How far are we from automatic grading of handwritten cloze form questions?

No Author Given

No Institute Given

Abstract. AI-based grading is increasingly used in education, yet its effectiveness in evaluating handwritten responses for cloze-form questions remains underexplored. This study investigates the reliability of state-of-the-art (SOTA) AI models in assessing 3,000 handwritten responses from 500 cloze-form questions, collected from students with diverse writing styles. We systematically group responses based on handwriting characteristics and compare AI-based grading with human evaluation. Our findings reveal a 49% average error rate for the best-performing AI model, with a 6.7% discrepancy between AI and human assessments. Additionally, AI performance varies significantly across handwriting groups, indicating potential biases in automated grading systems. These results highlight the challenges of AI-based assessment in real-world educational settings and the need for robust, fairness-aware grading models to ensure equitable evaluations.

Keywords: Automated assessment· Cloze-form questions

1 Introduction

Assessing handwritten responses in educational settings requires significant human effort, especially in large-scale exams. AI-based grading systems have gained traction as a potential solution, particularly for standardized assessments where students provide written responses [1,15,12]. While handwriting has been linked to cognitive processes and academic performance, it remains a persistent challenge for automated assessment [16,5,3,14]. The problem is more pronounced in cloze-form questions, where answers are short and lack contextual information, making accurate evaluation difficult.

Despite advancements in AI-based grading, little attention has been given to handwriting diversity and AI biases in evaluating cloze-form responses. Existing AI-driven assessment systems often overlook variations in handwriting styles, leading to systematic grading inconsistencies. These challenges are particularly concerning in regions with limited educational resources, where AI-based assessment tools could provide scalable solutions [8,6]. However, biases in AI models could reinforce inequities rather than mitigate them.

In this study, we evaluate the robustness of AI-based grading on a dataset of 500 cloze-form questions and 3,000 handwritten responses, collected from students with diverse handwriting styles. Our key contributions include:

- Comprehensive evaluation of SOTA AI models on handwritten cloze-form responses, analyzing their error rates and biases across different handwriting groups.
- Comparison of AI-based and human-based grading, highlighting key discrepancies and challenges in automated assessment fairness.

Our findings reveal that even the best-performing AI model exhibits a 49% error rate, with notable inconsistencies across handwriting styles. This underscores the need for more adaptive, fairness-aware AI models to ensure equitable grading in educational settings.

2 Related Work

AI-Based Handwriting Recognition in Education: Handwriting recognition plays a key role in AI-driven assessment systems. Optical Character Recognition (OCR) models, such as TrOCR[9] and Qwen-VL[17,4], leverage deep learning to improve handwritten text recognition. These models incorporate language modeling techniques to reconstruct difficult-to-read handwriting. However, their effectiveness varies based on handwriting quality, leading to errors in AI-based grading[20,15]. Despite advancements, research on AI susceptibility to handwriting variations remains limited[11,12].

AI for Cloze-Form Question Assessment: AI has been widely applied to automatically generate and evaluate cloze-form questions [10,13,19]. A notable system, Answers in Mind (AiM), evaluates Chinese handwritten responses but relies on examiner-provided correct answers [20]. More importantly, AiM does not examine grading bias due to handwriting differences, which is a focus of our study.

Fairness and Bias in AI-Based Grading: Fairness in AI-based grading is critical, as models can exhibit biases based on linguistic and visual factors [7]. However, the impact of handwriting diversity on grading accuracy remains underexplored. Our work addresses this gap by analyzing how AI models struggle with specific handwriting styles, raising concerns about bias and equity in automated grading.

3 Experimental Setup

To systematically evaluate AI-based grading in real-world educational settings, we designed an experimental framework to assess AI models' performance on handwritten responses to cloze-form questions. Our goal is to analyze how handwriting diversity affects grading accuracy and to identify AI limitations in recognizing handwritten responses. Below, we describe our dataset, response categorization, and the AI models used for evaluation.

