Learning from Scoring Disagreements: Contrastive Error Mining for Efficient and Robust LLM-based Assessment

Lei Chen[1], **Tengteng Cheng**[1], **BoYu Gao**[1 2*], **Zitao Liu**[1], **Weiqi Luo**[1]

[1]Guangdong Institute of Smart Education, Jinan University, Guangzhou, China
[2]College of Cyber Security, Jinan University, Guangzhou, China
leibnizchen@foxmail.com, chengt19980421@163.com, {bygao, liuzitao, lwq}@jnu.edu.cn

Abstract

Automated grading of student responses still faces numerous challenges, particularly when dealing with complex and ambiguous answers. In particular, large models are prone to scoring bias when handling uncertain responses, and few-shot reasoning methods often lack stability, which limits their applicability in real educational scenarios. To tackle these challenges, we propose the Contrastive Error Mining and Fine-Tuning (CEM-FT) framework, which automatically identifies high-value hard samples by analyzing scoring disagreements between a full fine-tuned model and a few-shot model. A lightweight LoRA adapter is then trained on these samples to refine model performance with minimal computational overhead. Experiments on the SciEntsbank, Beetle, and Mohler datasets show that CEM-FT can improve QWK by up to 3.9% compared to the fine-tuned Qwen model on SciEntsbank datasets, which is a significant improvement over the few-shot baseline. The proposed framework substantially enhances both scoring accuracy and consistency, providing a practical, robust solution for reliable automated assessment with large language models.

Code — https://github.com/leibnizchen/LLMScoring

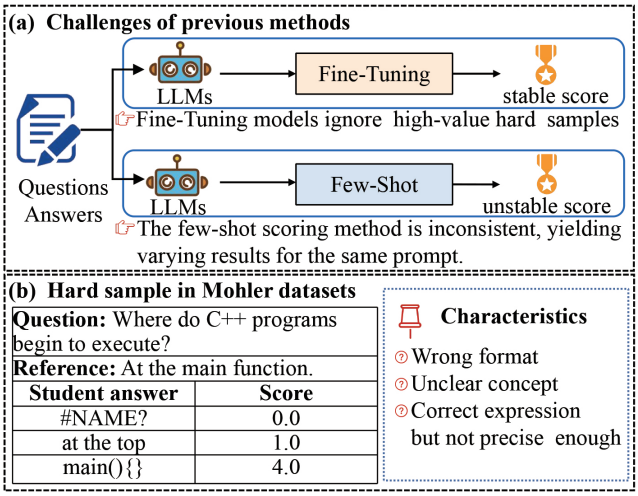

Figure 1: Challenges of current mainstream methods and examples of high-value hard samples

Introduction

With the rapid advancement of artificial intelligence, particularly in natural language processing and large language models, automated grading systems are becoming increasingly pivotal in educational contexts. (Kooli and Yusuf 2025; Liang 2025; Lee et al. 2024). Such systems not only substantially reduce teachers' workload in grading subjective responses but also enhance instructional efficiency and student engagement through timely, accurate feedback. In recent years, the powerful language understanding and generation capabilities of large models such as GPT-4 (Achiam et al. 2024), Qwen (Bai et al. 2023), and LLama (Touvron et al. 2023) have shown broad prospects in open-ended answer grading tasks. An increasing number of studies have focused on applying large language models to the automatic evaluation of student responses. For example, (Jiang and Bosch 2024) proposed using GPT-4o with carefully designed prompt-word templates for essay scoring, introducing a new scoring paradigm driven by prompt-word engineering. (Stahl et al. 2024) incorporated scoring criteria into prompts to enable large models to generate more accurate scores. (Chamieh, Zesch, and Giebermann 2024; Funayama et al. 2025) employed fine-tuning techniques to adapt large models for short answer scoring, aligning them more closely with expert-level evaluations.

However, despite the initial results of large language models on multiple educational scoring benchmarks, they still face two core challenges in practical applications. As shown in Figure 1(a), existing methods suffer from the following major issues: First, current mainstream fine-tuning strategies (Chang and Ginter 2024; Chamieh, Zesch, and Giebermann 2024; Latif and Zhai 2024) generally employ uniform sampling for training, neglecting "high-value hard samples" that critically influence model behavior, as shown in Figure 1(b). This strategy often results in models retaining specific systematic errors after fine-tuning, limiting the reliability and stability of their scoring results. Second, while prompt-based few-shot learning methods (Chamieh, Zesch,

*Corresponding author.

and Giebermann 2024; Latif and Zhai 2024) offer particular advantages in low-resource scenarios, their performance fluctuates significantly when faced with complex semantic structures (high-value hard samples). They also rely heavily on the design of prompt templates and the selection of examples, resulting in significant deficiencies in consistency and generalization. These two methods are prone to inconsistent or even biased scoring when dealing with complex, ambiguous, or novel answers, further weakening the fairness and reliability of the automatic scoring system in actual teaching scenarios.

To address the challenges above, we introduce an efficient and robust framework for contrastive error mining and fine-tuning (CEM-FT). The framework automatically identifies and collects key samples with high error potential by analyzing disagreements in scoring results between the fine-tuned model and the few-shot large model. On this basis, we use a lightweight LoRA to correct scoring biases with minimal training overhead systematically.

