# CAFES: A Collaborative Multi-Agent Framework for Multi-Granular Multimodal Essay Scoring

**Anonymous ACL submission**


## Abstract

Automated Essay Scoring (AES) is crucial for modern education, particularly with the increasing prevalence of multimodal assessments. However, traditional AES methods struggle with *evaluation generalizability and multimodal perception*, while even recent Multimodal Large Language Model (MLLM)-based approaches can produce *hallucinated justifications and scores misaligned with human judgment*. To address the limitations, we introduce **CAFES**, the first collaborative multi-agent framework specifically designed for AES. It orchestrates three specialized agents: an *Initial Scorer* for rapid, trait-specific evaluations; a *Feedback Pool Manager* to aggregate detailed, evidence-grounded strengths; and a *Reflective Scorer* that iteratively refines scores based on this feedback to enhance human alignment. Extensive experiments, using state-of-the-art MLLMs, achieve an average relative improvement of 21% in Quadratic Weighted Kappa (QWK) against ground truth, especially for grammatical and lexical diversity. Our proposed CAFES framework paves the way for an intelligent multimodal AES system. The code will be available upon acceptance.

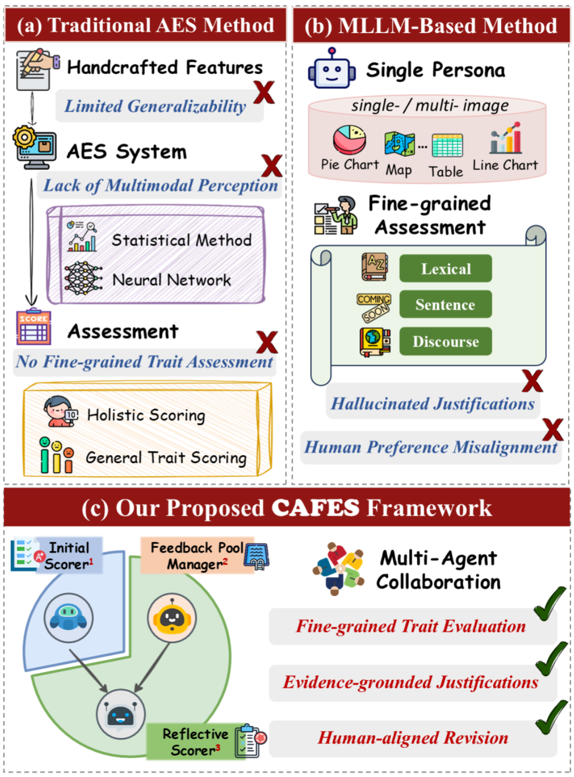




Figure 1: Comparisons among the traditional AES method (a), MLLM-based method (b), and our proposed multi-agent CAFES framework (c) on AES task.

## 1 Introduction

Automated Essay Scoring (AES) plays a crucial role in educational assessment today, offering efficient, fair, and scalable evaluation of student writing tasks (Ramesh and Sanampudi, 2022; Li and Liu, 2024; Wu et al., 2024; Xia et al., 2024). AES systems benefit both students by highlighting areas for improvement and educators by reducing manual grading workloads. As contemporary assessments increasingly emphasize students' abilities to integrate information from both text and images, multimodal writing tasks have become a key focus. Therefore, there is *a growing need for AES systems capable of precise, detailed, context-aware evaluations that effectively handle multimodal inputs* (Ye et al., 2025; Su et al., 2025; Li et al., 2024).

**Traditional AES methods** suffer from several critical limitations, as shown in Figure 1 (a). ❶ They rely heavily on handcrafted features like word frequency and essay length, limiting their generalizability across diverse topics (Yang et al., 2024; Jansen et al., 2024; Uto et al., 2020). ❷ They lack multimodal perception, making them unsuitable for multimodal inputs. ❸ They struggle to assess fine-grained traits, such as coherence and organizational structure (Lim et al., 2021; Uto, 2021; Wang et al., 2022). Recently, Multimodal Large Language Models (MLLMs) have been applied to AES, yet **MLLM-based methods** still introduce new challenges like ❹ hallucinated justifications and ❺ scoring misaligned with human preference (Su et al., 2025), as shown in Figure 1 (b).

However, the emergence of multimodal multi-agent systems offers a promising solution to these challenges (Wang et al., 2025; Chu et al., 2025). Specifically, **multi-agent frameworks** have the following advantages, as illustrated in Figure 1 (c): ✔ They enable fine-grained trait evaluation, providing detailed feedback across various writing traits. ✔ They can generate evidence-grounded justifications and engage in cross-agent collaboration, effectively mitigating hallucinations introduced by a single MLLM. ✔ The reflective mechanism of multi-agent systems ensures human-aligned revisions, dynamically adjusting scores to better align with human preferences.

Therefore, we propose **CAFES, the first-ever collaborative multi-agent framework designed specifically for AES**. In particular, CAFES decomposes the scoring process into three specialized modules: an *initial scoring agent* that provides fast trait-specific scores; a *feedback pool agent* that aggregates detailed strengths across writing traits; and a *reflective scoring agent* that iteratively updates scores based on the feedback pool.

Our contributions can be summarized as follows:

- We introduce CAFES, **the first multi-agent framework** for AES tasks, integrating three specialized agents including Initial scorer, Feedback Pool Manager, and Reflective Scorer to enable collaborative multi-granular essay scoring.
- We demonstrate **the essential impact of the Feedback Pool Manager, Reflective Scorer, and teacher-student MLLM collaboration mechanism** through ablation studies and case analyses.
- We evaluate CAFES framework with **state-of-the-art MLLMs** as student models, GPT-4o as the default teacher model, achieving an average improvement of 21% in Quadratic Weighted Kappa (QWK) against ground truth scores, especially for grammatical and lexical diversity traits.

By addressing the gaps in the existing AES approaches, CAFES pave the way for reliable, nuanced, and context-sensitive multi-agent AES systems driven by MLLMs in the AGI era.

