American Journal of Educational Research, 2024, Vol. 12, No. 4, 147-158
Available online at http://pubs.sciepub.com/education/12/4/4
Published by Science and Education Publishing
DOI:10.12691/education-12-4-4

The Theoretical and Practical Implications of OpenAI System Rubric Assessment and Feedback on Higher Education Written Assignments

LaJuan Perronoski Fuller[1,*], Christa Bixby[2]

[1]College of Business and Information Technology, South University, Savanah, USA
[2]Dissertation Department, Westcliff University, Irvine, USA
*Corresponding author: lfuller@southuniversity.edu

Received March 05, 2024; Revised April 07, 2024; Accepted April 14, 2024

Abstract Integrating artificial intelligence (AI) in teaching and assessment is becoming increasingly common. However, concern exists regarding the reliability and consistency of generative AI grading and feedback in higher education. This study aims to investigate AI chatbots, such as ChatGPT and Claude, and their ability to apply consistent grading and feedback. The research revealed implications of applying these OpenAI systems outside their intended purpose as language models for generating human-like text. The data collected reveal significant discrepancies in grading patterns, feedback rationale, and formatting of responses. These inconsistencies challenge some traditional theories of learning and assessment. For example, ChatGPT applied a 24-point difference between the lowest and highest scores (74%-98%) on the same assignment. Claude's lowest and highest scores revealed a 33-point difference. Each OpenAI system provided feedback that was less likely to promote learning due to inconsistent rationale per rubric item. Educators are encouraged to exercise caution when utilizing OpenAI as a grading and feedback system. Educators should rely on the expertise of a subject matter expert to ensure accuracy and fairness in assessment practices. By understanding the limitations of OpenAI grading and feedback, educators can mitigate potential unfair and inconsistent assessments to optimize student success and learning outcomes.

Keywords: *OpenAI, rubric, assessments, feedback, grading*

Cite This Article: LaJuan Perronoski Fuller, and Christa Bixby, "The Theoretical and Practical Implications of OpenAI System Rubric Assessment and Feedback on Higher Education Written Assignments." *American Journal of Educational Research*, vol. 12, no. 4 (2024): 147-158. doi: 10.12691/education-12-4-4.

1. Introduction

Generative artificial intelligence (AI) relies on large language models to imitate human-like language using machine learning (ML) algorithms. One such ML text-generating model is OpenAI, which processes terabytes of data from textbooks, news articles, and websites [1,2]. The information from these ML data sources creates elegant poems, stories, and human-like responses. The spark of this new technology has created an urgency for educational institutions to maximize OpenAI in student learning. However, university-level educators report that OpenAI systems can introduce hallucinations and biases into student written assignments, even generating false citations [1,2,3]. Hallucinations and biased responses are less likely to impact creativity but more likely to produce consistent rubric grading and feedback on student assignments.

Educators may adopt OpenAI for grading and giving feedback based on the Technology Acceptance Model (TAM) [4]. TAM explains that technology deemed convenient and expedient has a significant positive relationship with user acceptance [5]. Because OpenAI can process vast amounts of information faster than humans and imitate human-like responses, educators may see it as convenient and faster. TAM may explain why educators will likely accept OpenAI as appropriate for administering grades or feedback on written assignments. However, TAM does not account for information accuracy or margin of error in technology outputs.

OpenAI technologies are creative but do not give accurate information reports or display a margin of error. The lack of reporting a margin of error can create levels of uncertainty in OpenAI generation. For example, users asked Bard by Google (known as Gemini) to "create" an image of a specific historical people group. However, Gemini's creativity did not always align with the user's actual or factual source for right or wrong. This level of uncertainty is troubling as educators may unknowingly deem OpenAI output as right rather than creatively a-moral. Therefore, educators who use OpenAI outside of its intended creative purposes may introduce margins of error into a student's assignment assessment. Hallucinations are appropriate for creativity but problematic for administering grades and quality feedback to account for student learning outcomes.

Student learning outcomes align with the behaviorist theory (BT). BT suggests that student learning occurs by reinforcing behaviors that promote a positive outcome, known as operant conditioning. Operant conditioning aims to shape student learning through consistent feedback based on what is right or wrong [6,7]. OpenAI's intended purpose is to be creative and imitate human-like responses. The OpenAI community widely accepts that hallucinations and biases exist within this technology. Hallucinations and biases are not presented as a margin of error and may camouflage what is right or wrong due to designing OpenAI not replicate itself [8,9]. Therefore, it is plausible that student grading and feedback will be inconsistent, which may degrade the metrics for student learning and successful outcomes.

Higher education institutions typically rely on course pass rates to measure student outcomes. Additionally, student pass rates indicate courses that may require revisions or updates. However, the lack of reporting an OpenAI system's margin of error may lead to irrelevant feedback, impact effective course redesign, and not accurately reflect student proficiency. The cognitive load theory (CLT) emphasizes adapting design instruction according to the student's expertise levels. However, OpenAI inconsistencies, hallucinations, and biases may not accurately assess student knowledge, resulting in inaccurate reporting of student success outcomes.

Additionally, most higher educational institutions apply Bloom's Taxonomy of Educational Objectives to deliver knowledge to students based on levels of learning. Consequently, OpenAI creativity, hallucinations, and biases may give inaccurate assessments of student assignments due to variations in grading and feedback each time [9]. Inconsistent OpenAI grading and feedback is problematic for institutions that use levels of learning similar to Bloom’s Taxonomy. These revelations present three potential issues relating to OpenAI grading and feedback in higher education institutions.

The first issue is that consistency of OpenAI rubric grading has yet to be established. The research posits there is a margin of error in accuracy, and lack of consistency can impact confidence in OPenAI's response output. Secondly, OpenAI hallucinations can produce irrelevant or less accurate rubric feedback. This second issue may hinder student learning outcomes, lead to unnecessary course redesign, and impact course rewrites using Bloom’s Taxonomy. Thirdly, OpenAI creativity may apply variations in rubric grading, which may alter student success metrics such as course pass rates. Therefore, further investigation is needed to determine to what extent OpenAI systems can apply consistent and accurate rubric grading and feedback on written assignments. The intent is to fill gaps in the literature on OpenAI hallucinations and false affirmatives in higher education [10] and to what extent these systems will give false affirmatives [1,2,3].

This research examines OpenAI's consistency in human-like responses (feedback) and grading on undergraduate and graduate-level written assignments. The OpenAI systems included in this research are ChatGPT, Claude, and Bard by Google. The goal is to observe these OpenAI systems' ability to apply consistent grades and feedback using a rubric. Therefore, if ChaptGPT, Claude AI, and Bard by Google are designed to be creative, hallucinate, and give false affirmatives, then grading and feedback are less likely to be consistent and relevant. The data collected is intended to answer the following research questions.

1.1. Research Questions

RQ1: How consistently can each OpenAI system, ChatGPT, Claude, and Bard by Google, apply a grade on a written assignment using a rubric?

RQ2: How consistently can each OpenAI system, ChatGPT, Claude, and Bard by Google, apply feedback for each rubric item based on the grade?

RQ3: How much grading and feedback variance exists between ChatGPT, Claude, and Bard by Google on the same written assignment?

2. Literature Review

ChatGPT can apply deep learning algorithms to generate and understand human-like text. It enhances language capabilities over time by continuously learning through user interactions. ChatGPT uses a transformer-based neural network of large data text to generate human-like and grammatically correct text [11,12]. As a result, within the first two months, ChatGPT gained over 100 million users worldwide [13]. The increase in users confirms this technology's ease of use and acceptance.

