ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/oluwole-fagbohun/19

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 2 & Iss: 1

Research Article

Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices

Oluwole Fagbohun[1*], Nwaamaka Pearl Iduwe[2], Mustapha Abdullahi[3], Adeseye Ifaturoti[4] and Obinna Maxwell Nwanna[4]

[1]Tech Team Changeblock, London, SW1H 0HW, UK

[2]Department of IT, Microsystems Int'l (UK) Limited, UK

[3]Gen AI Unit, Readrly Limited, London, UK

[4]University of Greenwich, London, United Kingdom

[5]Tech team, RideNear LTD London, W1U 6AG, UK

Citation: Fagbohun O, Iduwe NP, Abdullahi M, Ifaturoti A, Nwanna OM. Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *J Artif Intell Mach Learn & Data Sci 2024*, 2(1), 1-8. DOI: doi.org/10.51219/JAIMLD/oluwole-fagbohun/19

Received: 30 January, 2024; **Accepted:** 03 February, 2024; **Published:** 05 February, 2024

***Corresponding author:** Oluwole Fagbohun, Tech Team Changeblock, London, SW1H 0HW, UK, E-mail: oluwole.fagbohun@changeblock.com

Copyright: © 2023 Fagbohun O. et al., This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

In this study, we examine the transformative role of large language models (LLMs) in redefining educational assessments. Traditional grading systems, characterized by their uniform and often manual approaches, face significant challenges in terms of scalability, consistency, and personalized feedback. The advent of LLMs heralds a new era of assessment, offering nuanced, scalable, and efficient solutions. This study explores the integration of LLMs in grading practices and examines their potential to revolutionize the assessment landscape.

We begin by analyzing the limitations of traditional grading methods and emphasizing the need for more sophisticated and adaptable solutions. This paper then introduces the concept of LLMs, outlining their advanced capabilities in natural language processing and machine learning, which are pivotal in understanding and evaluating student responses. We delve into the mechanisms by which LLMs process, analyze, and grade a wide range of responses, from short answers to complex essays, highlighting their ability to provide detailed feedback and insights beyond mere correctness.

The core of the discussion revolves around real-world applications and case studies in which LLMs have been implemented in educational assessments. These include automated grading systems and adaptive testing platforms, showing the effectiveness of LLMs in handling diverse and intricate responses. The outcomes of these implementations were analyzed, demonstrating LLMs potential in enhancing the accuracy, fairness, and efficiency of grading practices. However, the integration of LLMs into grading systems is challenging. This study critically examines issues such as potential biases in AI models, data privacy concerns, and the need to maintain ethical standards in automated grading. We propose strategies to mitigate these challenges by emphasizing the importance of human oversight and continuous model refinement.

This study offers a forward-looking perspective on the future of grading practices that use LLMs. We envision a paradigm shift towards more personalized, fair, and efficient assessment methods, facilitated by the ongoing advancements in LLM technologies. The integration of LLMs into grading systems promises a more adaptive and insightful approach to educational assessment, aligning with the broader goals of personalized learning and educational equity.

Keywords: Large Language Models, Educational Assessment, Grading Practices, Automated Grading, Adaptive Testing, Bias in AI, Data Privacy, Ethical Standards

1 Introduction

Grading is a fundamental aspect of educational assessment that serves as a crucial means of providing feedback from students to educators. Despite its importance, traditional grading methods, which rely on manual processes, face several challenges that hinder their effectiveness and ability to scale. Grading is an essential yet labor-intensive task that is prone to subjectivity and repetition, which poses significant challenges to assessment practices[1]. While traditional methods, such as rubric-based evaluation and norm-referenced grading, aim for objectivity, they often fall short due to subjective biases and inconsistencies. These methods frequently fail to provide personalized feedback at scale, which is a critical component in enhancing the instructional value of assessments[2-4].

As the academic world stands on the brink of a technological revolution, large language models (LLMs) have emerged as a groundbreaking force poised to revolutionize traditional grading systems. Fueled by advancements in natural language processing (NLP) and machine learning, LLMs demonstrate an extraordinary capacity to analyze, comprehend, and evaluate complex textual data[5-7]. Celebrated for their ability to process natural language with near-human proficiency, LLMs are reshaping various fields such as education being a prime example[8-10]. The transformative potential of LLMs in redefining grading is two-fold. First, they promise significantly enhanced efficiency and can rapidly and consistently evaluate vast volumes of student work. Second, they promise to provide individualized feedback custom-tailored to each student's unique responses, thereby fostering a more personalized and effective learning experience[2,11].

The incorporation of LLMs in the grading process marks a significant paradigm shift that transcends the constraints of conventional methods. This shift results in enhanced accuracy in assessments and fosters a more adaptable and expandable educational feedback mechanism. In the following sections, we will explore the intricacies of this transformative integration by analyzing the competencies of LLMs in educational assessment and assessing the extensive ramifications of their implementation.

