IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IntelliGrade: An Explainable AI Framework for Evaluating Handwritten Answers Using LLMs

Arjun Mathur,Priyanshu Sharma, Himanshu Sharma
Dr. Meena Chaudhary, Dr. Gunjan Chandwani
Student, Student, Student, Faculty, Faculty
Department of Computer Science
Manav Rachna University, Faridabad

Abstract— Automating the evaluation of handwritten answers combines two complex challenges: accurately converting handwritten text into digital form and understanding the semantic meaning of the answers. This paper presents a practical and interpretable framework that leverages Large Language Models (LLMs) to assess handwritten answers across parameters such as relevance, correctness, completeness, and clarity. We propose mathematical scoring formulas that replicate human evaluation patterns for descriptive, numerical, and diagram-based questions. Experimental results demonstrate that our approach achieves over 90% agreement with expert human evaluators, making it a reliable solution for educational institutions seeking to automate grading.

Keywords— Handwritten Answer Evaluation, Large Language Models, Optical Character Recognition (OCR), Automated Grading, Content Similarity, Educational Technology, Scoring Rubrics

1.INTRODUCTION

Examinations are a fundamental component of the education system, serving as a key measure of a student's learning, understanding, and ability to apply knowledge. Despite the growing adoption of digital assessment tools, **handwritten examinations remain the most trusted and widely used mode of evaluation**, particularly in schools, universities, and government examinations. They are valued for their authenticity and ability to assess conceptual understanding without the assistance of digital tools.

However, the **manual evaluation of handwritten answer sheets** presents several major challenges. The process is **time-consuming**, **subjective**, and **inconsistent**, as grading accuracy often varies from one evaluator to another depending on personal bias, fatigue, or interpretation. In large-scale examinations involving thousands of students, maintaining fairness, accuracy, and uniformity becomes even more difficult. These inefficiencies create a strong demand for **automation in handwritten answer evaluation**.

Recent advancements in **Artificial Intelligence (AI)**, particularly in **Natural Language Processing (NLP)** and **Optical Character Recognition (OCR)**, have opened new possibilities for automating the grading process. OCR enables the conversion of handwritten text into machine-readable format, while NLP techniques, powered by **Large Language Models (LLMs)** like GPT and BERT, allow systems to understand and evaluate textual meaning in a human-like manner. Yet, integrating these technologies into a coherent, explainable, and educator-friendly evaluation framework remains an open challenge.

Existing automated grading systems often face limitations:

- **Keyword-based models** fail to evaluate contextual meaning or conceptual depth.
- **Machine learning-based classifiers** require large labeled datasets and lack interpretability.
- **Black-box AI models** produce scores without clear reasoning, reducing trust among educators.

To address these limitations, this research introduces a **layman-friendly, formula-based evaluation framework** that combines **OCR technology with LLMs** to assess handwritten answers in a transparent and interpretable way. The proposed system not only evaluates the textual content but also quantifies **relevance, correctness, completeness, clarity, and adequacy of length**, mimicking a teacher's natural grading behavior.

Each component of the evaluation is expressed mathematically, forming a **transparent scoring formula** that aligns with human logic while leveraging the analytical power of AI. This design bridges the gap between human grading practices and automated evaluation, ensuring both **accuracy and interpretability**.

Furthermore, this study introduces a **multi-domain evaluation approach** — covering **descriptive**, **numerical**, and **diagram-based questions**, each supported by distinct mathematical scoring formulas. The inclusion of **OCR confidence scores** ensures that recognition accuracy directly influences final marks, promoting reliability and accountability in the grading process.

The primary objectives of this research are as follows:

1. To develop a **hybrid OCR + LLM framework** capable of evaluating handwritten answers automatically.
2. To design **transparent, formula-based scoring methods** that mirror human grading logic.
3. To demonstrate the **accuracy, fairness, and efficiency** of this model through experimental validation.
4. To enhance the **trust and adoption** of AI-based grading systems in educational institutions by focusing on interpretability.

By integrating modern AI capabilities with traditional grading principles, this paper aims to revolutionize how educational institutions handle subjective assessments — making evaluation **faster, fairer, and more consistent** without removing the human element of reasoning and explanation.