## 2 Related Work

### 2.1 AES Datasets

Existing AES datasets have been widely used to support research on writing assessment (more details are shown in the Appendix A.1). In terms of modality, these datasets can be categorized into text-only and multimodal datasets.

Among **text-only datasets**, $\text{ASAP}_{\text{AES}}$ (Cozma et al., 2018a) is widely used due to its large scale. Its extended version, ASAP++, adds trait-level annotations, but merges key content traits into a single "CONTENT" label (Mathias and Bhattacharyya, 2018). Both of them only have few topics. The CLC-FCE dataset provides detailed annotations of grammatical errors (Yannakoudakis et al., 2011). The TOEFL11 dataset uses only coarse-grained proficiency labels (low / medium / high) (Lee et al., 2024a). The ICLE (Granger et al., 2009) and ICLE++ (Li and Ng, 2024c) datasets offer more detailed and multi-granular annotations. Nevertheless, their topic coverage is highly limited. Similarly, the AAE corpus focuses solely on argumentative structure (Stab and Gurevych, 2014). The CREE corpus is designed to evaluate sentence understanding and error types (Bailey and Meurers, 2008). In summary, existing text-only AES datasets face two key limitations: (1) limited topic diversity, and (2) a lack of fine-grained trait-level annotations (Ke and Ng, 2019; Li and Ng, 2024b,a).

EssayJudge is the only publicly available **multimodal AES dataset** (Su et al., 2025), with ten fine-grained trait annotation. Lexical-level traits include *lexical accuracy* (LA) and *lexical diversity* (LD). Sentence-level traits include *coherence* (CH), *grammatical accuracy* (GA), *grammatical diversity* (GD), and *punctuation accuracy* (PA). Discourse-level traits include *argument clarity* (AC), *justifying persuasiveness* (JP), *organizational structure* (OS), and *essay length* (EL).

### 2.2 AES Systems

AES systems are typically classified into three types: heuristic, machine learning, and deep learning approaches (Li and Ng, 2024a; Kamalov et al., 2025; Atkinson and Palma, 2025; Xu et al., 2025; Song et al., 2025a). **Heuristic methods** assign overall scores by combining rule-based trait scores such as organization, coherence, and grammar. For instance, the organization can be assessed using templates like the three-paragraph essay format (Attali and Burstein, 2006). **Machine learning**

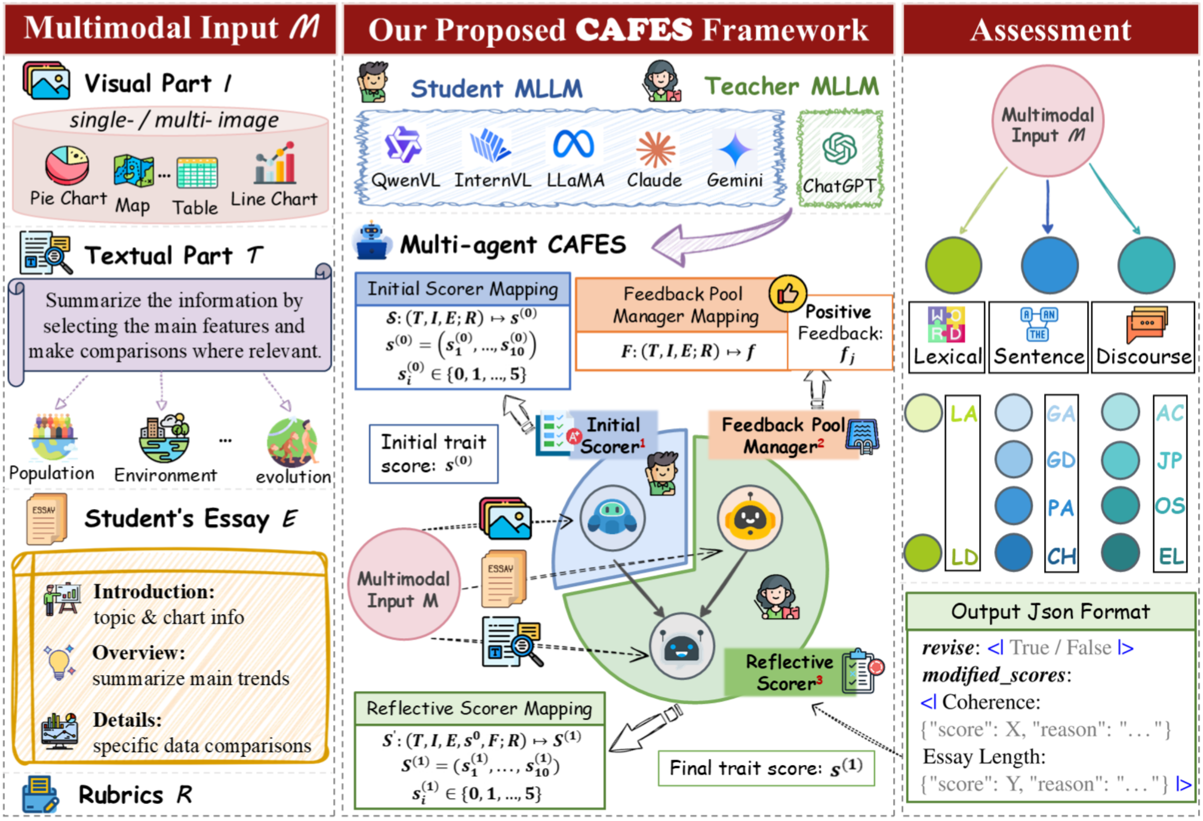



Figure 2: The framework of our proposed CAFES. The system follows a three-stage process: ❶ Initial scoring via the student MLLM; ❷ Feedback generation for each trait via the teacher MLLM; and ❸ Final reflective scoring with justification-based revision via the teacher MLLM.