Extensive experiments across multiple real-world scoring datasets, including SciensBank, Beetle, and Mohler, show that CEM-FT significantly outperforms existing fine-tuned models and achieves the highest QWK improvement of 20.4% over few-shot methods. Our research shows that automatic scoring should not only focus on improving performance, but also systematically identify and correct model weaknesses. CEM-FT injects precise, controllable capabilities into the automatic scoring of large models, taking a key step toward transitioning from generalization to deployment reliability.

Related Work

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in language understanding and generation, offering a novel paradigm for automated scoring. Unlike earlier encoder-based models, such as the Transformer family of grading (Liu et al. 2019, 2020), LLMs with billions to trillions of parameters can perform complex natural language tasks under few-shot or even zero-shot conditions. In addition, they can be adapted to specific scoring rubrics through parameter-efficient fine-tuning. In few-shot scenarios, prompt engineering has emerged as a critical factor. Studies such as (Chang and Ginter 2024) systematically evaluated the Few-Shot capabilities of GPT-3.5 (Brown et al. 2020) and GPT-4 (Achiam et al. 2024) in scoring Finnish-language responses. Their findings indicate that well-crafted prompts, incorporating clear criteria and diverse examples, allow LLMs to achieve high consistency with human expert evaluations. Similarly, (Jiang and Bosch 2024) used the gpt-4-1106-preview model and proposed two prompting approaches: a direct scoring mode and a two-step method that splits the prompt into a direct-scoring and a reasoning-analysis component. A systematic comparison revealed that incorporating a reasoning step significantly improves score accuracy. Furthermore, (Stahl et al. 2024) conducted extensive experiments to validate the feasibility of zero-shot and few-shot prompting strategies to score long-form answers, while (Henkel et al. 2024) investigated various versions of GPT and prompt engineering techniques across domains (science and history) and age groups (5 to 16 years). Their results highlight that FewShot GPT-4 achieves a Kappa value of 0.70, closely aligning with expert human graders. Additional works, including (Zhao, Silva, and Poulsen 2025; Wei et al. 2025; Gao et al. 2025), have also leveraged prompt-based scoring methodologies.

Despite these advances, Few-Shot scoring methods struggle with student responses that contain spelling errors or slightly irregular but acceptable content, often producing vague outputs rather than precise scores. To mitigate such issues, researchers have shifted focus toward fine-tuning LLMs to standardize outputs and improve evaluation reliability. These efforts not only explore zero-shot and few-shot scenarios, but also benchmark fine-tuned models against supervised upper bounds. For example, (Sethi and Singh 2022) proposed an efficient fine-tuning approach for automated essay scoring, based on the transformer model (Vaswani et al. 2017) with parameter-efficient fine-tuning techniques to improve the grammatical accuracy of the assessment. Similarly, (Latif et al. 2025) introduced an enhanced shared-backbone architecture that incorporates lightweight LoRA adapters for task-specific fine-tuning. This approach achieved competitive performance (raising the average QWK from 0.848 to 0.888 compared to full fine-tuning), while reducing GPU memory consumption by 60% and inference latency by 40%, thus improving deployment efficiency.

In the context of Chinese-language science education, works in (Yang et al. 2025) fine-tuned GPT, demonstrating its ability to evaluate students' written explanations after domain-specific adaptation accurately. Other studies, such as (Chamieh, Zesch, and Giebermann 2024; Latif and Zhai 2024) have also employed fine-tuning to boost scoring precision. Although LLMs exhibit exceptional scoring performance, their deployment on resource-constrained devices remains challenging. To address this, (Latif et al. 2024) applied knowledge distillation (KD) techniques (Gou et al. 2021) to compress fine-tuned LLMs into smaller student models. The distilled models retained scoring capabilities comparable to those of their teacher counterparts while being more suitable for real-world deployment.

Framework

In the field of automatic scoring, student answers show highly diverse characteristics. For example, when faced with unanswerable questions, some students provide disjointed words or symbols instead of complete sentences. However, these answers may still contain partially correct components under lenient scoring criteria. Large language models (LLMs) often struggle to interpret such answers using a few-shot prompt. Although standard full fine-tuning (FFT) methods (Lv et al. 2024) use all training data, the predominance of simple samples leads the model to overfit to easy patterns while failing to learn from genuinely complex and critical cases. These challenging samples often reflect the core difficulties of the task, including ambiguous expressions, conceptual confusion, and borderline-scoring cases.

