MaPLe: Marker-guided Partial Labeling

Anonymous ACL submission

Abstract

In recent years, using large language models (LLMs) as evaluators has emerged as a new evaluation paradigm. However, when reasoning processes are complex, models often struggle to determine appropriate analysis directions, and unguided evaluation may lead to erroneous judgments. To address this, we propose a novel training strategy called MaPLe (Marker-guided Partial Labeling). This method explicitly triggers the model's implicit reasoning paths by randomly masking prompt information, thereby guiding the reasoning direction and enhancing evaluation accuracy. To validate the method's cross-lingual and multi-scenario adaptability, we constructed an automatic question-answering scoring chinese dataset for second language learners, Chinese-L2. Experimental results demonstrate that MaPLe achieves superior performance across multiple benchmarks and exhibits strong generalization capabilities in cross-lingual and multi-scenario data environments.Our method and related resources are released at `https://anonymous.4open.science/r/MaPLe1-60D6`

1 Introduction

In recent years, LLMs have experienced rapid development, demonstrating formidable capabilities across numerous natural language processing tasks(Bonthu et al., 2021; Mann et al., 2020; Ouyang et al., 2022; OpenAI, 2023). However, their performance as evaluators in automated scoring tasks—which demand fine-grained reasoning and precise numerical calibration—remains unsatisfactory, sometimes even underperforming models with smaller parameter scales(Ferreira Mello et al., 2025; Henkel et al., 2024a; Carpenter et al., 2024) (Figure1).

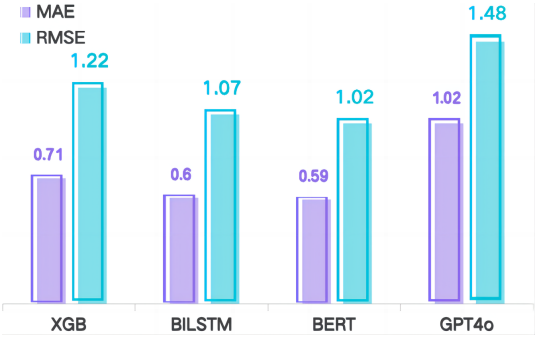

Figure 1: For the Mohler dataset, both MAE and RMSE metrics (where lower values are better) show that small models outperform large models. (Ferreira Mello et al., 2025).

Research using LLMs as evaluators can be broadly categorized into two types: the first involves directly invoking LLM for zero-shot or few-shot prompt scoring(Yuan et al., 2023); the second entails supervised fine-tuning of LLM for specific scoring tasks(Wei et al., 2022; Kojima et al., 2022; Rosset). However, both approaches exhibit inherent limi-

Dataset	Reference Answer	Scoring Criteria	Maximum Length	Average Length	Score Range	Number of Questions (essays)
Chinese-L2	×	✓	218	52	0-8	181
Mohler	×	×	1031	163	0-5	80
AES ENEM	✓	×	2004	4571	0-1000	2263

Table 1: Key information table for the Chinese-L2, Mohler and AES ENEM datasets

tations. Prompt-based approaches heavily rely on closed-source and expensive commercial models, posing significant challenges in cost control, assessment transparency, and data privacy protection (OpenAI, 2023). Supervised fine-tuning methods, while achieving efficient and accurate scoring on specific datasets, suffer from severe limitations in generalizability and flexibility. Consequently, neither paradigm fully unleashes the inherent reasoning potential of LLMs to achieve robust and generalizable automated scoring.

We contend that the core issue lies in the absence of explicit guidance for the model's implicit reasoning process during scoring (though methods like Chain of Thought can guide LLMs toward deep reasoning, they cannot guarantee the correctness of the reasoning process). To address this, we propose the MaPLe training strategy. By masking key words in the prompt template, we guide the LLM to perform implicit reasoning within "information gaps" while specifying the direction of reasoning. This approach transforms masking into an internal reasoning chain, activating the model's latent scoring potential. Simultaneously, to evaluate large models' generalization capabilities across languages, we constructed Chinese-L2—the first multi-scenario question-answering scoring dataset tailored for Chinese as a second language learners.

Our main contributions are listed below:

• We independently constructed Chinese-L2, the first multi-scenario question-answering grading dataset focused on Chinese second-language learners.

• We proposed MaPLe, which promotes thinking through masking, introducing analogical reasoning chains without requiring additional annotations, significantly improving LLM grading accuracy.

• We systematically evaluated the performance of mainstream LLMs in cross-language and cross-scenario automatic grading tasks, revealing both challenges and opportunities in multilingual environments.

2 Related Work

This paper focuses on automatic scoring tasks, bringing together two major mainstream scenarios: Automated Essay Scoring (AES) and Automated Short Answer Grading (ASAG) under a unified perspective. It systematically reviews and compares the latest research advancements of large models in these areas.

LLMs for AES. The emergence of LLMs and their versatility across various downstream tasks (Wei et al., 2022; Kojima et al., 2022) has drawn attention to their potential for essay grading. This has spurred the development of zero-shot and few-shot automated essay grading techniques based on LLMs. Concurrently, research has shown that incorporating context learning and the Chain of Thought (COT) into prompt design for automated grading can enhance the effectiveness of automated essay scoring (Lee et al., 2024a,b). Other studies have explored prompt engineering, investigating how different prompt templates influence automatic scoring tasks (Zhou et al., 2022; Ye et al., 2024; Chen et al., 2024; Xue et al., 2025). However, LLMs employing these strategies still perform poorly, with results comparable to classifiers relying solely on predicting text

length.

LLMs for ASAG. The application of LLMs for ASAG has become a current research focus. Recent studies have explored the potential of LLMs in grading short-answer questions, comparing their performance through zero-shot and few-shot settings (Henkel et al., 2024b,a). Additional research has focused on utilizing various prompt templates to reflect LLM performance in ASAG tasks (Yang et al., 2023; Kamesh, 2024; Duong and Meng, 2024; Zhao et al., 2025; Wang et al., 2025; Ferreira Mello et al., 2025). However, these studies have not demonstrated the advantages of LLMs.

3 Why Do LLMs Underperform on Scoring Tasks?

Many researchers have observed that current LLMs exhibit limitations in numerical reasoning (scoring tasks) (Lai et al., 2025). In this section, we will explore potential reasons for LLMs' poor performance on scoring tasks and propose recommended approaches.

3.1 The Pre-training Objectives of LLMs

Common LLMs primarily optimize the "next-word prediction" objective during pretraining, granting them significant advantages in generating coherent long-form text. However, this objective predominantly learns superficial statistical correlations rather than interpretable discriminative criteria. Consequently, models often exhibit scale instability and bias in scoring tasks requiring numerical calibration and multidimensional trade-offs, making it challenging to achieve stable alignment with human evaluations.