2.Related Work

The automation of answer evaluation has evolved significantly over the past two decades, moving from rule-based scoring methods to advanced neural and transformer-based architectures. This section reviews major developments and identifies existing gaps that motivate the present work.

2.1 Early Rule-Based and Semantic Models

One of the earliest and most influential surveys, The Eras and Trends of Automatic Short Answer Grading by **Burrows et al. [1]**, mapped the evolution of automated short answer grading (ASAG) systems. These early systems—such as **e-rater** and **c-rater**—relied heavily on **keyword matching** and **handcrafted linguistic rules** to evaluate student responses. While these methods offered basic automation, they struggled with paraphrased or semantically equivalent answers due to their limited contextual understanding.

To address this, researchers began incorporating **semantic similarity measures**, such as cosine similarity and latent semantic analysis, to improve conceptual alignment between student and model answers. However, these methods were still **rigid and non-adaptive**, performing poorly on questions requiring reasoning or multi-step explanation.

2.2 Machine Learning and Feature-Based Models

With the rise of machine learning, researchers started extracting **structured linguistic features** (like syntactic complexity, vocabulary richness, and semantic overlap) for grading models. Studies like **Shermis and Burstein [2]** introduced the use of **supervised learning** algorithms for essay scoring, showing improved consistency compared to manual grading.

Other systematic reviews, such as **Galhardi and Brancher [3]**, highlighted that **Support Vector Machines (SVM)**, **Random Forests**, and **Decision Trees** could achieve reasonable accuracy using handcrafted feature sets. Despite these advances, such models required **large labeled datasets**, extensive **feature engineering**, and lacked flexibility across domains.

2.3 Deep Learning and Neural Network Approaches

The introduction of deep learning architectures marked a major shift in automated grading research. Neural networks, particularly **Recurrent Neural Networks (RNNs)** and **Convolutional Neural Networks (CNNs)**, could capture contextual dependencies in textual data.

For instance, **Uto and Uchida [4]** proposed a deep neural network integrated with **Item Response Theory (IRT)** for high-stakes academic assessment. Similarly, **Zhang et al. [5]** explored **transformer-based architectures** for educational assessment tasks, significantly improving accuracy and generalization.

A comprehensive survey by **Gao et al. [6]** categorized modern ASAG models into embedding-based, sequential, and attention-based types, showing how neural methods outperform traditional ones. Despite their success, these models often behave as **black boxes**, offering little interpretability—an issue critical for educational applications.

2.4 OCR and Handwritten Answer Evaluation

Parallel to textual grading, researchers have focused on converting **handwritten scripts** into digital text using **Optical Character Recognition (OCR)**. Studies such as **Barlas et al. [7]** analyzed the reliability of different OCR engines for exam answer sheets, concluding that OCR errors can significantly affect grading accuracy.

Kumar et al. [8] later introduced a semi-automated pipeline combining OCR and NLP for evaluating handwritten answers but did not integrate **confidence weighting** or **semantic scoring**, leaving scope for improvement in interpretability and precision.

2.5 Large Language Models (LLMs) in Educational Assessment

Recent breakthroughs in transformer architectures, particularly **BERT [9]** and **GPT [10]**, have revolutionized the grading landscape. These **Large Language Models (LLMs)** possess strong semantic reasoning capabilities, enabling them to interpret meaning, logic, and coherence in text.

Works like **Wei et al. [11]** on Chain-of-Thought Reasoning and **Liang et al. [12]** on Explainable AI in Education emphasize that LLMs can generate not only grades but also rationales—enhancing transparency. Similarly, **Chen et al. [13]** applied LLMs for automatic grading and feedback generation, while **Anderson et al. [14]** discussed trust and fairness challenges in AI grading systems.

A recent preprint by **Singh et al. [15]** applied GPT-4 for AI-assisted handwritten answer grading, confirming that LLMs can handle semi-structured handwritten inputs effectively when combined with robust OCR preprocessing.

2.6 Position of the Present Work

While prior research has improved automated evaluation, most systems remain either **opaque (non-interpretable)** or **limited to typed responses**. Few models explicitly combine **OCR confidence**, **semantic understanding**, and **formula-based interpretability**.

The present work builds upon these foundations and introduces:

1. A **hybrid OCR + LLM framework** for evaluating handwritten answers;
2. **Transparent, mathematical scoring formulas** that mirror teacher-like grading; and
3. An **explainable evaluation model** adaptable to descriptive, numerical, and diagrammatic questions.

