Overview of CCL25-Eval Task 11: Evaluation of the Quality of Handwritten Chinese Characters

Meng Wang[1] Shicong Lu[1] Zhidan Hu[1] Chen Su[2] Yujie Cao[2]
School of Humanities, Jiangnan University[1]
School of Artificial Intelligence and Computer Science, Jiangnan University[2]
{wangmengly, zsdlsc, hu_zd}@163.com
suchen@stu.jiangnan.edu.cn, m15800273662_3@163.com

Abstract

As an important means of disseminating Chinese cultural heritage, the development of Chinese handwriting skills faces dual challenges in the digital era: insufficient pedagogical resources and a lack of personalized feedback. At the 24th China National Conference on Computational Linguistics (CCL 2025), we organized a handwritten Chinese character evaluation task focusing on writing quality grading and comments generation. This benchmark utilized an expert-annotated calligraphic dataset to enhance task efficacy. Eight teams participated in the evaluation, three of which submitted valid entries. In the character grading subtask, the top-performing team achieved an F1-score of 90.5%, whereas the optimal system in the comments generation subtask attained a score of 52.8%.

Keywords: Handwritten Chinese characters, Writing quality evaluation, Multimodal large language model, Comments generation

1 Introduction

Chinese characters are an important bearer of the cultural heritage of the Chinese nation. Enhancing people's competence in Chinese character writing constitutes a fundamental pathway to advancing cultural cultivation, ultimately holding profound significance for transmitting the nation's distinguished cultural legacy. While in the information age, the widespread use of computers has led to a noticeable decline in university students' overall Chinese character writing ability. Concurrently, calligraphy education in higher institutions faces challenges due to limited teaching resources and evaluation methods, resulting in insufficient post-class practice and feedback, which hinders effective improvement in students' writing skills. In the field of Chinese character writing quality evaluation, traditional deep learning methods still fall short in providing fine-grained and personalized textual evaluations. Large Language Models (LLMs), with their powerful natural language understanding and generation capabilities, offer a novel solution to this issue. LLMs can generate detailed, personalized feedback based on input features, emulating the evaluation style of human experts. The task of " Evaluation of the Quality of Handwritten Chinese Characters" aims to leverage multimodal large language models (MLLMs) for image comprehension and text generation, addressing the limitations of existing evaluation methods in delivering personalized fine-grained feedback. This initiative bridges the gap from manual evaluation to intelligent personalized evaluation, enabling adaptive and actionable insights tailored to individual writing proficiency.

In recent years, artificial intelligence (AI) technologies have been widely applied in this field, significantly advancing the automation of Chinese character handwriting quality evaluation systems. First, Wang Kailing (Wang, 2021) emphasized, from an educational practice perspective, the necessity of constructing a comprehensive intelligent assessment system for Chinese character writing. Concurrently, Xu Bing (Xu, 2021) systematically reviewed advancements in applying machine learning techniques to calligraphy evaluation in primary and secondary education, identifying dynamic feedback and personalized guidance as core future research directions.

Currently, mainstream research methodologies can be categorized into two technical approaches:1、Skeleton extraction techniques to evaluate structural similarity between characters;2、Predefined standard-based evaluation systems for scoring character forms. Sun Mingwei et al.(Sun, 2022) proposed a deep learning model incorporating aesthetic perception, marking the first integration of calligraphic aesthetic elements—such as inter-frame structure and ink density gradients —into a quantifiable assessment framework. Rongju Sun(Sun and Lian and Tang and Xiao, 2015) introduced a topology-based skeleton matching algorithm. Chang Qinghe et al.(Chang and Wu and Luo, 2020) refined the ZS thinning algorithm to enhance handwritten character skeleton extraction precision. Zhang Zijun(Zhang, 2023) achieved end-to-end modeling of calligraphic character skeletons using deep learning. Notably, Wang Zhaoyi(Wang, 2022) innovatively employed LSTM networks to capture dynamic features of the writing process (e.g., stroke velocity variations and stroke coherence), overcoming limitations inherent in traditional static image analysis.

Driven by the continued advancement of deep learning and the computational power revolution, large language models (LLMs) have achieved breakthrough progress in natural language processing. These models demonstrate exceptional performance across diverse tasks—including text generation, machine translation, text classification, question answering, and sentiment analysis—thereby creating new possibilities for solving complex multimodal language challenges.