The TAM framework suggests that performance expectancy, effort expectancy, social influence, and facilitating conditions can improve technology acceptance. TAM found that educators are likely to use ChatGPT-type systems because they perform tasks more effectively and efficiently [14].

Effort expectancy significantly influences technology adoption and usage [15]. Therefore, as educators increasingly perceive ChatGPT-type systems as easy to use, they are more likely to integrate them into routine grading and feedback on assignments. However, researchers found that ChatGPT-3 gave affirmative responses even though the essays and reports were not real [16]. This revelation suggests that performance and effort influence usage but cannot account for factual accuracy and consistency.

The need to maximize ChatGPT-type systems in grading and feedback can result from social influence. Previous research revealed that colleagues can significantly influence the use of a specific technology [17]. These findings suggest that the more educators recommend ChatGPT-type systems, the more accepting educators are to deploy them for grading and feedback. Therefore, the more colleagues recommend AI, the more educators may view responses as correct or accurate. However, ChatGPT-type systems can spread misinformation, biased content, or harmful recommendations, which may negatively impact student learning and perpetuate stereotypes [14]. This is problematic as AI relative reasoning may misalign students from the program or course learning objectives based on grading and feedback.

2.1. Behaviorist Theory and AI Grading and Feedback

Students are human and, by default, have a reactive attitude. These attitudes are formulated based on interpersonal relationships. However, if interpersonal expectations are not met, it is natural to recognize reactive attitudes, such as resentment toward a person [18]. These attitudes suggest that AI hallucinations that present false or inaccurate feedback can degrade the student-educator relationship within the behaviorist learning paradigm.

Education is essential for social development and personal growth. Behaviorist Theory (BT) was recently applied to digital educational games. The research tested the impact of feedback and scores as a stimulus to measure student motivation. The goal was to determine an educator's ability to set the motivating factors for students. Digital game technology was able to give grades and feedback that motivated student learning using the BT framework [19]. However, the study did not apply ML or large language models. Additionally, digital learning does not appear to hallucinate and produce misinformation or false claims. Therefore, the ChatGPT-type AI system hallucinations produce varying results that may harm students' learning interests and motivation.

BT explains how behavioral outcomes and reward mechanisms influence student motivation [19]. BT suggests that external stimuli such as grading and feedback influence student achievement. Students can achieve course and lesson objectives through repetitive practices and consistent educator reinforcement. These results were validated and revealed that giving a grade is a form of external stimulus that acts as a form of feedback [20]. Grades can motivate students to achieve learning objectives, but it is still being determined to what extent ChatGPT-type systems can apply consistent grades and feedback. Therefore, inconsistent ChatGPT-type grading or feedback may hinder students from achieving designed learning outcomes.

2.2. Cognitive Load Theory and AI Grading and Feedback

Cognitive Load Theory (CLT) can provide insight into the presentation of information that encourages intellectual performance. First, CLT is the foundation of design education and competency-based instructional design. In this learning design, students acquire cognitive skills, measured by their ability to apply them in new situations [21]. This is similar to undergraduate and graduate courses requiring students to produce written assignments using a rubric. This is consistent with CLT, which connects working memory with long-term memory [22,23].

CLT consists of causal and assessment factors. Causal factors are cognitive abilities. Comparatively, assessment factors include mental load, effort, and performance. The focus will be on OpenAI consistency and fairness by focusing on CLT's assessment factors. The mental load is the cognitive load imposed by the assignment. Mental effort is the cognitive capacity students allocate toward completing the assignment. Finally, performance reflects mental load and effort represented by the rubric grade and feedback [21]. CLT can bridge the theoretical with practical activation of a student's working memory based on the grading and feedback activation process [24,25,26].

Educators may feel pressure to adhere to the institution's fast pace in grading and feedback requirements. These pressures may make educators rush through new and essential content faster [27]. This behavior can lead to students having issues retaining new content as educators may accept OpenAI for feedback and grading to keep up with a faster pace. Additionally, CLT research reveals that a faster pace may lead to students learning incomplete or inaccurate information [26], potentially creating another set of issues for educators and educational institutions. For example, an educator may rely on a ChatGPT-type system to grade and give feedback on 30 assignments to meet a short grading deadline. However, AI hallucinations may present misinformation, false information, and biases that produce incomplete or inaccurate information, similar to research findings, according to the above reference [26]. Therefore, it is essential to use a rubric as the evaluation method to determine the level of consistency and accuracy of ChatGPT-type systems on written assessments.

2.3. Rubric Grading and Feedback

Educators rely on the primary trait, holistic, and analytical rubrics for enhancing objectivity when grading written assignments. The primary trait rubric is analytical, allowing value judgments for specific writing tasks. It focuses on the essential elements of that writing task [28]. Educators assist students with cognitive overload associated with primary trait rubrics. The primary trait rubric is less likely applicable across genres and needs to focus on scientific argument-based explanations [29]. Therefore, observing AI value judgments on primary trait rubrics is inappropriate for this study.

Holistic rubrics measure multiple criteria representing specific levels of quality and are more reliable and valid for writing assignments [29]. Holistic rubric value judgments vary based on how well students met all, some, or none of the written assignment descriptors. Holistic rubrics require coding and are suitable for large-scale assessments and in evaluating open-ended higher-order skills [30,31]. However, these rubrics may need to be more diagnostic in pinpointing what is involved in student thinking [32,33]. Educational institutions integrate descriptive analytical rubrics with holistic rubrics to account for this limitation.

Descriptive-analytical rubrics provide a separate score for each specific criterion. This type of rubric outlines student performance by assigning a point for each task criterion. Analytic rubrics ensure clear articulation of standards and measures to what degree students meet a desirable outcome [29]. Additionally, these rubrics can recognize students' strengths and weaknesses, which is essential in providing feedback and improving student outcomes [34]. This study will observe holistic descriptive analytical rubrics to better understand OpenAI value judgments and feedback. There are three types of rubrics [35].

Analytic rubrics apply shared score sets to analyze the sub-skills in samples of each student's writing [36]. Analytic scoring is based on properties/components (handwriting, sentences, objectives, etc.). Primary trait

rubrics assess basic writing skills relating to particular writing tasks. Holistic rubrics assess student work by rating the categories with a score in line with the determined properties and define different performance levels superficially [35,36,37].

Holistic descriptive analytical rubrics are two-dimensional. The first dimension depicts the level of achievement as columns. The second dimension accounts for assessment criteria in rows. Educators assess assignments using weights (values) to each criterion, presented in table form. This rubric is best for assessing assignment objectives. These descriptive analytical rubrics serve as powerful tools in academic assessments, requiring an investment of time to properly assess student achievements of several criteria on a single rubric. However, ChatGPT, Claude AI, and Bard by Google may apply different weights (grades) to the same criteria in the assessment stage of the analysis.

Scientists agree that analytic scoring surpasses other methods in its wide-ranging, comprehensive, and practical evaluation [36]. This comprehensive assessment tool proves advantageous for both educators and learners. Faculty members can better understand assigning scores based on specific criteria, enabling more precise and targeted evaluation. Simultaneously, students benefit from this clarity, empowering them to craft high-quality work aligned with the specified criteria and objectives. Nonetheless, it is still being determined how Open AI systems such as ChatGPT, Claude AI, and Bard by Google will produce consistency in the rationale and the clarity of feedback assigned to each rubric item to justify that value judgment.

