Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Error Tolerance in Automatic Short Answer Grading with Large Language Models: The Case of Handwriting Recognition Errors

Ziqi TAN[a], Yingbin ZHANG[b*] & Su MU[b]
[a]*School of Information Technology in Education, South China Normal University, China*
[b]*Institute of Artificial Intelligence in Education, South China Normal University, China*
*zyingbin@m.scnu.edu.cn

Abstract: Assessing students' answers is a labor-intensive task that could significantly burden educators. Technological advances have led to the development of automatic scoring systems. However, handwriting recognition errors significantly impacts the deployment of automatic scoring in the context of paper-pencil assessment, which is still the most widely used format of academic exams. Large language models (LLMs) have proven to be effective in scoring tasks and are promising for automatic short answer grading. Nevertheless, LLMs' capability to grade the short answers with handwriting recognition errors is underexplored. The current study addressed this issue in the context of paper-pencil Chinese tests in elementary schools. The LLM used was ERNIE 4.0 because of its outstanding capability in Chinese language comprehension. We compared the grading accuracy of the model on the raw data extracted from handwritten answers by optical character recognition with that on the preprocessed data where the recognition errors were corrected. We found a substantial accuracy difference between raw data and preprocessed data, indicating that LLMs have not yet achieved the precision for grading short answers with handwriting recognition errors. Nevertheless, LLMs showed an interesting characteristic during grading: awarding points to incorrect answers as an acknowledgment of student's effort.

Keywords: Large Language Models, paper-pencil assessment, automatic short answer grading, handwritten text recognition

1. Introduction

Despite the increasing adoption of computer-based assessment, the paper-pencil assessment is still the most widely used format of academic exams. In this context, grading students' answers, particularly for large-scale assessment, is a laborious task. Additionally, grading a large number of students' handwritten answers is prone to errors and inefficiency (Nagarajan & Jayasurya, 2021). Implementing machine grading can alleviate the burden on graders and enhance the accuracy and consistency of the grading process. Multiple-choice and short answer questions are two of the most popular question formats. The former can be easily scored using computational methods due to their fixed answers, while automatic short answer grading (ASAG) is much more challenging due to the complexities involved in understanding natural language (Burrows et al., 2015). In addition, errors in handwritten text recognition (HTR) are unavoidable, adding another challenge to ASAG in paper-pencil assessment.

The advent of large language models (LLMs), such as GPT-3.5 and Llama-2, has created new opportunities for ASAG because of their advanced capabilities in language comprehension. Generally, the scoring rubrics for short answer questions focus on the semantics of the answer rather than the syntax (Chang & Ginter, 2024). In other words, the grading of short answer questions prioritizes the content of the response over the perfection of its expression. Schneider et al. (2023) found that ChatGPT tended to assign higher scores to responses from students with poor language skills, while human raters might awarded lower scores due to the difficulty in understanding the answers or the presence of grammatical

errors. This suggests that LLMs may have the capability to accurately grade handwritten short answers, even when comprehension is challenging due to recognition errors. Nevertheless, this LLMs' capability is underexplored. The current study aimed to investigate this topic in the context of paper-pencil Chinese test in elementary schools. We compared the grading accuracy of LLMs on the raw data extracted from handwritten answers by optical character recognition (OCR) and on the preprocessed data where OCR errors were corrected.

