Large Language Models vs. Human Expert: Benchmarking Automated Short Answer Grading on Traditional Chinese Reading Comprehension Examinations Using *PIRLS-HK* Dataset

Anonymous ACL submission

Abstract

Large language models (LLMs) are rapidly finding their way into automated short-answer grading (ASAG) systems, yet we still lack a realistic benchmark for evaluating their reliability on Traditional Chinese K-12 reading-comprehension tasks. Existing ASAG benchmarks either emphasize encyclopaedic or STEM knowledge, are multiple-choice rather than open-response, or target Simplified-Chinese or English, leaving traditional Chinese ASAG task underexplored. We address this gap with *PIRLS-HK*, a dataset distilled from 15 years of Hong Kong in International Reading Literacy Study (PIRLS) materials. The first release contains 2,352 expert-graded question–answer pairs (25 questions, 4 passages) written in Traditional Chinese by 292 fourth-grade students, each accompanied by the official marking scheme. Using *PIRLS-HK* we benchmark 11 LLMs under zero-shot and few-shot settings. Performance is measured with Quadratic Weighted Kappa (QWK), Tolerance-Adjusted Accuracy (TAA) and Relative Merit Consensus (RMC). Results show Few-shot mid-sized models (e.g. qwq-32b, BM: 0.674) rival or surpass much larger variants. Full-size models show only marginal gains across prompting modes. Agreement and accuracy with human graders remain modest: the best QWK is 0.383 (deepseek-V3 few-shot) and the highest TAA ($\tau = 0$) is 71.71% (deepseek-r1 zero-shot). These findings indicate that LLMs that excel on mainstream NLP leaderboards may still lack consistency and fairness when confronted with authentic, culturally embedded assessment data. *PIRLS-HK* provides the first open benchmark for advancing ASAG research in Traditional Chinese; dataset and code will be released under CC-BY-NC 4.0.

1 Introduction

Grading short answer reading comprehension questions is a time-consuming and subjective task for teachers, requiring careful judgment of factual accuracy and reasoning (Sadler, 2009). Study show they often devolve into fact recall, with grading reliability remaining a challenge (Palmer and Devitt, 2007). Could LLMs offer a solution? Recent research demonstrates that LLMs like ChatGPT can grade university exams with moderate agreement to human scores, though inconsistencies and overcautious scoring persist (Flodén, 2025). Advances in AI-driven grading promise efficiency and consistency, but ethical concerns and domain-specific challenges remain (Gnanaprakasam and Lourdusamy, 2024).

For Hong Kong's K12 teachers, automating Traditional Chinese short-answer grading is particularly complex due to linguistic and cultural nuances (Li, 2017). While LLMs excel in multilingual tasks (Hagos et al., 2024), their application to culturally sensitive assessments is underexplored. Our study bridges this gap by benchmarking 11 LLMs against human graders using the *PIRLS-HK* dataset, evaluating performance across model sizes and prompt designs. By addressing these challenges, we aim to determine whether LLMs can become reliable, fair, and scalable grading assistants for educators (Yan et al., 2024; Xie et al., 2024).

2 Related Work

2.1 Large Language Models

Transformer architecture revolutionized NLP with self-attention, enabling efficient parallelization & superior translation performance (Vaswani et al., 2017). Building on this, BERT introduced bidirectional pre-training, achieving state-of-the-art results across tasks like GLUE (80.5%) with minimal

fine-tuning (Devlin et al., 2019). GPT-3 scaled this further, demonstrating strong few-shot learning, rivaling fine-tuned models (Brown et al., 2020). These advancements underpin LLMs' potential for automated grading, though their application to culturally sensitive tasks like Traditional Chinese assessments remains underexplored.

2.2 Automated Short Answer Grading

ASAG has evolved from early statistical approaches to advanced transformer-based methods. Early systems relied on feature engineering, such as Bag-of-Words and TF-IDF, achieving moderate performance (F1-score: 0.72) but struggling with semantic depth (Ripmiatin et al., 2023). Text-mining frameworks incorporated lexical diversity and structural features, aiding teachers but requiring broader validation (Vairinhos et al., 2022). Later, models like CBOW-LSTM improved grading by capturing word sequences, though they demanded large, labelled datasets (Zhang et al., 2022). The shift to transformer-based models, such as BERT, marked a breakthrough. Fine-tuning BERT with domain-specific textbooks or QA pairs boosted performance but limited generalizability (Sung et al., 2019). Techniques like Semantic Feature-wise Transformation Relation Networks enhanced grading by modelling question-reference-answer triples, achieving up to 11% improvement on benchmarks (Li, 2021). Back-translation further refined BERT-based systems, outperforming state-of-the-art models (Lun et al., 2020). Hybrid architectures, combining BERT with Bi-LSTM and Capsule networks, achieved near-human correlations (Pearson's r: 0.897) (Zhu et al., 2022).

Recent LLMs, such as GPT-4, demonstrate near-human grading accuracy (QWK: 0.91) on low-resource datasets like ROARS, even with minimal prompt engineering (Henkel et al., 2024). Similarly, GPT-4 outperformed baselines in Finnish undergraduate grading, though longer answers posed challenges (Chang and Ginter, 2024). Multilingual benchmarks like SciEx show LLMs surpassing student performance, with LLM-as-a-judge achieving a 0.948 correlation (Dinh et al., 2024). Zero-shot analysis of educational feedback further highlights LLMs' versatility (Parker et al., 2024).

However, ASAG systems face vulnerabilities. Adversarial inputs can trick models into accepting 60% of incorrect answers, necessitating robust countermeasures (Ding et al, 2020). Systems like AutoSAS and ESAS improved grading by 8–7.8% but relied on general datasets (Kumar et al., 2019; Goenka et al., 2020). While BERT outperforms Word2Vec, non-embedding features (e.g., lexical overlap) remain prevalent (Putnikovic and Jovanovic, 2023). Interpretability is another concern; methods like InputXGradient are needed to align model attention with human judgment (Poulton and Eliens, 2021; Zeng et al., 2022). Despite progress, ASAG systems dominate STEM fields but lack support for complex reasoning (Gao et al., 2024). Template-based systems offer precise feedback but require significant setup (Sychev et al., 2020). Frameworks like AVA, leveraging peer attention, show promise for scalable evaluation (F1: 74.7%) (Vu and Moschitti, 2020).

2.3 Related Datasets & Benchmarks

ASAG systems rely on diverse datasets, from early benchmarks like *SciEntsBank* and *Beetle* (Dzikovska et al., 2017) to modern challenges like *SQuAD 2.0*'s unanswerable questions (Rajpurkar et al., 2018). Recent surveys highlight the evolution from feature-based methods to transformer models across these datasets (Haller et al., 2022). For low-resource languages, *ScAA* provides Hindi/Marathi answers (Agarwal et al., 2020), while *CESA* and *ASAP-ZH* address Chinese segmentation (Ding et al., 2020). Reading comprehension datasets like *RACE* (Lai et al., 2017) and *RACE-C* (Liang et al., 2019) test reasoning skills, and vision-language benchmarks like *VisTW* evaluate Traditional Chinese in Taiwanese contexts (Tam et al., 2024). Knowledge-focused benchmarks include *TMMLU+* (Tam et al., 2024) for Traditional Chinese and *CMMLU* for Mandarin (Li et al., 2023), while the *multitask suite* (Hendrycks et al., 2020) spans 57 subjects.