Analytic scoring extends its influence beyond the grading process [38]. It serves as a guiding force for educators, offering valuable insights into effective teaching strategies. For students, analytic rubrics illuminate the writing process, provide structured and comprehensive feedback derived from analytic scoring, and provide a discerning analysis of strengths and weaknesses in students' assignments. Nonetheless, there is a need to study assessment practices of OpenAI as academia champions cutting-edge advancement in rubric grading. The goal is to ensure that the integration of OpenAI does not introduce inaccuracies, stereotypes, or biases in value judgments before its use or integration into educational institutions [10-39], ensuring that it does not lead to educational inequities.

Educational institutions typically integrate a student-centric learning approach. Further, a standard practice to advance inter-rater reliability amongst scorers when grading based on an analytical rubric is calibration. The increase in reliability reduces error margins between student grades and enhances the objectivity and fairness of grading [40]. OpenAI systems do not inherently calibrate for reliability and instead rely on natural language processing (NLP) to predict text similar to human-like speech.

2.4. Artificial Intelligence and Holistic Analytical Rubrics

AI, or a machine's capacity to simulate a human's intelligence, has advanced into the paradigm of generative AI, in which new content is created using systems that have been trained [41]. Under the umbrella of AI is ML and NLP. ML, whether supervised, semi-supervised, or unsupervised, refers to software and the algorithms that inform it that can learn from experience instead of following a series of programmed commands [42]. NLP refers to computational techniques that automatically analyze and represent human language, enabling the functionality of sensemaking of the written word or oral speech. For OpenAI large-scale model-based chatbots to work, NLP is required to make sense of the users' requests, and ML is required to produce responses that become more accurate over time with increased frequency of interactions. However, how the same OpenAI systems will apply weight to identical analytical rubrics still needs to be discovered.

As institutions of higher learning begin to incorporate AI into the field of education, where machines have begun taking on roles previously only held by humans, its incorporation as a nonhuman agent has been referred to as the inclusion of a machine teacher [43,44]. One such area in which AI, as a machine teacher, is beginning to be used is with assessments, both in developing and evaluating formative and summative assessments [45]. However, as the NLP algorithm is the basis for AI chatbots, a student could write, for example, Organization B is an exceptional workplace. The NLP algorithm may contain only some essential contextual information about the organization or be fully presented to the student in class to determine the accuracy of this statement [46]. Limitations such as these can be problematic should educators use AI for student analytical rubric assessments without the ability to correctly identify or train the system similar to faculty calibration [47].

AI scoring systems have seen enhancements with the addition of NLP. While benefits have been noted regarding efficiency with a reduction in both time and effort for scorers/educators [48], discrepancies in machine-human score agreements (MHA) have been noted as an area of concern in tandem with offering an opportunity to explore why such discrepancies exist [42]. Limited research has been done on the accuracy of training AI to achieve MHA. Research studies focused on using AI with assessments in the scientific fields, where the need for three-dimensional thinking is a primary focus [42]. When using Cohen's kappa [k] in science assessments to determine MHAs, a 0.09 to 1.0 range was present, indicating a statistically significant gap. This research intends to fill this gap by focusing on analytical rubrics and evaluating scores using Cohen's kappa based on assignment value.

Before using AI for assessment purposes, the Smarter Balanced Assessment Consortium suggests a minimum of 1,500 on-grade responses for training purposes before evaluating reliability [49]. Thus, studies that have tested the reliability and accuracy of scoring have done so after training on a specific prompt to train the system of what is considered accurate or correct when evaluating essays. However, it is plausible that academic leaders have yet to consider that OpenAI systems cannot be trained for their specific assignments without additional layers of models and specific language models and datasets explicitly curated for their institution.

AI can calculate large data sets, recall information, and create complicated texts faster than human intelligence.

The researcher inserted a prompt asking ChatGPT to describe itself and what it is designed to do, which generated the following text:

I am ChatGPT, a large language model created by OpenAI. My purpose is to assist users like you by generating human-like text responses to a wide range of prompts and questions. I have been trained on vast amounts of text data from the internet, books, articles, and other sources up until January 2022. My design allows me to understand and generate text in natural language, enabling me to engage in conversations, provide information, answer questions, offer suggestions, and even generate creative content like stories or poems.

While I don't possess consciousness, emotions, or personal experiences like a human, I can process and analyze text input to generate relevant and coherent responses. My capabilities include understanding context, recognizing patterns, and providing helpful information based on the input I receive. I'm here to assist you with whatever questions or tasks you have to the best of my abilities within the scope of my training data and algorithms (OpenAI, 2024).

The text generated by ChatGPT suggests that ML models can allow for human-like engagement. However, it did not indicate that it is meant to be used for evaluation purposes or for analyzing information to provide a consistent or fair score. Nonetheless, ChatGPT's ability to apply consistent grading on student knowledge, ideas, observations, and imagination using a rubric needs to be clarified. Therefore, the following hypotheses are proposed.

H1: There is no difference between rubric grading on one undergraduate level and one graduate-level written assignment uploaded multiple times in ChatGPT.

H1a: There is a difference between rubric grading on one undergraduate-level and one graduate-level written assignment uploaded multiple times in ChatGPT.

The second OpenAI system was Claude. Claude responded to a prompt asking what it is designed to do and what its purpose is, which generated the following text:

I'm Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest. I don't have a singular purpose. I am an AI assistant created by Anthropic to be generally helpful, harmless, and honest through natural language conversations.

Claude is designed to be useful, harmless, and honest. However, rubric scores and feedback identify student strengths and weaknesses, which will impact student perception and progress and vary in interpretation of usefulness or honesty [50]. Claude is harmless, but the definition of harm is unclear as it relates to applying grades to student learning. If Claude is generally harmless, observing its ability to apply a fair and consistent grading to the same assignment uploaded multiple times is beneficial. Therefore, the following hypotheses are presented in this study.

H2: There is no difference between rubric grading on one undergraduate level and one graduate-level written assignment uploaded multiple times in Claude.

H2a: There is a difference between rubric grading on one undergraduate level and one graduate-level written assignment uploaded multiple times in Claude.

Bard was the third AI system. Bard is a conversational AI tool by Google. Bard responded to a prompt asking it to explain what it is designed to do and its purpose, which generated the following text:

I am designed to be a large language model capable of communicating and generating human-like text responding to various prompts and questions. Here are some of my key functionalities: Answer your questions in an informative way: I have access to a vast amount of information and can use it to answer your questions in a comprehensive and informative way, even if they are open-ended, challenging, or strange. Generate different creative text formats: I can generate different creative text formats of text content, like poems, code, scripts, musical pieces, emails, letters, etc. I will try my best to fulfill all your requirements. Follow your instructions and complete your requests thoughtfully: I will try my best to follow your instructions and complete your requests thoughtfully, even if they are complex or require multiple steps. My purpose is multifaceted, evolving, and influenced by the needs and expectations of those interacting with me. Here are some key aspects of my purpose: Advancing research and development: I am used by researchers and developers to explore the possibilities of large language models, improve their capabilities, and discover new applications. This contributes to developing more advanced AI technologies and their potential benefits for society. Fostering learning and understanding: By answering your questions and providing information, I can help you learn and understand new things. I can also be used to create educational materials and tools.

Bard suggests assisting with research and development and providing informative discussions. Nonetheless, the decision-making algorithm needs more transparency and may contribute to inaccuracies or inconsistencies in value judgments and feedback [10-39]. Bard by Goole gave an affirmative that it could grade a written assignment using a rubric. Upon submitting the assignment, Bard stated that it could not grade the assignment using a rubric and to consult with an academic professional based on three reasons: a) not being able to understand context, b) not being able to apply subjective judgment, and c) not being able to ensure fairness and accuracy. Therefore, Bard was not included in the data collection.