3.2 Implicit Reasoning Capabilities Remain Underutilized

Although existing research has endowed LLMs with some automated scoring capabilities through techniques like prompt engineering, few-shot learning, and supervised fine-tuning (Zhao et al., 2023; Henkel et al., 2024a; Fan et al., 2024), these approaches often remain confined to outcome-level alignment. They fail to sufficiently activate and constrain the model's implicit reasoning processes. Without explicit guidance on reasoning pathways, models tend to rely on superficial linguistic cues or heuristic associations during scoring. This leads to unstable scoring scales and drifting dimension weights, making reliable alignment with human evaluation criteria difficult. This issue is particularly pronounced in dimensions requiring deep semantic understanding, such as content quality and logical consistency (Seßler et al., 2024).

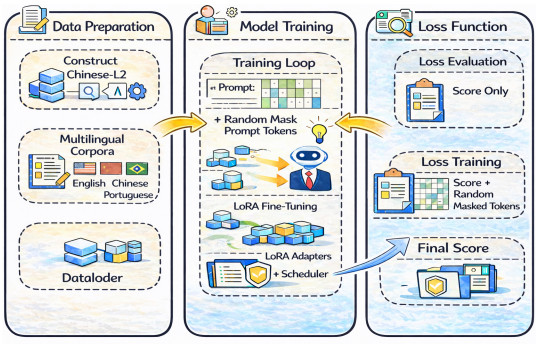

Figure 2: Data construction workflow and MaPLe method (we used LoRA Fine-Tuning with a rank of 64).

3.3 MaPLe

MaPLe is a straightforward plug-and-play training strategy designed to enhance the reasoning capabilities of LLMs. This approach processes input prompts (as definition in the theorem below), randomly selects a starting masking position, and subsequently masks all subsequent tokens (including the ground truth label) starting from that position. The model must then infer the content of the masked tokens. By converting the single score label into partial labeling for tokens and scores, and in-

corporating masked tokens and scores into the loss function calculation, MaPLe provides explicit direction for the model's reasoning process. Masked tokens are treated as the model's "thought trail" or "reasoning path," enhancing contextual understanding during inference. This activates latent reasoning capabilities, significantly improving inference accuracy.

Definition MaPLe Given a prompt template P, whose token sequence is represented as $P = (t_1, t_2, \ldots, t_n)$, a starting index q_i is randomly generated, where $q_i \in \mathrm{U}\{i_1, i_2, \ldots, i_n\}$, and U denotes a discrete uniform distribution. Each batch uses a different random q to better stimulate LLMs' implicit reasoning capabilities and generalization. For the original sequence $P = (t_1, t_2, \ldots, t_n)$, applying masking yields: $M_q(P) = (t_1, \ldots, t_{q-1}, [Mask], [Mask], \ldots, [Mask])$. The partial labeling consisting of $[Mask]$ is where we need to compute the loss, providing clear direction for model inference.

Corollary 1 (Progressive Masking Property) When $q = 1$, the entire prompt is masked, requiring the model to generate content entirely autonomously. When $q = n$, the task approximates traditional language modeling. As q decreases from n to 1, reconstruction difficulty monotonically increases.

Corollary 2 (Structural Integrity Constraint) Unlike traditional random sparse masking, this method forces the model to learn the structural integrity of the prompt template. Specifically, it must correctly infer the structure and semantics of the entire suffix $(t_q, \ldots, t_n)$ based on the prefix $(t_1, \ldots, t_{q-1})$.

3.4 Loss Function

To support our MaPLe, we modified the components involved in loss calculation. During training: we compute loss for both the random masking portion and the scores simultaneously. During validation: we mask only the scores, aiming to reduce interference from random selection and identify the model that most accurately outputs standard scores. Specifically, for the training phase, we first construct labels. We define tokens for which no loss is computed as γ. Then, for the input to the model x_i=[$x_{i1}, \ldots, x_{iL}$], we randomly determine the start position a_i of the masking. From a_i onwards, all subsequent tokens are included in loss computation. We denote the interval where loss is computed as k_i. Thus, the labels for the training phase are:

$$y_{it} = \begin{cases} y, & t \in k_i \\ x_{it}, & t \notin k_i \end{cases} \tag{1}$$

For the validation phase, we design an identifier g_i to separate the prompt content from the final output score. We do not compute loss for any content preceding g_i; only the loss for the score is calculated. Thus, the labels for the validation phase are:

$$y_{it} = \begin{cases} \gamma, & t \leq g_i \\ x_{it}, & t > g_i \end{cases} \tag{2}$$

Therefore, the final loss function is:

$$\mathcal{L}_i = -\frac{1}{|\Omega_i|} \sum_{t \in \Omega_i} \log \frac{\exp(z_{i,t,y_{i,t}})}{\sum_{v=1}^{V} \exp(z_{i,t,v})} \tag{3}$$

where Ω_i denotes the valid set for loss calculation (equation 4), z_{itv} represents the logits for token t at position v in the vocabulary, and y_{it} is the target token.

$$\Omega_i = \{t \mid y_{i,t} \neq \gamma\} \tag{4}$$

Models	Mohler			AES ENEM			Chinese-L2		
	$\Delta\downarrow$	$\lvert\Delta\rvert\downarrow$	$\rho\uparrow$	$\Delta\downarrow$	$\lvert\Delta\rvert\downarrow$	$\rho\uparrow$	$\Delta\downarrow$	$\lvert\Delta\rvert\downarrow$	$\rho\uparrow$
Praetor(Leng et al., 2025)	1.84	2.16	0.41	-	-	-	1.08	1.93	0.74
MTS(Lee et al., 2024c)	-	-	-	193.71	215.94	0.58	2.25	2.79	0.73
Qwen3-8B(ours)	0.32	**0.54**	0.86	59.56	87.42	0.83	0.61	1.11	0.91
Qwen3-14B(ours)	**0.30**	0.55	**0.86**	**54.86**	**81.25**	**0.85**	**0.58**	**1.05**	**0.92**
Llama-3.1-8B-Instruct(ours)	0.36	0.59	0.83	59.31	88.15	0.83	0.62	1.10	0.92

Table 2: Comparison results with existing methods across three datasets. Δ represents MAE, $|\Delta|$ represents RMSE, ρ represents Pearson.

4 Experiment

Currently, most research on LLMs heavily relies on massive computational resources (see Appendix A.1 for comparison), posing significant challenges for universities with limited resources. Due to insufficient computational resources, we introduced LoRA fine-tuning technology into model training and designed a universal prompt template (see Appendix A.2). Figure 2 illustrates our data construction workflow and training methodology.

4.1 Evaluation Indicators

To assess the performance of our proposed method, we employ Pearson's correlation coefficient (Pearson) to measure the trend correlation between the scores generated by LLMs and the true labels. Additionally, we calculate the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to quantify the gap between the LLMs' output scores and the true labels.

4.2 Baseline

We selected the state-of-the-art open-source LLMs as baseline models: Qwen3-8B, Qwen3-14B (Yang et al., 2025), and Llama-3.1-8B-Instruct (Dubey et al., 2024). (Detailed experimental settings are provided in A.3.) Additionally, we compared with recent scoring task studies: MTS (Lee et al., 2024c) and Praetor (Leng et al., 2025). (Detailed experimental settings for Praetor and MTS are provided in Appendix A.4. Furthermore, we selected additional recent relevant studies for comparison, as detailed in Appendix A.5.)