2.4 Gaps in Current Research

Existing ASAG datasets largely originate from translated English materials, limiting their cultural and linguistic authenticity for Chinese assessments, motivating our *PIRLS-HK* dataset. Moreover, vulnerabilities to adversarial inputs remain unaddressed. Addressing these issues is beyond the scope of our work, leaving important challenges for future research.

3 Research Objectives & Questions

Grading short-answer questions in Traditional Chinese reading comprehension is time-consuming for

teachers. Our research aims to simplify this by using LLMs to grade answers automatically. We'll introduce a new dataset, *PIRLS-HK*, with answers from Hong Kong students in the 2006 PIRLS study. We want to see how well different LLMs grade compared to human experts do, making grading faster and fairer.

The research objective for the proposed study is to create and use *PIRLS-HK* to test how well LLMs can grade Traditional Chinese reading comprehension answers, and there are 3 research questions for the proposed study:

RQ1) How well do LLMs grade short answers in *PIRLS-HK*? What differences do we see between them?

RQ2) Do small-sized LLMs grade as well as medium-sized or large-sized ones? Can smaller models work well enough for grading?

RQ3) How does the design of prompts (e.g., zero-shot vs. few-shot) impact the grading accuracy of LLMs on *PIRLS-HK* for Traditional Chinese reading comprehension?

4 *PIRLS-HK* Dataset

4.1 About PIRLS

PIRLS [1] conducted every five years since 2001, evaluates fourth-grade reading comprehension globally, assessing literary and informational text skills alongside contextual data from students, teachers, and schools.

4.2 Dataset Description

PIRLS-HK contains 1,282 scanned Traditional Chinese answer booklets (2006–2016). The 2006 subset provides 2,352 expert-graded question–answer pairs from 292 students across 25 questions, with 4 passages that include marking schemes for few-shot prompt testing. All sensitive personal data have been removed; instead, a unique StudentID is used to anonymize the data. See Appendix A for details. Annotators processed and verified OCR outputs, ensuring both data reliability and privacy.

5 Methodology

5.1 LLMs Selection & Classification

We selected 11 LLMs from LiveBench based on their recent popularity and performance (i.e., global average, reasoning average, language average, etc.), grouped by parameter size. Hosted as per Table 1, the selected LLMs graded *PIRLS-HK* answers to compare accuracy against human experts. Refer to Appendix B for the creator of the artifacts.

Model	GPU	Grp	Host
gpt-4o-mini	--	A	Microsoft Azure
gpt-4o	--		
deepseek-r1:671b	--		
deepseek-V3:671b	--		
deepseek-r1:70b	~40hr	B	ollama hosted by: - Mac Mini (M4 PRO/64GB RAM); - NVIDIA RTX3050 (quantized to 4-bit precision)
llama3.3:70b	~10hr		
qwen2.5:72b	~10hr		
qwq:32b	~38hr		
glm4:9b	~4hr	C	
deepseek-r1:8b	~6hr		
qwen2.5:7b	~2hr		

Table 1: LLMs selected for the proposed study.

5.2 Prompt Design

Two prompt types were tested: zero-shot (basic instructions) and few-shot (including marking scheme examples). See Appendix C for details.

5.3 Evaluation Metrics

To compare how well LLMs grade short answers against human expert in *PIRLS-HK*, we use 3 metrics, inspired by a recent study in Finland (Chang and Ginter, 2024):

1) **Quadratic Weighted Kappa** (QWK) is a standard ASAG metric (Bonthu et al., 2021), which ranges from -1 to 1, this measures agreement between LLM and human scores (1 means perfect agreement), giving bigger penalties for larger disagreements. The weight matrix W is defined as

$$W_{x,y} = \left(\frac{x-y}{k-1}\right)^2 \quad (1)$$

where x and y are the score graded by human and LLM respectively, k is number of score categories. QWK can then be calculated by:

$$QWK = 1 - \frac{\sum_{x,y} W_{x,y} O_{x,y}}{\sum_{x,y} W_{x,y} E_{x,y}} \quad (2)$$

where O is a matrix contains the scores observed. $O_{x,y}$ corresponds to the adoption records that have a rating of x and predicted a rating of y.

2) **Tolerance-Adjusted Accuracy** (TAA) checks how often LLM scores are close to human scores, within a small tolerance τ. TAA ranges from 0 to 100, where 100 means all scores are

[1] https://www.iea.nl/studies/iea/pirls

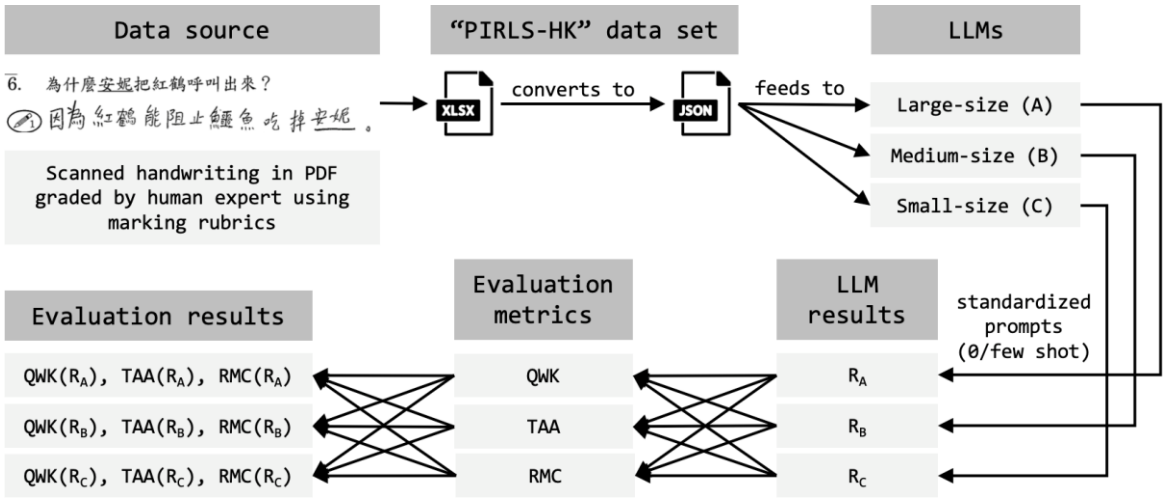

Fig. 1: The overview of the experimental setup.

within tolerance. It's a simple way to see how precise LLMs are. First, we need to define the correctness of a prediction c_i as

$$c_i = \begin{cases} 1 & \text{if } |x_i - y_i| \leq \tau \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

TAA is defined as $\overline{C}$, while $\{c_{1\cdots n}\} \in C$:

$$TAA = \overline{C} = \frac{1}{n}\sum_{i=1}^{n} c_i \times 100\% \quad (4)$$

where n is the total number of answers.