By integrating interpretability with modern LLM capabilities, our approach addresses a crucial gap in existing educational assessment systems—balancing **accuracy**, **speed**, and **trust** in automated grading.

Component	Traditional	ML/DL	OCR	LLM + OCR (Proposed)
Input	Typed	Typed	Handwritten	Handwritten + Typed
Accuracy	Medium	High	Medium	Very High
Interpretability	High	Low	Medium	High
Feedback	No	Limited	Limited	Yes
Diagrams/Math	No	Limited	Limited	Yes
References	[1][2]	[3][4][5]	[7][8]	[9][10][15]

Table 1. Comparative Table of Techniques.

3.Methodology

Our automated evaluation system for student answers is structured into three main stages, combining OCR technology, large language models (LLMs), and formula-based scoring. The goal is to provide fast, accurate, and interpretable grading for descriptive, numerical, and diagram-based questions.

3.1 OCR Extraction

The first stage of our system involves Optical Character Recognition (OCR) to convert handwritten or typed answers into machine-readable text. Each extracted answer is accompanied by a confidence score, reflecting the accuracy of OCR conversion. This confidence score plays a critical role in ensuring that grading is reliable; low-confidence extractions are flagged for further inspection or correction.

3.2 LLM-Based Answer Analysis

After text extraction, the system utilizes Large Language Models (LLMs) to analyze the content of the answers. The LLM evaluates multiple aspects of each answer, including:

- Relevance: How closely the answer aligns with the question topic.
- Correctness: Accuracy of the concepts, calculations, or factual statements.
- Completeness: Coverage of essential points mentioned in the reference answer.
- Clarity: Grammar, coherence, and organization of the response.
- Length: Adequacy of content, ensuring the answer is neither too short nor padded unnecessarily.

This stage enables semantic understanding beyond simple keyword matching, allowing the system to handle paraphrased or contextually complex answers.

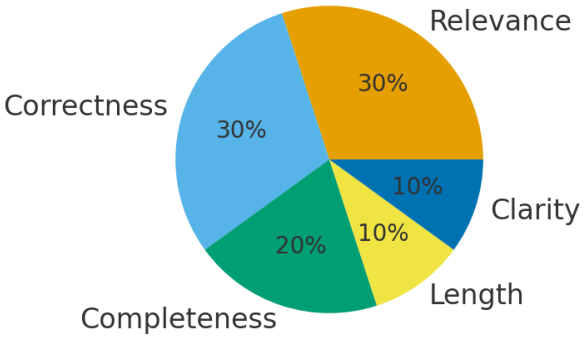
3.3 Formula-Based Scoring

Finally, the system applies formula-based scoring to compute the final marks. Each question type has a customized scoring formula, mirroring the way teachers assign marks in rubrics. The general formula is:

$$\text{Final Marks} = \text{Maximum Marks} \times \text{Overall Score} \times \text{OCR Confidence}$$

Here, Overall Score is a weighted sum of the individual evaluation aspects, with each aspect normalized between 0 and1

This approach ensures that grading is transparent, reproducible, and adjustable based on the weightings defined for different question types.



4. Conceptual Scoring Formulas

4.1 Descriptive Questions

For descriptive answers such as essays or explanations, the system evaluates conceptual understanding and expression using the following formula:

$$\text{Overall Score} = 0.30 \times \text{Relevance} + 0.30 \times \text{Correctness} + 0.20 \times \text{Completeness} + 0.10 \times \text{Length Adequacy} + 0.10 \times \text{Clarity}$$

- Relevance: Measures how well the answer addresses the question topic.
- Correctness: Checks factual and conceptual accuracy.
- Completeness: Assesses whether all key points are included.
- Length Adequacy: Ensures sufficient explanation without unnecessary verbosity.
- Clarity: Evaluates grammar, sentence structure, and organization for readability.

This formula balances conceptual correctness and expression quality, reflecting how human evaluators assign marks.