2. Related Work

2.1 LLMs for Automatic Short Answer Grading

Early ASAG models and transfer-learning based approaches are challenged by technical complexity and limited generalization capabilities (Henkel, Roberts, Hills et al., 2024). The recent generation of LLMs offer a distinct advantage by reducing the need for task-specific training and can be used "out of the box" (Kortemeyer, 2023). The potential and advantages of LLMs for ASAG have been demonstrated. Henkel, Boxer, Hills et al. (2024) investigated the effectiveness of LLMs in grading open-ended short answer questions and found that the performance of GPT-4 with few-shot learning was close to human performance (Cohen's kappa = 0.75). Similarly, Cohn et al. (2024) demonstrated that GPT-4 could accurately grade short answer questions in science courses. The model achieved a strong agreement (quadratic weighted kappa $\geqslant$ 0.8) with human raters, and in some questions, the agreement was almost perfect (quadratic weighted kappa > 0.9). Overall, LLMs have shown the potential as capable graders. Their ease of use and flexibility make automatic grading a feasible option for educators without technical backgrounds. However, the performance of LLMs on ASAG in paper-pencil assessment is underexplored, where the answers need to be extracted from handwritten text via such techniques as OCR.

2.2 Handwritten Text Recognition

HTR is a challenging subtask in OCR. Despite the advancements in OCR techniques, errors in recognition cannot be avoided. The greatest challenge in HTR lies in accurately recognizing text of varying styles and sizes (Rahaman & Mahmud, 2022). Particularly, handwritten Chinese character recognition has always been considered as a difficult problem due to the multitude of character types, the similarity between characters, and the diversity of handwriting styles (Zhang et al., 2017). Due to the handwritten recognition errors, the scoring performance of automatic scoring systems may be significantly inferior in the case of handwritten answers than the typed answers (Gold & Zesch, 2020). It is evident that the quality of HTR is a critical factor influencing the accuracy of automatic scoring. However, it is unclear whether the HTR errors may also damage the grading performance of LLMs.

3. Method

3.1 Data Collection

The data were collected from paper-pencil Chinese exams administered to 203 fourth graders and 207 fifth graders at an elementary school in a large coastal city in China. The question formats included multiple-choice questions, short answer questions, and essay writing. For this study, we focused on short answer questions, with responses typically ranging from a single sentence to a short paragraph (no more than three lines). Specifically, we focused on making sentence and reading questions, which are prevalent in Chinese language exams in the elementary school. In this study, making sentence questions required students to construct sentences that had a particular structure and were grammatically correct and logical. Reading questions required students to reason and make judgments based on the information extracted from a short article (670 and 715 Chinese characters for the fourth and fifth grades). There were four making sentence questions and two reading questions in the exams for both

grades. The maximum scores ranged from two to four. The school teachers had manually graded students' answers to these questions before the current study.

3.2 Experimental Setup

3.2.1 Data Preprocessing

Students' handwritten Chinese characters in elementary schools often appears scribbled and is prone to misspellings. Consequently, recognizing their handwritten texts poses a significant challenge during grading the answers of Chinese exams. Despite advancements in OCR technology, a number of challenges persist (Awel & Abidi, 2019). For instance, in the current study, the raw dataset extracted from students' handwritten text by OCR specialized in HTR contained the following errors: frequent misrecognition of Chinese characters, omissions, and the inclusion of irrelevant information (e.g., question stems or crossed-out answers). Given that this study aimed to investigate the impact of these OCR errors on the grading performance of LLMs, we corrected these errors to create a preprocessed dataset that exactly represented students' actual answers. By comparing the grading performance of LLMs on the raw dataset and the preprocessed dataset, we could infer the impact of OCR errors. The preprocessing included three steps (see Figure 1): (1) removing any irrelevant information to improve text comprehensibility; (2) identifying and correcting misrecognized characters; (3) adding characters omitted by OCR. We used the OCR tool for handwriting by Baidu AI Cloud[1].