To address this, we propose the Learning from Scoring Disagreements: Contrastive Error Mining for Efficient and

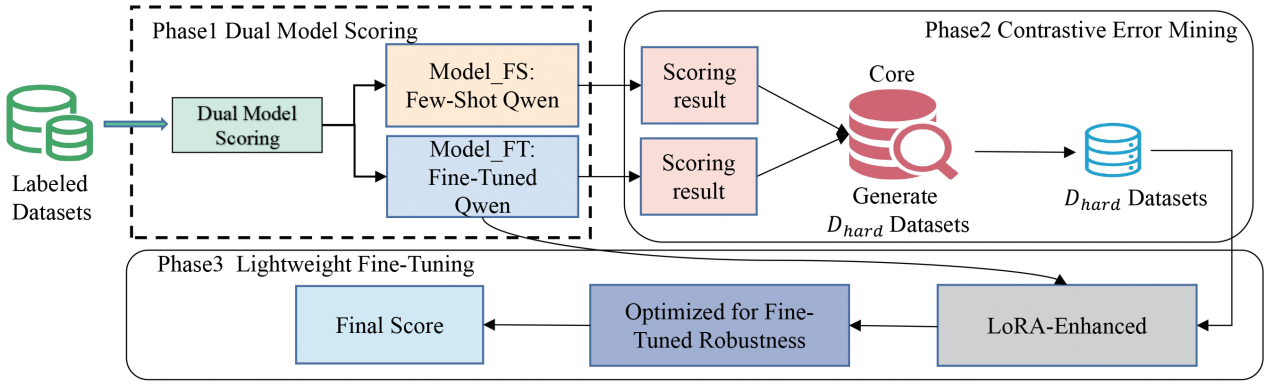

Figure 2: Overview of the CME-FT framework, which is divided into three phases: Dual Model Scoring, Contrastive Error Mining, and Lightweight Fine-Tuning.

Robust LLM-based Assessment framework, shown in Figure 2. The core idea is to proactively and automatically identify these difficult samples using two key signals:

1. Disagreements between models: When the models based on full-scale fine-tuning and few-shot give significantly different scores to the same answer, it indicates that the answer itself is ambiguous, complex, or in the "gray area" of the scoring rules. Different models make judgments based on the various "perspectives" or "biases" they have learned. Such samples naturally represent scoring difficulties.
2. Model error: Regardless of the size of the disagreements between models, as long as there is a significant deviation between the model prediction and the actual score, it is a clear error signal. Such samples expose the model's systematic biases or knowledge gaps.

We collect samples with considerable disagreements between models or large model errors (denoted as D_{hard}), and then build a high-value training set that condenses the task challenges, model weaknesses, and scoring uncertainties. This subset represents the most common types of error within the model's current capabilities. On this basis, this framework only performs lightweight LoRA adjustments based on the complete fine-tuned model on this subset, achieving local performance enhancement and generalization improvement at the lowest cost. The framework effectively addresses the credibility challenge in automatic scoring through a closed-loop process comprising three consecutive stages.

Dual Model Scoring

We begin by quantifying the discrepancies in scores between the full fine-tuned model (2 epochs) and the few-shot model. Specifically, we construct two scoring models to establish a multi-perspective scoring mechanism: the fine-tuning model M_{FT}, and the few-shot model M_{FS} based on few-shot prompts. Given a set of training samples $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i denotes the student answers text and y_i is the manual expert score. We compute the predicted scores from both models: $s_{FT}^{(i)}$ and $s_{FS}^{(i)}$:

$$s_{FT}^{(i)} = M_{FT}(x_i; \theta_{FT})$$
$$s_{FS}^{(i)} = M_{FS}(x_i; \mathcal{P}_K)$$

where $\mathcal{P}K$ denotes a K-shot prompt template ($K = 3$ in our experiments), θ_{FT} represents the model trainable parameters

Contrastive Error Mining

Based on inter-model score discrepancies and deviations from reference scores, we identify representative samples with a high risk of scoring errors to construct a hard sample set D_{hard}. For each response, we define the indicators:

$$\delta_d^{(i)} = |s_{FT}^{(i)} - s_{FS}^{(i)}|$$

quantifies the magnitude of disagreement between the two models;

$$\delta_e^{(i)} = |s_{FT}^{(i)} - y_i|$$

Quantifies deviation from expert judgment We then construct the hard sample set as:

$$D_{hard} = \{(x_i, y_i) | \delta_d^{(i)} > \tau_d \text{ or } \delta_e^{(i)} > \tau_e\}$$

To determine hard samples more effectively, we introduce a dynamic threshold mechanism to adaptively set screening criteria based on the global statistical characteristics of the samples:

$$\tau_d = \mu_d + \alpha \cdot \sigma_d \, , \tau_e = \mu_e + \beta \cdot \sigma_e$$

Where μ_d, μ_e are the means of the score disagreement δ_d and the score deviation δ_e respectively; σ_d, σ_e are their corresponding standard deviations; α and β are sensitivity hyperparameters that control the strictness of the sample selection process.

In our experiments, we varied the hyperparameters α and β across datasets to evaluate the robustness of the proposed complex sample mining strategy. In addition, these two parameters can be further fine-tuned based on the validation

set performance to adapt to specific task requirements. In the experimental part, we will systematically analyze the impact of α and β on model performance and explore their optimal configuration.

