role in grading and help refine strategies for its optimal integration into higher education assessment processes. By exploring how AI and human grading can coexist and complement each other, educators can develop more effective, fair, and efficient grading systems for the future of education.

7. REFERENCES

Colonna, L. (2024). Teachers informedin the loop? An analysis of automatic assessment systems under Article 22 GDPR. *International Data Privacy Law*, 14(1), 3–18. https://doi.org/10.1093/idpl/ipad024

Defrijin, S.; Mathijs, E.; Gulinck, H. & Lauwers, L. (2007). Facilitating and evaluating farmer innovations toward more sustainable energy and material flows: A case study in Flanders. 8th European IFSA Symposium, 6-10 July 2008, Clermont-Ferrand (France). Available online: https://www.researchgate.net/publication/228785130_Facilitating_and_evaluating_farmer_innovations_towards_more_sustainable_energy_and_material_flows_case-study_in_Flanders [accessed Sep 17 2024].

Funda, V. &Piderit, R. (2024). A review of the application of artificial intelligence in South African Higher Education, 2024 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2024, pp. 44-50, doi: 10.1109/ICTAS59620.2024.10507113.

GameDevNews (2023). The Evolution of AI Writing Models: From GPT-2 to the Future, Open AI & ChatGPT News, Linkedin, https://www.linkedin.com/pulse/evolution-ai-writing-models-from-gpt-2-future-open-ai-gpt-news-95zlc/

Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., & Schütt, S. (2024). Beyond human subjectivity and error: A novel AI grading system. ArXiv. /abs/2405.04323

Huriye, A. Z. (2023). The Ethics of Artificial Intelligence: Examining the Ethical Considerations Surrounding the Development and Use of AI. American Journal of Technology, 2(1), 37 44. Retrieved from https://gprjournals.org/journals/index.php/AJT/article/view/142

Kamalov F, Santandreu Calonge D., &Gurrib I. (2023). New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability*. 15(16):12451. https://doi.org/10.3390/su151612451

Kharbach, M. (2024). A Timeline of The Evolution of ChatGPT. https://www.educatorstechnology.com/2024/06/the-evolution-of-chatgpt.html

Khurana, D., Koli, A., Khatter, K. *et al.* (2023). Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* **82**, 3713–3744 https://doi.org/10.1007/s11042-022-13428-4

Kortemeyer.,GNöhl.,J. &Onishchuk., D. (2024). Grading Assistance for a Handwritten Thermodynamics Exam usingArtificial Intelligence: An Exploratory Study. doi: 10.48550/arxiv.2406.17859

Kurzhals, H. D. (2022). Challenges and approaches related to AI-driven grading of open exam questions in higher education: Human in the loop, Computer Science, Education. Available online: https://essay.utwente.nl/90957/1/Kurzhals_BA_BMS.pdf

Minaee, S., Mikolov, T., Chenaghlu, N., Socher, M., Amatriain, R., X. & Gao., J. (2024). Large Language Models: A Survey, https://doi.org/10.48550/arXiv.2402.06196

Opesemowo, O., &Adekomaya, V. (2024) Harnessing Artificial Intelligence for Advancing Sustainable Development Goals in South Africa's higher education system: A Qualitative Study.

International Journal of Learning, Teaching and Educational Research, 23. 67-86. 10.26803/ijlter.23.3.4.

Patel, S., & Ragolane, M. (2024). Implementing Artificial Intelligence in Higher Education Institutions in South Africa: Opportunities and Challenges. Technium Education and Humanities, 9, 51–65. https://doi.org/10.47577/teh.v9i.11452

Ragolane, M., & Patel, S. (2024). Transforming Educ-AI-tion in South Africa: Can AI-Driven Grading Transform the Future of Higher Education?. *Journal of Education and Teaching Methods*, *3*(1), 26–51. https://doi.org/10.58425/jetm.v3i1.267

Schleicher, A. (2018). Educating Learners for Their Future, Not Our Past. ECNU Review of Education, 1(1), 58-75. https://doi.org/10.30926/ecnuroe2018010104

Stoica, E. (2022). A Student's Take on Challenges of AI-driven Grading in Higher Education.TScIT 37, July 8, 2022, Enschede, The Netherlands. https://essay.utwente.nl/91784/1/Stoica_BA_EEMCS.pdf

VSO. (2019). THE ACTION RESEARCH GUIDEBOOK, 'Progress is only possible by working together.' Available at: https://www.vsointernational.org/sites/default/files/2020-04/vso-cambodia-action-research-guidebook-english.pdf

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *Int J EducTechnol High Educ* **21**, 15. https://doi.org/10.1186/s41239-024-00448-3

Walvoord, B.E. & Johnson Anderson., V. (1998). Effective Grading: A Tool for Learning and Assessment. San Francisco: Jossey-Bass.

Zuber-Skerritt, O. (Ed.). (1991). Action Research for Change and Development (1st ed.). Routledge. https://doi.org/10.4324/9781003248491