Multimodal large language models (MLLMs) can leverage input Chinese character image features to generate fine-grained, personalized feedback that emulates human-expert styles. This approach not only delivers richer and more targeted instructional insights but also significantly enhances evaluation efficiency and consistency, ultimately providing students with actionable learning guidance of greater pedagogical value.

This task presents significant challenges, primarily manifested in three aspects: First, the structural complexity of Chinese characters and the diversity of their stroke forms pose difficulties for models in capturing fine-grained features, which substantially impacts character-level classification accuracy. Second, the assessment of handwriting aesthetics involves subjective judgment, particularly for borderline cases near grade boundaries. This poses a substantial challenge for generating comments in the second subtask. Third, generating personalized and detailed comments requires models to not only identify structural and stroke-related issues but also articulate them in a professional and instructive manner—constituting a challenging multimodal reasoning task.

2 Task and Data Description

In this section, we will provide a detailed description of the two subtasks along with their corresponding datasets. The Chinese character handwriting samples used in this task are sourced from assignments submitted by pre-service teachers at our university as part of their participation in a handwriting training practicum project.

Subtask 1: Handwritten Chinese Characters Grading The establishment of evaluation criteria for handwritten Chinese characters has long been a challenging task. The evaluation of a handwritten work must consider aesthetic acceptance at multiple levels including calligraphic principles and spiritual dimensions, while the evaluation results are also influenced by both subjective and objective factors. Considering that the writing assignments involved in this project belong to Chinese character writing rather than calligraphy creation in nature, we will conduct evaluations solely from the perspective of writing techniques. The quality of Chinese character writing is divided into three levels: Excellent, Average, and Unqualified, with specific descriptions as follows:

Excellent: Proper structural proportions with stable balance, harmonious and aesthetically pleasing form. Correct stroke forms with mutual coherence, clear and precise strokes demonstrating pressure variations. No significant flaws in character composition.

Average: Generally reasonable structural proportions with basically stable balance and a relatively harmonious form. Essentially correct stroke forms with clear execution but lacking pressure variations. Systematic flaws exist in one or multiple categories of writing elements.

Unqualified: Disproportionate structure or unstable balance with irregular character form. Arbitrary stroke execution with unclear or sloppy lines. Serious defects in character composition.

The grade categories of handwritten Chinese characters are all manually annotated by professional calligraphy teachers. The data distribution is shown in Table 1. All images of handwritten Chinese characters are in JPG format with 224×224 pixels.

Grade	**Training Set**	**Test Set**
Excellent	500	100
Average	500	100
Unqualified	500	100

Table 1: Data distribution of Subtask 1.

Subtask 2: Comments Generation of Handwritten Chinese Characters This task aims to generate personalized and fine-grained comments and feedback on the quality of handwritten Chinese characters based on the given Chinese character images. Pen technique and structure are the two core technical elements of calligraphic art. High-quality writing work often harmonizes the aesthetic integrity of structure with the finesse of penwork. Therefore, for the feedback of writing quality, this task mainly focuses on two primary dimensions—structural composition and stroke morphology—to deliver targeted Sevaluations and descriptions. The evaluation criteria are detailed in Table2.

Tier-1 Indicator	Structure				Stroke	
Tier-2 Indicator	Proportion	Gravity	Inter-stroke Spacing	Stroke Configuration	Stroke motion	Tip lifting Down-stroke pressing

Table 2: Dimensions of Calligraphic Evaluation Metrics

The Chinese character writing quality comments for this task were manually annotated by professional calligraphy teachers, with a dataset of 700 samples partitioned into a training set (600 samples) and a test set (100 samples).