2. The Limitations of Traditional Grading Methods

Traditional grading practices, a staple in educational assessment for generations, face numerous challenges that hinder their effectiveness in today's dynamic academic environments. Although manual grading has been a longstanding norm across various disciplines, its limitations, particularly concerning scalability, consistency, and personalized feedback, are becoming increasingly apparent as shown in **Figure 1**.

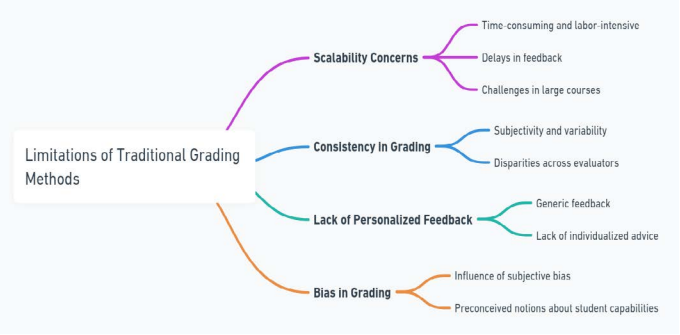

Figure 1: Limitations of traditional grading methods.

2.1 Scalability Concerns

The scalability of manual grading systems has been severely tested in the context of growing class sizes and the corresponding increase in the volume of assignments. Grading is inherently time-consuming and requires substantial human effort for a detailed evaluation of student work[2,12-15]. Instructors often struggle to provide prompt feedback to many students, a situation exacerbated in under-resourced institutions or where teaching assistants are scarce. This limitation often leads to delays in feedback, potentially hindering the learning process.

Grading text-based questions, especially in large courses, is particularly laborious and time-consuming, posing significant challenges for instructors in both formative and summative assessments[16]. This challenge highlights the need for innovative solutions in the field of educational technology, specifically using AI in education to facilitate large-scale teaching and provide real-time personalized feedback[16,17].

2.2. Consistency in Grading

Another significant limitation of manual grading is the difficulty of achieving consistency. Subjectivity in the grading process can lead to disparities in the assessment of student work even when detailed rubrics are in place. Variability may occur not only across different evaluators but also within the assessments of a single instructor over time, undermining the reliability and fairness of the grading system[2-4].

2.3. Lack of Personalized Feedback

Personalized feedback plays a crucial role in the educational process by providing students with specific guidance tailored to their individual needs. However, manual grading systems often fail to deliver personalized attention. Instructors faced with large volumes of student work may resort to generic feedback that lacks the specificity necessary for effective learning enhancement. Consequently, students may be deprived of individualized advice, which is crucial for their academic development[18].

2.4. Bias in Grading

Bias in grading is another critical issue, with studies indicating that subjective bias can influence grading outcomes[13,19]. Such biases can stem from a range of factors, including preconceived notions about student capabilities or unconscious preferences, further compromising the fairness of the traditional grading methods.

The limitations outlined above highlight the urgent need for alternative assessment mechanisms to overcome these challenges. This necessity opens the door for the integration of LLMs in educational assessments, offering solutions to the drawbacks of traditional grading methods. LLMs hold the promise of providing robust, scalable, and personalized evaluation approaches that address the key challenges of scalability, consistency, personalized feedback, and bias inherent in manual grading systems.

3. Large Language Models in Educational Assessment

Large language models (LLMs) have ushered in a new era in the field of NLP, particularly in educational assessment. These advanced models are built on deep learning architectures and trained on wide-ranging textual data, enabling them to effectively understand, generate, and analyze the human language. This capability positions them as ideal tools for various applications in educational assessment, such as processing and analyzing student responses[20-22].

3.1 Mechanisms of LLMs in Processing and Analyzing Student Responses

LLMs represent a departure from traditional keyword matching techniques. By leveraging algorithms and pre-trained language patterns, they can interpret and evaluate student responses in a context-aware manner. This makes them particularly adept at handling complex assessment tasks, such as essay grading, where they can assess not only the content but also the coherence, structure, argumentation, grammatical correctness, and usage of domain-specific terminology.

In educational settings, LLMs can be fine-tuned to accommodate specific content and grading rubrics, thus aligning their general language understanding capabilities with the specific needs of a given educational context. Additionally, attention mechanisms enable these models to focus on the most relevant parts of a student's response, thus mimicking the approach of human graders.

3.2 Comparison with Traditional Grading Methods

Traditional grading methods are often limited by human bias, inconsistencies, and significant time consumption. By contrast, LLMs offer a more consistent and objective assessment mechanism. However, it is crucial to acknowledge the limitations of LLMs, particularly their potential to perpetuate biases present in their training data. Therefore, it is essential for educators to apply human oversight to complement and verify LLM-based assessment.

3.3. Analytics and Insights

LLMs in educational assessments can also generate valuable analytics and insights into student performance. These insights20 can help educators understand learning patterns, identify common misconceptions, and make data-driven instructional decisions[21].

