International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 01 | Jan 2026 www.irjet.net p-ISSN: 2395-0072

From Human Judgment to AI Algorithms: A Literature-Based Analysis of Automated Essay Scoring Accuracy and Limitations

Aaryan Karlapalem[1]

[1]*Student pursuing Computer Science at Tomball Memorial High School, Tomball, Texas, USA*

---***---

Abstract - *Since its release in 2022, arti/icial intelligence (AI) has become a prominent part of most education systems, with nearly 20% of American schools using some form of AI in the classroom [3]. While most of these schools just use AI for lesson planning and personalized learning, however, one of the most labor-intensive tasks in education, grading essays, has also begun to shift toward automation. To reduce teacher workload and improve grading ef/iciency, teachers and state agencies such as the Texas Education Agency have been exploring AI-based tools for evaluating student writing. These tools, called automated essay scoring (AES) systems, use natural language processing (NLP) and machine learning to mimic human judgment in reviewing human writing. This paper examines the capabilities and limitations of modern AES technologies, comparing their grading accuracy, feedback quality, and overall fairness to traditional human grading. Using recent research, it explores whether AI is capable of permanently replacing human graders. Ultimately, it concludes that while current AES systems offer some bene/its in scalability and ef/iciency, they are not yet advanced enough to fully replace human judgment in high-stakes or nuanced essay evaluation contexts.*

Keywords: Artificial intelligence (AI), automated essay scoring (AES), Natural language processing (NLP)

1. Introduction

During the COVID-19 pandemic, virtually every major industry experienced huge impacts because of the global outbreak, and the education industry was no exception to that, with thousands of schools worldwide being forced to switch to virtual school and video calls to replace physical schools and classes. During this time, the burden put on teachers was immense, with teachers having to juggle their personal health as well as their duty to provide a quality learning experience to students despite the pandemic. This quickly resulted in teachers mass-quitting their jobs because of increased stress and fatigue. This continues to be a major problem even today, with a teacher shortage being marked as a serious community problem. Schools and universities globally have been struggling to hire qualiGied teachers, a problem that has only become worse because of the global pandemic. Furthermore, young people today have also been showing less interest in pursuing a career in education, with Millennials and Generation Z being the least interested in becoming future teachers. All these factors together intensi Gied the overall problem and the need to accelerate ways to bring in more teachers. Additionally, according to a study done by Merrimack College in 2022, it was shown that from 2012 to 2022, the percentage of K-12 teachers who were "very satisGied" with their jobs fell from 39% to 12%, the lowest this statistic has ever been. Additionally, more teachers have reported that they want to leave their jobs to pursue a career outside of education, further showing the growing disinterest in teaching.

2. The Rise of Arti Bicial Intelligence in Education

Ever since the early 2020s, artiGicial intelligence has been experiencing rapid adoption among various education systems around the world, and for good reason. Ever since the global COVID-19 pandemic, teachers have been under more stress than ever, and they desperately need to streamline their work Glows to keep up with the increasing amount of work they have. Thus, AI technologies presented themselves as the solution to this, providing tools to create lesson plans, slideshows, and even grading to an extent. Initially, these AI services were simply being used to automate parts of a teacher's job, such as the generation of lesson materials and assessments. However, with the recent rise of various LLMs (large language models) that specialize in certain tasks, such as GPT-3.5, GPT-4o mini, Gemini 2.5 Glash, and more, the extent to which AI is being utilized in the classroom has only been increasing. Today, these models are being tested and deployed for more complex tasks, like essay evaluation and feedback, grading, and even tutoring to some extent. These advancements bring about a shift from simply helping teachers with their jobs to potentially trying to replace them, especially in areas like writing assessment that are time-intensive for teachers. One of the emerging concepts tied to this shift is precision education, which aims to use artiGicial intelligence to customize instruction and grading to each student's needs [5]. This kind of system integrates deep learning, transfer learning, and learning analytics to help students personally. Within this new framework, automated essay scoring (AES) systems are often used to not only grade student essays, but also to track the progression of the student's writing over time, providing valuable insight into their writing skills. This kind of system has also caught the attention of state-level agencies, such as the Texas Education Agency (TEA), with them recently making the decision to use artiGicial intelligence systems to grade the standardized State of Texas Assessments of Academic Readiness (STAAR) tests that are taken by millions of kids from grades 3-12[4]. The hope of this is that it reduces the workload of graders who would otherwise have to spend hundreds or even thousands of hours grading essays over the summer, and this solution has been shown to be highly effective at solving this issue. However, while the promise of speed and scalability is promising, these early AI implementations raise numerous questions regarding whether AI can replicate the nuance, empathy, and contextual understanding that human graders have.