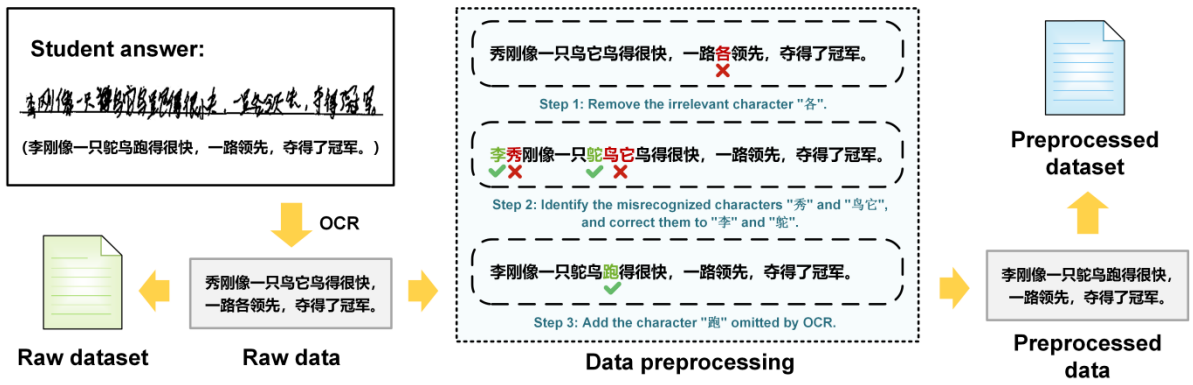

Figure 1. An example of data preprocessing.

3.2.2 Model

To assess the performance of LLMs in scoring Chinese short answer questions, we utilized ERNIE 4.0. This model was selected because of its outstanding capability in Chinese semantic comprehension (Hou et al., 2024; Zhang et al., 2023). The model version used is ERNIE-4.0-8K-0329, the most advanced and recent version available.

3.2.3 Prompt Design

A previous study demonstrated that few-shot learning achieved an average accuracy 12.6% higher than zero-shot learning in automatic scoring, and Chain-of-Thought (CoT) was particularly effective in improving scoring accuracy when paired with question stems and scoring rubrics (Lee et al., 2024). Therefore, we employed a strategy that combines few-shot learning with CoT reasoning. Specifically, ERNIE 4.0 was prompted to learn scoring from labeled examples and analyze student responses before assigning scores. For each question, we randomly selected students' responses from the preprocessed data and their grades by the school teachers as the labeled examples. The selected examples for a question covered the full range of its scores. The prompt included three parts: (i) role and task; (ii) question

[1] Baidu AI Cloud: intl.cloud.baidu.com/product/ocr.html. OCR API: aip.baidubce.com/rest/2.0/ocr/v1/handwriting.

stems, maximum scores, scoring rubrics, examples, and students' responses; (iii) the format of expected output.

3.3 Evaluation Metrics

We evaluated the scoring performance of the model on both raw and preprocessed data, using metrics bias, Kendall correlation coefficient, and mean absolute error (MAE). Students' responses selected for few-shot learning were excluded from the analysis. Bias was defined as the mean difference between the scores from ERNIE 4.0 and humans, which measure the direction of the discrepancy. A positive bias indicated that the model tended to give scores to students' responses higher than humans, while a negative bias indicated the opposite. The Kendall correlation coefficient assessed the consistency between the rankings of scores by the model and humans, with the mean score difference omitted. The MAE gauged the absolute accuracy of model grading, with a lower value suggesting a closer approximation to human grading.

4. Results and Discussion

4.1 Raw vs. Preprocessed Data

The aim of this study is to compare the scoring performance of LLMs when students' responses were not preprocessed and when they were preprocessed. Table 1 displays the evaluation results. ERNIE 4.0 performed substantially better in the preprocessed dataset than in the raw dataset. Across grades and question types, preprocessing improved Kendall correlation by 11% and MAE by 27%. Preprocessing also changed the negative bias of the model to a relatively smaller positive bias. The negative bias in the preprocessed dataset was expected as the text recognition errors could make the model assign scores lower than humans. However, it was interesting that after preprocessing, LLMs tended to give scores higher than human, indicating that LLMs might be more lenient than humans in grading. Section 4.3 returned to this point. Overall, the results indicate that the raw text extracted by OCR from students' handwritten answers may be not yet for automatic scoring by LLMs. Additional data preprocessing can enhance the accuracy of LLMs, but this places extra burden on human raters and teachers. Therefore, further investigation is necessary to improve the accuracy of OCR in elementary school students' handwritten Chinese characters as well as mitigate the impact of text recognition errors on the scoring accuracy of LLMs.