3.4 Enhanced Automatic Scoring

Automatic scoring systems have become more precise, adaptive, and context-aware with the introduction of LLMs, such as Chat GPT, which provide an extensive knowledge base and contextual understanding of the table, supplemented by domain-specific expertise[9]. LLMs have shown remarkable effectiveness in grading complex text-based responses, historically a challenge for machine evaluation. Their capabilities extend to various domains, including translation, where they have demonstrated an ability to understand linguistic nuances[23-25].

Moreover, LLMs have shown potential in evaluating divergent thinking, an area that automated systems have traditionally found challenging. This advancement indicates their ability to assess creativity and originality beyond simple semantic analysis[26]. The advent of LLMs in educational assessment marks a significant step forward, providing tools that offer greater accuracy, consistency, and insight than traditional methods. While they are not without limitations, their potential in transforming educational assessment is considerable. Educators and institutions should harness these technologies judiciously, ensuring that the benefits of LLMs are fully realized while mitigating their inherent limitations.

4. Related Work

4.1. Traditional methods

The development of automatic scoring algorithms began over a decade ago, primarily using token-based models in student responses, often referred to as the 'bag of words' approach[15,27]. These early models served as the foundation for current automatic scoring systems. The integration of convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) has led to a substantial improvement in automatic scoring capabilities, significantly enhancing the accuracy of score predictions. Studies by[28-30] provide evidence of this effectiveness.

However, it is essential to acknowledge the limitations of traditional machine-learning methods in addressing the nuanced contexts necessary for successful educational interactions. Recent research has highlighted the need for AI models that are more attuned to educational contexts, as indicated by[31-34]. Researchers have explored various approaches to enhance the automated assessment of student responses using NLP techniques[35]. investigated an e-learning system that utilized NLP to generate standard answers for grading essay questions.

[36]Created QuizBot, an interactive learning system that outperforms conventional methods in enhancing student learning. [37]Developed an automated rating system using key phrases and machine learning, laying a foundation for future improvements despite its limitations in application-based responses and lack of training data.

[38]Developed a comprehensive electronic exam assessment system that utilizes semantic and document similarity, which is notable for its structured approach, although it does not address syntactic errors in keywords or study high-quality documentation.[39]Implemented LSTM-RNN and GloVe vectors in their NLP-based system, which is an advanced application of deep learning in education that requires optimization to reduce computation time[40] demonstrated the effectiveness of evaluating long or descriptive answers in small class scenarios using methods such as TFIDF and LDA, offering adaptability in resource-limited settings. Finally[41] used WordNet charts for short-answer evaluation, focusing on semantic relationships, an approach that shows promise in understanding student responses, although it has limitations with incorrectly spelled words.

4.2. LLM Based approaches

The development of LLMs like BERT[42] and GPT variants[43] has led to significant breakthroughs in the field of automatic scoring, which is rapidly evolving. According to studies by[44] and Rodriguez, et al.[1], BERT has been successfully used to evaluate essays and short answers, showcasing the potential of transformer-based models in comprehending, and analyzing intricate textual content. Fine-tuning these models, as demonstrated by Wang, et al.[45] and Yang, et al.[46], has led to the creation of more refined and accurate scoring systems. The convergence of AI and education is a crucial juncture that foreshadows the advent of personalized and inclusive learning experiences[47-49]. LLMs are poised to offer promising results in terms of accuracy, adaptability, and contextual comprehension. The employment of transformer-based models to assess lengthy textual responses, as demonstrated by Ormerod, et al.[24] and Rodriguez, et al.[1], exemplifies this capability.

LLMs have been found to be valuable in various domains, such as science education and translation, where a deep understanding of language is essential[21,25,31]. Moreover, the development of specialized models, like Math BERT, for mathematics education[50], further demonstrates their versatility.

Kasneci, et al.[9] have highlighted the potential of utilizing LLMs such as ChatGPT, supplemented with domain-specific expertise, to achieve more precise scoring. LLMs, with their extensive knowledge base, adaptability, and context awareness, set themselves apart from conventional AI models. They have demonstrated remarkable potential in evaluating and scoring intricate text-based responses. This task has traditionally been challenging for automated systems due to the richness and variability of human expressions.

In the study conducted by Gao, et al.[51], the efficacy of LLMs in grading short-answer questions within mechanical engineering courses was explored. This research not only compared LLMs to NLP systems but also assessed the potential for improving grading accuracy through keyword detection integrated into LLM frameworks. The study's key contribution lies in the precision achieved by LLMs in binary evaluations. However, it also highlights the need for further research to enhance the reliability and applicability of LLMs in diverse educational settings.

Lin, et al.[52] focused on enhancing the capabilities of human tutors through LLMs by generating real-time explanatory feedback during online tutoring sessions. This research, particularly in the context of effective praise strategies, leveraged named entity recognition facilitated by LLMs to identify key elements in tutor responses. The study's contribution is significant as it showcases the potential of integrating advanced NLP techniques with educational practices to refine tutor training and enhance feedback mechanisms.