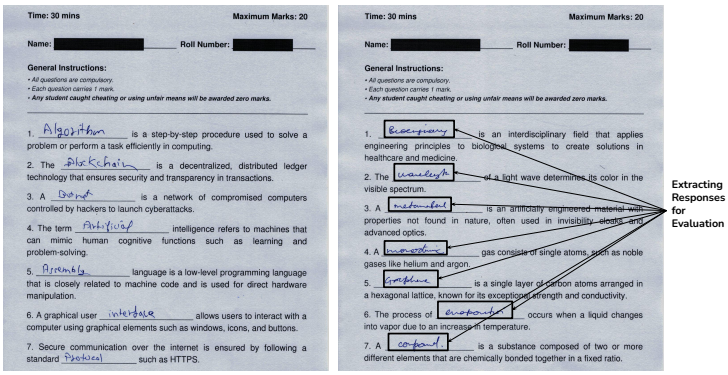

Fig. 1: Samples from the prototype implementation of exams to be evaluated by the AI Systems. Right: sample with the location of responses identified by AI.

Fig. 2: Binarized version of different handwriting samples.

3.1 Prototype Exam for AI-Based Grading

We constructed a prototype exam consisting of 500 cloze-form questions with 3,000 handwritten responses. The questions are based on science and technology topics, requiring students to provide single-word answers in designated spaces. The response sheets were scanned at 600 dpi to ensure high-quality OCR recognition and preserve handwriting details. Fig. 1 illustrates the structured exam format, while Fig. 2 highlights handwriting variations across responses. By using a structured exam format, we ensure controlled evaluation conditions while maintaining handwriting variability, making this dataset well-suited for assessing AI-based grading systems.

3.2 Dataset Construction and Handwriting Categorization

To analyze AI susceptibility to handwriting diversity, we grouped responses into six handwriting categories based on key features such as legibility, stroke vari-

ations, and letter distortions [7]. Since AI grading accuracy should not be influenced by student mistakes, we carefully curated a subset of correct responses (Fig. 2), ensuring that AI errors are purely due to handwriting recognition challenges. The dataset includes responses from students with diverse educational backgrounds, regional writing styles, and gender balance. This diversity ensures that our findings are representative of real-world classroom scenarios and highlight potential fairness concerns in AI-based grading.

3.3 AI Models for Handwriting Recognition

We evaluated four state-of-the-art AI models for handwritten response recognition, selected based on their architectures and relevance to OCR-based grading:

- **Qwen2-VL-7B** [17,4] – A vision-language model optimized for OCR tasks, leveraging contextual reasoning for better text recognition.
- **Deepseek-vl2-small** [18] – A lightweight OCR model, chosen to assess performance trade-offs between model size and accuracy.
- **LLaMA-3.2-11B-Vision** [2] – A multi-modal model capable of processing both text and images, useful for complex handwriting patterns.
- **TrOCR-Large-Handwritten** [9] – A dedicated OCR model fine-tuned for handwriting recognition, serving as a strong baseline.

These models were selected to provide a comprehensive performance evaluation across general-purpose and handwriting-specific AI models, helping us analyze how different architectures handle handwriting recognition challenges.

4 Evaluation of AI-Based Grading

We evaluate the reliability of AI-based grading for handwritten responses, analyzing recognition accuracy, bias across handwriting groups, and comparison with human evaluators. Our results reveal that current AI models struggle with handwriting diversity, leading to significant grading inconsistencies.

4.1 Overall AI Model Performance

Table 1 presents the precision, recall, F1-score, and error rate for each model. Qwen2-VL-7B [17,4] achieves the highest F1-score of 51.60 with an error rate of 49%, outperforming other models. However, despite being the most effective, this model still exhibits substantial grading errors. The consistently high error rates across all models indicate that handwriting recognition remains a major challenge in AI-driven assessments.

These results highlight the significant limitations of current AI models in handling handwritten responses. AI grading performance is particularly affected by handwriting diversity, which we analyze in the next section.

AI Model	Precision ↑	Recall ↑	F1-score ↑	Error Rate (%) ↓
Qwen2-VL-7B	**51.39**	**51.81**	**51.60**	**49.00**
Deepseek-vl2-small	42.70	45.37	44.00	60.72
LLaMA-3.2-11B-Vision	29.87	30.36	30.11	70.29
TrOCR-Large-Handwritten	14.47	26.07	18.61	82.18

Table 1: Performance of AI models on handwritten responses.