**methods** (*e.g.,* Logistic Regression, SVMs) rely on handcrafted features (Chen and He, 2013; Yannakoudakis and Briscoe, 2012), including length-based attributes (Vajjala, 2016; Yannakoudakis and Briscoe, 2012). Thus, their cross-topic generalization is limited. **Deep learning methods**, especially Transformer models like BERT (Wang et al., 2022), improve AES by learning semantic features directly from text (Jiang et al., 2023; Cao et al., 2020), enabling multi-trait and cross-topic scoring. **LLM-based methods** further advance AES (Mizumoto and Eguchi, 2023; Choi et al., 2025; Cai et al., 2025). They support zero-shot scoring using rubrics alone (Lee et al., 2024a), or few-shot settings with minimal examples (Mansour et al., 2024; Xiao et al., 2024), offering better performance and adaptability in low-resource scenarios.

### 2.3 Multi-Agent Collaboration

Recent studies suggest that multi-agent collaboration, by organizing and coordinating multiple LLMs, enables more effective handling of complex tasks (Tran et al., 2025; Yan et al., 2025b). Systems like CAMEL (Li et al., 2023) and AutoGen (Wu et al., 2023) assign roles such as planner, coder, and critic, allowing agents to interact through multi-turn dialogue and perform better in reasoning, generation, and self-revision (Liang et al., 2024). This approach offers key benefits: improved task decomposition and control through role division, reduced bias via mutual verification, and enhanced adaptability and modularity. It is increasingly adopted in areas such as decision-making (Liu et al., 2024b), code generation (Yuan et al., 2024), and automated evaluation (Lifshitz et al., 2025). More related work on MLLMs can be found in Appendix A.2.

## 3 Methodology

Figure 2 illustrates the overall architecture of our multi-agent AES framework, which consists of **three core agents:** ❶ Initial Scorer, ❷ Feedback Pool Manager, and ❸ Reflective Scorer. To execute these agents, we introduce **two types of models**: ❶ a student MLLM and ❷ a teacher MLLM. The student model, which has relatively weaker capabilities, handles the Initial Scorer by giving an initial score for each of the ten fine-grained traits. The teacher model, with stronger **reasoning abilities**, executes the Feedback Pool Manager to generate positive feedback comments and applies the Reflective Scorer to revise the student model's initial scores based on the feedback pool. This collaborative setup mirrors the human-in-the-loop process of "scoring → feedback → revision" in real-world scoring assessment. In the following sections, we describe each module in detail.

### 3.1 Initial Scorer

The Initial Scorer module is responsible for producing preliminary scores across the ten fine-grained traits. Given the text of the essay topic $T$, the corresponding image $I$, the student's essay $E$, and the detailed scoring rubrics $\boldsymbol{R} \in \mathbb{R}^{10}$, the student MLLM assigns an initial score $s_i^{(0)}$ for each trait $d_i$. Formally, the Initial Scorer defines a mapping:

$$\mathcal{S} : (T, I, E; \boldsymbol{R}) \mapsto \boldsymbol{s}^{(\mathbf{0})} \in \mathbb{R}^{10}$$

where $\boldsymbol{s}^{(\mathbf{0})} = (s_1^{(0)}, \ldots, s_{10}^{(0)})$ denotes the preliminary scores with $s_i^{(0)} \in \{0, 1, \ldots, 5\}$.

This step can be viewed as the student model independently answering an exam based on its own understanding. The subsequent modules, executed by the teacher MLLM, are responsible for reviewing the student MLLMs' answers, providing feedback, and refining the initial judgments.

### 3.2 Feedback Pool Manager

The Feedback Pool Manager module is responsible for generating positive feedback for the student's essay based on the ten traits. Prior studies have indicated that MLLMs tend to adhere to the rubrics more strictly than human raters, often assigning lower scores during essay scoring (Su et al., 2025; Kundu and Barbosa, 2024). To address this tendency, we design the Feedback Pool Manager to focus exclusively on extracting positive feedback, emphasizing the strengths demonstrated in the essay. Formally, the Feedback Pool Manager defines a mapping:

$$\mathcal{F} : (T, I, E; \boldsymbol{R}) \mapsto \boldsymbol{f} \in \mathbb{R}^{10}$$

where $\boldsymbol{f} = (f_1, \ldots, f_{10})$ denotes a set of positive feedback entries, each associated with a specific trait $d_i$. For each trait, MLLMs return the extracted positive comments highlighting well-performed aspects of the essay.

The positive feedback generated by the teacher MLLM provides crucial and structured guidance for the Reflective Scorer in determining whether the initial scores require revision.

### 3.3 Reflective Scorer

The Reflective Scorer module is responsible for revising the student's initial scores by integrating positive feedback information. Formally, the Reflective Scorer defines a mapping:

$$\mathcal{S}' : (T, I, E, \boldsymbol{s}^{(\mathbf{0})}, \boldsymbol{f}; \boldsymbol{R}) \mapsto \boldsymbol{s}^{(\mathbf{1})} \in \mathbb{R}^{10}$$

where $\boldsymbol{s}^{(\mathbf{1})} = (s_1^{(1)}, \ldots, s_{10}^{(1)})$ denotes the revised scores across the ten traits. The teacher MLLM outputs a revised JSON object, and an example is shown in Figure 3:

***revise***: <| True / False |>
***modified_scores***:
<| Coherence: {"score": X, "reason": "..."}
Essay Length: {"score": Y, "reason": "..."} |>

Figure 3: Reflective scorer's JSON output format.

This reflective revision mechanism ensures that the final assessment fairly incorporates the strengths recognized in the essay, while avoiding unnecessary or overly aggressive adjustments.