Lightweight Fine-Tuning

The hard sample set D_{hard} is used to perform lightweight LoRA-based (Hu et al. 2022) fine-tuning on the complete fine-tuned model of the first phase, thereby achieving targeted corrections to the model scoring accuracy. Specifically, for a given intermediate weight matrix W in the Transformer model, we introduce two low-rank matrices, $A \in \mathbb{R}^{k\times r}$ and $B \in \mathbb{R}^{d\times r}$, to construct the update term:

$$W' = W + \Delta W = W + BA^{\top}$$

The rank parameter $r \ll \min(d, k)$ ensures that the added parameters are minimal, reducing overfitting and enhancing generalization.

Experiments

Datasets

This study used three representative automated short answer scoring (ASAG) datasets, namely SciEntsbank, Beetle (Dzikovska et al. 2013), and Mohler (Mohler and Mihalcea 2009), covering a variety of scoring dimensions and evaluation scenarios and effectively supporting the comprehensive verification of the proposed method.

- SciEntsbank and BEETLE datasets: These two datasets are standard benchmark datasets that are widely used in short-answer automatic scoring tasks. They are designed to test model robustness across various generalization scenarios, including:
  - Unseen Answers: Test answers do not appear during training;
  - Unseen Questions: Test questions are unseen during training;
  - Unseen Domains: Test questions originate from domains not present in training.

  (Note: The Beetle dataset does not have the Unseen Domains task.)

  The student answers in the dataset are annotated into one of the following 5-way: Correct, Partially Correct, Incomplete, Contradictory, Irrelevant, and Non-domain.
- Mohler datasets: Mohler is a classic dataset compiled by Mohler et al. at the University of Texas. It contains responses to 80 short-answer questions from 12 course assignments, contributed by 2,442 students. Two domain experts independently score each response on a 0–5 scale. The final label is the arithmetic mean of the two scores. To ensure consistency, we use the same 80/20 train-test split strategy as in BEETLE and ensure that the test set covers both "unseen answers" and "unseen questions," providing a realistic evaluation of generalization in educational settings.

Baselines

To evaluate the effectiveness of the proposed CEM-FT framework in mining complicated cases from scoring disagreements and enhancing the stability and accuracy of large language model (LLM) based scoring, we selected the model of BERT (Devlin et al. 2019), DeBERTa-v3 (He, Gao, and Chen 2021) and Qwen series (Bai et al. 2023), including Qwen2-0.5B, Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B and Qwen3-8B. Each model was evaluated in both fine-tuning and few-shot inference settings.

Evaluation Metrics

To comprehensively evaluate model performance, we use a set of metrics that cover both classification and regression. Specifically, for ordered classification tasks, we use accuracy (ACC), quadratic weighted Kappa (QWK) (Brenner and Kliebsch 1996), and Mean Absolute Error (MAE), and for continuous value score prediction, we use tolerance-aware accuracy (TAA) and relative matching coefficient (RMC) (Chang and Ginter 2024). These metrics collectively assess prediction accuracy, inter-rater agreement, and the model's ability to preserve ranking information.

Overall Performance

To verify the versatility and robustness of the proposed method, we conducted systematic experiments on three datasets and compared the thoroughly fine-tuned model with the proposed CEM-FT framework. The results are summarized in Tables 1-3.

SciEntsbank datasets analysis: In the SciEntsbank dataset, CEM-FT (Qwen3-4B) also showed excellent performance, especially on the Unseen Answers task, achieving 0.7685, 0.7671, and 0.7466 in the ACC, QWK, and RMC indicators, respectively. In the interdisciplinary transfer task Unseen Domains, the method also achieved high scores of 0.6357 (ACC) and 0.6309 (QWK). In the Unseen Questions scenario, CEM-FT once again leads all the comparison methods. The overall results show that this method has significant effectiveness in uncertainty modeling and hard sample mining strategies.

Beetle datasets analysis: On the Beetle dataset, the CEM-FT framework (based on Qwen3-4B) achieved the best results in all subtasks, especially in the Unseen Answers and Unseen Questions tasks, reaching 0.7995/0.6899 and 0.7899/0.6554 (Accuracy and QWK Metrics, respectively), which is significantly better than the few-shot version (Qwen3-4B_FS) and all baseline models. This shows that the proposed framework has strong generalization and consistency optimization capabilities. Compared with the DeBERTa-V3 and Qwen2 series, CEM-FT significantly improves RMC and demonstrates greater robustness against semantic ambiguity and diversity in student answers.

Mohler datasets analysis: On the Mohler dataset, CEM-FT (based on Qwen3-1.7B) achieved RMC of 0.6931 and TAA of 0.7150 in the Unseen Answers task, outperforming all baseline models; its MAE dropped significantly from BERT's 0.5538 to 0.4019, further approaching the