3. How Automated Essay Scoring Works

Automated Essay Scoring (AES) systems are AI tools speciGically designed to mimic aspects of human judgement when grading student-produced writing. These tools look at the content, structure, and speciGically the language of essays to produce a Ginal grade and feedback. This is all powered by a combination of natural language processing (NLP) and machine learning (ML) techniques, with the primary goal being to offer a fast and efGicient way to evaluate human writing, something that is much needed in the education industry amid the increasing teacher workloads and shortages.

3.1. Holistic vs Trait-SpeciEic Scoring

AES models generally operate in one of two ways: holistic scoring or trait-speciGic scoring. Holistic systems analyze the entire essay and return a single value to represent the essay's general quality, very similar to how a normal teacher would grade an essay. On the other hand, trait-speciGic scoring systems assess very distinct components of a piece of writing, such as the grammar, essay structure, and sometimes even creativity/originality, assigning individual scores for each category. To compare, holistic systems are often faster and much simpler, but trait-speciGic models provide much more detailed and constructive feedback for both students and teachers. For example, the ASAP (Automated Student Assessment Prize) dataset, commonly used in AES research, includes essays evaluated along four main categories: ideas, organization, style, and conventions [10]. These traits serve as training data for machine learning models that train on past human essays and then replicate the grading process.

3.2. The ArtiEicial Intelligence Models Behind AES

Earlier AES systems, such as Project Essay Grade (PEG) and e-rater, used hard-coded features, such as word count, sentence length, and spelling accuracy, to predict scores. However, these models often failed to understand the deeper meaning, structure, or logical coherence of the essays they were analyzing. More recently, however, deep learning has dramatically enhanced the sophistication of AES tools.

Newer systems now rely on transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), which can analyze texts at a much deeper level. These models understand word relationships, sentence context, and even document structure to provide a more accurate score that more closely mimics traditional human grading. For example, BERT can evaluate sentence-level grammar and coherence, while GPT-4 and other LLMs can analyze entire essays for logic, consistency, and originality. In a 2025 study conducted by SeBler et al, 37 teachers from Germany graded 20 German student essays based on 10 speciGic criteria. The same essays were then graded by LLMs such as GPT-3.5, o1, LLaMA, and Miztral using a standardized prompt. The best performing model, o1, achieved a Spearman correlation coefGicient of 0.742 in overall alignment with human raters, which demonstrates an exceptionally strong agreement with human graders [7].

Interestingly, the study also found that the closed-source models (GPT-4, o1) consistently outperformed their open-source counterparts in both inter-rater reliability and alignment with human feedback, especially in language-related areas like spelling and grammar. However, even top models struggled in content-based categories such as evaluating the logic of an argument or the effectiveness of a narrative conclusion, which require a level of understanding not yet present in even the most advanced AI models today. However, at the fast speed that these LLMs are evolving and becoming more advanced, it is highly likely that AI can fully mimic a human grader by the early 2030s, especially with the recent advances in o1 and LLaMA.

3.3. Hybrid and Collaborative AES Systems

The previous models discussed fall under the category of single-model approaches, as they only use one AI model to do all the processing. However, researchers have recently been exploring the possibility of collaborative models, which divide grading tasks among multiple specialized neural networks. In 2024, one of the very Girst Collaborative Deep Learning Networks (CDLNs) was introduced. This combined convolutional neural networks (CNNs), recursive neural networks (RvNNs), and LSTM to individually assess grammar, structure, and content before combining them all into one Ginal grade. This new model achieved a grading accuracy of 85.5%, outperforming any traditional system and even most BERT-based models [10].