4.3 Datasets

Currently, automatic scoring research primarily focuses on English-language scenarios, while data resources for other languages remain relatively scarce. To address this imbalance and validate the effectiveness of the proposed method in multilingual settings, we constructed the Chinese-L2 dataset, with its construction details outlined in Appendix A.6. Additionally, we incorporate the English Mohler corpus (Mohler et al., 2011) and the Portuguese AES ENEM dataset (Silveira et al., 2024). Key information for the three datasets is summarized in Table 1, with a more detailed introduction provided in Appendix A.6.

4.4 Results: Comparison of Open-Source LLMs

Through systematic comparisons of various LLMs, we found Qwen3-14B to deliver the most competitive overall performance on automated scoring tasks. Its MAE on the Mohler, AES ENEM, and Chinese-L2 benchmark datasets was 0.30, 54.86, and 0.58, respectively. Notably, under comparable parameter counts, Qwen3-8B consistently outper-

Models	Method	Mohler			AES ENEM			Chinese-L2		
		$\Delta\downarrow$	$\lvert\Delta\rvert\downarrow$	$\rho\uparrow$	$\Delta\downarrow$	$\lvert\Delta\rvert\downarrow$	$\rho\uparrow$	$\Delta\downarrow$	$\lvert\Delta\rvert\downarrow$	$\rho\uparrow$
Qwen3-8B	Base	0.34	0.62	0.82	60.92	91.06	0.82	0.69	1.21	0.90
	NoMask	0.34	0.60	0.84	60.18	89.30	0.82	0.65	1.18	0.90
	Fixed_Random	0.35	0.59	0.84	78.01	103.86	0.78	0.66	1.20	0.90
	MaPLe	**0.32**	**0.54**	**0.86**	**59.56**	**87.42**	**0.83**	**0.61**	**1.11**	**0.91**
Qwen3-14B	Base	0.32	0.56	0.86	63.52	88.04	0.82	0.62	1.14	0.90
	NoMask	0.35	0.62	0.81	59.93	87.00	0.83	0.59	1.10	0.91
	Fixed_Random	0.35	0.59	0.84	58.69	85.04	0.83	0.60	1.12	0.91
	MaPLe	**0.30**	**0.55**	**0.86**	**54.86**	**81.25**	**0.85**	**0.58**	**1.05**	**0.92**
QLlama-3.1-8B-Instruct	Base	0.39	0.70	0.81	59.46	95.88	0.79	0.66	1.15	0.91
	NoMask	0.37	0.60	0.82	60.55	93.29	0.80	0.68	1.23	0.89
	Fixed_Random	0.39	0.62	0.82	59.31	89.97	0.83	0.69	1.16	0.90
	MaPLe	**0.36**	**0.59**	**0.83**	**59.31**	**88.15**	**0.83**	**0.62**	**1.10**	**0.92**

Table 3: Results of ablation experiments across three datasets. Δ represents MAE, $|\Delta|$ represents RMSE, ρ represents Pearson.

formed Llama3.1-8B-Instruct. Furthermore, while larger models within the same series typically exhibit superior performance across most metrics, exceptions were observed: for instance, on the Mohler dataset, Qwen3-8B's RMSE correlation coefficient was marginally lower than Qwen3-14B's. Performance gains from model scale were less pronounced than anticipated—e.g., a mere 0.02 difference in MAE on Mohler and 0.03 difference on Chinese-L2—likely due to our limited dataset size. Therefore, in practical deployment scenarios, selecting models with smaller parameter counts remains viable when balancing inference efficiency and resource costs.

4.5 Results: Method Comparison

As shown in Table 2, our method demonstrates superior competitiveness compared to existing approaches such as Praetor and MTS, particularly on the AES ENEM dataset. This indicates that existing methods are often constrained by insufficient transferability, making it challenging to effectively generalize to similar tasks. In contrast, our method can be directly applied to common rating tasks without requiring task-specific tuning, while still achieving excellent performance.

5 Analysis

This section provides an in-depth analysis of each module within MaPLe, exploring the factors behind its outstanding performance and their impact on the model. To this end, we conducted three additional comparative experiments: (1) Base: Removing the random masking mechanism and using standard cross-entropy loss; (2) NoMask: Retaining the proposed loss function but not applying random masking to the prompt; (3) Fixed_Random: Using a fixed masking region for each sample during training. The experimental results are shown in Table 3.

5.1 Synergy Between MaPLe and Loss Functions

To investigate the synergistic effect of MaPLe and the loss function in enhancing model performance, we designed ablation experiments by simultaneously removing both components for evaluation. The experimental results are shown in Table 3. After removing both MaPLe

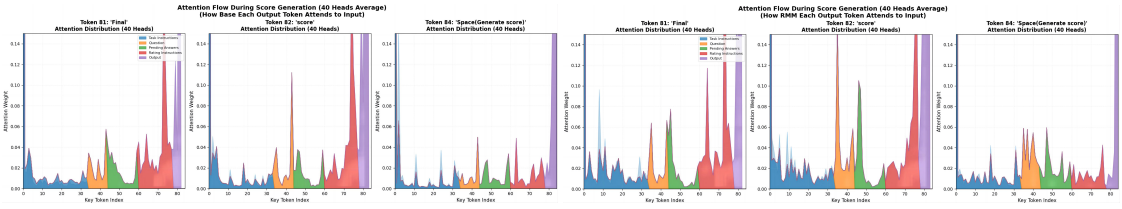

Figure 3: Comparison between Base and MaPLe. We divided the prompt into distinct regions to observe the distribution of attention weights during generation scoring, with colors representing different regions.

and the loss function, evaluation metrics significantly decreased across all three datasets. For instance, on the Mohler dataset, the RMSE increased by 12.62%. This result indicates a synergistic effect between MaPLe and the loss function, not only validating the advantages of MaPLe but also emphasizing the critical role of selecting an appropriate loss function during training and validation for model stability.

5.2 Role of MaPLe

NoMask and Fixed_Random are two supplementary control experiments we designed to further validate the effectiveness of the MaPLe module. As shown in Table 3, removing the MaPLe module resulted in a 6.61% increase in root mean square error (RMSE) on the AES ENEM dataset. These experiments clearly demonstrate the role of introducing random masking into prompts: the model no longer blindly memorizes prompt content but instead learns to "answer the structure." This forces the model to engage in internal reasoning: to correctly predict masked tokens and fill the information gaps we created, the model must perform implicit inference internally. This drives the model to more deeply understand and integrate unmasked contextual information, building stronger logical connections and semantic associations within the neural network. Simultaneously, masked tokens function as a "chain of reasoning" during training, simulating intermediate thought steps in human inference. This enhances the model's ability to reason and evaluate complex tasks.