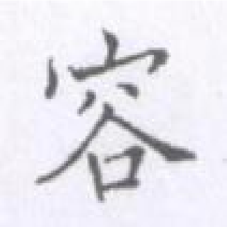
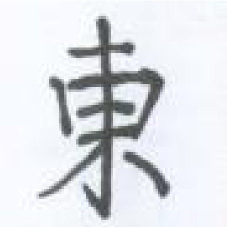
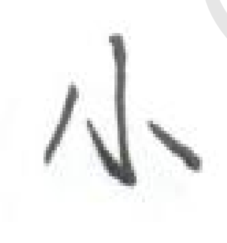
Image	Grade	Comments
	Excellent	The structure demonstrates ingenious conception with meticulously executed strokes, exhibiting natural and fluid writing flow. The characters display neat elegance—steadily poised yet dynamically alive—manifesting distinctive artistic individuality.（结构精妙，笔画精到，行笔自然流畅。字体工整秀丽，稳健灵动，具有艺术个性。）
	Average	The structure lacks symmetry with the visual center of gravity positioned excessively low, while the na-stroke morphology deviates from standard execution. Unsteady brush manipulation results in deficient weight modulation throughout the stroke sequences.（结构不够匀称，重心偏下，捺笔形态不正确。行笔不稳，笔画缺少轻重变化。）
	Unqualified	Structurally disproportionate with inconsistent stroke length and angular deviation, the script suffers from irregular density distribution and destabilized equilibrium, revealing an absence of foundational stroke intentionality.（结构不匀称，笔画长短、斜度不合理，字体疏密不匀，重心不稳，无笔画意识。）

Table 3: presents the samples of grade categories and corresponding comments.

3 Evaluation Metrics

Subtask 1:Handwritten Chinese Characters Grading

We use three metrics for each category: Precision (P), Recall (R) and F1 score (F1).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{3}$$

In equation(1) and (2), True Positives (TP) represents the number of samples correctly predicted as the target class, False Positives (FP) denotes the number of samples incorrectly predicted as the target class and False Negatives (FN) indicates the number of targeted samples missed by the system. F1-score constitutes the harmonic mean of Precision and Recall.

Subtask 2: Comments Generation of Handwritten Chinese Characters

For subtask2, we mainly focus on the following indicators: ROUGE-1, ROUGE-2, ROUGE-L.

$$ROUGE - 1 = \frac{|S \cap R|}{R} \tag{4}$$

$$ROUGE - 2 = \frac{|S_2 \cap R_2|}{R_2} \tag{5}$$

$$ROUGE - L = \frac{LCS(S, R)}{len(R)} \tag{6}$$

In equation(4)-(6), S represents the set of single characters in the generated comments, R represents the reference comments from professional calligraphy teachers, represents the number of overlapping single characters between the system-generated text and the reference text, represents the set of all consecutive two-character tuples (2-grams) in the system-generated text, represents the set of all consecutive two-character tuples (2-grams) in the reference text, and represents the number of overlapping 2-grams between the system and the reference text. LCS(S,R) represents the length of the longest common subsequence between the system-generated text S and the reference text R. The longest common subsequence refers to the longest character sequence in two texts that have the same order but does not require continuity. Len(R) represents the total character length of the reference text R. The final score is calculated as: SCORE=0.3×ROUGE-1+0.3×ROUGE-2+0.4×ROUGE-L.

4 Evaluation Results

4.1 Participants and Team Scores

A total of 8 teams registered for the evaluation, 5 of which were from academic institutions and 3 from commercial organizations. In the end, 3 teams submitted their results. We released the training dataset for each task on February 13, all participating teams submitted their result sets by May 20, and the final scores and rankings were announced on June 1. The scores of participating teams are shown in Table 4-Table 6.

Class	Metrics	Task 1		
		ZZUNLP	HZXD	ouchnai
Class A(Excellent)	Precision	0.939	0.930	0.933
	Recall	0.930	0.930	0.700
	F1	0.935	0.930	0.800
Class B(Average)	Precision	0.861	0.849	0.608
	Recall	0.870	0.840	0.930
	F1	0.861	0.844	0.732
Class C(Unqualified)	Precision	0.920	0.911	0.917
	Recall	0.920	0.920	0.660
	F1	0.920	0.916	0.768
Total Score		0.905	0.897	0.767

Table 4: Scores of Subtask 1: Handwritten Chinese Characters Grading

Metrics	Task 2		
	ZZUNLP	HZXD	ouchnai
ROUGE-1	0.598	0.575	0.632
ROUGE-2	0.412	0.403	0.345
ROUGE-L	0.562	0.548	0.565
Score	0.528	0.513	0.519

Table 5: Scores of Subtask 2: Comments Generation of Handwritten Chinese Characters

Rank	Team	Organization	Task 1	Task 2	Total Score
			F1 Score	Score	
1	ZZUNLP	Zhengzhou University	90.5	52.8	64.1
2	HZXD	East China Normal University	89.7	51.3	62.8
3	ouchnai	The Open University of China	76.7	51.9	59.3

Table 6: Final Rankings and Scores of Participating Teams in the Evaluation (Unit: %). Total Score is calculated with a weighting of 3:7.