ChatGPT and Claude can assign a grade to an assignment put into the system, provide feedback, and administer instructions [10]. It is plausible that AI can streamline the evaluation process and ensure consistent assessments for all but may omit inherent disparities often associated with inclusivity in student learning. Educational advancements include intelligent and adaptive web-based systems that learn instructor and learner behavior to adjust accordingly to enrich the educational experience [51]. While Reinforcement Learning from Human Feedback (RLHF) trains big language models, ChatGPT and Claude can modify altered responses so the student perceives the feedback as from a human. However, the consistency in value judgments on text assignments

and relevant feedback across various rubric domains has yet to be discovered, and thus, the following hypotheses are proposed.

H3: No differences exist between rubric grading on the same written assignment uploaded on ChatGPT and Claude.

H3a: Differences exist between rubric grading on the same written assignment uploaded on ChatGPT and Claude.

AI chatbots apply ML but have been questioned for potential information bias [52]. Additionally, AI systems that use ML still require calibration and create urgency in determining the consistency and reliability of the overall weight (value) applied to written assignments. This area requires exploration as generative AI can rewrite responses to ensure originality, exposing the possibility of differing scores and feedback dependent upon the repetition of submission. Therefore, there needs to be more knowledge on to what extent AI systems that apply ML, specifically since it is unknown assessment training, can apply consistent value judgments to analytic rubric-assessed assignments.

Due to the need for more transparency in decision-making algorithms [39], these differences in ML types may create inconsistencies. Additionally, a lack of training to accommodate ML may explain the potential evaluation bias that can influence students' negative feelings toward grading [55]. OpenAI systems may have consistent understanding in responding to queries; however, the uniqueness of each query suggests variations may exist in applying value judgments on analytical rubrics on written assignments.

Previous research revealed limitations with OpenAI systems and the ability to provide value judgments and feedback. ChatGPT sometimes suggests inaccurate information [10]. Additionally, some instructions may be derived from biased material. As a result, it is less likely to verify the truthfulness of the information. These limitations can lead to ethical considerations as faculty rely on ChatGPT for student feedback. Consequently, biases in the chatbot training dataset model's output can result in incorrect feedback [10]. Therefore, to advance institutional perspectives on AI grading and feedback accuracy [53]. This research study will apply a holistic descriptive analytical rubric and conduct an observation theme analysis of feedback to answer the following research question.

H4: There is no difference in the rubric feedback given by ChatGPT and Claude AI.

H4a: There is a difference in the rubric feedback given by ChatGPT and Claude AI.

The goal is to determine if AI value judgments can support educators in student success [54]. The following section provides the methodology to test these hypotheses and answer research questions.

3. Methodology

The research aimed to test whether AI systems can use a rubric to grade university-level written assignments consistently and accurately. The first step was to select the appropriate rubric and a written assignment that aligned with that rubric. The research included two undergraduate and two graduate-level written assignments anonymized from different universities. Second, the researchers inserted the rubric and each assignment into ChatGPT, and Claude but Bard, as aforementioned, was not included) was added five times sequentially per chatbot. Third, an independent evaluation of the grade and feedback was conducted to determine consistency, fairness, and accuracy. The researchers independently assessed the AI grading and compared the overall scores and feedback. Fourth, the researchers identified and discussed areas where assessment discrepancies existed.

Two rubrics were used for this study. Rubric 1 consists of 0 = Unsatisfactory and 3 = Exemplary. Assignments 1 and 2 were anonymized and measured using this rubric. The rubric criteria consist of eight categories. The categories include orienting paragraphs, coverage and relevance, synthesis and integration, critical analysis, assignment organization, literature relevance, writing clarity and style, structure, and citations.

Rubric 2 consists of 0 = No Submission to 5 = Exemplary. Assignments 3 and 4 were anonymized and measured using this rubric. These rubric criteria consist of five categories. The categories include the company's reasoning for market entry, analysis, impact analysis, product quality, and APA writing. The four assignments were written according to the appropriate rubric items and included APA formatting instructions.

While the research was intended to be completed with three AI systems, Bard only conducted the value judgments once per the established research design. Interestingly, when Bard was initially asked, "Can you grade this student's paper with this rubric and provide feedback on the rationale for this score?" The system responded that it was able and indicated, "I am confident that my ability to analyze text and apply rubrics will provide you with valuable insights into the paper's quality and help you make informed decisions about grading or providing feedback to the author." It provided feedback and, when further prompted, scores. The second time the paper was submitted following an identical process, the system stated, "I'm unable to assign a specific numerical score to your paper as it goes against my principles of not being used for assessment purposes. Scores can be subjective and vary depending on the rubric and instructor's criteria." The third time an attempt was made, the system stated, "I'm not able to help with that, as I'm only a language model." When asked a week later if it could score a paper, the system stated that it could analyze text. Still, it cannot provide a grade due to the lack of context, subjectivity in providing judgments, and the issue of fairness and accuracy. Therefore, the Bard system was removed from the collection phase.

Rubric 1 was used in ChatGPT to grade Assignment 1. Assignment 1 was reintroduced to ChatGPT four additional times to evaluate consistency. The five scores and feedback were recorded to determine consistency across value judgments within the same system using the same assignment. The value judgment ranges were noted, and a comparative analysis was done on feedback consistencies between the highest and lowest scores. This process was replicated for Assignment 2. Additionally, Rubric 2 was used in ChatGPT to grade Assignment 3 on

five separate occasions. Each rubric score and feedback were recorded to determine the value judgments' consistency using the same assignment. A comparative analysis was done on value judgment scores and feedback. A comparative analysis between the highest and lowest scores was essential to evaluate consistency in these judgments. This process was replicated for Assignment 4.

The research applied this methodology with Claude's AI system to compare value judgment consistency and identity potential similarities with ChatGPT assessments of the four assignments. The researcher attempted to replicate the steps for ChatGPT; however, the steps were adjusted to account for errors in the overall value judgment process. Steps 1 through 6 were presented to Claude at one time. The Claude AI system then conducted value judgments on Assignments 1 and 2 using Rubric 1 and Assignments 3 and 4 using Rubric 2.

An assessment of inter-rater reliability is required to promote validity to the results. The second researcher compared scores assigned by the same AI system to understand the rubric scores. Cohen's kappa measurement was used to quantify inter-rater agreement. Then, a cross-system comparison was conducted using the two rubrics. The researchers replicated the value judgment criteria across ChatGPT 3.5 and Claude AI systems.

The research calculated means, standard deviations, and other relevant statistical measures for assessments conducted using each software system. Additionally, the results were compared, and any significant differences in mean scores or variability were noted. Then, the researchers performed statistical tests (e.g., ANOVA) to determine if statistically significant differences exist between the mean scores between AI systems.

ANOVA's first assumption is to check for normality. This check will ensure scores within each group are approximately normally distributed. Secondly, the researchers analyzed the Homogeneity of Variance (HOV). HOV determines if the variances in the different groups are roughly equal. Third, the researchers ran a one-way ANOVA to calculate the F-statistic and corresponding p-value. Fourth, the researchers would reject the following null hypotheses if the p-value exceeds the significance level (α). Therefore, value judgments were the overall rubric score for each assignment.