3) **Relative Merit Consensus** (RMC) looks at whether LLMs rank answers in the same order as humans, which is the % of answer pairs where LLM scores match human score rankings. It ensures LLMs respect the relative quality of answers, which is key for fair grading. For a set of answers $\{a_{1\cdots n}\} \in A$ to be evaluated, $\{x_{1\cdots n}\} \in X$, and $\{y_{1\cdots n}\} \in Y$ are graded score by human and LLM respectively. For every pair of answers (a_i, a_j) where $i \neq j$, there must be at least two distinct values within X and Y. The correctness of the pair of scores $s_{i,j}$ is defined as:

$$s_{i,j} = \begin{cases} 1 & (x_i \geq x_j \text{ and } y_i \geq y_j) \\ 1 & x_i < x_j \text{ and } y_i < y_j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

RMC is defined as the fraction of correctly scored pairs out of all possible pairs:

$$RMC = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} s_{i,j}}{\binom{n}{2}} = \frac{2\sum_{i=1}^{n}\sum_{j=i+1}^{n} s_{i,j}}{n(n-1)} \quad (6)$$

6 Experiments & Results

6.1 Experimental Setup

We processed 658 scanned answer booklets from the 2006 PIRLS, converting them into an excel file with student responses and human-assigned grades. 4 passages with marking schemes were selected, and their 2,352 valid question-answer pairs were saved as a JSON file. We tested 11 LLMs, grouped by size as suggested by Table 1, using python to run zero-shot and few-shot prompts. Each LLM's output was saved as a JSON file. They were evaluated via python using the 3 proposed metrics to compare LLMs performance against human.

6.2 Quantitative Results

1) **QWK**: Fig. 2, 3, 4 show how different LLMs grade students' answers are compared to human graders using QWK (-1 to 1). Higher QWK means better agreement with humans. Solid lines mean the model used a marking scheme (/w), dashed lines mean it didn't (w/o). **A line that's higher and stretches further right means the model grades more like a human.** At QWK = 0.8:

Group A: deepseek_v3(w/) is the best (20.5%), deepseek_r1(w/) is the worst (14.5%).

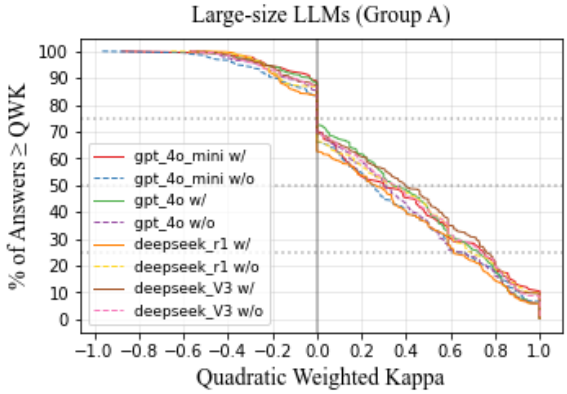

Fig. 2: % of Questions ≥ QWK (Group A).

Group B: qwq_32b(w/) is the best (19%), deepseek_r1_70b(w/o) is the worst (9.5%).

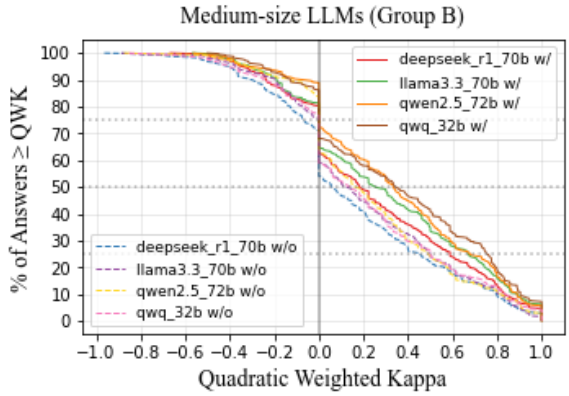

Fig. 3: % of Questions ≥ QWK (Group B).

Group C: glm4_9b(w/) is the best (11%), qwen2.5_7b(w/o) is the worst (3%).

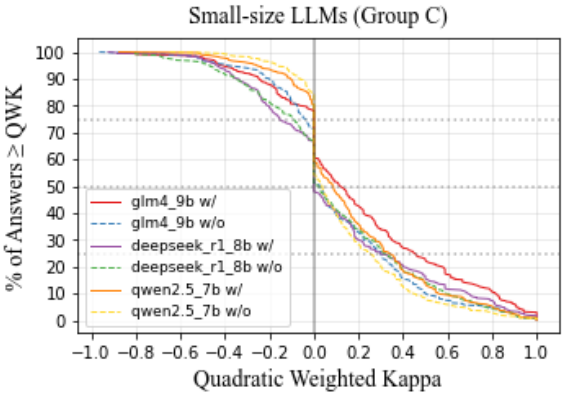

Fig. 4: % of Questions ≥ QWK (Group C).

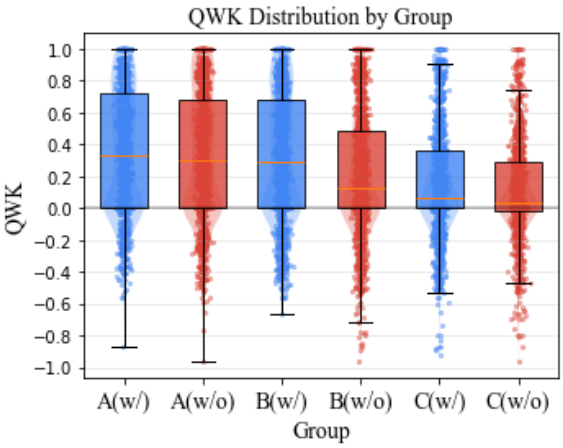

Fig. 5: QWK distribution by group.