4.2 Methodology and Analysis

We ultimately received system reports from 3 participating teams. This section will summarize and analyze the methods used in those teams.

The ZZUNLP team took the Qwen2.5-VL-7B model as the foundation and optimized computational efficiency in resource-constrained scenarios through a LoRA-based fine-tuning strategy. To address the issue of insufficient training data, they introduced gradient checkpointing, BF16 mixed-precision training, and a linear fusion cross-entropy loss function, which significantly reduced video memory occupation. Their dynamic validation algorithm validates every ten training steps and saves models that perform excellently in both Subtask 1 and Subtask 2, which is of great help for improving the effect of Subtask 2 with strong subjectivity. In addition, to further enhance the stability of predictions, they adopted the Voting Ensemble in ensemble learning, integrating the outputs of different models through majority voting or weighted voting to determine the final prediction results. This strategy can not only reduce the overfitting risk of a single model on a specific data distribution but also make full use of the advantages of different models to improve the generalization ability on unknown data. Through the above methods, the ZZUNLP team performed excellently in both Subtask 1 and Subtask 2.

The HZXD team (Chinese Character Squad) proposed a hybrid framework that integrates the vertical-domain small model ACBAM-VGG16 with multimodal large models. The ACBAM-VGG16 model is enhanced by FGSM adversarial training and CBAM attention mechanism, acting as an expert system to guide large model inference, which improves the model's robustness to input perturbations and generalization ability and makes the score of Subtask 1 increase by 7.3%. For Subtask 2, they adopted hierarchical training, using separate models for inference on the two types of excellent and unqualified with high internal comments similarity, and using multi-model collaborative inference for the low-similarity comments of the qualified type. Although this strategy performed well in the comments generation of excellent and unqualified Chinese characters, it had poor effects on the comments generation of qualified Chinese characters, which also led to the team's Subtask 2 score being only 51.3%.

The OUCHNAI team used LoRA to fine-tune the Qwen2.5-VL-72B model and utilized the in-context learning strategy: selecting the k most similar examples to the test image from the training data and sorting them by the cosine similarity of image embeddings, so that Subtask 1 achieved good results. For Subtask 2, they took the classification labels generated in Subtask 1 as training data and input them into the model to narrow the range of comments generation, which improved the comments generation effect to a certain extent. However, during the training process, the team tried to introduce external data (CHAED) other than the task-provided data to increase the training scale. Although this strategy reduced the training difficulty, it had a great impact on the training results of Subtask 1, thus indirectly leading to the decline of the effect of Subtask 2.

Rank	Team	Model Used
1	ZZUNLP	Qwen2.5-VL-7B
2	HZXD	ACBAM-VGG16,Qwen2.5-VL Series Multimodal Large Model
3	ouchnai	Qwen2.5-VL-72B

Table 7: Models Used

Table 7 shows that all participating teams in Task 11 of CCL25-Eval employed models from the Qwen2.5 series, with differences primarily reflected in their model fine-tuning methods. The more effective approaches were: the LoRA fine-tuning strategy, hierarchical training strategy, and multi-model collaborative inference strategy.Table 7 shows that all participating teams in Task 11 of CCL25-Eval adopted models of the Qwen2.5 series, and the differences mainly lie in the model fine-tuning methods. The more effective methods include the LoRA fine-tuning strategy, FGSM adversarial training, and in-context learning strategy. Among them, LoRA is an efficient large model fine-tuning strategy. By performing low-rank decomposition on the weight matrix of the pre-trained model and only training the newly added low-rank adapter parameters (usually accounting for less than 0.1% of the original model parameters), while freezing the original weights, it can greatly reduce the video memory requirements and calculation costs, and maintain the effect close to full-parameter fine-tuning. It is especially suitable for scenarios such as domain adaptation and task customization of billion-level parameter models under single-card or limited computing power, and is one of the mainstream technical solutions for lightweight applications of large models at present. FGSM generates adversarial samples that can make the model misjudge by calculating the gradient of the loss function for the input sample and adding a small perturbation along the gradient sign direction, and adds such samples to the training set, forcing the model to learn more robust feature representations. In this way, the model continuously contacts adversarial samples during the training process, thereby enhancing the defense ability against adversarial attacks and improving the generalization and robustness in the real scene. Finally, the in-context learning strategy guides the model to reason and generate responses based on the input-output mode of the example by embedding the task-related example (Demo) in the input prompt. Its core mechanism is to use the implicit pattern generalization ability of the pre-trained model. Without parameter fine-tuning, only by constructing a prompt containing task examples, the model can realize the prediction of new samples by matching the example features and semantic structure in the inference stage.