Table 1. *Bias, Kendall Correlation Coefficient (τ) and MAE*

		Bias		τ		MAE	
		Raw	Preprocessed	Raw	Preprocessed	Raw	Preprocessed
All data		-0.14	0.07	0.63	0.70	0.48	0.36 ***
Grade	Four	-0.17	0.05	0.63	0.69	0.49	0.36 ***
	Five	-0.11	0.09	0.60	0.66	0.48	0.36 ***
Question type	Reading Question	-0.02	0.23	0.58	0.63	0.64	0.54 ***
	Making Sentence Question	-0.20	-0.01	0.62	0.71	0.40	0.27 ***

Note. ***p < 0.001. Paired tests were applied to examine the MAE differences of raw and preprocessed data.

The improvement in LLMs' performance by preprocessing was the same between grades four and five. It is worth noting that the negative bias in the raw dataset of grade four was higher than that of grade five. The reason was that fourth-grade students' handwriting was generally worse than that of fifth-grade students. The improvement by preprocessing also had no difference between making sentence and reading questions. But there was a significant difference in the bias between the two question types, detailed below.

4.2 Making Sentence vs. Reading Questions

The results in Table 1 indicate that the scoring accuracy of ERNIE 4.0 was higher in making sentence questions than in reading questions. This discrepancy may be attributed to the difference in the rubrics and the maximum scores. The rubrics of making sentence questions were simpler and more specific than that of reading questions, and the maximum scores of most making sentence questions were smaller than reading questions. In addition, grading reading questions required the model understanding the reading material. These differences made the scoring on reading questions more complex, and thus, the model performed worse in the reading questions than making sentence questions.

It is also interesting that, for reading questions, ERNIE 4.0 did not exhibit a substantial negative bias in the raw dataset. However, after data preprocessing, the model showed substantial positive bias. After inspecting students' answer where the model showed positive biases, we found that the reason for the positive bias might be that human raters and the model understood the rubrics of reading questions differently. The rubrics of reading questions did not explicitly require that the answers should be related to the reading material, but, by default, human raters would consider whether the answers were relevant to the reading material. By contrast, the model might not consider or emphasize the relevance of students' answers to the reading material. Consequently, when students' answers were correct in terms of the rubrics but irrelevant to the material, the model would give scores higher than human raters. Below is an answer where the model awarded 0.5 points, but the human rater gave 0 point:

- The reading question: Please constructing a problem based on the reading material and giving two problem solutions.
- The rubric: Constructing a problem is worth 1 point, and one solution is worth 1 point.
- The scoring explanation provided by the model: "The student constructed a problem, 'Can the cells in our bodies move?' Although this problem is not relevant to the topic of the reading material 'Is water edible?' and deviates from the core content of the material, it is still a legitimate problem, thereby earning 0.5 points…"

4.3 The Human Touch in LLM Grading

We thoroughly reviewed all of the model's scoring explanations and found that the model effectively assumed the role of a human teacher. It adeptly recognized students' attempts and awarded points, a stark contrast to the typically *cold* machine learning models. This behavior demonstrates a humanistic concern for students, which is particularly vital for those in elementary school. Below is an example where the LLM awarded points to an answer without correcting the OCR errors:

> "The student's answer was not blank; however, the content was nearly incomprehensible, with only a few words barely recognizable as relevant to the topic. According to the rubric, the answer did not effectively capture the father's mental activity and thus merited only the lowest point (i.e., 0). The 0.5 points served primarily as an acknowledgment of the student's effort, rather than an endorsement of the answer correctness."