## 4 Experiments and Analysis

### 4.1 Experimental Setup

| Statistic | Number |
|---|---|
| Total Multimodal Essays | 1,054 |
| Image Type | |
| - Single-Image | 703 (66.7%) |
| - Multi-Image | 351 (33.3%) |
| Multimodal Essay Type | |
| - Flow Chart | 305 (28.9%) |
| - Bar Chart | 211 (20.0%) |
| - Table | 153 (14.5%) |
| - Line Chart | 145 (13.8%) |
| - Pie Chart | 71 (6.7%) |
| - Map | 62 (5.9%) |
| - Composite Chart | 107 (10.2%) |

Table 1: Key statistics of the EssayJudge dataset.

**Dataset.** We evaluate our agent-based AES system CAFES on the EssayJudge dataset. It consists of 1,054 multimodal essays written at the university level. Each essay requires students to analyze and construct arguments based on visual inputs such as line charts and flow charts, posing significant challenges for MLLMs in terms of visual-textual understanding and reasoning. What's more, it covers 125 distinct essay topics across diverse domains including population, education, environment, production, and evolution. More details about the dataset are shown in Table 1. The diversity in both topics and visual formats increases the complexity of the scoring task and provides a strong foundation for evaluating the robustness and generalizability of AES systems under varied multimodal scenarios.

**Basic Settings.** In the CAFES AES framework, **GPT-4o is assigned as the default teacher model** throughout all experiments to guide and refine student MLLM's outputs, given its strong performance in AES (Hurst et al., 2024; Su et al., 2025). To evaluate CAFES's generalization ability, we systematically assign a wide range of leading MLLMs to the student model, grouped as follows: (i) **Open-source MLLMs**: InternVL2.5 (2B, 4B, 8B, 26B) (Chen et al., 2025b), Qwen2.5-VL (3B, 32B) (Chen et al., 2025c), and LLaMA-3.2-Vision (11B, 90B) (Dubey et al., 2024); (ii) **Closed-source MLLMs**: Claude-3.5-Sonnet (Anthropic, 2024), Gemini-2.5-Flash (DeepMind, 2025), and GPT-4o-mini (OpenAI, 2024). Since no existing AES model is designed for multimodal settings, we use the initial scores produced by each student model — before any teacher MLLM's feedback or reflection — to serve as the baseline for comparison. This setup ensures that any observed improvements can be fully attributed to the multi-agent process introduced in the CAFES framework. Detailed rubrics, prompts and model sources are listed in Appendix B.1, B.2, and B.3.

**Evaluation Metric.** After extensively reviewing previous AES studies (Song et al., 2024; Lee et al., 2024b; Mathias and Bhattacharyya, 2018; Cozma et al., 2018b; Wang et al., 2018), we select QWK as our evaluation metric, which is the most commonly used for assessing model alignment with ground truth scores. Its formula is expressed as:

$$k = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}},$$

where $w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$ is the element of the weight matrix penalizing larger differences between $i$ and $j$, $\boldsymbol{O}_{i,j}$ is the observed agreement, and $\boldsymbol{E}_{i,j}$ is the expected agreement under random chance. QWK values range from -1 (complete disagreement) to 1 (perfect agreement). Higher values are expected.

### 4.2 Main Results

**Our proposed CAFES framework yields consistent and significant improvements of QWK across each student MLLM on most traits.** Compared to the initial scores of single MLLMs, it improves the average QWK score by 0.07, from 0.29 to 0.36, representing a 21% relative improvement. Notably, the Qwen2.5-VL-3B achieves a remarkable 0.25 QWK improvement on the Grammatical Accuracy trait, which is shown in Table 2). These results clearly demonstrate the robustness and effectiveness of our framework.

**The CAFES framework yields the most significant improvements in grammatical and lexical diversity.** As shown in Figure 4, these two traits show the largest improvements under the CAFES framework compared to the initial scores. This is because single student MLLMs tend to focus on surface-level errors of grammar and vocabulary in initial scoring while overlooking the positive aspects of diversity (Su et al., 2025; Kundu and Barbosa, 2024), leading to underestimation compared to human raters (as shown in Appendix B.4). With the help of the Feedback Pool Manager, the agent highlights strengths and passes them to the Reflective Scorer, enabling better recognition of diverse expression and more aligned score revisions.

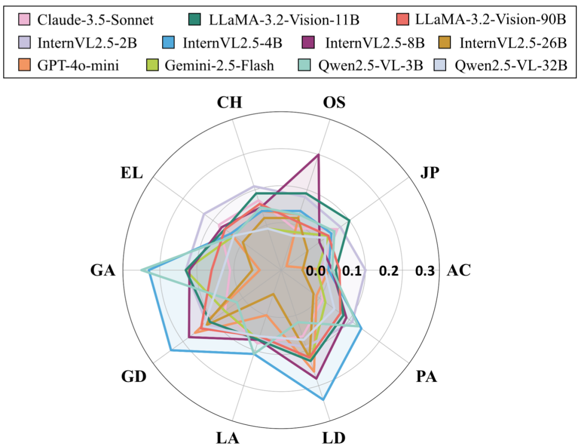
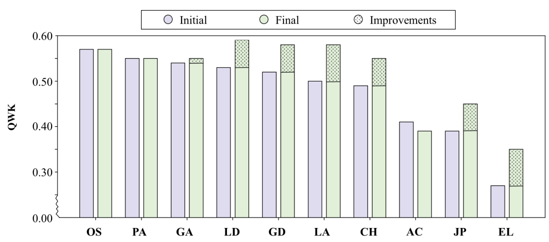



Figure 4: Trait-level score improvements after reflection via CAFES across different student MLLMs.