Their technical focuses differed significantly: ZZUNLP emphasized training optimization and dynamic evaluation; the HZXD Team concentrated on adversarial robustness and model collaboration; while OUCHNAI leveraged API-driven semi-supervised approaches. Results demonstrated that in data-scarce scenarios, hybrid architectures integrating domain-specific expert models with multimodal large language models (MLLMs) exhibited distinct comparative advantages.

5 Conclusion and Future Work

This paper presents an overview of the evaluation of the Quality of Handwritten Chinese Characters. In this task, we established the evaluation criteria for handwritten Chinese characters from the perspective of stroke execution techniques and structures, and released the datasets with manual annotation. The evaluation is divided into two subtasks: Handwritten Chinese Characters Grading and Comments Generation of Handwritten Chinese Characters. A total of 8 teams registered for the competition, with 3 teams submitting valid results and system reports.

All the competing teams employ multimodal language models (e.g., Qwen2.5-VL-72B), optimizing them through LoRA fine-tuning, multi-model collaboration, and semi-supervised approaches to adapt to the rating and evaluation tasks for Chinese calligraphic writing. In the Handwritten Chinese Characters Grading task, the best performing system achieved an F1-score of 90.5% , while in the Comments Generation of Handwritten Chinese Characters task, the best performances reached ROUGE-score of 52.8%. Overall, the challenges inherent in Chinese character evaluation tasks stem from the structural complexity of handwritten Chinese characters and the aesthetic expressiveness of expert commentaries. Enhancing models' capabilities for systematic character structural analysis and nuanced evaluative feedback generation continues to be critically important.

From the evaluation results of the Chinese character handwriting quality, it indicates that the model exhibits a significantly higher accuracy for “Excellent” compared to the other two categories. This discrepancy arises because samples graded “Excellent” demonstrate greater consistency in both character structures and stroke patterns. Conversely, samples graded “unqualified” and “average” with relatively poor structures and stroke patterns exhibit significant variation in the nature of their calligraphic flaws and encompass a diverse range of styles. Consequently, the precision is relatively low. In the future, data collection focusing on these categories is essential to enhance the data comprehensiveness and improve the overall recognition performance.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (62307021); Minjiang University 2024 Talent Recruitment Project (ZD202401).

References

Kailing Wang. 2021. On the Application and Exploration of Artificial Intelligence in Traditional Calligraphy Art [J]. *Jiangsu Education*, (30): 7 - 9.

Zhaoyi Wang. 2022. Deep Learning Calligraphy Evaluation Modeling Incorporating Stroke Motion Capture [D]. *Jiangnan University*. DOI: 10.27169/d.cnki.gwqgu.2022.000516.

Bing Xu. 2021. A Literature Review and Prospect of Calligraphy Evaluation in Primary and Middle Schools Based on Machine Learning [J]. *Jiangsu Education*, (30): 17 - 21.

Mingwei Sun. 2022. Research and Application of Intelligent Calligraphy Aesthetic Evaluation Algorithm [D]. *Central South University*.

Rongju Sun, Zhouhui Lian, Yingmin Tang, Jianguo Xiao. 2015. Aesthetic Visual Quality Evaluation of Chinese Handwritings [C]. In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI’15). AAAI Press, 2510–2516.

Mu Li. 2013. Research on Computer - Aided Evaluation of Calligraphy Copying [D]. *South China University of Technology*.

Chuzhou Wu. 2017. Research and Implementation of Calligraphy Copying Evaluation System [D]. *South China University of Technology*.

Qinghe Chang, Minhua Wu, Liming Luo. 2020. Skeleton Extraction for Handwritten Chinese Characters Based on Improved ZS Thinning Algorithm [J]. *Computer Applications and Software*, 37(07): 107 - 113.

Zijun Zhang. 2023. Research on Skeleton Extraction Algorithm for Calligraphy Characters Based on Deep Learning [D]. *East China University of Science and Technology*.