A repeated analysis of variance (ANOVA) is the most suitable statistical method used to determine if there are any statistically significant differences between the means of unrelated groups. In grading written assignments, the mean value across three AI system categories will be assessed using a repeated one-way ANOVA using IBM Statistical Package for Social Sciences (IBM SPSS) version 27. The results of the one-way ANOVA will be used to test the null hypotheses. The significance level (α) of < 0.05 is the set threshold for determining statistical significance. The p-value of >0.05 will suggest that the researcher will test and consider rejecting the null hypotheses.

If the null hypothesis is rejected, further analyses (post hoc tests) may be needed to identify which specific groups differ. These Post Hoc Tests will consist of Tukey's HSD or Bonferroni to identify specific group differences. Additionally, Post Hoc Tests will include an effect size calculation to determine the practical significance of the observed differences.

The researcher will collect rationale and feedback from written assignments uploaded into the AI-based systems. The research will conduct a calibration process based on the lowest and highest score of each OpenAI group. For example, the researcher will collect the lowest and highest scores from ChatGPT on Assignment 1, Rubric 1. Then, the researcher will replicate this action four additional times to determine consistency in value judgments. Next, the researcher will apply this logic to Claude AI. Finally, the research will categorize each system's lowest and highest scores and compare the feedback according to the rubric.

4. Findings

Summary statistics were calculated for each interval and ratio variable. A frequency of written assignment letter grades was necessary to examine grade consistency based on the letter grade issued on the same assignment coupled with the same rubric. The frequency distribution of letter grades was based on value judgments 100-90 = A; 89-80 = B; 79-70 = C; 69-60 = D; < 59 = F.

The most frequently observed category of ChatGPT Assignment 1 was F (n = 5, 100.00%). Assignment 2 was C (n = 3, 60.00%). Assignments 3 were B and C, each with an observed frequency of 2 (40.00%). Assignments 4 were C and A, each with an observed frequency of 2 (40.00%).

The most frequently observed category of Claude Assignment 1 was F (n = 4, 80.00%). Assignment 2 was C (n = 2, 40.00%). Assignments 3 were A and B, each with an observed frequency of 2 (40.00%). Assignment 4 was B (n = 4, 80.00%). Each testing group's lowest and highest scores were observed to evaluate the grading range between scores within the same assignment and identical rubric.

ChatGPT grading consisted of variations in scoring on the same assignment. Assignment 1 scoring was between 29% to 37%. Assignment 2 grading range was between 66% to 77%. Assignment 3 lowest to highest score by ChatGPT was between 58% to 82%. Assignment 4 was given a low score of 74% and the highest score of 98%. ChatGPT applied different grades when asked to apply the appropriate rubric to the same assignment. Claude revealed results similar to ChatGPT's in that it relates to variations in assignment grades. Claude scored Assignment 1 between 62% and 70%. Assignment 2 grading ranged between 58% and 91%. Assignment 3 scores were between 78% and 96%. Assignment 4 received 84% to 90%. These variations in grading suggest that these OpenAI systems are less likely to present consistent grading.

Each AI system assignment grade was annotated and recorded via Microsoft Excel. The purpose of calculating score ranges is to determine the variance between scores per assignment in ChatGPT. ChatGPT value judgments for Assignment 1 were at .34, with a 4-point difference between the lowest and highest score (29%-37%). Assignment 2 value judgments average score was at .70, with a 4-point difference between the lowest and highest score (66%-70%). Assignment 3 value judgment was at .74, with a 32-point difference between the lowest and highest score (58%-82%). Assignment 4 value judgment

was at .85 with 24 24-point difference between the lowest and highest score (74%-98%).

The differences in several assignment categories suggest that ChatGPT may apply inconsistent value judgments on replicated assignments. ChatGPT for grading needs to be more consistent and likely to create disparities when assessing criteria due to the variance between an A or C letter grade. Based on these findings, the research team rejects the null hypothesis H1: There is no difference between rubric scores on one undergraduate-level and one graduate-level written assignment uploaded multiple times in the OpenAI system ChatGPT.

Claude's systems grading for Assignment 1 was .64, with an 8-point difference between the lowest and highest score (62%-70%). Assignment 2 average score was at .75, with 33 point difference between the lowest and highest score (58%-91%). Assignment 3 was at an 86% average with a 12-point difference between the lowest and highest scores (78%-90%). Assignment 4 was a .84 with a 6-point difference between the lowest and highest score (84%-90%). The Claude system revealed results similar to those of ChatGPT. The results revealed that some assignments in five replications had a point difference of approximately 33 points. This low-to-high point difference suggests that an assignment can receive either an A or an F value judgment, and the research team rejects the null hypothesis H2: There is no difference between value judgments (rubric score) on one undergraduate-level and one graduate-level written assignment uploaded multiple times in the OpenAI system Claude.

It was necessary to compare each assignment's value judgment across AI systems. The logic is to determine the impact of using more than one AI system to apply value judgments. For example, educator A may use ChatGPT to assign value judgments to Assignment 1. Comparatively, educator B may use a different AI system like Claude to apply value judgments.

For Assignment 1, ChatGPT's average score was 34%, and Claude's average score was 64%. For Assignment 2, ChatGPT's average score was 70%, and Claude's average score was 75%. For Assignment 3, ChatGPT's average score was 74%, and Claude's average score was 86%. For Assignment 4, ChatGPT's average score was 85%, and Claude's average score was 84%. The averages suggest that differences exist across AI systems, and students may receive inconsistent feedback from AI systems used for this purpose. Therefore, the research team rejects the null hypothesis H3: No differences exist between value judgments (rubric score) on written assignments uploaded across ChatGPT and Claude.

A two-tailed paired samples t-test was conducted to examine whether the mean difference between ChatGPT and Claude grades differed significantly from zero. First, a normality test was conducted using a Shapiro-Test to determine if a normal distribution could have produced any differences.

The results of the Shapiro-Wilk test for ChatGPT and Claude Assignment 1 were not significant based on an alpha value of .05, $W = 0.96$, $p = .814$. ChatGPT and Claude Assignment 2 tests were not significant based on an alpha value of .05, $W = 0.95$, $p = .736$. ChatGPT and Claude Assignment 3 were not significant based on an alpha value of .05, $W= 0.93$, $p = .627$. ChatGPT and Claude Assignment 4 tests were not significant based on an alpha value of .05, $W = 0.91$, $p = .490$. These results suggest the possibility that the differences in assignment grading between ChatGPT and Claude produced by a normal distribution cannot be ruled out, indicating the normality assumption is met.

Levene's test assessed whether the variances of ChatGPT and Claude's graded assignments were significantly different. ChatGPT and Claude Assignment 1 result was not significant based on an alpha value of .05, $F(1, 8) = 0.00$, $p = 1.000$. ChatGPT and Claude Assignment 2 were not significant based on an alpha value of .05, $F(1, 8) = 3.12$, $p = .116$. ChatGPT and Claude Assignment 3 were not significant based on an alpha value of .05, $F(1, 8) = 0.13$, $p = .728$. ChatGPT and Claude Assignment 4 were not significant based on an alpha value of .05, $F(1, 8) = 5.29$, $p = .050$. These results suggest that it is possible that ChatGPT and Claude's grades were produced by distributions with equal variances, indicating the assumption of homogeneity of variance was met.

A repeated measures analysis of variance (ANOVA) with one within-subjects factor was conducted to determine whether significant differences exist among ChatGPT and Claude assignments. The usual sphericity assumption does not apply when only two repeated measurements exist. For example, ChaptGPT Assignment 1 and Claude Assignment 1. The results were examined based on an alpha of .05. The main effect for the within-subjects factor was significant, $F(1, 4) = 225.09$, $p < .001$, indicating significant differences between the ChatGPT and Claude Assignment 1. The mean contrasts utilized Tukey comparisons based on an alpha of .05. Tukey comparisons were used to test the differences in the estimated marginal means for each combination of within-subject effects.ChatGPT Assignment 1 was significantly less than Claude Assignment 1, $t(4) = -15.00$, $p < .001$.