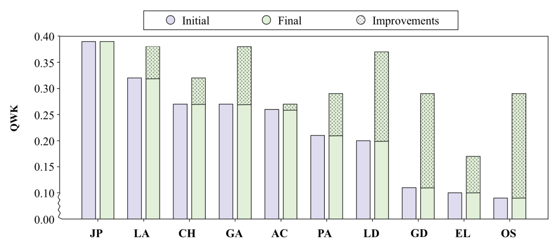
**In general, the lower the QWK of initial score generated by Initial Scorer, the greater the QWK improvement brought by CAFES.** This trend appears both within and across student MLLMs. Within a student MLLM, traits with lower initial QWK tend to improve more with CAFES (as shown in Figure 5). More examples are shown in Appendix B.5. Across different student MLLMs, those with weaker initial performance benefit more from CAFES framework, which is clearly demonstrated in Figure 6). This is likely because lower-performing traits or MLLMs have

| MLLMs | Lexical Level | | Sentence Level | | | | Discourse Level | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LA | LD | CH | GA | GD | PA | AC | JP | OS | EL |
| ***Open-Source MLLMs*** | | | | | | | | | | |
| InternVL2.5-2B (Chen et al., 2025b) | 0.04 | 0.06 | 0.07 | 0.02 | 0.08 | 0.05 | 0.03 | 0.01 | 0.05 | 0.04 |
| **+ CAFES (Ours)** | **0.11** | **0.16** | **0.18** | **0.15** | **0.20** | **0.16** | **0.12** | **0.08** | **0.13** | **0.17** |
| **Improvements** | ↑0.07 | ↑0.10 | ↑0.11 | ↑0.13 | ↑0.12 | ↑0.11 | ↑0.10 | ↑0.07 | ↑0.08 | ↑0.13 |
| InternVL2.5-4B (Chen et al., 2025b) | 0.12 | 0.14 | 0.35 | 0.08 | 0.11 | 0.13 | 0.24 | **0.29** | 0.37 | 0.30 |
| **+ CAFES (Ours)** | **0.23** | **0.37** | **0.40** | **0.31** | **0.35** | **0.27** | **0.24** | 0.24 | **0.33** | **0.34** |
| **Improvements** | ↑0.11 | ↑0.24 | ↑0.04 | ↑0.23 | ↑0.24 | ↑0.14 | - | ↓0.04 | ↑0.04 | ↑0.04 |
| InternVL2.5-8B (Chen et al., 2025b) | 0.32 | 0.20 | 0.27 | 0.27 | 0.11 | 0.21 | 0.26 | 0.39 | 0.09 | 0.10 |
| **+ CAFES (Ours)** | **0.38** | **0.37** | **0.32** | **0.38** | **0.29** | **0.29** | **0.27** | **0.39** | **0.29** | **0.17** |
| **Improvements** | ↑0.07 | ↑0.18 | ↑0.05 | ↑0.12 | ↑0.18 | ↑0.09 | ↑0.01 | - | ↑0.20 | ↑0.07 |
| InternVL2.5-26B (Chen et al., 2025b) | **0.48** | 0.26 | 0.28 | **0.46** | 0.23 | **0.33** | **0.31** | **0.33** | 0.32 | 0.30 |
| **+ CAFES (Ours)** | 0.42 | **0.38** | **0.30** | 0.40 | **0.35** | 0.31 | 0.25 | 0.29 | **0.34** | **0.31** |
| **Improvements** | ↓0.06 | ↑0.12 | ↑0.02 | ↓0.05 | ↑0.12 | ↓0.02 | ↓0.07 | ↓0.04 | ↑0.02 | ↑0.01 |
| Qwen2.5-VL-3B (Chen et al., 2025c) | 0.19 | 0.28 | 0.34 | 0.19 | 0.29 | 0.20 | 0.26 | 0.29 | 0.34 | 0.32 |
| **+ CAFES (Ours)** | **0.30** | **0.30** | **0.39** | **0.44** | **0.32** | **0.34** | **0.27** | **0.35** | **0.37** | **0.35** |
| **Improvements** | ↑0.11 | ↑0.02 | ↑0.05 | ↑0.25 | ↑0.02 | ↑0.13 | ↑0.01 | ↑0.05 | ↑0.03 | ↑0.03 |
| Qwen2.5-VL-32B (Chen et al., 2025c) | 0.43 | 0.40 | **0.50** | 0.48 | 0.39 | 0.38 | 0.26 | 0.35 | **0.46** | 0.22 |
| **+ CAFES (Ours)** | **0.49** | **0.47** | 0.49 | **0.51** | **0.51** | **0.43** | **0.26** | **0.38** | 0.43 | **0.25** |
| **Improvements** | ↑0.06 | ↑0.07 | ↓0.01 | ↑0.03 | ↑0.13 | ↑0.05 | - | ↑0.02 | ↓0.03 | ↑0.03 |
| LLaMA-3.2-Vision-11B (Dubey et al., 2024) | 0.25 | 0.16 | 0.22 | 0.22 | 0.17 | 0.21 | 0.11 | 0.16 | 0.20 | 0.14 |
| **+ CAFES (Ours)** | **0.32** | **0.29** | **0.31** | **0.35** | **0.28** | **0.29** | **0.14** | **0.27** | **0.29** | **0.20** |
| **Improvements** | ↑0.07 | ↑0.13 | ↑0.10 | ↑0.13 | ↑0.11 | ↑0.08 | ↑0.02 | ↑0.10 | ↑0.09 | ↑0.06 |
| LLaMA-3.2-Vision-90B (Dubey et al., 2024) | 0.40 | 0.29 | 0.38 | 0.40 | 0.30 | 0.32 | 0.21 | 0.30 | 0.35 | 0.16 |
| **+ CAFES (Ours)** | **0.45** | **0.42** | **0.43** | **0.46** | **0.44** | **0.39** | **0.25** | **0.33** | **0.37** | **0.22** |
| **Improvements** | ↑0.06 | ↑0.12 | ↑0.06 | ↑0.06 | ↑0.14 | ↑0.07 | ↑0.03 | ↑0.03 | ↑0.01 | ↑0.06 |
| ***Closed-Source MLLMs*** | | | | | | | | | | |
| Claude-3.5-Sonnet (Anthropic, 2024) | 0.50 | 0.53 | 0.49 | 0.54 | 0.52 | 0.55 | **0.41** | 0.39 | 0.57 | 0.27 |
| **+ CAFES (Ours)** | **0.58** | **0.59** | **0.55** | **0.55** | **0.58** | **0.55** | 0.39 | **0.45** | **0.57** | **0.35** |
| **Improvements** | ↑0.08 | ↑0.06 | ↑0.07 | ↑0.01 | ↑0.06 | ↑0.01 | ↓0.02 | ↑0.06 | - | ↑0.08 |
| Gemini-2.5-Flash (DeepMind, 2025) | 0.33 | 0.26 | 0.47 | 0.28 | 0.30 | 0.36 | **0.31** | 0.32 | **0.51** | 0.23 |
| **+ CAFES (Ours)** | **0.40** | **0.39** | **0.47** | **0.41** | **0.36** | **0.38** | 0.28 | **0.34** | 0.50 | **0.26** |
| **Improvements** | ↑0.07 | ↑0.13 | - | ↑0.13 | ↑0.06 | ↑0.02 | ↓0.03 | ↑0.03 | ↓0.01 | ↑0.03 |
| GPT-4o-mini (OpenAI, 2024) | 0.51 | 0.34 | 0.48 | **0.64** | 0.38 | **0.50** | 0.37 | **0.55** | 0.45 | 0.24 |
| **+ CAFES (Ours)** | **0.51** | **0.50** | **0.52** | 0.57 | **0.54** | 0.49 | **0.37** | 0.44 | **0.48** | **0.28** |
| **Improvements** | - | ↑0.16 | ↑0.04 | ↓0.07 | ↑0.15 | ↓0.01 | - | ↓0.11 | ↑0.03 | ↑0.04 |