5.3 Implicit Reasoning Analysis Capability

To verify whether our method genuinely enhances the implicit reasoning capabilities of large language models, we visualized the final layer attention weights of the Qwen3-14B model on the Mohler dataset, with results shown in Figure 3. By comparing the performance of the baseline model and the MaPLe model, we observe that our approach enables the model to focus not only on the output segment but also on multiple regions within the prompt. Specifically, by comparing the first three subfigures (Base model) with the last three subfigures (MaPLe model) in Figure 3, we can clearly observe this shift. This indicates that MaPLe enables the model to synthesize information from multiple parts during reasoning to arrive at the final inference result, which requires robust reasoning capabilities. For a more detailed comparative analysis, please refer to Appendix A.7.

5.4 Score Distribution Analysis

To further validate that our method genuinely activates the model's implicit reasoning capabilities rather than merely falling within the inherent scoring preference range of large language models(Zheng et al., 2023), we compared the scoring distributions generated by Qwen3-14B across three datasets with manually annotated distributions. Figure 4 demon-

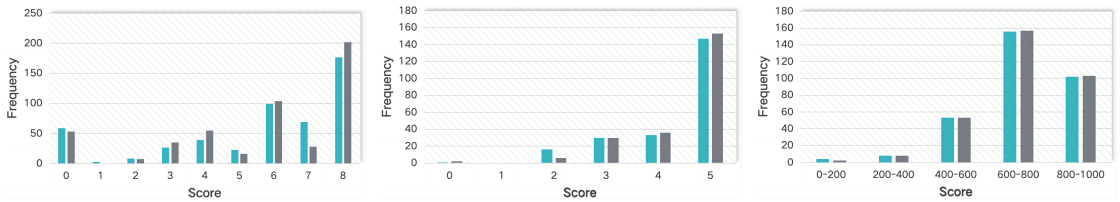

Figure 4: Distribution figures of actual versus predicted scores across the AES ENEM, Mohler, and Chinese-L2 datasets. In each figure, the left side shows human-labeled scores, while the right side displays model-predicted scores.

strates high consistency in distribution shapes between the two, indicating that the model is no longer constrained by its inherent scoring preferences and can produce results more aligned with human judgments. Furthermore, this result demonstrates that MaPLe provides the model with explicit "alignment direction" during scoring inference, prompting its scoring behavior to more stably align with human standards. This highlights its robustness and effectiveness in automated scoring tasks.

5.5 Evaluation Feedback Analysis

Current research also focuses on enabling LLMs to generate scoring feedback to further elucidate their interpretability. While this is not the primary focus of this paper, we additionally verified whether our method possesses the potential to generate scoring feedback. We performed zero-shot reasoning using our frozen MaPLe checkpoint by incorporating relevant prompt words to guide the model in generating scoring rationale. Appendix A.8 presents our generated prompt templates and feedback examples, demonstrating our method's capability for feedback generation.

6 Conclusion

The MaPLe method we propose is simple and efficient. It does not require designing task-specific prompts or introducing additional multi-dimensional labels, and is compatible with existing large language models. This method randomly masks tokens in the input prompt, with the masked tokens forming partial labeling, providing clear direction for reasoning. It also simulates "thinking chains" and "reasoning paths," activating the implicit reasoning abilities of LLMs and significantly improving evaluation accuracy. Furthermore, MaPLe is task-agnostic and can be widely applied to various scoring tasks, from essay scoring to short-answer question evaluation, showing outstanding performance across multiple languages and scenarios.

Limitations

Domain Extension Potential. Although this study has validated the effectiveness of the MaPLe method across multiple scoring task datasets, its potential for extension to broader domains remains to be further explored. As LLMs increasingly penetrate scientific domains, MaPLe holds promise for migration to complex tasks such as drug discovery (Zheng et al., 2024), disease diagnosis (Zhou et al., 2024), and weather forecasting (Wang and Karimi, 2024). These directions constitute research topics worthy of exploration in future work.

Capability to Generate Detailed Feedback. While our method demonstrated preliminary feedback generation on the Chinese-L2 dataset, its feedback generation mechanism requires further training and validation on more complex tasks. Generating high-quality, inter-

pretable feedback remains a key research focus (Nguyen et al., 2023; Ouyang et al., 2022; Leng et al., 2025), and we are committed to advancing the practical application of LLMs in this domain.

Prompt Design and Model Coverage. This study employed uniformly designed prompts for the evaluation task and has not yet systematically validated the method's dependence on different prompts. Furthermore, constrained by computational resources, experiments were conducted solely on two open-source models Qwen3 and Llama3, without extending to larger-parameter models or closed-source models. The original intent of this work was to propose an efficient, generalizable, and resource-conserving research approach. Future efforts will further validate the method's universality across a broader range of model scales and architectures.

Ethics Statement

We confirm that all authors of this study have adhered to the ACL ethical guidelines and recommended code of conduct. All code and datasets used in this research are publicly accessible, and we have fully cited the sources of all datasets. We believe this study poses no potential risks. It should also be noted that all research content, experimental design, data results, conclusion analysis, and academic viewpoints in this paper were independently completed and are the sole responsibility of the author. GPT-5.2 (OpenAI, 2025) was used solely to enhance the quality of textual expression and did not participate in any creative research processes nor automatically generate any substantive academic content.

References

Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*, pages 61–78. Springer.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. Proceedings of the 19th Workshop on Innovative Use of NLP for Building ....

Tzu-Lin Chang, Keng-Pei Lin, and Zong-Shun Chen. 2024. Automatic short-answer grading with a pseudo-siamese neural network.

Weizhe Chen, Sven Koenig, and Bistra Dilkina. 2024. Reprompt: Planning by automatic prompt engineering for large language models agents. *arXiv preprint arXiv:2406.11132*.

Niharika Dadu, Harsh Vardhan Singh, and Romi Banerjee. 2025. Grade guard: A smart system for short answer automated grading. *arXiv preprint arXiv:2504.01253*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ta Nguyen Binh Duong and Chai Yi Meng. 2024. Automatic grading of short answers using large language models in software engineering courses. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–10. IEEE.

Zhiyuan Fan, Weinong Wang, Debing Zhang, and 1 others. 2024. Sedareval: Automated evaluation using self-adaptive rubrics. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16916–16930.

Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 93–103.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024a. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 300–304.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024b. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 300–304.

. Kamesh, R. 2024. Think beyond size: Adaptive prompting for more effective reasoning.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Peichao Lai, Kexuan Zhang, Yi Lin, Linyihan Zhang, Feiyang Ye, Jinhao Yan, Yanwei Xu, Conghui He, Yilei Wang, Wentao Zhang, and 1 others. 2025. Sas-bench: A fine-grained benchmark for evaluating short answer scoring with large language models. *arXiv preprint arXiv:2505.07247*.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024a. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024b. Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv preprint arXiv:2404.04941*.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024c. Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv preprint arXiv:2404.04941*.

Yongqi Leng, Renren Jin, Yue Chen, Zhuowen Han, Ling Shi, Jianxiang Peng, Lei Yang, Juesi Xiao, and Deyi Xiong. 2025. Praetor: A fine-grained generative llm evaluator with instance-level customizable evaluation criteria. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10386–10418.

Ben Mann, Nick Ryder, Melanie Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1(3):3.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 752–762.

Huy A Nguyen, Hayden Stec, Xinying Hou, Sarah Di, and Bruce M McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning*, pages 278–293. Springer.