The following results relate to ChatGPT and Claude Assignment 2. The results were examined based on an alpha of .05. The main effect for the within-subjects factor was not significant, $F(1, 4) = 0.69$, $p = .453$, indicating the values of both assignments were similar.

The results were examined for ChatGPT and Claude Assignment 3 based on an alpha of .05. The main effect for the within-subjects factor was significant, $F(1, 4) = 11.61$, $p = .027$, indicating there were significant differences between these values.

The mean contrasts utilized Tukey comparisons based on an alpha of .05. Tukey comparisons were used to test the differences in the estimated marginal means for each combination of within-subject effects. ChatGPT Assignment 3 was significantly less than Claude Assignment 3, $t(4) = -3.41$, $p = .027$.

The results were examined for ChatGPT and Claude Assignment 4 based on an alpha of .05. The main effect for the within-subjects factor was not significant, $F(1, 4) = 0.03$, $p = .865$, indicating the values of Chat_GPT_Assign_4 and Claude_Assign_4 were all similar. As a result, the findings concluded that the research team rejects each null hypothesis and accepts the alternative hypotheses below:

H1a: There is a difference between rubric grading on one undergraduate-level and one graduate-level written assignment uploaded multiple times in ChatGPT.

H2a: There is a difference between rubric grading on

one undergraduate level and one graduate-level written assignment uploaded multiple times in Claude AI.

H3a: Differences exist between rubric grading on the same written assignment uploaded on ChatGPT and Claude AI.

To better understand the differences in value judgments that account for a full letter grade difference between low and high scores, the rubric items' feedback was evaluated to investigate themes. This allowed the researchers to better understand how an item could receive a different grade for the same assignment but at different upload times. Each assignment that received an inconsistent score was recorded, and feedback on the highest and lowest scores was noted. Inconsistencies in the responses were categorized into two predominant patterns: (1) rationale for the same criteria of the same paper was similar to near identical, but the score differed between a minimum of 1 point to a maximum of 3 points; (2) rationale for the same criteria of the same paper was different, and the score differed between a minimum of 1 point to a maximum of 4 points.

ChatGPT scored the writing criterion on a work at opposite ends of the rubric spectrum based on the rationale differences. The work was rated as exemplary for one submission, then as no submission during a later resubmission of the same paper. The rationale for the exemplary paper was "writing is clear and organized with coherent arguments," whereas the rationale for the no submission was" submission contains a random presentation of ideas, which prevents understanding." Claude AI system Assignment 1 includes a value judgment range between 62% and 70%. This score range is significantly different from ChatGPT.

Additionally, these differences reveal that specific rubric items may be valued vastly differently, as shown in Table 1.

Table 1. Feedback Variance Assignment 1

Highest Score Feedback	Lowest Score Feedback
Citations are present and **mostly properly formatted** according to APA 7th edition. (proficient)	Citations are present, but APA **formatting is inconsistent or missing**. (satisfactory)
Substantial critical analysis is provided. (exemplary)	**More depth** of analysis on the advantages/disadvantages of AI **could be provided**. (satisfactory)
Paper **effectively** conducts a SWOT analysis. (exemplary)	Analysis **lacks** depth and credible research support. (emerging)
Follows APA formatting guidelines. (exemplary)	**Lacks** proper APA formatting. (proficient)

Table 2 contains another observation of grading feedback variations by ChatGPT's highest and lowest scores.

H4a: There is a difference in the rubric feedback given by ChatGPT and Claude.

5. Discussion

The data from this study confirms that using OpenAI systems outside of their intended purpose can result in unreliable and inconsistent grading and feedback. The system's default to produce authentic, original text resulted in variations of grades, rationale, and formatting of the responses to fulfill its designed purpose. This study proposed that creativity within OpenAI systems would lead to inconsistencies in grading and feedback on similar assignments and across OpenAI systems.

Table 2. ChatGPT Assignment 1 Feedback Variance Assignment 1

Rubric Item	Highest Score Feedback	Lowest Score Feedback
Orienting Paragraph	The orienting paragraph provides some clarity regarding the purpose and focus of the assignment by introducing the concept of neuromarketing and its applications. However, it could be more concise and precise in communicating the overall purpose. (proficient)	The student's paper lacks a clear, orienting paragraph that effectively communicates the purpose and scope of the assignment. It briefly mentions neuro-marketing but fails to provide a clear overview of what the assignment will cover. (satisfactory)
Organizational Structure	N/A	The document lacks a clear structure and logical flow. It fails to organize the content effectively, leading to confusion for the reader. (proficient)
Overall Comment	Overall, the paper shows potential but requires significant improvements in coverage, synthesis, critical analysis, organization, writing clarity, and citation formatting to achieve a higher score.	The paper demonstrates some understanding of the topic but falls short in terms of coverage, critical analysis, organization, and adherence to academic writing standards. Additionally, there are several grammar and style errors that need to be addressed for clarity and coherence.

5.1. Theoretical Implications

The first theoretical implication is based on BT, which suggests that students require consistent behavior as an external stimulus to promote learning. ChatGPT and Claude's grading and feedback inconsistencies are ineffective as external stimuli. The findings presented a frequency that revealed a divergence in grading patterns on the same assignment within and across the two systems. ChatGPT not only applied variations in grading the same assignment but assigned lower grades than Claude. For instance, while all students received an F (37%-29%) in ChatGPT Assignment 1, Claude AI grades were more varied, with a mix of C (70%) and F (62%) grades. This suggests a potential discrepancy in the assessment criteria or algorithms the two systems deploy when grading the same assignment multiple times and across systems.

Additionally, the research findings revealed that ChatGPT and Claude provided variations in feedback on the same rubric item. These inconsistencies in feedback go beyond a stimulus-response association based on a student's prior knowledge, learning strategies, and personal experiences when writing assignments. This outcome suggests that students would be less likely to learn through consistent feedback regardless of how an educator perceives the usefulness, ease, or convenience of

the OpenAI systems [7].

Overall, the inconsistencies in grading and feedback challenge the assumptions of behaviorism. For example, Claude's grading had a 33-point difference between the lowest and highest scores (58%-91%) on Assignment 2, which may negatively impact the stimulus-response relationship between the student and educator. Therefore, educators using OpenAI-type systems like ChatGPT and Claude to grade assignments are less likely to improve student learning outcomes and may negatively impact the instructor-student relationship built on trust and fairness.

The second theoretical implication applies to CLT. CLT posits that the cognitive load imposed on a student's working memory positively influences learning. Intrinsic cognitive load relates to the complexity inherent in the material being learned, while extraneous cognitive load refers to the impact of the instructional design through grading. The intrinsic load should remain consistent, which is not seen with ChatGPT, which applies a 24-point difference between the lowest and highest scores (74% - 98 %) on the same assignment.

The summary statistics unveil not just differences in grades but also variations in mean scores and standard deviations. Additionally, statistical tests confirm significant differences in the grading outcomes when the same assignment is graded multiple times within the same OpenAI system. This suggests that ChatGPT and Claude apply unknown decision-making methodologies during each assignment evaluation, which may have led to inconsistent grading of the same assignment. CLT emphasizes the importance of managing extraneous cognitive load to prevent cognitive overload and promote effective learning. However, ChatGPT and Claude's hallucinations, misinformation, and false affirmatives may be responsible for inconsistencies in grading and feedback. For example, ChatGPT applied exemplary for a reference page in APA format. However, the assignment did not contain a reference page.