Jiang, et al.[53] conducted an evaluation of LLMs' effectiveness in assessing the accuracy of Chinese language writing. By comparing LLM assessments to human ratings, the study highlighted the potential of LLMs in language learning assessments. It emphasized the accuracy, efficiency, and capability of LLMs in identifying common error types. However, the study also identified challenges related to contextual understanding and overcorrection, underscoring the continued importance of human oversight in educational assessments.

Meyers, et al.[54] introduced a pioneering pipeline for creating and assessing educational materials using LLMs. The system's contribution lies in its development of an end-to-end solution that seamlessly integrates with Learning Management Systems (LMS). It aims to streamline the assessment creation process, making it more efficient for educators across a spectrum of subjects. This innovation holds the promise of transforming how educational materials are generated and evaluated.

While the integration of advanced AI tools in educational settings has shown promise, there are concerns about potential biases and ethical implications that require careful evaluation[22,55,56]. Recent studies by Han, et al.[11] and Mizumoto & Eguchi[57] suggest that the implementation of these tools in automatic scoring systems still faces challenges in outperforming existing benchmarks. In conclusion, the use of advanced AI models like BERT and GPT variants has advanced automatic scoring, but there are still challenges that need to be addressed. Future research should focus not only on enhancing these models' technical capabilities but also on addressing ethical concerns and ensuring their effectiveness and adaptability in various educational contexts.

5. Case Studies and Real-World Applications of Large Language Models in Education

Open-ended questions are integral to educational settings for assessing students' understanding and fostering critical thinking. However, providing personalized feedback for such responses is often time-consuming, leading instructors to opt for simpler formats such as multiple-choice questions. This section introduces a tool utilizing large language models (LLMs) guided by instructor-defined criteria to automate feedback for open-ended questions, presenting a significant leap in educational assessment methodologies[58].

5.1 Case study one: Automating feedback for open-ended questions

Matelsky, et al.[58] conducted a study to examine the potential of LLMs in evaluating student responses against predefined criteria established by instructors. The process involves instructors formulating questions and criteria, which the LLM then assesses to provide tailored feedback on student responses. The study's results demonstrate that this tool delivers prompt and customized feedback, facilitates knowledge assessment, and identifies areas for improvement. This research highlights the potential of LLMs to improve educational outcomes and teaching methodologies, particularly in addressing the challenges associated with grading open-ended questions.

5.2 Case study two: Chain-of-thought in automatic scoring

In a study conducted by Lee, et al.[22], the application of GPT-3.5 and GPT-4 models in tandem with chain-of-thought (CoT) reasoning for the automated scoring of science assessments was explored. By utilizing prompt engineering techniques and evaluating zero-shot and few-shot learning approaches, the research revealed that the implementation of CoT with few-shot learning resulted in increased scoring accuracy[59,60]. This finding underscores the importance of incorporating contextual information and appropriate prompting strategies when incorporating LLMs into educational assessment systems.

5.3 Case study three: Scoring divergent thinking

Organisciak, et al.[26] aimed to improve the automated evaluation of divergent thinking tasks by fine-tuning LLMs on human-evaluated responses. The study showed a significant advancement compared to conventional semantic distance methods, as the LLMs' assessments were found to be more in line with human judgments. This finding underscores the potential of LLMs to accurately gauge creative thinking abilities.

5.4 Case study four: Enhancing math self-explanation scoring

The authors of Nakamoto, et al.[61] propose a novel approach that integrates human-labeled and synthetic data generated by a LLM to achieve a semi-supervised learning method. This method proves to be highly effective in significantly improving the accuracy of assessing mathematical self-explanations, which is a critical aspect in evaluating complex cognitive abilities. This breakthrough in automated evaluation holds great promise for future advancements in this field.

5.5 Case study five: Fine-tuning ChatGPT for automatic scoring

Latif and Zhai[21] investigated the application of fine-tuned ChatGPT (GPT-3.5) for automatic scoring in their study titled "Fine-tuning ChatGPT for Automatic Scoring." Through the adaptation of GPT-3.5 on a diverse dataset of middle- and high-school student responses and comparing its performance to that of BERT, the study found a considerable improvement in scoring accuracy. This observation highlights the effectiveness of fine-tuning LLMs for domain-specific applications in education.

These case studies underscore the transformative potential of LLMs for educational assessment and grading. Institutions that implement these technologies have reported improvements in grading efficiency, accuracy, and the provision of personalized feedback, indicating a significant advancement in automated educational technologies.

6. Challenges and Ethical Considerations of Using LLMs in Educational Assessments

The adoption of large language models (LLMs) in educational grading systems presents various challenges and ethical considerations. This section delves into the complexities associated with implementing LLMs in educational settings, highlighting the need for careful consideration of biases, data privacy, ethical implications, and feedback quality.

6.1 Data Privacy and Security

Incorporating LLMs in educational settings necessitates strict data privacy and security measures to ensure the protection of sensitive student information from unauthorized access. Compliance with regulations, such as the General Data Protection Regulation (GDPR) and the Family Educational Rights and Privacy Act (FERPA), is crucial[63].