Fig. 5 used a combination of boxplots, violin plots, and scatter points to show how different groups (w/ & w/o a marking scheme) grade students' answers compared to human using QWK. A(w/ & w/o) have similar median (0.333 & 0.300), but A(w/) are slightly more consistent. The median of B(w/) is 0.298, but B(w/o) varies more, indicating less consistency with human grader. C(w/) (0.067) performs better when compared with C(w/o) (0.030), and more consistency. **Overall, using a marking scheme leads to better and more consistent scores, especially for Group B.**

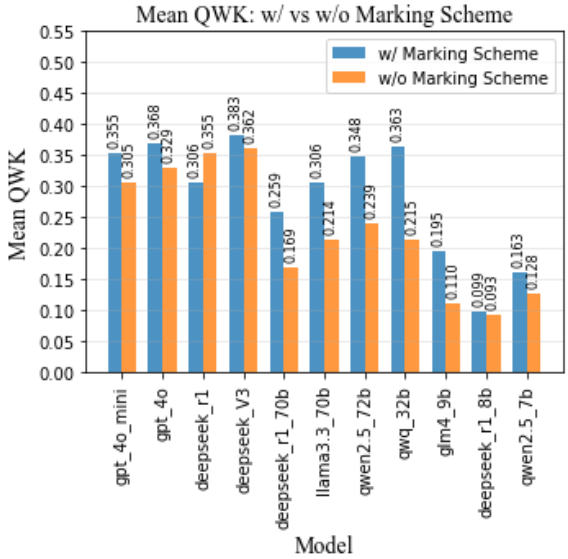

Fig. 6: Mean QWK: w/ vs w/o marking scheme.

The bar chart in Fig. 6 shows the mean QWK for each LLM (w/ & w/o). **For nearly all models, w/ is greater than w/o, meaning with marking scheme improves QWK.** deepseek_V3 scores highest (w/:0.383 & w/o:0.362). deepseek_r1_8b scores lowest (w/:0.099 & w/o:0.093). **The marking scheme helps a little across all models, but it makes a big difference in Group B** (qwq_32b: 0.215 to 0.363 (+68.9%); deepseek_r1_70b: 0.169 to 0.259 (+53.2%)).

2) **TAA:** Fig. 7, 8, 9 evaluates how different LLMs score students' answers compared to TAA (0 to 100), with higher value means better performance. **A line that stays higher and stretches further right shows more questions scored accurately**. At TAA(τ=0) = 80:

Group A: deepseek_r1(w/o) is the best (44%), gpt_4o (w/o) is the worst (21.5%).

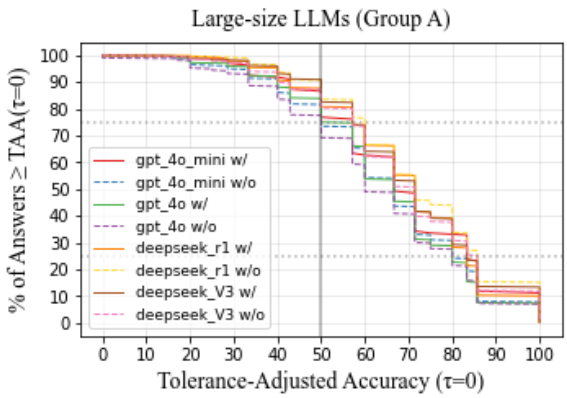

Fig. 7: % of Questions ≥ TAA(τ=0) (Group A).

Group B: qwq_32b(w/) is the best (28%), deepseek_r1_70b (w/o) is the worst (9.7%).

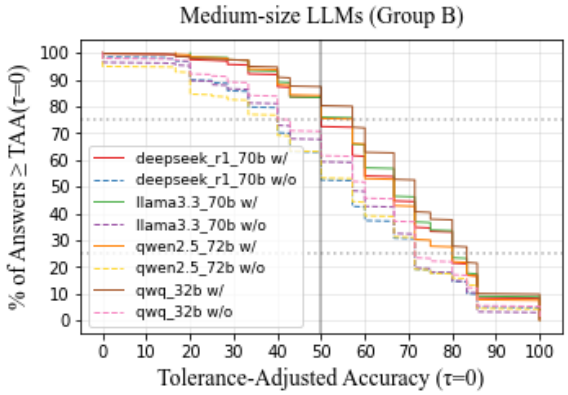

Fig. 8: % of Questions ≥ TAA(τ=0) (Group B).

Group C: glm4_9b(w/) is the best with 24%, deepseek_r1_8b (w/o) is the worst with 5%.

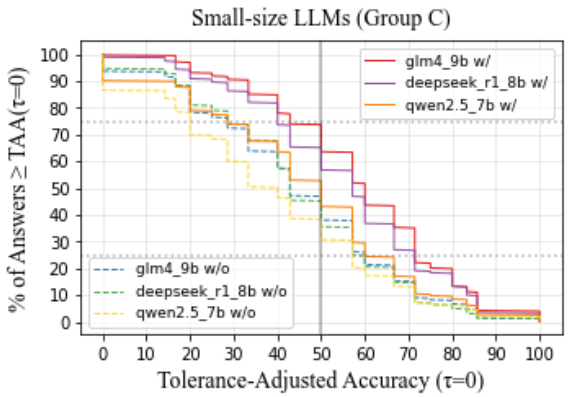

Fig. 9: % of Questions ≥ TAA(τ=0) (Group C).

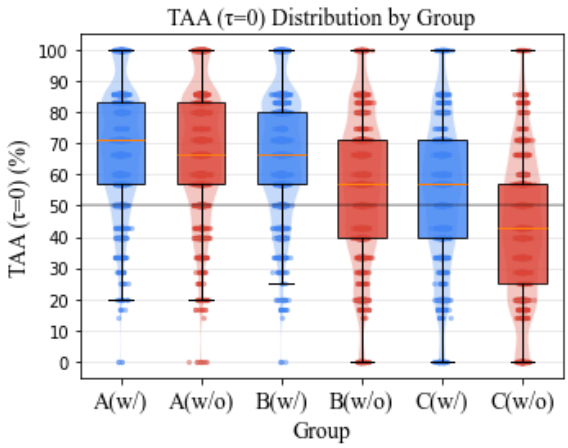

Fig. 10: TAA (τ=0) distribution by group.

Fig. 10 illustrates the distribution of TAA (τ=0) across different groups (w/ & w/o a marking scheme). A(w/ & w/o) exhibits median 71.4 & 66.7, though A(w/) shows slightly greater consistency with a narrower interquartile range, while A(w/o) displays more variability. Group B reveals a stark contrast: B(w/) maintains a median TAA of 66.7, but B(w/o) is far less consistent, with scores ranging widely, indicating significant divergence. Group C demonstrates the most pronounced improvement with a marking scheme, as C(w/) achieves a higher median TAA of 57.1 and greater consistency, while C(w/o) drops to a median of 32.9 with more spread. **Overall, the use of a marking scheme consistently enhances TAA performance and reduces variability across all groups, particularly for Group C.**

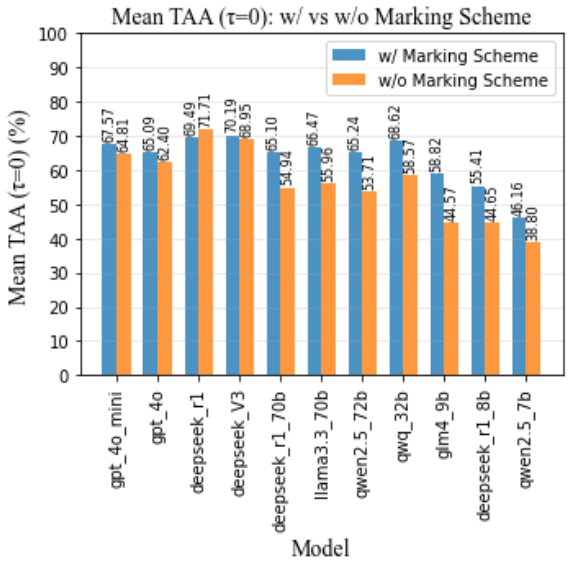

Fig. 11: Mean TAA (τ=0): w/ vs w/o.