OpenAI technologies offer scalability in assessment processes, which is significant to education systems accepting the grading based on the TAM. Nonetheless, these inconsistencies prompt questions about the reliability and trustworthiness of grading and feedback within higher education. The study concluded that the rationale for the rubric judgment needed to be more consistent. These outcomes suggest that students are less likely to be able to apply feedback to improve learning outcomes. By examining and understanding these disparities, educators and developers can work towards improving the transparency, reliability, and alignment of OpenAI grading and feedback algorithms with human standards.

The data sheds light on the intricate interplay between AI and human judgment in educational assessment. While AI systems seemingly offer promise in streamlining grading processes, their effectiveness hinges on their ability to accurately reflect human grading standards. By testing the consistency in grading and applying feedback on the same assignment using the same rubric multiple times, the findings suggest that educators are less likely to ensure fairness, promote consistency, and align feedback to improve student learning. Additionally, OpenAI hallucinations and false affirmatives may lead to irrelevant course redesigns and degrade Bloom’s Taxonomy. This exploration catalyzes ongoing dialogue and collaboration between educators, curriculum developers, and institutional stakeholders to advance the field of AI-driven grading and feedback responsibly and ethically.

5.1.2. Practical Implications

The results revealed significant discrepancies in responses, such as inconsistencies in grading, that can negatively impact student learning outcomes for five reasons. First, inconsistent grading within the same OpenAI system can create confusion about expectations. Inconsistent grading standards may require students to request additional clarification about educator expectations. For example, if ChatGPT gives a high grade for Assignment 1 on the first attempt but on the second attempt gives a low grade for Assignment 1, a student may not understand what constitutes success, nor would an educator be able to defend the rationale. These grading inconsistencies and feedback variations confirm that OpenAI hallucinations can generate inaccurate or misinformation [1,2,3].

Secondly, inconsistent grading can lead to a misalignment with course or lesson learning objectives. The results revealed that grades can be inflated or deflated on the same assignment and when compared across OpenAI systems. These findings suggest that universities may find themselves relying on inaccurate course pass rates to measure student success. Additionally, students are less likely to receive the most accurate feedback on their progress toward meeting the course learning objectives, which is presented in the variations in the rationale and feedback amongst high and lower scores. For example, ChaptGPT gave affirmative feedback on rubric items that were not present in the written assignment. These inaccuracies are consistent with previous literature that ChatGPT-3 can give affirmative responses even though the information may not be present or real [16].

Third, inconsistent grading and feedback can impact student motivation. Suppose student A receives a 90% for work they believe to be of high quality and applies the feedback on the next assignment. In that case, the inconsistency in grading may result in a lower grade and serve to demotivate the student. Fourth, variations in grading and feedback can lead to unfair comparisons. For example, some written assignments require students to collaborate to meet diversity and inclusion initiatives. However, if students upload the assignment separately in the learning management system, professors who rely on OpenAI-type systems for grading may unknowingly apply different grades on that group assignment, creating a sense of unfairness. Consequently, these inconsistencies or perceptions of unfairness may explain why students experience negative feelings associated with AI feedback [55].

Fifth, inconsistent grading and feedback can hinder student development of skills and knowledge. Students who do not receive accurate feedback on assignment strengths and weaknesses are less likely to know where to improve on future higher-level courses. This can prevent progress in their learning and educational journey. To address these five issues and promote proper student learning, educators need to strive for consistency in grading. Ultimately, OpenAI systems such as ChatGPT

and Claude cannot produce a consistent and fair grading system, which is crucial for fostering an environment where students can thrive academically. Therefore, educators should be mindful of relying solely on OpenAI grading and feedback, as hallucinations may lead to inconsistency, variations, and unfair grading as education institutions attempt to maximize AI in education [9].

There are some limitations to this research study. First, the sample size included only two OpenAI systems, ChatGPT and ClaudeAI. Additional studies should investigate the ability of other OpenAI and AI-assisted grading systems to produce consistent grading. For example, if a researcher uploads the same assignment and rubric into one OpenAI or AI-assisted grading system, does it apply a similar or consistent grade during multiple iterations? Additionally, the results from this research are not generalized to all OpenAI systems. Therefore, future research is required using control variables to account for potential confounding factors that may impact inclusive learning environments. For example, demographics, cultural background, and socio-economic factors on the OpenAI language model and the ability to read context language in written assignments.

6. Conclusions

In conclusion, using AI chatbots, such as ChatGPT and Claude AI, outside their intended purpose results in unreliable and inconsistent information. The findings reveal that these systems generate authentic and original text, which can lead to variations in assignment grading, rationale, and feedback. These findings support significant theoretical and practical implications for educational assessment and the integration of OpenAI technology in learning environments.

The study challenges assumptions derived from behaviorism theory. This theory emphasizes the importance of consistent stimulus-response associations for effective learning. The findings support that there are significant inconsistencies in grading patterns and feedback across ChatGPT and Claude AI. Consequently, the behaviorist approach may not apply in AI-driven assessment. Furthermore, the findings raise questions about the validity of applying CLT in situations where OpenAI hallucinations introduce unpredictable variations in grading outcomes, potentially disrupting the cognitive processes essential for learning.

Furthermore, inconsistencies within and across OpenAI systems can confuse students regarding expectations and learning objectives. This can lead to a misalignment between grades and actual student progress, impacting motivation, hindering skill development, and potentially degrading the creation of an inclusive learning environment. As a result, these variations in grading and feedback may result in unfair comparisons among students and impede their academic growth.

Finally, this study confirms the need for collaboration between educators, developers, and stakeholders to address these challenges. By understanding and addressing the limitations of using OpenAI systems for grading, educators can enhance the quality of assessment practices and promote optimal learning outcomes for students.

References

[1] Scharth, M. (2022). The ChatGPT chatbot is blowing people away with its writing skills. The University of Sydney. https://www.sydney.edu.au/news-opinion/news/2022/12/08/the-chatgpt-chatbot-is-blowing-people-away-with-its-writing-skil.html

[2] Motlagh, N. Y., Khajavi, M., Sharifi, A., & Ahmadi, M. (2023). The impact of artificial intelligence on the evolution of digital education: A comparative study of openAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie. arXiv preprint arXiv:2309.02029.

[3] Gleason, N. (2022). ChatGPT and the rise of AI writers: How should higher education respond? Times Higher Education. https://www.timeshighereducation.com/campus/chatgpt-and-rise-ai-writers-how-should-higher-education-respond

[4] Davis, F. D., Bagozzi, R. P. & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. Management Science, 35, 982–1003.

[5] Hsu, H. H., & Chang, Y. Y. (2013). Extended TAM model: Impacts of convenience on acceptance and use of Moodle. Online Submission, 3(4), 211–218.

[6] Skinner, B. F. (1984). An operant analysis of problem solving. Behavioral and brain sciences, 7(4), 583-591.

[7] Clark, S. M., Leonard, M. T., Cano, A., & Pester, B. (2018). Beyond operant theory of observer reinforcement of pain behavior. Social and Interpersonal Dynamics in Pain: We Don't Suffer Alone, 273-293.

[8] Rawas, S. (2023). ChatGPT: Empowering lifelong learning in the digital age of higher education. Education and Information Technologies, 1–14.