OpenAI. 2025. Introducing gpt-5. https://openai.com. August 7, 2025.

N OpenAI. 2023. Gpt-4 technical report: Tech. rep.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Helia Hashemi Jason Eisner Corby Rosset. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts: Helia hashemi jason eisner corby rosset benjamin van durme chris kedzie.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2024. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *CoRR*, abs/2411.16337.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024. A new benchmark for automatic essay scoring in portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 228–237.

Hanling Wang, Banghao Chi, Yufei Wu, Kexin Chen, Di Wu, Songning Liu, Yiwei Li, Hanyan Niu, and Xiaohui Zhu. 2025. Llmarking: Adaptive automatic short-answer grading using large language models. In *Proceedings of the Twelfth ACM Conference on Learning@ Scale*, pages 105–115.

Yang Wang and Hassan A. Karimi. 2024. Exploring large language models for climate forecasting. *CoRR*, abs/2411.13724.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Mingfeng Xue, Yunting Liu, Xingyao Xiao, and Mark Wilson. 2025. Automatic prompt engineering for automatic scoring. *Journal of Educational Measurement*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xianjun Yang, Wei Cheng, Xujiang Zhao, Linda Petzold, and Haifeng Chen. 2023. Dynamic prompting: A unified framework for prompt tuning. *ArXiv*, abs/2303.02909.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454*.

Haifeng Zhao, Yuguang Jin, and Leilei Ma. 2025. Dynamic prompt adjustment formulti-label class-incremental learning. In *International Conference on Brain Inspired Cognitive Systems*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T. May, Geoffrey I. Webb, Shirui Pan, and George Church. 2024. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *CoRR*, abs/2409.04481.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, Liqiao Xia, Jeremy Yeung, Daochen Zha, Genevieve B. Melton, Mingquan Lin, and Rui Zhang. 2024. Large language models for disease diagnosis: A scoping review. *CoRR*, abs/2409.00097.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*.

A Appendix

A.1 Comparison of Computing Resources

This study systematically reviews the computational resources required for current large language model-based automatic scoring tasks and compares our proposed MaPLe method with existing representative approaches. The specific results are summarized in Table 4. The data indicates that mainstream approaches

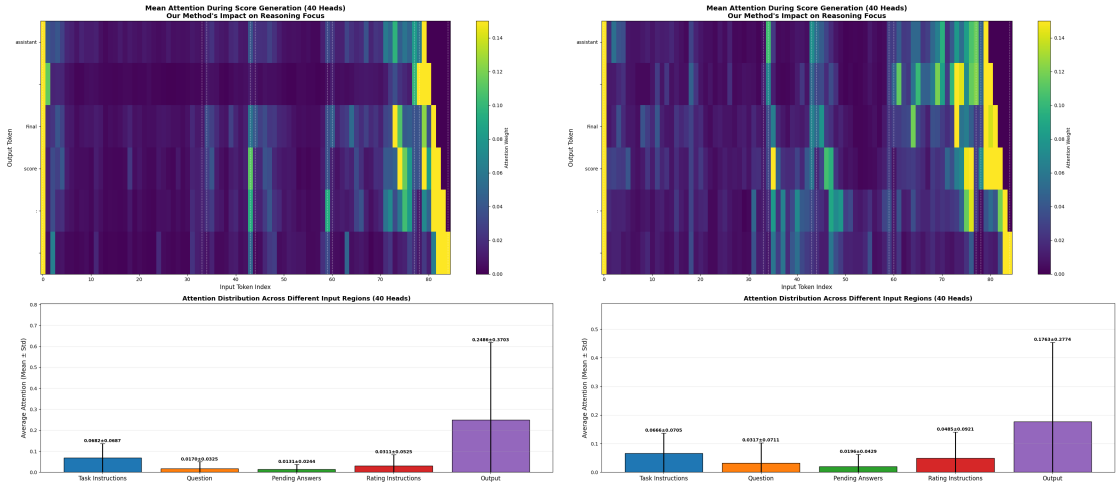

Figure 5: Comparison between Base and MaPLe. The left figure shows the Base method, while the right figure illustrates the MaPLe method. The upper section displays which tokens were focused on during inference-based score generation, while the lower section presents the mean and standard deviation of attention weights across regions (where a smaller standard deviation is preferable).

Method	computing resources
Zheng et al., 2023	8 A100 (80GB) GPUs
Duong and Meng, 2024	GPT-3.5-Turbo API
Fan et al., 2024	128 H100 (80GB) GPUs
Lee et al., 2024c	GPT-4 API
Lee et al., 2024c	1 A40 (48GB) GPUs
Lai et al., 2025	4 A800 (80GB) GPUs
Leng et al., 2025	8 A100 (80GB) GPUs
RMM(ours)	2 A40 (48GB) GPUs

Table 4: Comparison of computing resources

typically rely on large-scale computational infrastructure, such as multiple high-end GPUs (e.g., A100/H100) for parallel training, or high-capacity GPU memory to support full-parameter fine-tuning and large-scale inference. This imposes significant barriers for universities and research institutions with generally constrained computational resources. In contrast, the proposed MaPLe method integrates parameter-efficient fine-tuning techniques (e.g., LoRA) with lightweight prompt design. It completes all training and evaluation tasks using only two A40 GPUs, substantially reducing hardware configuration and VRAM occupancy requirements. This design enables the method to maintain comparable or even superior performance to existing approaches while substantially enhancing feasibility and practical value in resource-constrained environments, providing a viable technical pathway for broader research teams to pursue related explorations.

A.2 Prompt Template

This study proposes and designs a prompt template with strong versatility. To ensure the method's reusability and transferability, we directly apply the same template across different tasks and datasets without additional task-specific rewriting or customization. Structurally, the template integrates both comprehensive scoring criteria and non-scoring criteria formats. This enables the model to adhere to unified evaluation standards while maintaining stable comprehension of task-critical information and output constraints. Experimental results demonstrate consistent and significant

applicability and robustness across three experimental datasets, reflecting strong cross-task generalization capabilities. Detailed design specifications and template composition are presented in Tables 5 and 6.

A.3 Experimental Details

LLMs are deployed on 2 NVIDIA A40 GPUs (48GB VRAM) to achieve high-performance real-time processing. GPU acceleration is implemented using Python 3.12+, PyTorch 2.6, and CUDA 12.4. Asynchronous inference is enabled via vLLMs to enhance throughput (Kwon et al., 2023). Transformers is leveraged for optimized integration of pre-trained models, facilitating model management. For more experimental details from our study in Table 7.

A.4 Detailed Experimental Setup for Praetor and MTS

A.4.1 Detailed Experimental Setup for Praetor

Praetor A fine-grained generative LLMs evaluator trained using multi-task learning, supporting instance-level customizable evaluation criteria. It can evaluate LLMs through single-sentence scoring or pairwise comparisons, supporting both English and Chinese languages, while offering high flexibility in setting evaluation standards.