The bar chart in Fig. 11 presents the mean TAA (τ=0) for individual LLM (w/ & w/o). Across nearly all models, w/ surpasses w/o, indicating that the marking scheme consistently enhances TAA. deepseek_V3(w/o) model achieves the highest score at 71.71, outperforming its w/ counterpart at 69.49. In contrast, qwen2.5_7b records the lowest scores (w/:46.16 & w/o:38.8). The marking scheme provides modest improvement for most models, but its impact is particularly notable in certain cases (glm4_9b: 47.5 to 58.8 (+32%); deepseek_r1_8b: 44.6 to 55.4 (+24.1%). **These results highlight the marking scheme's overall benefit, especially for medium-sized models in Group C.**

3) **RMC:** Fig. 12, 13, 14 illustrate how well different LLMs maintain the relative ranking of students' answers compared to human graders using the RMC (0 to 1), with higher values indicating better alignment with human rankings. **A line that remains higher and stretches further right shows the model consistently ranks answers like a human across more questions.** At RMC=0.6:

Group A: deepseek_V3(w/) is the best (68%), gpt_4o_mini (w/o) is the worst (37%).

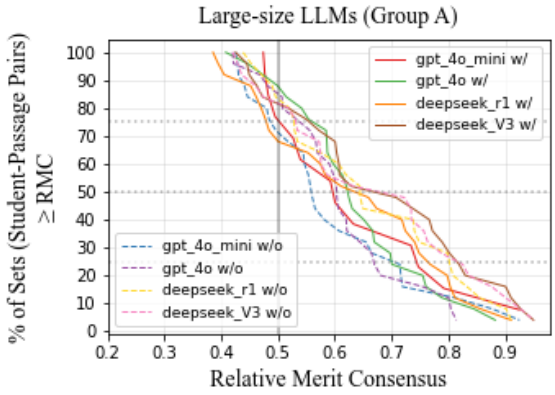

Fig. 12: % of Questions ≥ RMC (Group A).

Group B: llama3.3_70b(w/) is the best (56%), deepseek_r1_70b (w/o) is the worst (20%).

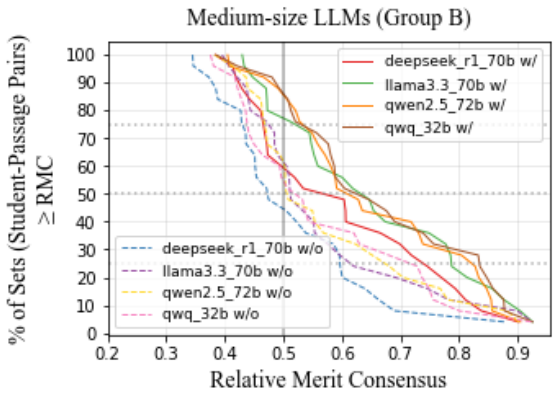

Fig. 13: % Questions ≥ RMC (Group B).

Group C: qwen2.5_7b(w/o) is the best (25%), deepseek_r1_8b (w/o) is the worst (6%).

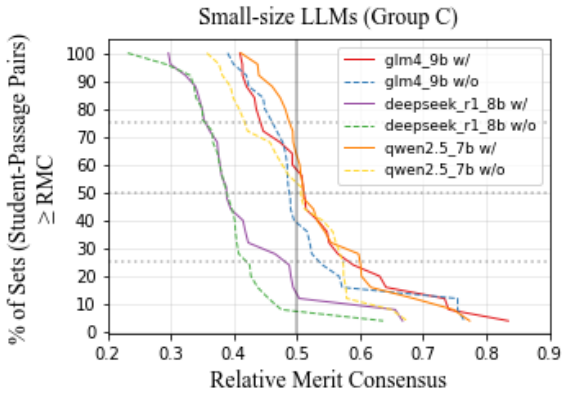

Fig. 14: % Questions ≥ RMC (Group C).

Fig. 15 illustrates the distribution of RMC across different groups (w/ & w/o a marking scheme). A(w/ & w/o) exhibit median at 0.611 & 0.600, though A(w/) shows slightly greater consistency with a narrower interquartile range, while A(w/) has more variability. Group B demonstrates the most significant improvement with a marking scheme, as B(w/) achieves a higher median (0.602) than B(w/o) (0.505) with more spread. C(w/) achieves a higher median of 0.492, while C(w/o) drops to a median of 0.449. **Overall, the use of a marking scheme consistently enhances RMC and reduces variability across all groups, particularly for Group B.**

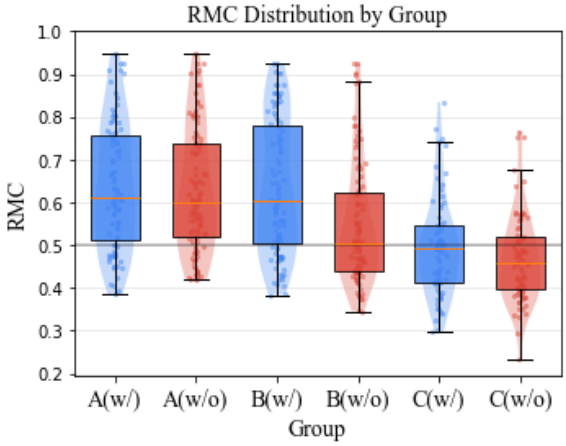

Fig. 15: RMC distribution by group.

The bar chart in Fig. 16 illustrates the mean RMC for individual LLMs (w/ & w/o). For most models, w/ is slightly greater than w/o, indicating that the marking scheme generally improves RMC. The deepseek_V3 achieves the highest score (w/:0.674 & w/o:0.662). In turn, deepseek_r1_8b records the lowest scores (w/:0.415 & w/o:0.390). The marking scheme's impact is most pronounced in models like qwq_32b (0.559 to 0.655 (+17.1%)), and deepseek_r1_70b (0.507 to 0.589 (+16.2%)). **These results underscore the marking scheme's consistent, though sometimes modest, benefit across models, particularly for Group B.**

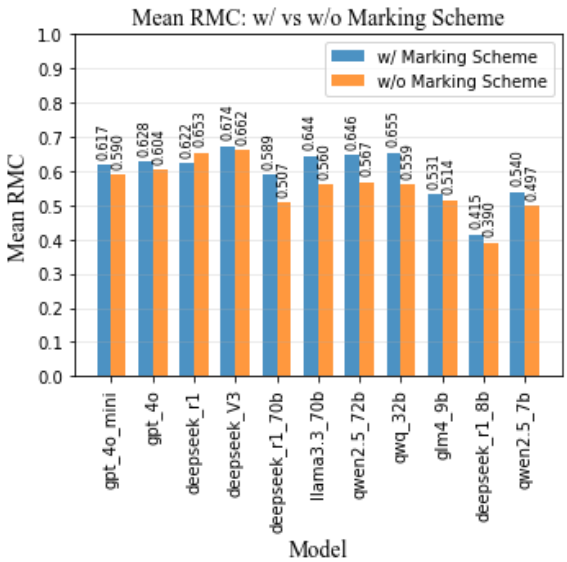

Fig. 16: Mean RMC: w/ vs w/o marking scheme.