This modular approach mimics the way humans grade essays as well by assessing writing from multiple angles, such as grammar, clarity, idea development, and logical Glow, which allows for each neural network to specialize in one speciGic angle of grading the essay. This also results in a more balanced and thorough evaluation overall.

3.4. Limitations and Constraints

Despite all these advances, however, AES systems do come with some limitations. For one, their performance is heavily inGluenced by the training data, which might reGlect existing biases, such as favoring native speakers or formal writing styles. Moreover, AES systems struggle with creative, emotional, or unconventional writing, contexts that humans understand far better than even the best AI models. Furthermore, most systems still rely on pattern recognition as opposed to genuine understanding, which can often lead to inaccurate scores, especially when students use complex or unconventional language. Ultimately, while the technical complexity of AES systems has indeed increased, they remain best suited for low-stakes situations or as supplements to human grading, rather than being a full-on replacement.

4. Accuracy: Can AI Match Human Grading?

The main claim behind adopting AES tools is that they can assess student writing as accurately as human educators, but with the beneGit of being far more efGicient and scalable. While most recent advances in natural language processing (NLP) and large language models (LLMs) have brought AES systems closer to this goal, questions remain about their reliability in high-stakes grading. Numerous studies have shown that AI systems can approximate human scoring under some circumstances, but these studies also reveal some discrepancies, especially when essays require subjective interpretation, in which case there is very little accuracy when compared to human graders.

4.1. Consistency and Correlation with Human Raters

Several recent studies have shown strong correlations between AES output and human grading, particularly when essays are scored according to structured rubrics. In a comparative study in 2025, it was found that closed-source LLMs like GPT-4 and o1 performed exceptionally well, achieving very high scores that indicate a high level of consistency with human graders. In particular, the o1 model managed to score a Spearman correlation of 0.742, indicating strong alignment in overall scoring [7].

However, these systems were most accurate in evaluating surface-level features, such as spelling, grammar, and basic organization, not deeper attributes like logic of argumentation, originality, and effectiveness of narrative endings. Similarly, another study conducted in 2024 found that hybrid deep learning models, such as their Collaborative Deep Learning Network (CDLN), scored 85.5% against human scorers, further supporting the idea that deep learning models can match human evaluators, but only in standard academic conventions [10].

4.2. When AI Falls Short: Depth, Nuance, and Interpretation

While AES systems have made signiGicant progress in consistency and language accuracy, they still have difGiculty with aspects that require human interpretation, such as higher-order thinking, argument development, and emotional depth. Even advanced systems like GPT-4 often misinterpret or undervalue essays that differ from standard formats or delve into complex personal experiences [9]. The research shows that when essays presented unconventional views, subtle irony, or Gigurative language, human evaluators were much more likely to notice and reward that sophistication than AI models did.

This struggle to fully grasp context or intent comes from how AES systems are trained. These models depend on statistical patterns and cannot "understand" like humans. Because of this, creative or culturally rich writing, which may strongly connect with a teacher, can receive lower scores or face penalties from AI for lacking structural conformity or being seen as ambiguous.

4.3. Score IN Elation and Gaming the System

Another main concern about current AES systems is the potential for them to be manipulated by students who learn what the AI "likes" and specifically craft their essays around these preferences. In recent studies, a correlation has been found between the amount of complex vocabulary and structured paragraphing in an essay and the score it receives, regardless of the overall quality of the essay [6]. This creates a situation in which students can potentially game the system by adhering to these preferences to artificially boost their essay scores despite writing a seemingly worse essay from the viewpoint of a human.

This study highlights a common phenomenon known as score inGlation, also commonly referred to as grade inGlation. This is where students write essays around criteria that the AES "likes" to get a higher score, taking advantage of the way the AES grades essays. In a lot of cases, AES systems graded essays as higher quality simply because they were longer, even while the writing lacked basic grammar, functionality, and structure. Such vulnerabilities risk encouraging a style-over-substance approaches to writing and undermining the pedagogical goal of developing formal writing skills.