In this paper, we applied the prompt design methodology from Praetor to construct corresponding prompts on the Mohler and Chinese-L2 datasets, with specific designs detailed in Table 8,9. Based on the code provided by Praetor, we conducted comparative experiments. The reason for selecting only the Mohler and Chinese-L2 datasets for experimentation is that this method currently supports only English and Chinese.

A.4.2 Detailed Experimental Setup for MTS

MTS A zero-shot prompting framework that automatically decomposes questions/essay into distinct dimensions and generates scoring criteria for each dimension. Subsequently, through multi-round dialogues, it guides LLMs to extract scores for each dimension. Each dialogue round completes the scoring for its corresponding dimension based on its specific criteria, ultimately yielding a total score.

This paper utilizes the prompt provided by MTS to invoke GPT-5.2 (OpenAI, 2025), automatically generating scoring criteria for each dimension of the AES ENEM and Chinese-L2 datasets (experiments were not conducted on the Mohler dataset as scoring dimensions were not identified). The dimension-specific scoring criteria for the AES ENEM dataset are presented in Tables 10, 11, 12, 13, and 14 (these dimensions are provided in the AES ENEM paper), while the scoring criteria for each dimension of the Chinese-L2 dataset are shown in Tables 15 and 16 (these dimensions are extracted from the overall scoring criteria).

A.5 Comparison of MaPLe with Existing Methods

To comprehensively evaluate MaPLe's competitiveness, we systematically compared it against multiple representative methods proposed in recent years. As shown in Table 17 (Mohler dataset) and Table 18 (AES ENEM dataset), we included not only state-of-the-art approaches based on large language models but also traditional methods leveraging lightweight models (e.g., BERT, XGBoost) to conduct a cross-paradigm, cross-model-scale performance assessment.

Experimental results demonstrate that MaPLe exhibits significant advantages in automated short-answer grading tasks. On the Mohler dataset, our method substantially out-

Prompt Template

###Task Description###
You are a professional {subject}. You will assign a reasonable score to answers regarding {subject}-related questions/essay, with the score not exceeding {score_range} points.
###Evaluation Materials###
Question content: {question/essay}
Rated answers: {answer}
###Output format###
Directly provide the final score. The output format is: Final score: {score}
###End marker and Target output###
Final score:

Table 5: Prompt template without scoring criteria

Prompt Template

###Task Description###
You are a professional {subject}. You will assign a reasonable score to answers for {subject}-related questions/essay based on the scoring criteria, with the score not exceeding {score_range} points.
###Evaluation Materials###
Scoring criteria: {scoring criteria}
Question content: {question/essay}
Rated answers: {answer}
###Output format###
Directly provide the final score. The output format is: Final score: {score}
###End marker and Target output###
Final score:

Table 6: Prompt template with scoring criteria

Parameters	Details
Batch_size	24
Epoch_num	8
Learning_rate	3e-04
Lora_rank	64
Lora_alpha	16
Lora_dropout	0.05
Temperature	0
Top_k	1
Top_p	1

Table 7: Experimental details across all datasets

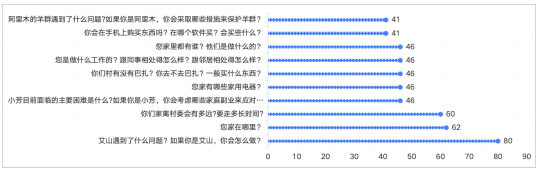

Figure 6: Question content distribution figure (showing only the top 10 most frequent questions).

performs all comparison baselines—including high-performing models like PSNN and large-model approaches such as GPT-4o—on key metrics like RMSE. This confirms that MaPLe's effectiveness in enhancing scoring accuracy stems not only from the powerful capabilities of large models but also from its unique training design. Furthermore, on the AES ENEM dataset, MaPLe achieved the best results, even surpassing the inter-rater reliability levels reported in the original paper for this dataset, further highlighting the method's robustness and practicality in complex scoring tasks.

A.6 Detailed Implicit Reasoning Analysis

In Section 5.3, we conducted a preliminary analysis of the information regions the model focuses on when generating scores. Here, we will elaborate on our analytical process in detail. First, we extracted the attention weights from the final layer of both the Base and MaPLe methods (using the Qwen3-14B model as the foundation) on the Mohler dataset. This layer comprises 40 attention heads. We computed the mean and standard deviation of attention across these 40 heads at their respective positions. Subsequently, we segmented the model input into five distinct regions: Task Instructions, Question, Pending Answers, Rating Instructions, and Output. By quantifying the model's attention distribution across these regions during final score generation, we examined whether it relies solely on the immediate context tokens or can effectively gather information across regions.

To delve deeper into the analysis, we further plotted heatmaps of the model's reasoning process during score generation and calculated the standard deviation of the final attention layer (as shown in Figure 5). The lower half of Figure 5 demonstrates that our method consistently exhibits a stable low standard deviation across generated score regions. This indicates that the 40 attention heads maintain consistent reasoning paths, enabling the model to clearly identify key areas. The heatmap in the upper part of Figure 5 indicates that during scoring inference, the model simultaneously focuses on multiple information regions rather than being confined to local context, thereby supporting a more comprehensive reasoning process. These findings further validate that our approach effectively activates the latent reasoning capabilities of LLMs.

A.7 Dataset Overview

A.7.1 Construction Process of the Chinese-L2 Dataset

We selected 5,000 data entries from our research group's projects. These entries primarily aim to assess second-language learners' proficiency through question-answering tasks, covering multiple scenarios: "single-question

Prompt Template
Task Description ### You are a lecturer for a data structures course (computer science major). The system will present you with questions related to the data structures course and students' responses. Please evaluate the accuracy of the responses and grade them based on their correctness. 1. Write a detailed commentary on the answer. This commentary will be used to assess the answer's quality. 2. After completing the review, provide a decimal score between score_range representing your assessment of the answer. A higher score indicates better quality. 3. Output format: Your review of the response Overall score: Your rating for the answer 4. Generate content strictly adhering to this format. Do not add any additional opening, closing, or explanatory text. ### User question ### question ### Answer to be evaluated ### answer

Table 8: Using the Praetor method on the prompt template for Mohler dataset

Prompt Template
Task Description ### You are a Chinese linguistics instructor. The system will present you with an open-ended question and a second language learner's response to it. Evaluate the response according to the grading criteria and assign a score based on its accuracy. 1. Write a detailed commentary on the answer. This commentary will be used to assess the answer's quality. 2. After completing the review, provide a decimal score between score_range representing your assessment of the answer. A higher score indicates better quality. 3. Output format: Your review of the response Overall score: Your rating for the answer 4. Generate content strictly adhering to this format. Do not add any additional opening, closing, or explanatory text. ### User question ### question ### Answer to be evaluated ### answer ### Scoring Criteria ### scoring criteria

Table 9: Using the Praetor method on the prompt template for Chinese-L2 dataset

Formal Language
0: The language does not comply with written language standards at all, contains serious grammatical errors, frequent spelling mistakes, is extremely casual and informal, and does not meet academic or formal writing requirements in any way. **1-40**: The language contains quite a few grammatical or spelling errors, some word choices are inaccurate or unsuitable for a formal style, the expression is somewhat chaotic or irregular, giving an overall unprofessional impression and lacking clear logic. **41-80**: The language meets basic written language standards, with occasional grammatical or spelling errors that have little impact on overall understanding; word choices are relatively formal but not precise, and some expressions may appear unclear or lack professionalism. **81-120**: The language is fairly standard, with occasional minor grammatical or spelling errors; overall expression is clear and basically meets the requirements of formal written language, with accurate, concise, and relatively professional word choices; the language style is consistent, though some parts may lack greater depth or refinement. **121-160**: The language is very standard, with almost no grammatical, spelling, or logical errors; word choices are appropriate and accurate, fully meeting the requirements of written and academic language; expression is concise and clear, with a consistent and professional style, demonstrating a strong level of formality. **161-200**: The language perfectly meets written language standards, with no grammatical or spelling errors; every word choice and sentence structure is extremely precise, showing a high degree of formality and academic quality; logic is tight, expression is clear and fluent, fully complying with written cultural standards, demonstrating exceptional language ability.