Models	Unseen Answers			Unseen Questions			Unseen Domains		
	ACC ↑	QWK ↑	RMC ↑	ACC ↑	QWK↑	RMC↑	ACC ↑	QWK↑	RMC↑
BERT	0.4407	0.1997	0.4033	0.4461	0.1321	0.3659	0.3652	0.0614	0.3473
DeBERTa-V3	0.4667	0.2932	0.4368	0.4911	0.2288	0.4142	0.4272	0.2468	0.4301
Qwen2-0.5B	0.7185	0.6896	0.6927	0.5583	0.4767	0.5496	0.5808	0.5735	0.5911
Qwen3-0.6B	0.7537	0.7115	0.7181	0.6276	0.5762	0.6082	0.6208	0.6035	0.6151
Qwen3-1.7B	0.7561	0.7272	0.7251	0.6344	0.5709	0.6189	0.6283	0.6241	0.6229
Qwen3-4B	0.7593	0.7481	0.7331	0.6248	0.5755	0.6024	0.6235	0.6183	0.6192
Qwen3-8B	0.7537	0.7285	0.7238	0.6521	0.6313	0.6364	0.6337	0.6208	0.6253
Qwen2-0.5B_FS	0.2944	0.0214	0.3342	0.2483	0.0228	0.3107	0.1565	0.0844	0.3027
Qwen3-0.6B_FS	0.3618	0.0336	0.3175	0.4175	0.0782	0.3572	0.2998	0.1901	0.2735
Qwen3-1.7B_FS	0.3811	0.3568	0.4879	0.4099	0.3673	0.4971	0.3856	0.3388	0.4807
Qwen3-4B_FS	0.5241	0.4211	0.5595	0.5143	0.4354	0.5663	0.4862	0.3235	0.5272
Qwen3-8B_FS	0.4981	0.4957	0.5545	0.5048	0.5438	0.5782	0.4579	0.4215	0.5404
CEM-FT(Qwen3-4B)	**0.7685**	**0.7671**	**0.7466**	**0.6968**	**0.5835**	**0.6398**	**0.6357**	**0.6309**	**0.6313**

Table 1: Performance of different models on SciEntsbank datasets, models with the _FS suffix represent few-shot models.

Models	Unseen Answers			Unseen Questions		
	ACC ↑	QWK ↑	RMC ↑	ACC ↑	QWK↑	RMC↑
BERT	0.6196	0.6317	0.5923	0.5726	0.547	0.5022
DeBERTa-V3	0.6834	0.6944	0.6607	0.6129	0.5625	0.5462
Qwen2-0.5B	0.7342	0.7688	0.7261	0.6412	0.6333	0.5888
Qwen3-0.6B	0.7543	0.7588	0.7261	0.6777	0.6453	0.6168
Qwen3-1.7B	0.7380	0.7475	0.7058	0.6768	0.6721	0.6148
Qwen3-4B	0.7677	0.7634	0.7303	0.6838	0.6549	0.6226
Qwen3-8B	0.7449	0.7402	0.7074	0.6923	0.6731	0.6203
Qwen2-0.5B_FS	0.3508	0.0345	0.2968	0.3101	0.1203	0.2747
Qwen3-0.6B_FS	0.3761	0.1334	0.3506	0.3394	0.0251	0.3032
Qwen3-1.7B_FS	0.2938	0.2345	0.4223	0.2772	0.2705	0.4388
Qwen3-4B_FS	0.5558	0.5547	0.5199	0.5558	0.5547	0.5199
Qwen3-8B_FS	0.5262	0.4967	0.5087	0.5287	0.4949	0.5047
CEM-FT(Qwen3-4B)	**0.7995**	**0.7899**	**0.7594**	**0.6899**	**0.6554**	**0.6281**

Table 2: Performance of different models on Beetle datasets, models with the _FS suffix represent few-shot models.

level of manual expert scoring, showing the advantage of this method in controlling scoring bias. This indicates that this method is more robust to scoring errors in question-answering scoring tasks. In the Unseen Questions task, CEM-FT maintained its leading RMC (0.6029), further confirming its strong adaptability to question-type changes and context transfer, and performing significantly better than traditional fine-tuning and few-shot reasoning strategies.

Judging from the overall experimental results, DeBERTa-V3 and multiple Qwen series models (ranging from 0.5B to 8B) across three datasets have achieved significant performance improvements over the BERT baseline under full-scale fine-tuning. In the few-shot scenario, the model performance generally declines significantly, indicating that the lack of targeted training makes it difficult to capture the complex semantic requirements of the scoring task. In comparison, the CEM-FT framework continues to lead across multiple generalization tasks and indicators, especially in terms of the relative matching coefficient (RMC) and tolerance-aware accuracy (TAA). These results demonstrate the effectiveness of the proposed scoring enhancement approach, which integrates hard-sample mining based on model disagreements with LoRA-based fine-tuning to address the challenges posed by complex educational assessment tasks. The method has strong practical potential and offers valuable insights for advancing research in automated student answer scoring.

Ablation experiments on different base models

To further verify the applicability and potential for performance improvement of the proposed framework across a variety of language models, we conducted systematic experiments on 5 Qwen series models.