6.2 Depersonalization of Education

The shift towards AI-driven grading risks depersonalizing the educational experience. AI systems may overlook the unique context, personal challenges, and subtleties in student work, potentially failing to provide constructive, personalized feedback that is crucial for effective learning[64].

6.3 Disincentive for Innovation

Utilizing LLMs for automatic scoring may discourage students from developing innovative answers. A rigid scoring system may overlook creative or unconventional solutions, thereby narrowing the scope of acceptable responses[65].

6.4 Training for Test-taking Over Learning

There is a risk of LLMs promoting a test-taking approach to genuine learning. Students might prioritize answering in a manner recognizable as correct by the system, potentially bypassing deeper engagement with the material[65].

6.5 Challenges in Feedback Quality

While LLMs can provide immediate feedback, the quality may not match the nuanced and context-sensitive feedback of human graders. This disparity highlights the difficulty in ensuring that automated feedback is as constructive as that from a human instructor[65].

6.6 Implementation Challenges

Exercise Context and Rubric Ambiguity: The diverse contexts of educational exercises and the ambiguity in scoring rubrics present challenges, particularly for abstract evaluation criteria like logical structure[2].

Specialized Domain Limitations: LLMs, due to their generic training, might not possess the specialized knowledge or pedagogical strategies needed in educational contexts[66].

Variability in Student Responses: The inherent variability in student responses can complicate fairness and bias elimination in scoring[29].

6.7 Ethical and Bias Concerns

Biases in AI-driven Grading: A significant concern with LLMs is the potential for the introduction of biases. These biases often reflect the composition of the training data and can result in unfair grading, especially affecting minority student groups or nonstandard language users. For instance, LLMs predominantly trained on data from native English speakers may inadequately address the nuances of responses from English language learners[62,67-69].

Black-box Nature and Transparency: The opaque nature of some LLMs challenges the need for transparency and accountability in educational assessments [9,70-72].

Hallucinations in large language models: LLMs are prone to generating plausible yet incorrect or misaligned responses, known as hallucinations, which pose significant challenges in educational settings [17,72-74].

6.8 Knowledge Cutoff

Finite knowledge of LLMs and the need for regular updates to maintain relevance in a constantly evolving educational landscape present additional challenges [75-78].

6.9 Safety Concerns

Ensuring the safe use of LLMs, particularly in sensitive areas such as education, is crucial. Risks associated with the generation of harmful or inappropriate content must be carefully managed[77,79].

6.10 High Stakes Decisions and Grading

The use of LLMs in high-stake decisions, such as grading, necessitates a high degree of accuracy and ethical considerations. Errors in automated grading can have far-reaching consequences on students (Schneider et al. 2023)[1]. Although LLMs offer significant potential to enhance grading practices, it is crucial to address these challenges and ethical concerns. This involves maintaining human oversight, ensuring diversity and fairness in the training datasets, and adhering to stringent data privacy and security standards. A balanced approach, in which LLMs augment rather than replace human educators, is key to harnessing their benefits while upholding ethical and fair educational practices.

7. The Future of Grading with Large Language Models

The evolution of grading practices with the growing influence of large language models (LLMs) heralds a new era in educational assessment. This future envisions a synergy between the computational efficiency of LLMs and the critical pedagogical insights of educators, leveraging advancements in NLP to transform grading into a more efficient and insightful educational tool as shown in **Figure 2**.

7.1 Evolution of Grading with LLMs

The trajectory of LLMs in grading is geared towards offering more than just assessment accuracy. Future developments will likely see LLMs providing formative feedback, personalized learning resources, and adaptive assessment pathways attuned to individual student profiles. Anticipated advancements include understanding cognitive patterns for personalized interventions and integrating multimodal inputs for comprehensive assessment.

7.2 LLMs as Autonomous Teaching Assistants

Emerging LLM solutions suggest a future where these models serve as autonomous teaching assistants, dynamically

adjusting curriculum materials to suit class mastery levels. The convergence of LLMs with augmented and virtual reality can create immersive grading environments, offer instant feedback, and enhance understanding of complex topics[42].

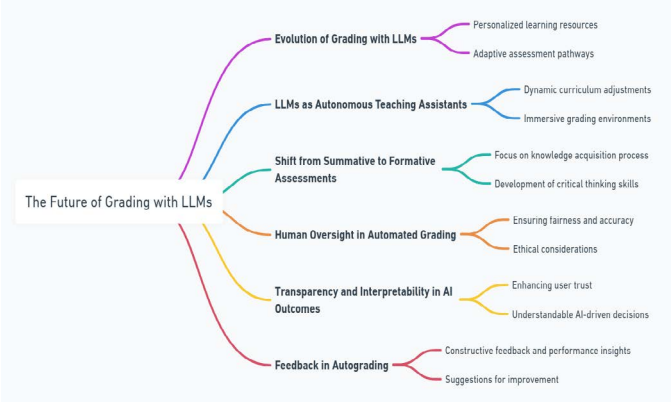

Figure 2: The Future of grading with LLMs.