Table 10: Scoring criteria for the formal language dimension using the MTS method on the AES ENEM dataset

Understanding the Task
0: Completely does not understand the task requirements, and the response content is entirely off-topic. No relevant knowledge is applied, the content is incoherent, and it lacks a reasonable logical structure. **1-40**: Has a shallow understanding of the task; part of the response is relevant, but there is a lack of deep understanding of key concepts. Knowledge application is insufficient, reasoning is confused, and the structure is loose or unclear, failing to effectively support the main argument. **41-80**: Has some understanding of the task, but still not deep enough. Able to apply some related knowledge and develop some reasonable arguments, but reasoning is inadequate, the structure is somewhat messy or not tight, and the organization is not optimal within the constraints of an essay. **81-120**: Has a good understanding of the task and can effectively analyze and argue using relevant knowledge. The theme is fairly well developed, reasoning is strong, the structure is reasonable, and content is organized well within the given framework, though there is room for improvement, such as in depth or comprehensiveness. **121-160**: Has a very deep understanding of the task and can accurately apply knowledge from various disciplines for analysis, fully developing arguments. Reasoning is clear and strong, structure is tight, showing strong logic and coherence. Completes the task within the essay structure while also deepening the discussion of the topic. **161-200**: Has an extremely thorough understanding of the task and can integrate concepts from multiple disciplines for profound analysis, developing arguments comprehensively from multiple angles. Reasoning is comprehensive and precise, logic is rigorous and deep, and the theme is perfectly presented within structural constraints. The text structure is flawless, content is rich and highly persuasive, fully demonstrating a high level of academic understanding and critical thinking skills.

Table 11: Scoring criteria for the understanding the task dimension using the MTS method on the AES ENEM dataset

Organization of Information
0: The selection of information is completely irrelevant or misleading, unable to effectively connect different pieces of information, and the content lacks structure. There is no clear argument or viewpoint, and the reasoning cannot support any perspective. **1-40**: The selection of information is relatively random, lacking focus, poorly connected, and some information is irrelevant or repetitive. The organization is loose, the content lacks clear hierarchy or logic, and some arguments are not strongly supported. **41-80**: The selection of information is basically reasonable, but there are still some irrelevant or redundant contents. The connection between information is preliminary, the structure is somewhat loose, some arguments are basically supported, but the overall reasoning is insufficient and there is room for improvement. **81-120**: The selection of information is clear and mostly relevant, connections are reasonable, and key information can be effectively organized and presented. The overall structure is clear, logic is coherent, and the arguments are fairly well supported, though some reasoning may still be insufficiently in-depth or have minor gaps. **121-160**: The selection of information is precise and relevant, connections are reasonable and tight, effectively supporting the arguments. The structure is clear, content is well-organized, reasoning is strong, fully supporting the position with information, facts, and viewpoints, and the expression is very clear. **161-200**: The selection of information is perfect, with all information closely connected and precisely supporting the arguments. The structure is rigorous, layered, with in-depth and persuasive reasoning. The integration and organization of information demonstrate strong logic and critical thinking. The arguments are strongly supported, showing high-level writing skills.

Table 12: Scoring criteria for the organization of information dimension using the MTS method on the AES ENEM dataset

Knowing Argumentation
0: Shows no understanding of the language mechanisms of argumentation; the argument lacks clear structure, logic is chaotic, reasoning is unsupported or contains obvious logical errors. Language mechanisms are used improperly, failing to effectively convey the argument. **1-40**: The argument structure is confusing or incomplete, reasoning lacks logic, and evidence is insufficient. The use of language mechanisms is somewhat stiff, reasoning or refutation is ineffective, and the argument lacks persuasiveness. **41-80**: The argument structure is basically clear, but there are still some logical gaps or imperfect reasoning. Able to use basic language mechanisms, such as causality or analogy, but not fully developed; the refutation part is still insufficient. **81-120**: The argument structure is relatively clear, and reasoning is fairly rigorous. Language mechanisms are used appropriately, effectively employing causality, comparison, etc., to support points. Able to respond to counterarguments, enhancing the depth and persuasiveness of the argument, though some parts may still lack depth. **121-160**: The argument structure is clear and rigorous, logical reasoning is solid, and evidence is strong. Able to skillfully use language mechanisms such as causality, analogy, comparison, and refutation, enhancing the persuasiveness and depth of the argument. Handles counterarguments very reasonably, demonstrating strong argumentation ability. **161-200**: The argument structure is extremely clear and rational, reasoning is flawless, points are fully supported by evidence, and logic is rigorous and perfect. Language mechanisms are used with ease, flexibly employing multiple mechanisms (such as refutation, causality, analogy, etc.) to effectively support the argument. Responses to opposing views are very strong, showing an exceptionally high level of argumentation.

Table 13: Scoring criteria for the knowing argumentation dimension using the MTS method on the AES ENEM dataset

Solution Proposal
0: The solution lacks innovation or is completely infeasible, ignores human values, has negative impacts on individuals or groups, completely disregards social and cultural diversity, and has a negligible effect on the problem. **1-40**: The solution lacks effectiveness or practical feasibility, has certain negative impacts or biases, fails to adequately respect human values or the diversity of social cultures, provides insufficient details, and is difficult to implement. **41-80**: The solution has some innovation and feasibility but may face challenges during implementation. It respects basic human values but does not fully consider social and cultural diversity. Implementation details are somewhat lacking and it may have adverse effects on certain groups. **81-120**: The solution is fairly innovative and considers practical feasibility while providing some implementation details. It adequately respects basic human values and makes efforts to consider social and cultural diversity. It can be effectively implemented in diverse cultural contexts, but may still require further optimization or improvement. **121-160**: The solution is innovative and highly feasible, with thorough details and consideration of potential implementation obstacles. It respects and embodies basic human values and has good adaptability to cultural diversity, allowing effective application across different cultural contexts. **161-200**: The solution is extremely innovative and practically feasible, with detailed and comprehensive consideration of all possible obstacles, and proposes concrete, workable countermeasures. It fully respects core human values and has high adaptability in global or multicultural contexts, effectively reflecting social and cultural diversity, and has a very significant impact on solving the problem.