Table 2: QWK scores of different student MLLMs on ten multi-granular essay traits. For each MLLM, the first row shows the baseline, the second shows the final result with CAFES, and the third shows the improvement. Improvements are marked in **green** arrow ↑ , while declines are indicated in **red** arrow ↓.




(a) Claude-3.5-Sonnet




(b) InternVL2.5-8B

Figure 5: Improvements of QWK score across all traits based on different student MLLMs.

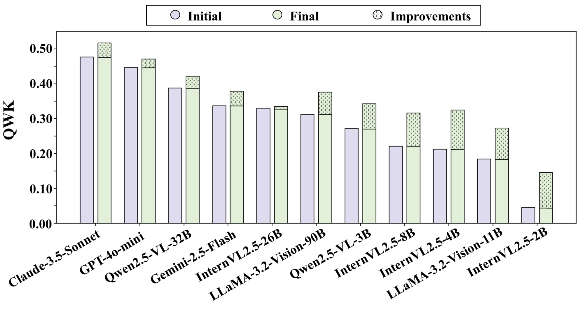



Figure 6: Improvements of average QWK score across ten traits of all student MLLMs.

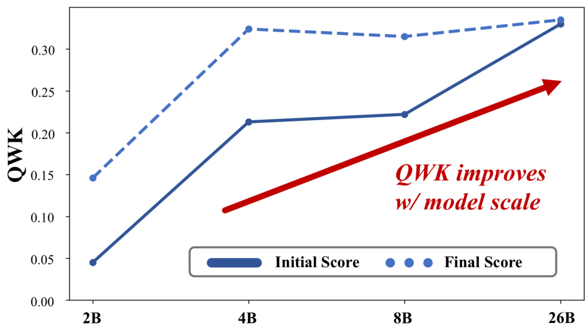



Figure 8: The average QWK scores of InternVL2.5 at different parameter scales (2B, 4B, 8B, 26B), evaluated before and after reflective feedback.

more room for improvement, while stronger ones are already close to the teacher MLLM's level.

### 4.3 Analysis of #image Setting

**CAFES framework yields greater improvements in the multi-image setting across most traits.** As shown in Figure 7, initial scores under multi-image settings tend to be more conservative, as MLLMs face greater difficulty in interpreting complex visual inputs. This conservative scoring provides more room for adjustment, allowing the CAFES framework to achieve more noticeable improvements. Notably, these more improvements of QWK in multi-image settings also supports the necessity of incorporating multimodal inputs.

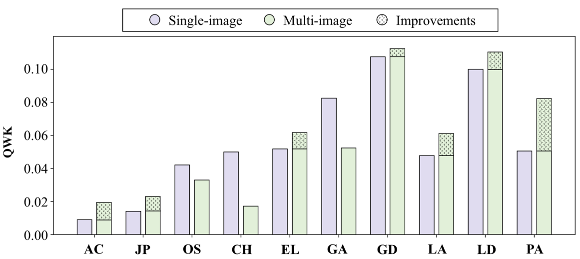



Figure 7: Improvements of average QWK of all student MLLMs under single-image and multi-image settings.

**For traits like Organizational Structure and Coherence, single-image topics yield greater improvements than multi-image topics.** Unlike most traits where multi-image topics lead to the greatest improvements, organizational structure and coherence exhibit higher QWK improvements in the single-image setting (as shown in Figure 7). This may be attributed to the fact that single-image topics typically provide less visual information, which lowers the demand for essay structure and coherence in students' essays. In such cases, models find it easier to evaluate structural clarity and coherence, resulting in more confident scoring and greater improvements after reflection.

### 4.4 Scaling Analysis

**The performance of the student MLLM consistently improves with the scale of MLLM parameters.** We observe a trend similar to the scaling law (Kaplan et al., 2020) in our setting. As shown in Figure 8, when the size of InternVL2.5 increases from 2B to 26B, the average QWK score rises from 0.045 to 0.33 in the initial scoring stage. After incorporating CAFES framework, the performance further improves, with the QWK increasing from 0.146 to 0.335. This result suggests that larger MLLMs exhibit stronger alignment with human judgment and greater reasoning ability.