7 *PIRLS-HK* Benchmark

The *PIRLS-HK* benchmark with or without marking scheme is defined as follows:

$$\left(\frac{\overline{QWK}+1}{2}+\frac{\overline{TAA}}{100}+\overline{RMC}\right)\div 3 \quad (7)$$

The overall benchmark is defined as follows:

$$BM=(BM_{w/}+BM_{w/o})\div 2 \quad (8)$$

#	Model	$BM_{w/}$	$BM_{w/o}$	BM
1	deepseek-V3	0.6891	0.6775	0.6833
2	deepseek-r1	0.6566	0.6825	0.6696
3	gpt-4o-mini	0.6567	0.6302	0.6435
4	gpt-4o	0.6543	0.6308	0.6426
5	qwq_32b	0.6742	0.5841	0.6292
6	qwen2.5_72b	0.6575	0.5745	0.6160
7	llama3.3_70b	0.6539	0.5755	0.6147
8	deepseek-r1_70b	0.6232	0.5470	0.5851
9	glm4_9b	0.5722	0.5049	0.5386
10	qwen2.5_7	0.5277	0.4830	0.5054
11	deepseek-r1_8b	0.5062	0.4610	0.4836

Table 2: *PIRLS-HK* benchmark.

Refer to Appendix D for the benchmark details.

8 Discussion

Our study used the *PIRLS-HK* dataset to test how well LLMs grade Traditional Chinese reading comprehension answers compared to human experts. Results directly address our 3 research questions, offering clear insights into LLMs' potential as grading tools for K12 teachers.

RQ1: How well do LLMs grade short answers in *PIRLS-HK*? What differences do we see between them?

The results show that LLMs can grade short answers with varying success. deepseek_V3 performed best (BM=0.6833). deepseek_r1_8b scored much lower (BM=0.4836). **This gap highlights that even top-performing LLMs are not reliable, not all models are equally effective, likely due to differences in training data and model design.**

RQ2: Do small-sized LLMs grade as well as medium-sized or large-sized ones? Can smaller models work well enough for grading?

Model size matters, but smaller models can still perform well. Medium-sized models like qwq_32b ($BM_{w/}$=0.6742) outperformed some large models with a marking scheme, suggesting that efficiency doesn't always require massive models. Small models like glm4_9b ($BM_{w/}$= 0.5722) also showed promise, particularly in Group C, but struggled without a marking scheme ($BM_{w/o}$=0.5049). **This indicates that smaller models can be practical for grading if paired with good prompt design, making them a cost-effective option for schools.**

RQ3: How does the design of prompts (e.g., zero-shot vs. few-shot) impact the grading accuracy of LLMs on *PIRLS-HK*?

Prompt design significantly boosts LLM performance. Across all groups, few-shot prompts with marking schemes improved scores compared to zero-shot prompts. For example, qwq_32b jumped 7.72% (from 0.5841 to 0.6742), and qwen2.5_72b rose 7.22% (from 0.5745 to 0.6575). Figures 5, 10, and 15 show that few-shot prompts led to higher medians and less variability, especially for Group B. **This suggests that providing examples in prompts helps LLMs better understand cultural and linguistic nuances, leading to fairer and more accurate grading.**

Interestingly, deepseek_r1 performed better without a marking scheme than with one (0.6566 vs. 0.6825). This may reflect an optimal chain-of-thought (CoT) length, where zero-shot prompts allow natural reasoning without rubric constraints, avoiding error accumulation from overly long CoT processes (Wu et al., 2025). Few-shot prompts might enforce suboptimal CoT lengths, reducing accuracy for nuanced grading tasks.

9 Conclusion & Future Work

Our study demonstrated that LLMs cannot effectively grade Traditional Chinese reading comprehension answers using the *PIRLS-HK* dataset. Top performers like deepseek_V3 with few-shot prompts barely matched human graders (QWK=0.383, TAA=71.71, RMC= 0.674). Medium-sized models like qwq_32b and smaller ones like glm4_9b when guided by marking schemes can perform like large-sized models, offering cost-effective solutions K12 teachers. Few-shot prompts significantly improved accuracy, highlighting the importance of prompt design.

For future work, we plan to expand the *PIRLS-HK* dataset to include more years and diverse question types, enhancing its robustness. We will fine-tune small- and medium-sized LLMs using this dataset to boost their grading performance. Additionally, we aim to identify which models excel at specific question types, paving the way for a multi-agent LLM-based ASAG platform that ensures reliable, efficient, and fair grading for educators.

Limitations

While our study provides valuable insights into the performance of LLMs on Traditional Chinese short-answer grading using the PIRLS-HK dataset, several limitations must be acknowledged:

1) Evaluation Limitations: We did not report key evaluation metrics for the human expert inter-rater reliability, such as kappa statistics. Although the experts involved in our study were well trained by PIRLS standards, including such measures could further strengthen the findings by quantifying the consistency among human graders.

2) Temporal Limitations: The current dataset is derived solely from the 2006 PIRLS assessments. Consequently, our evaluation reflects the performance of LLMs on reading comprehension tasks from that period. Future steps include extending the dataset to incorporate data from subsequent cohorts (2011, 2016, 2021), which will not only broaden the temporal scope but also serve as benchmarks for progressive iterations of the model. This study should thus be seen as an initial benchmark and baseline for further work.

3) Sampling Bias: The dataset comprises responses from 292 students, which may not fully represent the diversity of Hong Kong's student population. Any sampling bias present could limit the generalizability of our results, suggesting that further research is needed to ensure the dataset captures a wider range of student abilities and backgrounds.

4) Explainability and Feedback Evaluation: Our evaluation focused exclusively on the grading accuracy of LLMs and did not consider the quality of feedback or explanations provided alongside scores. As such, the study does not address the explainability of the grading decisions—an important factor for educators and students alike. Future work should explore methods that incorporate and evaluate the reasoning behind automated scores to ensure transparent and constructive feedback.

By elaborating on these limitations, we aim to provide a balanced view of the study's contributions while outlining clear pathways for future research.

Ethics Statement

All student identifiers—including names, school badges, and handwritten metadata—were removed from the *PIRLS-HK* dataset to ensure the privacy and confidentiality of all participants. In addition, all participants have provided consent for their data to be used for research purposes, and no personally identifiable information will appear in any published materials.