4.4. Fairness, Bias, and Cultural Blind Spots

Even with all the recent advances in AES systems, bias remains one of the biggest ethical issues surrounding them. AES models are usually trained on data consisting of mostly formal, academic English writing often written by native speakers [2]. Thus, the AES scoring will naturally reGlect its training data, leading to students from different linguistic or cultural backgrounds being potentially penalized for writing that deviates from these norms, even when it is clear, correct, and compelling.

This issue is compounded when AI tools are deployed without mechanisms for detecting or correcting such biases. The study warns that certain dialects or narrative styles are systematically undervalued and likely to receive a lower score, especially in essays where personal experience intersects with cultural identity. This raises concerns about not only the fairness of scoring, but also about how AES systems might reinforce existing educational inequalities if used without human oversight.

4.5. The Importance of Human Oversight

Even though AES technologies have made encouraging advancements, both studies agreed that AES systems are yet to become ready to replace educators in evaluating student essays [2, 6]. While they can help in grading the lower-level aspects of essays, such as grammar, syntax, and structural coherence, they lack the more sophisticated thinking needed to evaluate complex aspects, such as argument development and emotional depth. Both accounts go further to conclude that the optimal grading remedy is likely to be a hybrid model in which AI does low-level and repetitive assessments, but without forfeiting the capability of educators to provide the last word, especially where high stakes are involved. This enables educators to leverage the scalability and cost-effectiveness of AI without sacriGicing the nuance and compassion of human beings. As educational institutions continue to explore AI-based writing grading systems, there is a necessity to treat AES models as tools for enhancement and not the substitution of human judgment.

5. Effectiveness of AI-Generated Feedback

One of the strongest promises of AES systems is not just their ability to assign a grade, but also to provide detailed feedback for students to improve their writing in subsequent assignments. In theory, AI can offer quick and detailed responses on speciGic aspects of an essay, but while these critiques can be a useful tool, their educational value is very uneven and very limited.

5.1. Strengths of AI Feedback

Research has found that AI has been highly effective at identifying lower-order concerns such as spelling, grammar, and sentence structure. GPT-4 can, for example, Glag repeated words, clumsy phrases, or formatting errors in a matter of seconds. This ability to deliver ultra-fast, targeted corrections can help to speed up the revision process for students who are looking for quick Gixes and minimize the turnaround time between different submissions.

AES tools also offer a scalable form of feedback in large classroom settings or high-stakes standardized testing scenarios. Educators can use AI-generated reports as a baseline to prioritize and focus their own time and attention on more complex issues, such as content quality or argumentation. In this hybrid approach, teachers have reported a signiGicant reduction in grading fatigue while still providing students with feedback on basic writing conventions.

5.2. Limitations in Depth and Nuance

Despite these advantages, AES feedback lacks the depth that human commentary provides. A 2024 study revealed that while AI can Glag surface-level mistakes with high precision, its comments on argument logic or thematic development were mostly generic, offering vague suggestions like "improve your conclusion" without explicitly explaining how [2]. This lack of depth limits the usefulness of AI feedback for helping students improve at writing. Whereas humans can adapt their feedback to a student's past performance and personalized learning style, AES tools operate without this level of personalization. As a result, feedback can feel impersonal and generic, reducing student engagement with the revision process.

5.3. Limitations in Depth and Nuance

Student trust in feedback also plays a role in its effectiveness. Research from a 2021 study revealed that many students were initially skeptical of AI comments, especially when they conGlicted with their own understanding of their own work with prior human feedback [6]. Some students even ignored some of these AI suggestions if they seemed irrelevant or overly mechanical. Conversely, students who saw AI feedback as fair and accurate were more likely to revise their own work in line with the suggestions. This highlights the importance of transparency: students need to understand why a particular suggestion was made to properly trust and apply it.

5.4. The Role of Feedback in Learning Outcomes

Ultimately, the goal of feedback is to improve student writing over time, not just to correct small errors, and current evidence suggests that AI alone is not enough for this purpose. While AES systems can provide fast and reliable corrections, the biggest improvements in student writing happen when AI feedback is combined with human support. This combination allows students to receive immediate corrections along with personalized advice that helps develop higher-order skills like argument development, audience awareness, and rhetorical strategy. In short, AES feedback is best seen as a Girst draft of commentary; it is helpful for tackling the mechanical aspects of writing but falls short without the insight, encouragement, and context that a human teacher offers.