7.3 Shift from Summative to Formative Assessments

An essential transformation in LLM-integrated grading is the transition from summative to formative assessment. Instead of merely assigning scores, LLMs focus on understanding the process of knowledge acquisition and the development of critical thinking skills, thus contributing to a more nuanced view of student learning progression.

7.4 Human Oversight in Automated Grading

The integration of AI into education, particularly in automatic scoring, comes with challenges that require human oversight to ensure fairness, accuracy, and ethical considerations (Bozkurt & Sharma, 2023). A significant preference among students for a blend of instructors and LLM grading underscores the need for a collaborative approach to grading processes[80].

Balancing automated systems with human judgment is crucial to addressing nuances, contexts, and creativity in education. Educators' involvement in designing, implementing, and reviewing AI-driven systems ensures a comprehensive educational approach.

7.5 Transparency and Interpretability in AI Outcomes

Educators' concerns about the opacity of AI outcomes necessitate improved interpretability and explainability of AI applications. Enhancing user trust and ensuring that AI-driven decisions are understandable and justifiable in educational settings[9,62,81,82].

7.6 Feedback in auto grading

Feedback during auto grading plays a vital role in guiding student learning. In addition to providing assessments, AI-based systems also deliver constructive feedback, offer insights into performance, and suggest areas for improvement[2,13,83].

The future of grading with LLMs paints an optimistic scenario for educational assessment, where technology not only enhances the role of educators, but also enriches student learning experiences. In this emerging paradigm, the effectiveness of LLMs is measured by how well they complement indispensable human elements in education. Regular audits, updates to training datasets, and a balanced approach between automation and human oversight are key to realizing the full potential of AI in grading while upholding ethical standards.

8. Conclusion

The emergence and integration of large language models (LLMs) within the educational landscape herald a transformative era of grading practices. Throughout this paper, we discussed the limitations of traditional manual grading systems, including their inability to scale, maintain consistency, and provide meaningful, personalized feedback. The incorporation of LLMs in educational assessments presents a remarkable opportunity to overcome these challenges by harnessing their natural language processing and machine learning capabilities. Such systems can process extensive volumes of student responses with dexterity, offering nuanced and consistent evaluations that are nearly indistinguishable from human judgment.

LLMs have the potential to democratize education by providing high-quality feedback and evaluations without the traditional constraints of resource allotment and instructor availability. They not only enhance the efficiency and accuracy of academic assessments but also support the individual learning trajectories of students through customized feedback. As evidenced in the various case studies and real-world applications reviewed, early deployments of LLM-driven grading systems have already shown their promise in augmenting grading accuracy and operational bandwidth in educational institutions.

However, it is imperative to recognize that deploying these sophisticated models must be approached with a judicious balance between technological opportunities and ethical considerations. Concerns regarding potential biases embedded within AI models, data privacy, and security must be rigorously addressed. Moreover, the ethical implications of displacing human judgment with algorithmic determinations require careful scrutiny and ongoing oversight. Effective strategies, including transparent algorithmic processes, data governance protocols, and human-in-the-loop systems, must be crafted and implemented to safeguard against these risks. Looking forward, it is reasonable to posit that LLMs will become increasingly embedded within the fabric of educational assessment. Ongoing advancements in AI and machine learning will undoubtedly refine their precision and interactivity, further cementing their roles in academic grading. However, the future landscape will require harmonious collaboration between human insights and machine efficiency. This symbiosis will ultimately shape the trajectory for a reimagined grading paradigm that leverages the strengths of both educators and AI systems to foster a rich, equitable, and dynamic educational environment.

LLMs signify a watershed moment for educational assessment, offering a forward-looking lens through which we might reimagine and reshape traditional grading into a model that befits the dynamism and demands of 21st-century learning. Our collective responsibility educators, technologists, policymakers, and students is to navigate this transition responsibly, ensuring that the future of grading with LLMs is as equitable and beneficial as innovative.

9. References

1. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching, 2023;6: 342-363.