We averaged the evaluation indicator results across the three tasks to ensure they more accurately reflect the overall trends of different models across various generalization scenarios. Results in Figures 3 to 5 show that across all base models, the CEM-FT framework performs better than the corresponding complete fine-tuning model on multiple tasks, reflecting strong generalization and structural adaptability. Especially on the Mohler dataset, the TAA index of CEM-FT across multiple Qwen versions improves by 1.78%

Models	Unseen Answers			Unseen Questions		
	TAA↑	RMC↑	MAE↓	TAA↑	RMC↑	MAE↓
BERT	0.6449	0.6324	0.5358	0.6085	0.5802	0.5248
DeBERTa-V3	0.6308	0.6583	0.6014	0.5465	0.5485	0.6363
Qwen2-0.5B	0.4860	0.5903	0.675	0.4076	0.5327	0.7151
Qwen3-0.6B	0.6075	0.6545	0.5125	0.4341	0.5279	0.6787
Qwen3-1.7B	0.7009	0.6798	0.4216	0.6202	0.5974	0.5116
Qwen3-4B	0.6636	0.6618	0.4758	0.5543	0.5792	0.5845
Qwen3-8B	0.6682	0.6551	0.4811	0.5465	0.5603	0.5942
Qwen2-0.5B_FS	0.1333	0.3249	3.5839	0.1339	0.2988	4.0039
Qwen3-0.6B_FS	0.1756	0.5054	1.8652	0.1796	0.4074	1.9214
Qwen3-1.7B_FS	0.0871	0.4192	2.6019	0.1078	0.4051	2.5151
Qwen3-4B_FS	0.3636	0.5859	1.056	0.3636	0.5859	1.056
Qwen3-8B_FS	0.3353	0.5804	1.1519	0.3048	0.6072	1.0602
CEM-FT(Qwen3-1.7B)	**0.7150**	**0.6931**	**0.4019**	**0.6398**	**0.6029**	**0.5109**

Table 3: Performance of different models on Mohler datasets, models with the _FS suffix represent few-shot models.

α	β	SciEntsbank			Beetle			Mohler		
		ACC ↑	QWK ↑	RMC ↑	ACC↑	QWK↑	RMC	TAA↑	TMC↑	MAE↓
1	1	0.6891	0.6487	0.6593	0.7311	0.7104	0.6808	0.6721	0.6421	0.4584
	1.5	0.6808	0.6431	0.6523	0.7345	0.7123	0.6785	**0.6774**	**0.6480**	**0.4564**
	2	**0.7003**	**0.6605**	**0.6626**	**0.7447**	**0.7227**	**0.6938**	0.6644	0.6432	0.4571
1.5	1	0.6802	0.6502	0.6599	0.7312	0.7109	0.6801	0.6502	0.6311	0.4642
	1.5	0.6895	0.6473	0.6525	0.7345	0.7153	0.6753	0.6659	0.6359	0.4798
	2	0.6753	0.6538	0.6423	0.7283	0.7010	0.6870	0.6519	0.6395	0.4781
2	1	0.6964	0.6539	0.6528	0.7316	0.7101	0.6701	0.6512	0.6299	0.4705
	1.5	0.6951	0.6531	0.6486	0.7340	0.7123	0.6756	0.6538	0.6410	0.4843
	2	0.6913	0.6578	0.6467	0.7215	0.7038	0.6508	0.6582	0.6418	0.4667

Table 4: Results of screening data sets with different α and β hyperparameters. The bold font is the optimal value.

to 2.98% compared with the complete fine-tuning scheme. In addition, across the Beetle and SciEntsbank datasets, the framework shows a consistent performance gain trend for each base model, spanning a variety of task settings and evaluation dimensions.

Overall, the experimental results show that the CEM-FT framework can migrate across models and achieve stable optimization across a variety of scoring-based models. Its wide adaptability provides stronger theoretical and methodological support for the actual deployment and expansion of the task of automatic scoring of student answers.

Experiments with different screening thresholds

Setting a fixed threshold to measure score disagreements between the fine-tuned model and the few-shot inference model lacks sufficient objectivity. To address this issue, we introduce two hyperparameters, α and β, to dynamically adjust the screening threshold for a divergent sample, thereby enabling a more robust evaluation of the proposed risk sample mining strategy under varying conditions. To ensure consistency in the analysis across datasets, this strategy is also applied to the SciEntsbank and Beetle datasets.

Based on extensive preliminary experiments, the values of α and β are set within the range of [1, 1.5, 2]. This selection results in score differences between the models typically spanning 2 to 4 intervals, facilitating a more effective identification of challenging answer samples. Table 4 presents the corresponding experimental results on the Beetle, SciEntsbank, and Mohler datasets.

Overall, the model exhibits greater sensitivity to changes in β, while performance variations with respect to α are relatively stable. For example, on the SciEntsbank dataset, the optimal ACC (0.7003), QWK (0.6605), and RMC (0.6626) were achieved when $\alpha = 1$ and $\beta = 2$, significantly outperforming other configurations. This result suggests that a higher value of β facilitates more effective mining of potential semantic alignments, thereby enhancing the consistency of the scoring. On the Beetle dataset, the model also achieves the best ACC (0.7447) and RMC (0.6938) with $\beta = 2$. In the Mohler dataset, the best TAA (0.6774) and TMC (0.6480) are achieved with $\alpha = 1, \beta = 1.5$, This combination also yields the lowest MAE (0.4564), indicating its effectiveness in reducing error and improving scoring accuracy, making it the most balanced configuration for overall performance.