5. Conclusion and Future Work

This study is an initial effort to investigate the performance of LLMs in grading handwritten answers of Chinese exams in the wild. The aim was to compare the grading accuracy of LLMs in raw data with OCR errors and preprocessed data without these errors. Despite the use of few-shot learning and CoT reasoning, the accuracy of LLMs in the raw data was inferior than in the preprocessed data. This difference highlights the need for further research to address the challenges posed by errors in HTR and their impact on the implementation of automatic scoring by LLMs.

Regardless of the OCR errors, the grading of the LLM in this study did not approximate human raters' grading, it showed an interesting property: awarding points as an

acknowledgment of student's effort. Such properties suggest that LLMs can be a powerful tool for ASAG when their grading accuracies are practically acceptable.

The study exclusively utilized ERNIE 4.0 as the scoring model. Given the rapid evolution of LLMs, incorporating different models may yield different outcomes. Future research can explore the differences between various prompt strategies and that between LLMs. Moreover, it is noteworthy that the advent of multimodal LLMs (Yin et al., 2024) may create new possibility for ASAG in the context of paper-pencil assessment where the LLMs recognize and grade students' handwritten answers simultaneously.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2023YFC3305704) and Guangdong Philosophy and Social Science Foundation (GD24YJY06).

References

Awel, M. A., & Abidi, A. I. (2019). Review on optical character recognition. *International Research Journal of Engineering and Technology*, *6*(6), 3666-3669.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*, 60-117. https://doi.org/10.1007/s40593-014-0026-8

Chang, L., & Ginter, F. (2024). Automatic short answer grading for Finnish with ChatGPT. *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 38(21), 23173–23181. https://doi.org/10.1609/aaai.v38i21.30363

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 38(21), 23182–23190. https://doi.org/10.1609/aaai.v38i21.30364

Gold, C., & Zesch, T. (2020). Exploring the impact of handwriting recognition on the automated scoring of handwritten student answers. In *2020 17th International Conference on Frontiers in Handwriting Recognition* (pp. 252-257). IEEE. https://doi.org/10.1109/icfhr2020.2020.00054

Henkel, O., Boxer, A., Hills, L., & Roberts, B. (2024). Can large language models make the grade? An empirical study evaluating LLMs ability to mark short answer questions in K-12 education. *arXiv preprint*. https://doi.org/10.48550/arXiv.2405.02985

Henkel, O., Roberts, B., Hills, L., & McGrane, J. (2024). Can LLMs grade short-answer reading comprehension questions? An empirical study with a novel dataset. *arXiv preprint*. https://doi.org/10.48550/arXiv.2310.18373

Hou, J., Ao, C., Wu, H., Kong, X., Zheng, Z., Tang, D., Li, C., Hu, X., Xu R., Ni S., & Yang, M. (2024). E-EVAL: A comprehensive Chinese K-12 education evaluation benchmark for large language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2401.15927

Kortemeyer, G. (2023). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *arXiv preprint*. https://doi.org/10.48550/arXiv.2309.09338

Lee, G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Nagarajan, S., & Jayasurya, R. (2021). Evaluation of answer script using handwriting recognition and machine learning. *International Journal of Advanced Development in Science and Technology*, *3*(09).

Rahaman, M. A., & Mahmud, H. (2022). Automated evaluation of handwritten answer script using deep learning approach. *Transactions on Machine Learning and Artificial Intelligence*, 10(4). https://doi.org/10.14738/tmlai.104.12831

Schneider, J., Schenk, B., Niklaus, C., & Vlachos, M. (2023). Towards LLM-based autograding for short textual answers. *arXiv preprint*. https://doi.org/10.48550/arXiv.2309.11508

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2306.13549

Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., & Qiu, X. (2023). Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint*. https://doi.org/10.48550/arXiv.2305.12474

Zhang, X., Bengio, Y., & Liu, C. (2017). Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark. *Pattern Recognition*, 61, 348–360. https://doi.org/10.1016/j.patcog.2016.08.005