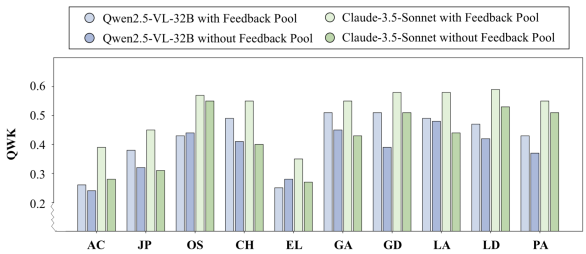



Figure 9: QWK changes w/ and w/o Feedback Pool for Qwen2.5-VL-32B & Claude-3.5-Sonnet.

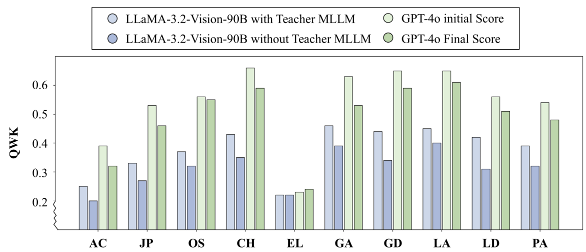



Figure 10: QWK scores across ten traits for two MLLMs (GPT-4o and LLaMA-3.2-Vision-90B), w/ and w/o the teacher-student collaboration mechanism.

### 4.5 Ablation Study

We conduct two ablation studies to test key components of CAFES. The first removes the Feedback

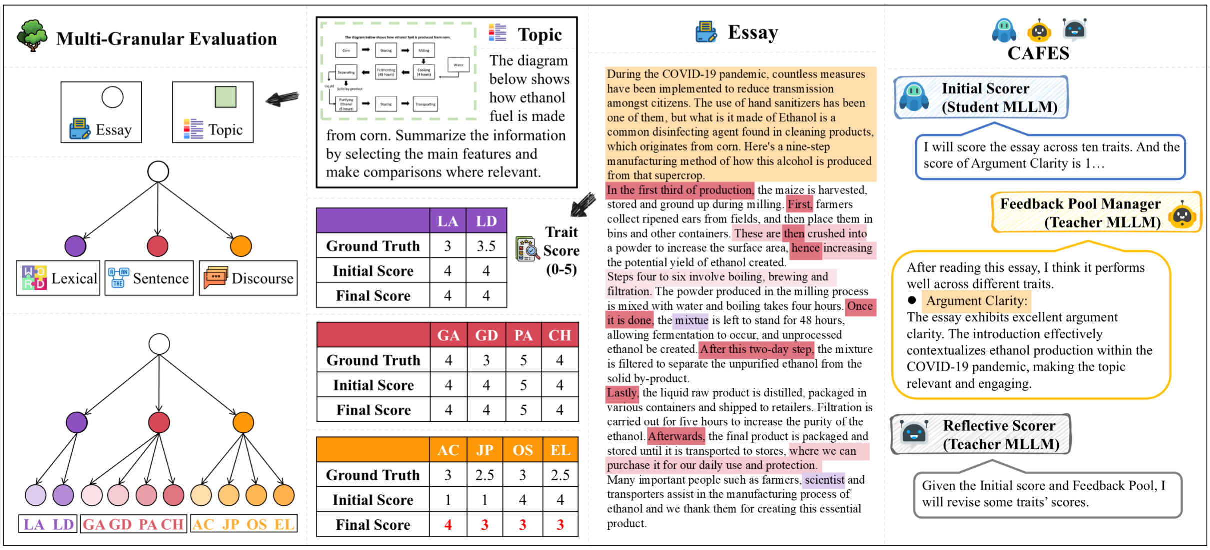



Figure 11: A representative case study illustrating CAFES's score revision process. The student MLLM is Claude-3.5-Sonnet, and the teacher MLLM is GPT-4o.

Pool Manager. The second uses the same MLLM for both student and teacher models.

**Removing the Feedback Pool Manager results in reflected scores worse than the initial ones.** We allowed the initial scores to be directly revised without positive feedback. We apply this setup to two strong MLLMs as the student model: Claude-3.5-Sonnet (closed-source) and Qwen2.5-VL-32B (open-source). In both cases, QWK scores drop after reflection without feedback pool (Figure 9). This may be because CAFES without positive feedback tends to over-focus on errors while ignoring strengths in essays, which underscores the importance of structured, trait-level positive feedback.

**Removing the teacher-student collaboration mechanism leads to a substantial drop in QWK.** We test two variants: using LLaMA-3.2-Vision-90B or GPT-4o for both student and teacher roles. As shown in Figure 10, QWK decreases notably compared to the original teacher–student setup, even when both roles use GPT-4o. These results suggest that different role assignment and independent reasoning between student and teacher MLLMs are essential for effective score revision in CAFES 's cross-agent collaborative framework.

### 4.6 Case Study

To demonstrate how our CAFES framework revises scores through feedback and reflection, we show an example using Claude-3.5-Sonnet as the student MLLM (as shown in Figure 11). More examples are shown in Appendix C. The essay explains how ethanol is produced, based on a flow chart. For example, we can find that the student MLLM initially gives a low argument clarity score of 1, while the ground truth score is 3. As mentioned in 4.2, this is because the MLLM focuses too much on surface-level errors and overlooks key strengths, such as the relevant introduction and logical structure. After receiving positive, trait-specific feedback from the teacher MLLM, the Reflective Scorer revises the score upward. The final score better matches the human judgment, showing that targeted feedback helps correct overly harsh assessments and highlight overlooked merits. This case shows how the CAFES framework uses structured feedback and subsequent reflection to refine the initial model output, leading to better alignment with human scores and preferences.

## 5 Conclusion

In this work, we present CAFES, the first collaborative multi-agent framework for AES task. It divides the essay scoring process into three core stages (*i.e.,* initial scoring, feedback generation, and reflective revision), enabling structured collaboration between three agents. Experiments across different student MLLMs show significant QWK improvements with CAFES framework, especially in grammatical and lexical diversity. Ablation studies further confirm the necessity of the Feedback Pool Manager and teacher-student collaboration mechanism. We hope CAFES can offer a new paradigm for building reliable and human-aligned AES systems and encourage the community to advance more effective and accurate scoring methods.