In summary, the best hyperparameter combination is not entirely consistent across datasets, indicating that the model has some generalization adjustment space across tasks. Notably, smaller α and larger β values tend to yield better performance, underscoring the dominant role of β in refining the sensitivity to semantic alignment. These findings demonstrate that appropriate hyperparameter tuning can significantly boost model performance, and automatic parameter

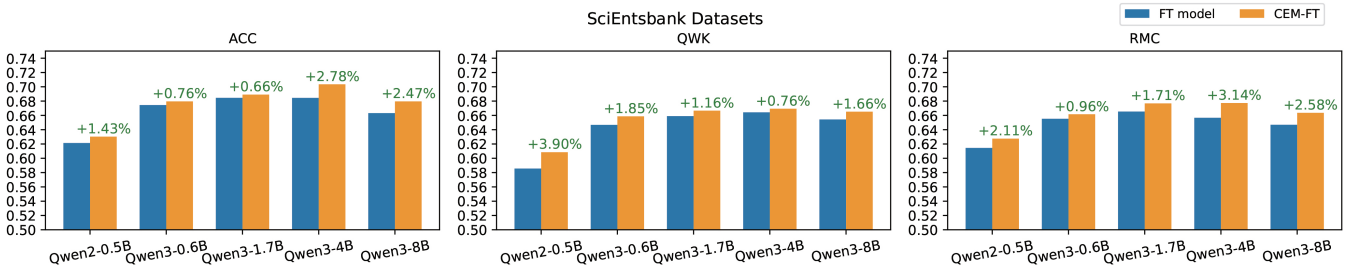

Figure 3: Improvement results of CEM-FT framework on different models on SciEntsbank datasets

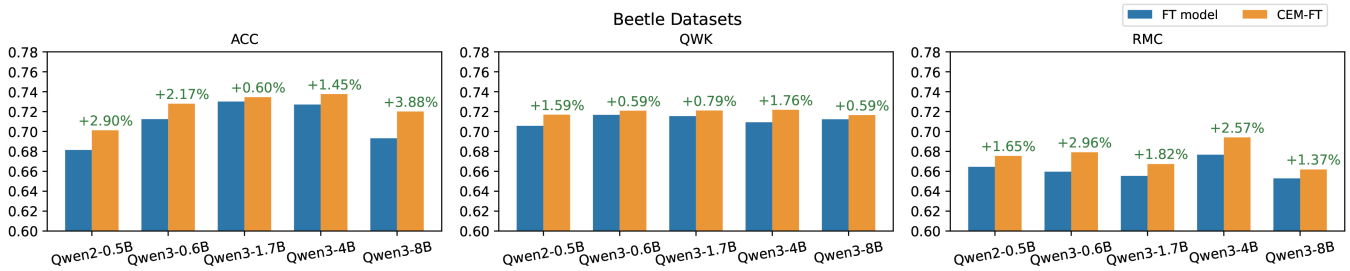

Figure 4: Improvement results of CEM-FT framework on different models on Beetle datasets

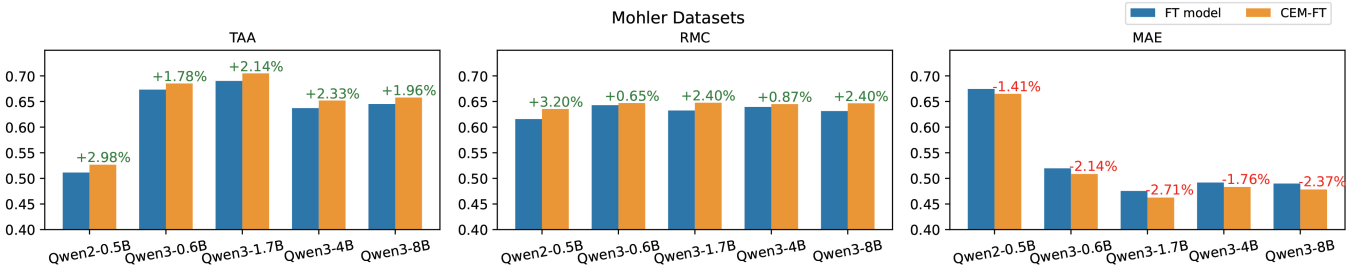

Figure 5: Improvement results of CEM-FT framework on different models on Mohler datasets

optimization guided by a validation set can be considered in future applications.

Limitation

The CEM-FT framework relies on the score difference between the fine-tuning model (FT) and the few-shot reasoning model (FS) to mine high-value samples. However, when both models exhibit systematic biases, their score Disagreements may stem from common blind spots or random noise rather than from truly challenging samples, thereby affecting the quality of the D_{hard} dataset. In addition, the threshold hyperparameters α and β introduced in CEM-FT need to be set empirically through pre-experiments, lacking a data-driven adaptive strategy. This mechanism, which relies on manual parameter adjustment, may lead to reduced robustness and difficulty in migration on different datasets or tasks.