Handwriting Group	Precision ↑	Recall ↑	F1-score ↑	Error Rate (%) ↓
Group 1 (Least Readable)	19.60	19.80	19.70	82.00
Group 2	44.11	44.55	44.33	57.00
Group 3	55.44	55.44	55.44	45.00
Group 4	57.42	57.42	57.42	43.00
Group 5	63.46	65.34	64.39	38.00
Group 6 (Most Readable)	**68.31**	**68.31**	**68.31**	**32.00**

Table 2: Performance of Qwen2-VL-7B across different handwriting groups.

4.2 Impact of Handwriting Diversity on Grading Accuracy

To assess how handwriting style impacts AI-based grading, we categorized responses into six groups based on readability and stroke complexity [7]. Table 2 presents the recognition performance of Qwen2-VL-7B across these groups. The model's error rate ranges from 32% for the most readable handwriting to 82% for highly stylized or cursive handwriting. These findings indicate a strong bias in AI grading, where students with difficult-to-read handwriting are disproportionately penalized, raising concerns about fairness in educational assessments.

The discrepancy in error rates suggests that AI models favor standardized, structured handwriting while struggling with more stylized or unconventional writing. This bias poses a risk of unfair assessment, especially in high-stakes testing environments where grading accuracy is critical.

4.3 Comparison with Human Evaluation

To further contextualize AI-based grading performance, we compare it with human assessments. Fig. 3 illustrates the error rates for both AI models and human evaluators. Overall, human graders exhibit a lower error rate of 42.24%, while the best AI model (Qwen2-VL-7B) reaches 49% (see rightmost bars in Fig. 3). The discrepancy is even more pronounced for handwriting groups with high structural variability, where AI models frequently misinterpret characters or fail to recognize words entirely. We conduct a Wilcoxon Signed-Rank test to compare human assessments with Qwen2-VL-7B. The null hypothesis, stating no performance difference, is rejected with a significance of 3.8% across six handwriting groups and supports our claim that humans are better evaluators than AI.

This performance gap highlights a critical limitation of AI-based grading while AI relies on character-level recognition, human evaluators leverage contextual understanding and adapt better to stroke variations, faded writing, and partial occlusions. These findings emphasize the importance of human oversight in AI-driven assessments, especially for handwritten responses with structural variability.

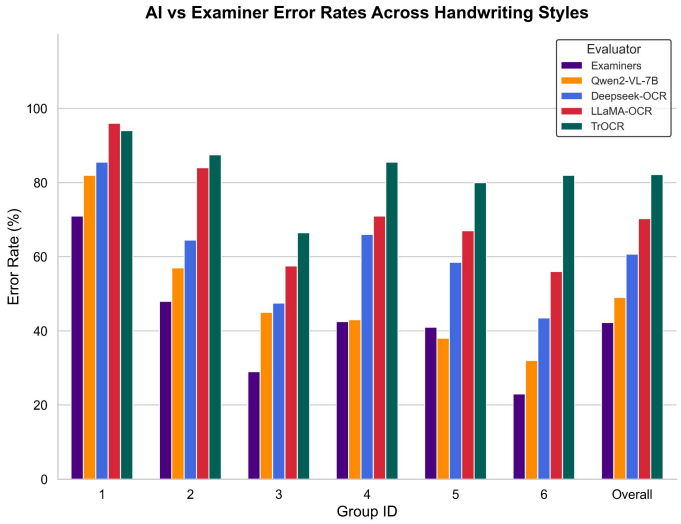

Fig. 3: Comparison of AI models and human examiners in terms of error rates across diverse handwriting groups.

4.4 Challenges in AI-Based Handwriting Grading

The evaluation results reveal multiple challenges in AI-based grading of handwritten responses. First, AI models demonstrate strong biases towards certain handwriting styles, which could disproportionately impact students with less structured or cursive handwriting. Second, despite improvements in OCR and vision-language models, AI still struggles with fine-grained text recognition, leading to frequent misclassifications. Finally, the large discrepancy between AI and human performance suggests that fully automated grading is not yet feasible for handwritten cloze-form questions.

These findings indicate that while AI-based assessment has potential for large-scale grading, it requires substantial improvements in robustness, bias mitigation, and contextual reasoning. In the next section, we discuss future directions to address these limitations.