2. Yan L, Sha L, Zhao L, et al. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. BJET, 2024;55: 90-112.

3. Bloxham S, Boyd P. Developing effective assessment in higher education: A practical guide. 2007, New York, NY: McGraw-Hill International.
4. Kasneci E, Sessler K, Kuchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 2023;103: 102274.
5. Chen Q, Wang X, Zhao Q. Appearance Discrimination in Grading? - Evidence from Migrant Schools in China. Economics Letters, 2019;181: 116-119.
6. Jiang Z, Xu Z, Pan Z, He J, Xie K. Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. Languages, 2023;8: 247.
7. Vij S, Tayal D, Jain A. A Machine Learning Approach for Automated Evaluation of Short Answers Using Text Similarity Based on WordNet Graphs. Wireless Personal Communications, 2019;111: 1271-1282.
8. Gao R, Thomas N, Srinivasa A. Work in Progress: Large Language Model Based Automatic Grading Study. Proceedings - Frontiers in Education Conference, FIE, 2023.
9. Kashi A, Shastri S, Deshpande AR, Doreswamy J, Srinivasa G. A Score Recommendation System Towards Automating Assessment In Professional Courses. Proceedings - IEEE 8th International Conference on Technology for Education, 2017; 140-143.
10. Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. A Systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 2020;30: 121-204.
11. Hattie J, Timperley H. The power of feedback. Review of educational research, 2007;77: 81-112.
12. Boeskens L. Not enough hours in the day: Policies that shape teachers' use of time. OECD, 2020.
13. Han J, Yoo H, Myung J, et al. FABRIC: Automated Scoring and Feedback Generation for Essays. arXiv, 2023.
14. Jonsson A, Svingby G. The use of scoring rubrics: reliability, validity, and educational consequences. Educational Research Review, 2007;2: 130-144.
15. Taghipour K, Ng HT. A Neural Approach to Automated Essay Scoring. Association for Computational Linguistics, 2016; 1882-1891.
16. Lin J, Thomas DR, Han F, et al. Using large language models to provide explanatory feedback to human tutors. arXiv, 2023.
17. Izacard G, Lewis P, Lomeli M, et al. Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 2022; 1-43.
18. Haque S, Eberhart Z, Bansal A, McMillan C. Semantic Similarity Metrics for Evaluating Source Code Summarization. IEEE International Conference on Program Comprehension, 2022; 36-47.
19. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Semantic Scholar, 2019.
20. Latif E, Zhai X. Fine-tuning ChatGPT for automatic scoring. arXiv, 2023.
21. Lee GG, Latif E, Wu X, Liu N, Zhai X. applying large language models and chain-of-thought for automatic scoring. arXiv, 2023.
22. Leacock C, Chodorow M. C-rater: Automated scoring of short-answer questions. Computers and the Humanities, 2003;37: 389-405.
23. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching, 2023;6: 342-363.
24. Organisciak P, Acar S, Dumas D, Berthiaume K. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity, 2023;49: 101356.
25. Wu X, He X, Liu T, Liu N, Zhai X. Matching Exemplar as Next Sentence Prediction (MeNSP): Zero-shot Prompt Learning for Automatic Scoring in Science Education. AIED, 2023.
26. Ram O, Levine Y, Dalmedigos I, et al. In-Context Retrieval-Augmented Language Models. arXiv, 2023.
27. Liao QV, Vaughan JW. AI Transparency in the age of llms: a human-centered research roadmap. arXiv, 2023.
28. Du M, Liu N, Hu X. Techniques for interpretable machine learning. Communications of the ACM, 2020; 63: 68-77.
29. Rodriguez PU, Jafari A, Ormerod CM. Language models and Automated Essay Scoring. arXiv, 2019.
30. Tahiru F. AI in education: A systematic literature review. In Journal of Cases on Information Technology 2021;23: 1-20.
31. Longo L, Brcic M, Cabitza F, et al. Explainable artificial intelligence (xai) 2.0: a manifesto of open challenges and interdisciplinary research directions. arXiv, 2023.
32. Ormerod CM, Malhotra A, Jafari A. Automated essay scoring using efficient transformer-based language models. arXiv, 2021.
33. Tossell CC, Tenhundfeld NL, Momen A, Cooley K, de Visser EJ. Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. IEEE. Transactions on Learning Technologies, 2024.
34. Zhai X. ChatGPT User Experience: Implications for education. SSRN Electronic Journal, 2023.
35. Fagbohun O, Harrison RM, Dereventsov A. An empirical categorization of prompting techniques for large language models: a practitioner's guide journal of artificial intelligence, Machine Learning and Data Science, 2023.
36. Saha SK, Rao CHD. Development of a practical system for computerized evaluation of descriptive answers of middle school level students. Interactive Learning Environments, 2019;30: 1-14.
37. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2023;2: e0000198.
38. Alrehily AD, Siddiqui MA, M Buhari S. Intelligent electronic assessment for subjective exams. Semantic Scholar, 2018; 47-63.
39. Ng DTK, Lee M, Tan RJY, Hu X, Downie JS, Chu SKW. A review of AI teaching and learning from 2000 to 2020. Education and Information Technologies, 2023;28: 8445-8501.
40. Schneider J, Schenk B, Niklaus C, Vlachos M. Towards LLM-based autograding for short textual answers. arXiv, 2023
41. Wang Y, Wang C, Li R, Lin H. On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. arXiv, 2022.
42. Dong F, Zhang Y. Automatic Features for Essay Scoring-An Empirical Study. Association for Computational Linguistics, 2016; 1072-1077.
43. Matelsky JK, Parodi F, Liu T, Lange RD, Kording KP. A large language model-assisted education tool to provide feedback on open-ended responses. arXiv, 2023.
44. Mao R, Chen G, Zhang X, Guerin F, Cambria E. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. arXiv, 2023.
45. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models Chain-of-Thought Prompting. arXiv, 2023.

46. Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv, 2023.
47. Huang L, Yu W, Ma W, et al. A Survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. Semantic Scholar, 2023.
48. Xiao C, Ma W, Xu SX, Zhang K, Wang Y, Fu Q. From automation to augmentation: large language models elevating essay scoring landscape. arXiv, 2024.
49. Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A Survey on hallucination in large language models. arXiv, 2023.
50. Smyrnaiou Z, Liapakis A, Bougia A. Ethical Use of Artificial Intelligence and New Technologies in Education 5.0. Journal of Artificial Intelligence, Machine Learning and Data Science, 2023;1: 119-124.
51. Gao R, Merzdorf HE, Anwar S, Hipwell MC, Srinivasa AR. Automatic assessment of text-based responses in post-secondary education: A systematic review. Computers and Education: Artificial Intelligence, 2024; 6: 100206.
52. Liu Z, He X, Liu L, Liu T, Zhai X. Context matters: a strategy to pre-train language model for science education. arXiv, 2023.
53. Johnson Anderson V, Walvoord BE. Effective Grading: A tool for learning and assessment. Jossey-Bass, 1998.
54. Mhlanga D. Open AI in Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning. SSRN Electronic Journal, 2023.
55. Ruan S, Jiang L, Xu J, et al. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. Conference on Human Factors in Computing Systems, 2019.
56. Yang R, Cao J, Wen Z, Wu Y, He X. Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. Association for Computational Linguistics, 2020.
57. Nakamoto R, Flanagan B, Yamauchi T, Dai Y, Takami K, Ogata H. Enhancing automated scoring of math self-explanation quality using LLM-generated datasets: A semi-supervised approach. Computers, 2023;12: 217.
58. Meyers P, Han A, Grewal R, Potnis M, Stamper J. Focal: A Proposed Method of Leveraging LLMs for Automating Assessments. 2023.
59. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv, 2020.
60. Wu D, Wang M, Li X. Automatic Scoring for Translations Based on Language Models. Computational Intelligence and Neuroscience, 2022;2022: 2171206.
61. Naveed H, Khan AU, Qiu S, et al. A Comprehensive Overview of Large Language Models. arXiv, 2023.
62. Zhao H, Chen H, Yang F, et al. Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 2023.
63. Alier M, Casañ Guerrero MJ, Amo D, Severance, C, Fonseca D. Privacy and e-learning: A pending task. Sustainability, 2021;13: 9206.
64. Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A Review of trustworthy and explainable artificial intelligence (XAI). IEEE, 2023;11: 78994-79015.
65. Latif E, Mai G, Nyaaba M, et al. AGI: Artificial General Intelligence for education. arXiv, 2023.
66. Mizumoto A, Eguchi M. Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2023;2: 100050.
67. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021; 610-623.
68. Caines A, Benedetto L, Taslimipoor S, et al. On the application of large language models for language teaching and assessment technology. arXiv, 2023.
69. Susanti MNI, Ramadhan A, Warnars HLHS. Automatic essay exam scoring system: A systematic literature review. Procedia Computer Science, 2022;216: 531-538.
70. Dumal PAA, Shanika WKD, Pathinayake SAD, Sandanayake T. Adaptive and automated online assessment evaluation system. In Proceedings of the 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2017.
71. Lun J, Zhu J, Tang Y, Yang M. Multiple data augmentation strategies for improving performance on automatic short answer scoring. Proceedings of the AAAI Conference on Artifical Intelligence, 2020;34: 13389-13396.
72. George N, Sijimol PJ, Varghese SM. Grading descriptive answer scripts using deep learning. IJITEE, 2019;8.
73. Zhai X. Advancing automatic guidance in virtual science inquiry: from ease of use to personalization. Educational Technology Research and Development, 2021;69: 255-258.
74. Zhao H, Chen H, Yang F, et al. Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 2023.
75. Holmes W, Tuomi I. State of the art and practice in AI in education. European Journal of Education, 2022;57: 542-570.
76. Jablonka KM, Ai Q, Al-Feghali A, et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. Digit Discov 2023;2: 1233-1250.
77. Lin J, Thomas DR, Han F, et al. Using large language models to provide explanatory feedback to human tutors. arXiv, 2023.
78. Ren P, Yang L, Luo F. Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis. Education and Information Technologies, 2023;28: 797-814.
79. Shen JT, Yamashita M, Prihar E, et al. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education. arXiv, 2021.
80. Usman Hadi M, al tashi Q, Qureshi R, et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv, 2023.
81. Chang Y, Wang X, Wang J, et al. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 2024.
82. Duan X, Pei B, Ambrose GA, Hershkovitz A, Cheng Y, Wang C. Towards transparent and trustworthy prediction of student learning achievement by including instructors as co-designers: a case study. Education and Information Technologies, 2023.
83. Alseddiqi M, AL-Mofleh A, Albalooshi L, Najam O. Revolutionizing online learning: The potential of chatgpt in massive open online courses. European Journal of Education and Pedagogy, 2023;4: 1-5.