For non-commercial research or educational use of PIRLS 2006 data and related materials, all publications and released items by PIRLS and IEA are explicitly made available only for these purposes. Users can confidently utilize the data and materials for non-commercial, educational, and research activities without concern for unauthorized commercial exploitation. Detail can refer to this website: https://timssandpirls.bc.edu/pirls2006/intl_rpt.html

This work adheres to the highest ethical standards by ensuring data anonymization and restricting use to non-commercial, academic contexts, thereby protecting the rights and privacy of all involved.

References

D Royce Sadler. 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2):159-179. https://doi.org/10.1080/02602930801956059.

Edward J Palmer and Peter G Devitt. 2007. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*, 7:1-7. https://doi.org/10.1186/1472-6920-7-49.

Jonas Flodén. 2025. Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51(1):201-224. https://doi.org/10.1002/berj.4069.

Johnbenetic Gnanaprakasam and Ravi Lourdusamy. 2024. The Role of AI in Automating Grading: Enhancing Feedback and Efficiency. *Artificial Intelligence and Education - Shaping the Future of Learning*. https://doi.org/10.5772/intechopen.100525.

David CS Li. 2017. *Multilingual Hong Kong: Languages, Literacies and Identities*. Springer. https://doi.org/10.1007/978-3-319-44195-5.

Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. 2024. Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. *IEEE Transactions on Artificial Intelligence*, 1(01):1-21. https://doi.org/10.1109/TAI.2024.3444742.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models

in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90-112. https://doi.org/10.1111/bjet.13370.

Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. 2024. Grade like a human: Rethinking automated assessment with large language models. *arXiv preprint arXiv:2405.19694*. https://doi.org/10.48550/arXiv.2405.19694.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. https://doi.org/10.18653/v1/N19-1423.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877-1901.

Endang Ripmiatin, Prima Dewi Purnamasari, and Anak Agung Putri Ratna. 2023. Comparing Classical Distance Measures and Word Embeddings for Automatic Short Answer Grading. In Proceedings of *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, 492-497. https://doi.org/10.1145/3638884.3638962.

Valter Martins Vairinhos, Luís Agonia Pereira, Florinda Matos, Helena Nunes, Carmen Patino, and Purificación Galindo-Villardón. 2022. Framework for classroom student grading with open-ended questions: a text-mining approach. *Mathematics*, 10(21):4152. https://doi.org/10.3390/math10214152.

Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu, and Fuzhen Zhuang. 2022. An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments*, 30(1):177-190. https://doi.org/10.1080/10494820.2019.1648300.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. In Proceedings of *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6071-6075. https://doi.org/10.18653/v1/D19-1628.

Zhaohui Li, Yajur Tomar, and Rebecca J Passonneau. 2021. A semantic feature-wise transformation relation network for automatic short answer grading. In Proceedings of *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6030-6040. https://doi.org/10.18653/v1/2021.emnlp-main.487.

Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In Proceedings of *Proceedings of the AAAI Conference on Artificial Intelligence*, 13389-13396. https://doi.org/10.1609/aaai.v34i09.7062.

Xinhua Zhu, Han Wu, and Lanfang Zhang. 2022. Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies*, 15(3):364-375. https://doi.org/10.1109/TLT.2022.3175537.

Owen Henkel, Libby Hills, Bill Roberts, and Joshua McGrane. 2024. Can LLMs Grade Open Response reading comprehension questions? An empirical study using the ROARs dataset. *International journal of artificial intelligence in education*, 1-26. https://doi.org/10.1007/s40593-024-00431-z.

Li-Hsin Chang and Filip Ginter. 2024. Automatic short answer grading for Finnish with ChatGPT. In Proceedings of *Proceedings of the AAAI Conference on Artificial Intelligence*, 23173-23181. https://doi.org/10.1609/aaai.v38i21.30363.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, and undefined others. 2024. SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading. In Proceedings of *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11592-11610. https://doi.org/10.18653/v1/2024.emnlp-main.647.

Michael J Parker, Caitlin Anderson, Claire Stone, and YeaRim Oh. 2024. A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*, 1-38. https://doi.org/10.1007/s40593-024-00414-0.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. 2020. Don't take "nswvtnvakgxpm" for an answer--The surprising vulnerability of automatic content scoring systems to adversarial input. In Proceedings of *Proceedings of the 28th International Conference on Computational Linguistics*, 882-892. https://doi.org/10.18653/v1/2020.coling-main.76.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Get it scored using autosas—an automated system for scoring short answers. In Proceedings of *Proceedings of the AAAI Conference on Artificial Intelligence*, 9662-9669. https://doi.org/10.1609/aaai.v33i01.33019662.

Palak Goenka, Mehak Piplani, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Esas: towards practical and explainable short answer scoring (student abstract). In Proceedings of *Proceedings of the AAAI Conference on Artificial Intelligence*, 13797-13798. https://doi.org/10.1609/aaai.v34i10.7170.

Marko Putnikovic and Jelena Jovanovic. 2023. Embeddings for automatic short answer grading: A scoping review. *IEEE Transactions on Learning Technologies*, 16(2):219-231. https://doi.org/10.1109/TLT.2023.3253071.

Andrew Poulton and Sebas Eliens. 2021. Explaining transformer-based models for automatic short answer grading. In Proceedings of *Proceedings of the 5th International Conference on Digital Technology in Education*, 110-116. https://doi.org/10.1145/3488466.3488479.

Zijie Zeng, Xinyu Li, Dragan Gasevic, and Guanliang Chen. 2022. Do Deep Neural Nets Display Human-like Attention in Short Answer Scoring?. In Proceedings of *Proceedings of the 2022 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies*, 191-205. https://doi.org/10.18653/v1/2022.naacl-main.14.

Rujun Gao, Hillary E Merzdorf, Saira Anwar, M Cynthia Hipwell, and Arun R Srinivasa. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6:100206. https://doi.org/10.1016/j.caeai.2024.100206.

Oleg Sychev, Anton Anikin, and Artem Prokudin. 2020. Automatic grading and hinting in open-ended text questions. *Cognitive Systems Research*, 59:264-272. https://doi.org/10.1016/j.cogsys.2019.09.025.

Thuy Vu and Alessandro Moschitti. 2021. AVA: an Automatic eValuation Approach for Question Answering Systems. In Proceedings of *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5223-5233. https://doi.org/10.18653/v1/2021.naacl-main.412.

Myroslava O Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In Proceedings of *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 263-274.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784-789. https://doi.org/10.18653/v1/P18-2124.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Dolly Agarwal, Somya Gupta, and Nishant Baghel. 2020. ScAA: a dataset for automated short answer grading of children's free-text answers in Hindi and Marathi. In Proceedings of *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 430-436.