Conclusion

This paper proposes a Contrastive Error Mining and Fine-Tuning (CEM-FT) framework that automatically identifies complex samples by comparing the full fine-tuned model with a few-shot model during scoring, and trains a lightweight adapter based on the LoRA architecture to efficiently correct systematic scoring bias. This method effectively solves three key problems in the current large model scoring: first, inconsistency and deviation when facing complex, ambiguous, or novel student answers, which limits the actual credibility; second, the instability of the few-shot reasoning method when dealing with atypical answers; and third, existing models often ignore representative and challenging complex samples. Extensive experiments across five variants of the Qwen LLM series on three public datasets (SciEntsbank, Beetle, Mohler) demonstrate that CEM-FT significantly enhances scoring accuracy and consistency, offering a practical approach for LLM-driven automated scoring systems.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2022YFC3303604), the National Natural Science Foundation of China (62372212), and the Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. arXiv:2309.16609.

Brenner, H.; and Kliebsch, U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 199–202.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Chamieh, I.; Zesch, T.; and Giebermann, K. 2024. Llms in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)*, 309–315.

Chang, L.-H.; and Ginter, F. 2024. Automatic short answer grading for Finnish with ChatGPT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23173–23181.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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.

Dzikovska, M.; Nielsen, R.; Brew, C.; Leacock, C.; Giampiccolo, D.; Bentivogli, L.; Clark, P.; Dagan, I.; and Dang, H. T. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In Manandhar, S.; and Yuret, D., eds., *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 263–274. Atlanta, Georgia, USA: Association for Computational Linguistics.

Funayama, H.; Matsubayashi, Y.; Asazuma, Y.; Mizumoto, T.; and Inui, K. 2025. Cross-prompt Pre-finetuning of Language Models for Short Answer Scoring. *International Journal of Artificial Intelligence in Education*, 1–22.

Gao, R.; Guo, X.; Li, X.; Narayanan, A. B. L.; Thomas, N.; and Srinivasa, A. R. 2025. Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering. In *Large Foundation Models for Educational Assessment*, 186–206. PMLR.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6): 1789–1819.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ICLR*, 23017.

Henkel, O.; Hills, L.; Boxer, A.; Roberts, B.; and Levonian, Z. 2024. 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*, 300–304.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 1(2): 3.

Jiang, L.; and Bosch, N. 2024. Short answer scoring with GPT-4. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, 438–442.

Kooli, C.; and Yusuf, N. 2025. Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading. *International Journal of Human–Computer Interaction*, 41(5): 3388–3399.

Latif, E.; Fang, L.; Ma, P.; and Zhai, X. 2024. Knowledge distillation of llms for automatic scoring of science assessments. In *International Conference on Artificial Intelligence in Education*, 166–174. Springer.

Latif, E.; and Zhai, X. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6: 100210.

Latif, E.; Zhou, Y.; Fang, L.; and Zhai, X. 2025. Efficient Multi-Task Inference with a Shared Backbone and Lightweight Task-Specific Adapters for Automatic Scoring. In Wang, Z.; Woodhead, S.; Ananda, M.; Mallick, D. B.; Sharpnack, J.; and Burstein, J., eds., *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, 212–220. PMLR.

Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; and Zhai, X. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6: 100213.

Liang, M. 2025. Leveraging natural language processing for automated assessment and feedback production in virtual education settings. *Journal of Computational Methods in Sciences and Engineering*, 25(3): 2502–2515.

Liu, T.; Ding, W.; Wang, Z.; Tang, J.; Huang, G. Y.; and Liu, Z. 2019. Automatic short answer grading via multiway attention networks. In *International conference on artificial intelligence in education*, 169–173. Springer.

Liu, Z.; Xu, G.; Liu, T.; Fu, W.; Qi, Y.; Ding, W.; Song, Y.; Guo, C.; Kong, C.; Yang, S.; et al. 2020. Dolphin: a spoken language proficiency assessment system for elementary education. In *Proceedings of The Web Conference 2020*, 2641–2647.

Lv, K.; Yang, Y.; Liu, T.; Guo, Q.; and Qiu, X. 2024. Full Parameter Fine-tuning for Large Language Models with Limited Resources. In Ku, L.-W.; Martins, A.; and Srikumar,

V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8187–8198. Bangkok, Thailand: Association for Computational Linguistics.

Mohler, M.; and Mihalcea, R. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 567–575.

Sethi, A.; and Singh, K. 2022. Natural language processing based automated essay scoring with parameter-efficient transformer approach. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 749–756. IEEE.

Stahl, M.; Biermann, L.; Nehring, A.; and Wachsmuth, H. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In Kochmar, E.; Bexte, M.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Tack, A.; Yaneva, V.; and Yuan, Z., eds., *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 283–298. Mexico City, Mexico: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei, Y.; Pearl, D.; Beckman, M.; and Passonneau, R. J. 2025. Concept-based Rubrics Improve LLM Formative Assessment and Data Synthesis. *CoRR*, abs/2504.03877.

Yang, J.; Latif, E.; He, Y.; and Zhai, X. 2025. Fine-tuning ChatGPT for Automatic Scoring of Written Scientific Explanations in Chinese. *Journal of Science Education and Technology*, 34(4): 719–736.

Zhao, C.; Silva, M.; and Poulsen, S. 2025. Language Models are Few-Shot Graders. In *Artificial Intelligence in Education*, 3–16. Springer Nature Switzerland. ISBN 978-3-031-98459-4.