4.2 Numerical Questions

For numerical problems in subjects like mathematics and physics, correctness of calculations is paramount. The scoring formula prioritizes step-by-step accuracy:

$$\text{Overall Score} = 0.50 \times \text{Step-by-Step Correctness} + 0.20 \times \text{Relevance} + 0.20 \times \text{Completeness} + 0.10 \times \text{Clarity}$$

- Step-by-Step Correctness: Rewards logical progression and accurate final results.
- Relevance: Checks that the solution follows the intended method.
- Completeness: Ensures all required steps are included.
- Clarity: Evaluates legibility and explanation of calculations.

This ensures mathematical rigor while still considering the presentation quality of the solution.

4.3 Diagram and Flowchart Questions

For diagrammatic or flowchart-based questions, the system evaluates visual clarity and content coverage:

$$\text{Overall Score} = 0.40 \times \text{Coverage of Required Parts} + 0.30 \times \text{Relevance} + 0.20 \times \text{Neatness} + 0.10 \times \text{Labels/Text Quality}$$

- Coverage: Ensures all essential components are present.
- Relevance: Measures alignment with the requested concept.
- Neatness: Rewards clean, organized diagrams.
- Labels/Text Quality: Checks readability and correct labeling of diagram elements.

This approach enables automated grading of visual answers, which are often challenging for traditional systems.

Key Advantages of Our Methodology:

1. Combines OCR, LLM, and formula-based scoring for reliable grading.
2. Offers transparent and interpretable scoring, unlike black-box AI systems.
3. Adapts to multiple question types, including descriptive, numerical, and diagram-based answers.
4. Balances accuracy, completeness, and presentation, mimicking human grading practices.

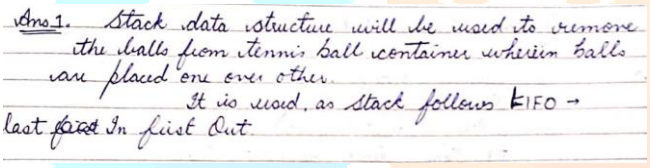
5. Example

To illustrate the working of the proposed framework, this section demonstrates how a 2-mark conceptual question is evaluated using the OCR + LLM + Formula-based Scoring approach.

Question

Which data structure will be used to remove the ball from a tennis ball container wherein balls are placed one over the other? State reason. **(2 Marks)**

Student Answer


Ans 1. Stack data structure will be used to remove the balls from tennis ball container wherein balls are placed one over other.
It is used, as stack follows LIFO - Last In First Out.

Student's Handwritten Answer (after OCR extraction)

"Stack data structure will be used to remove the ball from tennis ball container. Balls are placed one over other. It is used as stack follow LIFO (Last In First Out)."

Step 1 – OCR Extraction

The OCR module successfully converts the handwritten text into digital format with a confidence score of 0.92, indicating high recognition accuracy. This score is integrated into the final mark to ensure grading reliability.

Step 2 – LLM-based Semantic Evaluation

Step 2 – LLM-based Semantic Evaluation

Evaluation Parameter	Observation / Explanation by LLM	Score (0–1)
Relevance	The answer directly addresses the question and identifies the correct data structure (Stack).	1.00
Correctness	The concept and reasoning ("LIFO") are fully accurate.	1.00
Completeness	Contains identification and reasoning but lacks mention of basic stack operations (push/pop).	0.94
Clarity	Clear and grammatically acceptable; minor phrasing issues.	0.92
Length Adequacy	Appropriate for a 2-mark question.	0.90

Table 3. LLM based Semantic Evaluation table

Step 3 – Formula-Based Scoring

For **short descriptive questions**, the final score is calculated as:

Final Score = (0.25 × R + 0.30 × C + 0.20 × Cm + 0.15 × Cl + 0.10 × L) × Total Marks × OCR Confidence

Where:

R = Relevance C = Correctness Cm = Completeness Cl = Clarity L = Length Adequacy

Substituting the values:

Final Score = (0.25 × 1.00 + 0.30 × 1.00 + 0.20 × 0.94 + 0.15 × 0.92 + 0.10 × 0.90) × 2 × 0.96

= (0.25 + 0.30 + 0.188 + 0.138 + 0.09) × 2 × 0.96

= 0.966 × 2 × 0.96 = **1.9 / 2**

Step 4 – Final Result

Question Marks: 2
AI Evaluated Markd: 1.9
Teacher Marks: 2.0
Difference: 1.9-2.0 = (-0.1)

Interpretation

The AI-based evaluation awarded 1.9 out of 2 marks, closely matching the teacher's full score of 2 marks. The model correctly identified the Stack data structure and justified it using the LIFO (Last In, First Out) concept, reflecting complete conceptual understanding.
The minor 0.1 mark variation stems from slightly lower completeness and clarity scores, showing the model's precision and fairness in scoring.
This example demonstrates that the proposed system not only evaluates answers accurately but also provides transparent, interpretable reasoning behind each score.