Table 14: Scoring criteria for the solution proposal dimension using the MTS method on the AES ENEM dataset

Completeness of Response
0: Unable to answers or provides completely irrelevant answers. **1-3**: Barely answers some questions; answers are incomplete, sentences are disjointed or lack key information; frequent pauses and repetitions prevent clear communication of all key points. **4-6**: Answers all questions but with simple structure; answers lack detail or fail to fully cover some points. Expression is relatively concise, with occasional pauses, repetitions, or grammatical errors. **7-8**: Effectively answers all questions using one or more complete sentences, with substantial and detailed content.

Table 15: Scoring criteria for the completeness of response dimension using the MTS method on the Chinese-L2 dataset

Answer Completeness
0: The answer is completely off-topic or does not address the question at all. **1-3**: The answer has a weak connection to the question, contains some irrelevant information, and features significant pauses and repetition. **4-6**: The answer is generally on topic, conveying the main information, but may be somewhat off-topic or not fully aligned with the question. **7-8**: The answer is entirely on topic, clearly and accurately expressed, and closely aligned with the core of the question.

Table 16: Scoring criteria for the answer completeness dimension using the MTS method on the Chinese-L2 dataset

Models	**Mohler**		
	$\Delta \downarrow$	$\lvert\Delta\rvert \downarrow$	$\rho \uparrow$
GPT4o (Ferreira Mello et al., 2025)	0.69	1.01	-
Upstage Solar Minin (Dadu et al., 2025)	-	1.08	-
Gemini 1.5 Flash (Dadu et al., 2025)	-	1.12	-
GPT4o Minin (Dadu et al., 2025)	-	1.02	-
Upstage Solar Minin (Dadu et al., 2025)	-	0.93	-
PSNN (Chang et al., 2024)	-	0.705	0.774
BERT (Ferreira Mello et al., 2025)	0.64	1.11	-
XGB (Ferreira Mello et al., 2025)	0.69	1.18	-
Qwen3-8B (ours)	0.32	**0.54**	0.86
Qwen3-14B (ours)	**0.30**	0.55	**0.86**
Llama-3.1-8B-Instruct(ours)	0.36	0.59	0.83

Table 17: Results of comparing other methods with the Mohler dataset

Models	AES ENEM		
	$\Delta \downarrow$	$\lvert\Delta\rvert \downarrow$	$\rho \uparrow$
Grader A (Silveira et al., 2024)	-	264.40	-
Grader B (Silveira et al., 2024)	-	237.55	-
Qwen3-8B(ours)	59.56	87.42	0.83
Qwen3-14B(ours)	**54.86**	**81.25**	**0.85**
Llama-3.1-8B-Instruct(ours)	59.31	88.15	0.83

Table 18: Results of comparing other methods with the AES ENEM dataset

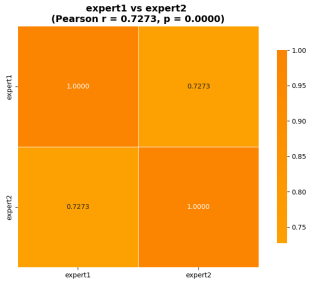

Figure 7: Heatmap of correlation among figure.

single-answer," "multiple-question multiple-answer," and "contextual Q&A." The ratio of these scenarios is 3:5:2. Each question is scored out of 8 points. The dataset comprises 181 distinct, fully open-ended questions spanning daily life, social culture, and other domains. Figure 6 displays the top 10 most frequently occurring questions.

This dataset was constructed through manual annotation. First, 5,000 candidate samples were selected from within the research team. Subsequently, the samples were randomly and independently assigned to two groups of domain experts for double-blind annotation based on a unified scoring standard. After annotation, the Pearson correlation coefficient between the two groups' scores was calculated and visualized in a heatmap (Figure 7). Results demonstrated a high positive correlation between expert evaluations. For samples with score discrepancies of $\geq$2 points, third-party experts were engaged for review and refinement, ultimately forming the final dataset. This dataset encompasses diverse scenarios and broad problem coverage, enabling comprehensive and nuanced assessment of examinees' overall capabilities.

Additionally, it should be noted that our dataset complies with the relevant regulations of the ethics review committee, does not involve any personal information, and paid data annotation experts $2,800.

A.7.2 Mohler Dataset Overview

The Mohler Dataset is an English-language automated short-answer grading dataset comprising 80 questions sourced from assignments in a data structures course. These questions span ten assignments and two exams, encompassing 2,273 student responses. Two graduate students in related fields independently graded these answers using a scale from 0 (completely incorrect) to 5.

A.7.3 AES ENEM Dataset Overview

The AES ENEM dataset is a Portuguese-language automated essay grading dataset comprising 3,586 essays sourced from the websites Educação UOL and Brasil Escola. Two experienced annotators independently scored these essays across five dimensions, each with a maximum score of 200 points, resulting in a total possible score of 1,000 points.

A.8 Feedback Prompt Template and Examples

To validate the potential of our method in generating feedback, we directly invoked the trained parameters for zero-shot reasoning. Tables 19 and 20 present the prompt templates employed, while Tables 21 showcase feedback examples generated across two datasets. The results demonstrate that the model outputs high-quality feedback without additional fine-tuning, fully illustrating how our approach

effectively taps into the implicit reasoning capabilities of LLMs.

Prompt Template

###Task Description###
You are a professional {subject}. You will assign a reasonable score to answers regarding {subject}-related questions/essay, with the score not exceeding {score_range} points. Not only should you provide the final score, but you should also explain the rationale behind it.
###Evaluation Materials###
Question content: {question/essay}
Rated answers: {answer}
###Output format###
Provide the final score and the rationale for the scoring. The output format is: Final score: {score}. Reasons for the scoring: {Reasons}
###End marker and Target output###
Final score: Reasons for the scoring:

Table 19: Feedback Prompt template without scoring criteria

Prompt Template

###Task Description###
You are a professional {subject}. You will assign a reasonable score to answers for {subject}-related questions/essay based on the scoring criteria, with the score not exceeding {score_range} points. Not only should you provide the final score, but you should also explain the rationale behind it.
###Evaluation Materials###
Scoring criteria: {scoring criteria}
Question content: {question/essay}
Rated answers: {answer}
###Output format###
Provide the final score and the rationale for the scoring. The output format is: Final score: {score}. Reasons for the scoring: {Reasons}
###End marker and Target output###
Final score: {score}. Reasons for the scoring: {Reasons}

Table 20: Feedback Prompt template with scoring criteria

Feedback Examples
Question: Have you been to Tiananmen Square? Answer: I went. Final score: 6 ###Reasoning for Scoring: ### The response is on topic, clearly expressed, with minimal pauses, repetitions, and linguistic errors. However, the content is incomplete, lacking subjects and objects. Simply stating "I went" is insufficient. If the question asked how one went to Tiananmen Square, the answer would be even less appropriate. Since the question only asks whether the person has been there, the content is fundamentally on topic and clearly expressed, warranting a score of 6.

Table 21: Feedback examples on the Chinese-L2 dataset