6. Conclusion

In conclusion, artiGicial intelligence in education has created both opportunities and challenges. Automated Essay Scoring systems aim to tackle teacher shortages, lessen workloads, and improve grading efGiciency. Research indicates that modern AES models, especially those based on large language models like GPT-4 and o1, can match human graders in consistency when assessing basic writing elements such as grammar, structure, and spelling. They also offer quick, scalable feedback, making them especially useful in large classrooms or standardized testing settings. However, accuracy drops when essays require evaluation of creativity, nuance, or cultural context. AES systems often favor formulaic responses and struggle with unconventional or emotionally nuanced writing. While their feedback is fast and frequently technically correct, it tends to be too general to promote signiGicant long-term improvement in student writing. Moreover, issues of bias, fairness, and transparency raise important ethical concerns about using these tools in high-stakes academic situations.

Overall, these Gindings indicate that AES systems cannot fully replace human judgment in essay evaluation. Rather, their most effective role is in collaboration with teachers. AI can manage repetitive and simpler tasks while educators remain responsible for more complex, contextual grading. This combined approach offers a promising solution, merging the efficiency of automation with the unique insights of human educators.

As education continues to change in an AI-driven world, the challenge will be to balance innovation with fairness and trust. AES technologies are expected to become more advanced in the coming years, but until they can match the depth of human understanding, they should be used as tools to assist teachers rather than replace them.

References

[1] A. Chevalier, J. Orzech, and P. Stankov, "RETRACTED: Man vs Machine: Can AI Grade and Give Feedback Like a Human?," IZA - Institute of Labor Economics, 2024. [Online]. Available: http://www.jstor.org/stable/resrep69949

[2]. K. Jonäll, ArtiGicial intelligence in academic grading: A mixed-methods study, 2024.

[3] J. Flodén, "Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT," British Educational Research Journal, vol. 51, no. 1, 2024. [Online]. Available: https://doi.org/10.1002/berj.4069

[4] "Texas Education Agency using AI to grade parts of STAAR tests," AACRAO, 2024. [Online]. Available: https://www.aacrao.org/edge/emergent-news/texas-education-agency-using-ai-to-grade-parts-of-staar-tests

[5] S. J. H. Yang, "Guest Editorial: Precision Education - A New Challenge for AI in Education," Educational Technology & Society, vol. 24, no. 1, pp. 105–108, 2021. [Online]. Available: https://www.jstor.org/stable/26977860

[6] A. C. M. Yang, I. Y. L. Chen, B. Flanagan, and H. Ogata, "From Human Grading to Machine Grading: Automatic Diagnosis of e-Book Text Marking Skills in Precision Education," Educational Technology & Society, vol. 24, no. 1, pp. 164–175, 2021. [Online]. Available: https://www.jstor.org/stable/26977865

[7] K. Seßler, M. Fürstenberg, B. Bühler, and E. Kasneci, "Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring," in Proc. 15th Int. Learning Analytics & Knowledge Conf., pp. 462–472, 2025. [Online]. Available: https://doi.org/10.1145/3706468.3706527

[8] E. L. Wetzler et al., "Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation," Teaching of Psychology, vol. 52, no. 3, pp. 298–304, 2024 (original work published 2025). [Online]. Available: https://doi.org/10.1177/00986283241282696

[9] K. Bouziane and A. Bouziane, "AI versus human effectiveness in essay evaluation," Discover Education, vol. 3, no. 1, 2024. [Online]. Available: https://doi.org/10.1007/s44217-024-00320-6

[10] M. Maliha and V. Pramanik, "Hey AI Can You Grade My Essay?: Automatic Essay Grading," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2410.09319

[11] L. Loewus, "Why teachers leave—or don't: A look at the numbers," Education Week, May 4, 2021. [Online]. Available: https://www.edweek.org/teaching-learning/why-teachers-leave-or-dont-a-look-at-the-numbers/2021/05