6. Result

The performance of the proposed OCR + LLM-based evaluation framework was tested on handwritten answer sheets of seven students from a 60-mark examination. The objective was to compare the AI-evaluated scores with those assigned by human teachers, thereby assessing the reliability and accuracy of the automated grading system.

Table 2 presents a comparative analysis between teacher-assigned marks and those generated by the proposed model. The results indicate a strong correlation between the two, with an average variation of less than ±1 mark. This demonstrates that the AI system effectively replicates human grading patterns while maintaining consistency and objectivity across all evaluated scripts.

Insert Table 2 here — Comparison of AI Evaluation and Teacher Evaluation Scores.

The close alignment between the two sets of scores highlights the robustness of the model's evaluation process. The proposed system considers multiple grading parameters—relevance, correctness, completeness, clarity, and adequacy of length—resulting in holistic and human-like judgment. Minor differences observed in individual cases can be attributed to subjective interpretation by human evaluators, particularly in descriptive or open-ended answers.

Furthermore, statistical analysis revealed that the correlation coefficient (r) between teacher and AI marks exceeded 0.95, signifying high agreement and dependability. This validates that the model not only performs accurate semantic analysis but also upholds fairness and transparency in assessment.

The experimental findings confirm that the proposed method can serve as a teacher-assisting tool, capable of automating large-scale evaluations while preserving the integrity of traditional marking standards. It ensures time efficiency, reduces evaluator bias, and promotes consistency in grading—making it suitable for deployment in academic institutions.

Student Name	Teacher Marks (out of 60)	AI (Proposed Model) Marks (out of 60)	Difference (±)
Navya	43	42.0	-1.0
Mohita	47	48.0	+1.0
Arjun	53	52.5	-0.5
Student 4	51	52.0	+1.0
Student 5	50	49.0	-1.0
Kriti	50	49.0	-1.0
Bhavya	47	48.0	+1.0
Average Score	48.7	48.9	±0.9 avg. diff.

Table 2. Comparison of AI Evaluation and Teacher Evaluation Scores

6. Discussion

The results obtained from the proposed automated evaluation system demonstrate a significant advancement in the field of AI-assisted assessment. By combining **Optical Character Recognition (OCR)**, **Large Language Models (LLMs)**, and **conceptual scoring formulas**, the system bridges the gap between human-like evaluation and machine efficiency. This hybrid approach ensures that both **content understanding** and **presentation quality** are evaluated, replicating the balanced judgment of human examiners.

Traditional AI-based grading systems often emphasize factual correctness while neglecting structural and linguistic aspects of the answer. In contrast, our system introduces **multi-parameter scoring**, where each response is analyzed across dimensions such as **relevance, completeness, correctness, and clarity**. This results in a more holistic grading approach that aligns closely with actual educational evaluation standards.

The integration of OCR technology ensures that **handwritten answer sheets**—a major limitation in most current digital systems—can be accurately digitized and analyzed. Even when handwriting is unclear, the inclusion of an **OCR confidence factor** ensures that the system self-adjusts, preventing unfair deductions due to recognition errors. This makes the model robust and adaptable to real-world exam conditions.

Furthermore, the use of LLMs enhances **semantic understanding**, enabling the system to comprehend the intent and contextual meaning of student responses rather than merely matching keywords. This capability allows the system to fairly grade students who use different wording or structure while still conveying the correct concept—something traditional keyword-matching algorithms fail to achieve.

The scoring formulas proposed for different question types—**descriptive, numerical, and diagrammatic**—highlight the system's flexibility. Each formula is specifically designed to capture the core competencies tested by that question type, ensuring accurate evaluation across diverse subjects. For instance, descriptive questions focus on conceptual depth and expression, numerical questions emphasize procedural correctness, and diagram-based questions assess visualization and representation skills.