Yuning Ding, Andrea Horbach, and Torsten Zesch. 2020. Chinese content scoring: Open-access datasets and features on different segmentation levels. In Proceedings of *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 347-357. https://doi.org/10.18653/v1/2020.aacl-main.37.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In Proceedings of *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 785-794. https://doi.org/10.18653/v1/D17-1082.

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In Proceedings of *Asian Conference on Machine Learning*, 742-757.

Zhi Rui Tam, Ya-Ting Pai, Yen-Wei Lee, and Yun-Nung Chen. 2025. VisTW: Benchmarking Vision-Language Models for Traditional Chinese in Taiwan. *arXiv preprint arXiv:2503.10427*.

Zhi Rui Tam, Ya Ting Pai, Yen-Wei Lee, Hong-Han Shuai, Jun-Da Chen, Wei Min Chu, and Sega Cheng. 2024. Tmmlu+: An improved traditional chinese evaluation suite for foundation models. In Proceedings of *First Conference on Language Modeling*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask

language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In Proceedings of *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17--20, 2021, Proceedings 5*, 61-78. https://doi.org/10.1007/978-3-030-84060-0_5

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When More is Less: Understanding Chain-of-Thought Length in LLMs. *arXiv preprint arXiv:2502.07266*.

A Details about *PIRLS-HK* Datasets

The related python code use in this paper, sample dataset with passage content, marking scheme, and 511 out of 2,352 answers (full dataset will be released once the paper is accepted), sample answer booklets for the 4 selected passages used for grading are shown in the following link: https://anonymous.4open.science/r/pirls-hk-dataset-benchmark-E4A9/

For the data structure of *PIRLS-HK*, please refer to Table 3.

Attribute	Description
StudentID	A unique identifier assigned to each student.
PassageID	A unique identifier assigned to each passage.
QuestionNumber	Question number in the passage.
QuestionContent	Question text.
AnswerContent	OCR result from student for the respective question.
FullMark	Full mark of the question.
Score	The score graded by the human.

Table 3: Data structure of *PIRLS-HK*.

Datasheet for *PIRLS-HK* is attached to the end of the paper.

B Creators of the Artefacts

Microsoft Azure:
https://azure.microsoft.com/
Ollama:
https://ollama.com/

C Grader Prompts

The grader prompts for the proposed study are shown in Fig. 17 and Fig. 18. (Temperature = 0.7)

D *PIRLS-HK* Benchmark in Detail

For details on the *PIRLS-HK* Benchmark results for 11 LLMs, please refer to Table 4.

Original Version (Zero-shot):

你是一名小學教師，請根據以下考試問題進行評分。你需要基於你的專業知識和對小學教育的理解來評估學生的回答。

[閱讀文章] {passage_content} [/閱讀文章]

[問題] {question_content} [/問題]

[學生答案] {answer_content} [/學生答案]

[滿分] {full_mark} [/滿分]

分數必須是整數，評分範圍在[0, {full_mark}]

請嚴格按照以下格式生成 JSON 響應，僅輸出 JSON 格式，不要添加額外的說明或解釋：

```
{
  "評分理由": "<評分理由>",
  "得分": "<得分>"
}
```

Translated Version (Zero-shot):

```
You are a primary school teacher. Please grade the following exam question
based on your professional knowledge and understanding of primary education.
[Reading Passage] {passage_content} [/Reading Passage]
[Question] {question_content} [/Question]
[Student Answer] {answer_content} [/Student Answer]
[Full Mark] {full_mark} [/Full Mark]
The score must be an integer, within the range [0, {full_mark}].
Please strictly follow the format below to generate a JSON response, output-
ting only the JSON format without additional explanations or comments:
{
   "Reason for the grade": "<Reason for the grade>",
   "Grade": "<Grade>"
}
```

Fig. 17: Original and translated versions of the zero-shot grader prompt.

Original Version (Few-shot):

你是一名小學教師，請根據以下考試問題進行評分。你需要基於你的專業知識和對小學教育的理解來評估學生的回答。

[閱讀文章] {passage_content} [/閱讀文章]

[問題] {question_content} [/問題]

[學生答案] {answer_content} [/學生答案]

[評分標準] {rubric} [/評分標準]

[滿分] {full_mark} [/滿分]

分數必須是整數，評分範圍在[0, {full_mark}]

請嚴格按照以下格式生成 JSON 響應，僅輸出 JSON 格式，不要添加額外的說明或解釋：

```
{
  "評分理由": "<評分理由>",
  "得分": "<得分>"
}
```

Translated Version (Few-shot):

```
You are a primary school teacher. Please grade the following exam question
based on your professional knowledge and understanding of primary education.
[Reading Passage] {passage_content} [/Reading Passage]
[Question] {question_content} [/Question]
[Grading Standard] {rubric} [/Grading Standard]
[Student Answer] {answer_content} [/Student Answer]
[Full Mark] {full_mark} [/Full Mark]
The score must be an integer, within the range [0, {full_mark}].
Please strictly follow the format below to generate a JSON response, output-
ting only the JSON format without additional explanations or comments:
{
   "Reason for the grade": "<Reason for the grade>",
   "Grade": "<Grade>"
}
```

Fig. 18: Original and translated versions of the few-shot grader prompt.

		Few-Shot				Zero-Shot				Overall	
#	Model	QWK	TAA	RMC	BM	QWK	TAA	RMC	BM	BM	*Diff*
1	deepseek-V3	0.383	70.19	0.674	0.6891	0.362	68.95	0.662	0.6775	0.6833	*0.86%*
2	deepseek-r1	0.306	69.49	0.622	0.6566	0.355	71.71	0.653	0.6825	0.6696	*-1.90%*
3	gpt-4o-mini	0.355	67.57	0.617	0.6567	0.305	64.81	0.590	0.6302	0.6435	*2.11%*
4	gpt-4o	0.368	65.09	0.628	0.6543	0.329	62.40	0.604	0.6308	0.6426	*1.86%*
5	qwq_32b	0.363	68.62	0.655	0.6742	0.215	58.57	0.559	0.5841	0.6292	*7.72%*
6	qwen2.5_72b	0.348	65.24	0.646	0.6575	0.239	53.71	0.567	0.5745	0.6160	*7.22%*
7	llama3.3_70b	0.306	66.47	0.644	0.6539	0.214	55.96	0.560	0.5755	0.6147	*6.81%*
8	deepseek-r1_70b	0.259	65.10	0.589	0.6232	0.169	54.94	0.507	0.5470	0.5851	*6.97%*
9	glm4_9b	0.195	58.82	0.531	0.5722	0.110	44.57	0.514	0.5049	0.5386	*6.67%*
10	qwen2.5_7	0.163	46.16	0.540	0.5277	0.128	38.80	0.497	0.4830	0.5054	*4.63%*
11	deepseek-r1_8b	0.099	55.41	0.415	0.5062	0.093	44.65	0.390	0.4610	0.4836	*4.90%*

Table 4: *PIRLS-HK* benchmark in detail.