[9] Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138, 107468.

[10] Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT tool towards ameliorating the education system. BenchCouncil Transactions on Benchmarks, Standards, and Evaluations, 3(2), 100115.

[11] Al Aziz M.M., Ahmed T., Faequa T., Jiang X., Yao Y., Mohammed N. (2021). Differentially private medical texts generation using generative neural networks. ACM Transactions on Computing for Healthcare;3(1):1–27.

[12] Manodnya K.H., Giri A. (2022). IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA) IEEE; 2022. GPT-K: A GPT-based model for text generation in Kannada; pp. 534–539.

[13] Saini, N. (2023). ChatGPT Becomes Fastest Growing App in the World, Records 100mn Users in 2 Month.

[14] Menon, D., & Shilpa, K. (2023). "Chatting with ChatGPT": Analyzing the factors influencing users' intention to use the open AI's ChatGPT using the UTAUT model. Heliyon, 9(11), e20962-e20962.

[15] Mogaji E., Balakrishnan J., Nwoba A.C., Nguyen N.P. Emerging-market consumers' interactions with banking chatbots. Telematics Inf. 2021;65.

[16] Anghelescu, A., Ciobanu, I., Munteanu, C., Anghelescu, L. A. M., & Onose, G. (2023). ChatGPT: "to be or not to be." in academic research. The human mind's analytical rigor and capacity to discriminate between AI bots' truths and hallucinations. Balneo and PRM Research Journal (Online. English Ed.), 14(Vol. 14, no. 4), 614.

[17] Venkatesh V., Morris M.G., Davis G.B., Davis F.D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly. 425–478.

[18] Ho, T. (2022). Moral difference between humans and robots: Paternalism and human-relative reason. AI & Society, 37(4), 1533-1543.

[19] Li, Y., Chen, D., & Deng, X. (2024). The impact of digital educational games on student's motivation for learning: The mediating effect of learning engagement and the moderating effect of the digital environment. PloS One, 19(1), e0294350-e0294350.

[20] Fokides E. (2018). Digital educational games and mathematics. Results of a case study in primary school settings. Education and Information Technologies, 23(2), 851–867.

[21] Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on learning design. Learning and Instruction, 12(1), 1-10.
[22] Baddeley, A. (1992). Working Memory. Science, 255(5044), 556–559.
[23] Baddeley, A. (2020). Working Memory. In Memory (pp. 71–111). Routledge.
[24] Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. Cognition and Instruction, 8(4), 351–362.
[25] Sweller, J. (2020). Cognitive load theory and educational technology. Educational Technology Research and Development, pp. 68, 1–16.
[26] Kennedy, M. J., & Romig, J. E. (2021). Cognitive load theory: An applied reintroduction for special and general educators. Teaching Exceptional Children, 4005992110482.
[27] Swanson, H. L., Lussier, C. M., & Orosco, M. J. (2015). Cognitive strategies, Working Memory, and growth in word problem-solving in children with math difficulties. Journal of Learning Disabilities, pp. 48, 339–358.
[28] Windschitl, M., Thompson, J., & Braaten, M. (2020). Ambitious science teaching. Harvard Education Press.
[29] Drew, S. V., Thomas, J. D., & Nagle, C. (2023). Rock out the rubric: Self-regulated strategy development to revise science writing. TEACHING Exceptional Children, 00400599231185846.
[30] Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. Journal of Science Education and Technology, 30(2), 150–167.
[31] Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.
[32] Gundlach, H., & Dawborn-Gundlach, M. (2020). Teacher perceptions of quality criterion referenced rubrics in practice. Literacy Learning: The Middle Years, 28(3), 64-75.
[33] Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., ... & Haudek, K. C. (2019). Deconstruction of holistic rubrics into analytic rubrics for large-scale assessments of students' reasoning of complex science concepts. Practical Assessment, Research, and Evaluation, 24(1), 7.
[34] Kennedy, E., & Shiel, G. (2022). Writing assessment for communities of writers: rubric validation to support formative assessment of writing in Pre-K to grade 2. Assessment in Education: Principles, Policy & Practice, 29(2), 127–149.
[35] Weigle, S. C. (2002). Assessing writing. Cambridge University Press.
[36] Babin, E., & Harrison, K. (1999). Contemporary composition studies a guide to theorist and terms. Portsmouth: Greenwood Publishing.
[37] Gunning, J. W. (2006). Budget support, conditionality, and impact evaluation. Budget Support as More Effective Aid? 295.
[38] Crehan, K. D. (1997). A Discussion of Analytic Scoring for Writing Performance Assessments.
[39] Cope, B., Kalantzis, M., Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies, Educational Philosophy and Theory, 53:12, 1229–1245.
[40] García Ros, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university context.
[41] Aydin, Ö., & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? Academic Platform Journal of Engineering and Smart Systems, 11(3), 118-134.
[42] Zhai, X., C Haudek, K., Shi, L., H Nehm, R., & Urban‐Lurain, M. (2020). From substitution to redefinition: A framework of machine learning‐based science assessment. Journal of Research in Science Teaching, 57(9), 1430-1459.
[43] Kim, J., Merrill Jr, K., Xu, K., & Sellnow, D. D. (2021). I like my relational machine teacher: An AI instructor's communication styles and social presence in online education. International Journal of Human–Computer Interaction, 37(18), 1760-1770.
[44] Kim, J., Merrill, K., Kun, X., & Sellnow, D. D. (2022). Embracing AI-based education: Perceived social presence of human teachers and expectations about machine teachers in online education. Human-Machine Communication, 4, 169–184.
[45] Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. Journal of Science Teacher Education, 27(1), 111-129.
[46] Chowdhary, K. (2020). Natural Language Processing. In: Fundamentals of Artificial Intelligence. Springer, New Delhi.
[47] Korteling, J.E.; van de Boer-Visschedijk, G.C.; Blankendaal, R.A.M.; Boonekamp, R.C.; Eikelboom, A.R. (2021). Human-versus Artificial Intelligence. Sec. AI for Human Learning and Behavior Change (4).
[48] Chen, J. (2021). Refining the teacher emotion model: Evidence from a review of literature published between 1985 and 2019. Cambridge Journal of Education, 51(3), 327–357.
[49] Schneider, C., & Boyer, M. (2020). Design and implementation for automated scoring systems. In Handbook of Automated Scoring (pp. 217-240). Chapman and Hall/CRC.
[50] Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. Diversity and inclusion in global higher education: Lessons from across Asia, 73–95.
[51] Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. Procedia Computer Science, 136, 16-24.
[52] Sudheesh, R., Mujahid, M., Rustam, F., Mallampati, B., Chunduri, V., de la Torre Díez, I., & Ashraf, I. (2023). Bidirectional encoder representations from transformers and deep learning model for analyzing smartphone-related tweets. PeerJ Computer Science, 9, e1432.
[53] Goel, A. K., & Joyner, D. A. (2017). Using AI to teach AI: Lessons from an online AI class. Ai Magazine, 38(2), 48-59.
[54] Braun, D., Rogetzer, P., Stoica, E., & Kurzhals, H. (2023). Students' Perspective on AI-Supported Assessment of Open-Ended Questions in Higher Education. In CSEDU (2)(pp. 73-79).
[55] Saplacan, D.; Herstad, J.; Pajalic, Z. (2018). Feedback from digital systems used in higher education: An inquiry into triggered emotions two universal design-oriented solutions for a better user experience. In Transforming Our World through Design, Diversity, and Education: Proceedings of Universal Design and Higher Education in Transformation Congress (256) 421–430.