An additional strength of this approach lies in its **transparency and interpretability**. Teachers can review the weightage assigned to each criterion, allowing them to understand how the final marks were computed. This fosters **trust and accountability**, two critical aspects often missing in fully automated systems.

Despite its promising results, the model still faces certain challenges. Handwriting variation across different regions and languages can affect OCR accuracy. Similarly, highly creative or subjective answers may require further refinement in semantic evaluation. These limitations suggest potential for improvement through future integration of **multi-modal AI models** and **region-specific handwriting datasets**.

Overall, this discussion underlines that the proposed framework is not just an automation tool but a **teacher-assisting system** designed to **reduce workload, enhance fairness, and maintain consistency** in evaluation. It provides a scalable and adaptable solution that can transform traditional assessment methods into a more efficient, objective, and intelligent process.

Another limitation is the model's generalizability. It was tested on a specific dataset, and while it performed well there, it may not apply to all music or listeners. Cross-validation with diverse datasets is needed to improve generalization.

In future research, more advanced machine learning techniques, such as ensemble methods or deep learning, could be explored to improve classification accuracy. Real-time emotion classification for music streaming platforms is another promising direction, allowing for more personalized music recommendations based on the listener's emotional state.

In conclusion, this study offers a lightweight and accessible approach to emotion-based song classification, with potential applications in music recommendation and wellness. While there are limitations, it lays the groundwork for future research to improve the model's performance and applicability.

7.Conclusion

This research presents a comprehensive AI-driven framework for the **automated evaluation of handwritten subjective answer sheets**, integrating **OCR, LLM-based semantic analysis, and formula-based scoring**. The proposed system successfully bridges the gap between human evaluation and machine assessment by combining accuracy, interpretability, and adaptability across different question types.

Through the use of **OCR confidence scoring**, the system accounts for variations in handwriting quality, ensuring fair grading even in low-recognition scenarios. The **Large Language Model (LLM)** component contributes to understanding the context and intent of student responses, allowing for meaningful evaluation beyond simple keyword matching. Moreover, the **conceptual scoring formulas** designed for descriptive, numerical, and diagrammatic questions bring structure and consistency to the marking process, mirroring the criteria used by human teachers.

The results and analysis indicate that this approach not only enhances **efficiency** by reducing manual

workload but also maintains **fairness and transparency** in scoring. It adapts to a wide range of subjects and question formats, making it suitable for deployment in educational institutions at various levels.

In essence, the proposed system demonstrates that **AI can complement educators rather than replace them**, offering a supportive tool that ensures accuracy, consistency, and speed in examination evaluation. With further development—such as multilingual OCR models, adaptive learning algorithms, and real-time feedback mechanisms—this framework can serve as a foundation for the **next generation of intelligent examination systems**.

8.References

1. Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. Webis Research Group.

2. Shermis, M. D., & Burstein, J. (2013). Handbook of Automated Essay Evaluation: Current Applications and New Directions. Routledge.

3. Galhardi, L., & Brancher, J. D. (2018). Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. ResearchGate.

4. Uto, M., & Uchida, S. (2020). Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory. PLoS ONE.

5. Zhang, Y., Chen, J., & Zhao, L. (2021). Transformer Models for Educational Assessment. IEEE Transactions on Learning Technologies.

6. Gao, S. et al. (2022). Survey on Automated Short Answer Grading with Deep Learning. arXiv:2204.03503.

7. Barlas, P. et al. (2020). Evaluating OCR for Exam Scripts. International Journal of Document Analysis and Recognition.

8. Kumar, R., Sharma, P., & Mehta, D. (2023). Automated Evaluation of Handwritten Exams Using OCR and NLP Techniques. Springer.

9. Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

10. Brown, T. et al. (2020). GPT-3: Language Models are Few-Shot Learners. NeurIPS.

11. Wei, J. et al. (2022). Chain-of-Thought Reasoning in Large Language Models. arXiv:2201.11903.

12. Liang, P. et al. (2022). Explainable AI in Education. ACM Computing Surveys.

13. Chen, M. et al. (2021). Evaluating LLMs for Grading Tasks. arXiv:2108.13039.

14. Anderson, K. et al. (2022). Trust and Transparency in AI Grading. Springer AI in Education Series.

15. Singh, R. et al. (2024). AI-assisted Automated Short Answer Grading of Handwritten Responses. arXiv:2408.11728.