5 Conclusion and Future Work

This study evaluated AI-based grading for handwritten cloze-form responses, revealing significant challenges in handling handwriting diversity. While Qwen2-VL-7B achieved the best performance (F1-score: 51.60), it still exhibited a high 49% error rate, higher than human graders (42.24%). AI models showed systematic biases, struggling with stylized or cursive handwriting, raising fairness concerns in educational assessments. Future work should focus on handwriting-aware AI models, context-aware post-processing, and hybrid AI-human grading to improve robustness. Ensuring fair, reliable, and unbiased AI grading is crucial for its effective deployment in education.

References

1. Deepgrade. https://smartail.ai/deepgrade/
2. AI, M.: Llama-3.2-11b-vision. Hugging Face Model Card (2024), https://huggingface.co/meta-llama/Llama-3.2-11B-Vision
3. Asci, F., Scardapane, S., Zampogna, A., D'Onofrio, V., Testa, L., Patera, M., Falletti, M., Marsili, L., Suppa, A.: Handwriting declines with human aging: A machine learning study. Frontiers in Aging Neuroscience **14**, 889930 (2022). https://doi.org/10.3389/fnagi.2022.889930
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Briggs, D.: A study of the influence of handwriting upon grades using examination scripts. Educational Review **32**(2), 186–193 (1980)
6. Hammond, D., Williams, K.: Addressing the challenges of rural students in education. Edutopia (2023), https://www.edutopia.org/article/addressing-challenges-rural-students/
7. Kingrani, S.K., Levene, M., Zhang, D.: Estimating the number of clusters using diversity. Artif. Intell. Res. **7**, 15– (2017), https://api.semanticscholar.org/CorpusID:2448156
8. Kraus, M.W., Stephens, N.M.: Socioeconomic status and education: A complex relationship. Annual Review of Psychology **64**, 1–22 (2012), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3902127/
9. Li, M., Lv, T., Cui, L., Lu, Y., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. ArXiv **abs/2109.10282** (2021), https://arxiv.org/abs/2109.10282
10. Matsumori, S., Okuoka, K., Shibata, R., Inoue, M., Fukuchi, Y., Imai, M.: Mask and cloze: Automatic open cloze question generation using a masked language model. IEEE Access **11**, 9835–9850 (2022), https://api.semanticscholar.org/CorpusID:248810716
11. Mondal, A., Mahadevan, V., Manmatha, R., Jawahar, C.V.: Icdar 2024 competition on recognition and vqa on handwritten documents. In: IEEE International Conference on Document Analysis and Recognition (2024), https://api.semanticscholar.org/CorpusID:272694571
12. Nguyen, H.T., Nguyen, C.T., Oka, H., Ishioka, T., Nakagawa, M.: Handwriting recognition and automatic scoring for descriptive answers in japanese language tests. In: International Conference on Frontiers in Handwriting Recognition (2022), https://api.semanticscholar.org/CorpusID:254127237
13. Pino, J., Heilman, M., Eskenazi, M.: A selection strategy to improve cloze question quality. In: Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada. pp. 22–32 (2008)
14. Ruusuvirta, T., Sievänen, H., Vehkamäki, M.: Negative findings of the handwriting legibility effect: the explanatory role of spontaneous task-specific debiasing. SN Social Sciences **1** (2021), https://api.semanticscholar.org/CorpusID:237753639
15. Sil, P., Chaudhuri, P., Raman, B.: Can ai assistance aid in the grading of handwritten answer sheets? In: International Conference on Artificial Intelligence in Education. pp. 291–298. Springer (2024)
16. Sishwashwa, K., Mwanza, D.S.: Factors affecting the quality of handwriting among grade five learners in selected public primary schools of mongu district, zambia.

International Journal of Research and Innovation in Social Science **7**(6), 283–291 (2023), https://www.researchgate.net/publication/382463167_Factors_Affecting_the_Quality_of_Handwriting_among_Grade_Five_Learners_in_Selected_Public_Primary_Schools_of_Mongu_District_Zambia

17. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
18. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., Ruan, C.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding (2024), https://arxiv.org/abs/2412.10302
19. Yang, A.C., Chen, I.Y., Flanagan, B., Ogata, H.: Automatic generation of cloze items for repeated testing to improve reading comprehension. Educational Technology & Society **24**(3), 147–158 (2021)
20. Zhang, Y., Li, Z., Zhou, Q., Liu, Z., Li, C., Ma, M., Cao, Y., Liu, H.: Aim: Taking answers in mind to correct chinese cloze tests in educational applications. arXiv preprint arXiv:2208.12505 (2022)