## Limitations

Despite the improvements we demonstrate in our CAFES framework, there are still minor limitations:

1. Our framework is evaluated on the EssayJudge dataset and achieves notable improvements over baseline models. However, EssayJudge — the only available multimodal essay dataset — focuses mainly on chart-based topics and does not cover more complex visual inputs such as film frames. We plan to include a broader range of multimodal essay types in future evaluations.

2. The reflection mechanism helps suppress hallucinations from a single MLLM, such as fabricated justifications or misinterpretation of charts, but hallucination-induced scoring errors still occur. In future work, we aim to further strengthen evidence grounding.

## References


Anthropic. 2024. Claude 3.5 sonnet.

John Atkinson and Diego Palma. 2025. An llm-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15(1):14551.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. journal of technology, learning, and assessment, 4(3). *Journal of Technology, Learning, and Assessment*, 4.

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–115.

Yida Cai, Kun Liang, Sanwoo Lee, Qinghan Wang, and Yunfang Wu. 2025. Rank-then-score: Enhancing large language models for automated essay scoring. *arXiv preprint arXiv:2504.05736*.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1011–1020. Association for Computing Machinery.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752.

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. 2025a. Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. *arXiv preprint arXiv:2502.12520*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Jaeyoon Choi, Tamara Tate, Daniel Ritchie, Nia Nixon, and Mark Warschauer. 2025. Anchor is the key: Toward accessible automated essay scoring with large language models through prompting.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and 1 others. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018a. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018b. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, and 1 others. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Google DeepMind. 2025. Gemini 2.5 flash.

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 preprint arXiv:2407.21783*.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English. Version 2. Handbook and CD-ROM*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024. Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models. *arXiv preprint arXiv:2410.04819*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. *arXiv preprint arXiv:2502.11051*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Thorben Jansen, Jennifer Meyer, Johanna Fleckenstein, Andrea Horbach, Stefan Keller, and Jens Möller. 2024. Individualizing goal-setting interventions using automated writing evaluation to support secondary school students' text revisions. *Learning and Instruction*, 89:101847.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470. Association for Computational Linguistics.

Firuz Kamalov, David Santandreu Calonge, Linda Smail, Dilshod Azizov, Dimple R Thadani, Theresa Kwong, and Amara Atif. 2025. Evolution of ai in education: Agentic workflows. *arXiv preprint arXiv:2504.20082*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders?

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024a. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024b. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Preprint*, arXiv:2303.17760.

Hang Li, Tianlong Xu, Chaoli Zhang, Eason Chen, Jing Liang, Xing Fan, Haoyang Li, Jiliang Tang, and Qingsong Wen. 2024. Bringing generative ai to adaptive learning in education. *arXiv preprint arXiv:2402.14601*.

Shengjie Li and Vincent Ng. 2024a. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888.

Shengjie Li and Vincent Ng. 2024b. Automated essay scoring: Recent successes and future directions. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8114–8122.

Shengjie Li and Vincent Ng. 2024c. Icle++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8458–8478.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Shalev Lifshitz, Sheila A McIlraith, and Yilun Du. 2025. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*.

Chun Then Lim, Chih How Bong, Wee Sian Wong, and Nung Kion Lee. 2021. A comprehensive review of automated essay scoring (aes) research and development. *Pertanika Journal of Science & Technology*, 29(3):1875–1899.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Zeyang Liu, Xinrui Yang, Shiguang Sun, Long Qian, Lipeng Wan, Xingyu Chen, and Xuguang Lan. 2024b. Grounded answers for multi-agent decision-making problem through generative world model. *Advances in Neural Information Processing Systems*, 37:46622–46652.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? *Preprint*, arXiv:2403.06149.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

SeungWoo Song, Junghun Yuk, ChangSu Choi, HanGyeol Yoo, Hyeonseok Lim, KyungTae Lim, and Jungyeul Park. 2025a. Unified automated essay scoring and grammatical error correction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4412–4426.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025b. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*.

Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

Jiamin Su, Yibo Yan, Fangteng Fu, Han Zhang, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025. Essayjudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th international conference on computational linguistics*, pages 6077–6088.

Sowmya Vajjala. 2016. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *CoRR*.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, and 1 others. 2025. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797, Brussels, Belgium. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Xuansheng Wu, Padmaja Pravin Saraf, Gyeong-Geon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. 2024. Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring. *arXiv preprint arXiv:2407.18328*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Wei Xia, Shaoguang Mao, and Chanjing Zheng. 2024. Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example. *arXiv preprint arXiv:2401.03401*.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. Human-ai collaborative essay scoring: A dual-process framework with llms. *Preprint*, arXiv:2401.06431.

Wenbo Xu, Muhammad Shahreeza, Wai Lam Hoo, and Wudao Yang. 2025. Explainable ai for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, and 1 others. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025a. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. 2025b. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024c. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017.

Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guangliang Chen. 2024. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22466–22474.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.

Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Zenglin Xu, Irwin King, Philip S Yu, and Qingsong Wen. 2025. Position: Llms can be good tutors in foreign language education. *arXiv preprint arXiv:2502.05467*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yuan Yuan, Zhaojian Li, and Bin Zhao. 2025. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*.

Zhiqiang Yuan, Weitong Chen, Hanlin Wang, Kai Yu, Xin Peng, and Yiling Lou. 2024. Transagent: An llm-based multi-agent system for code translation. *arXiv preprint arXiv:2409.19894*.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, and 1 others. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606.

## A More Related Work

### A.1 AES Datasets

Table 3 summarizes widely used AES datasets in terms of dataset size, number of essay topics, modality, and trait-level annotations. Most existing datasets (*e.g.,* ASAP, CLC-FCE, TOEFL11) are unimodal and offer either holistic scores or a limited number of traits, primarily focusing on text-based prompts. Recently, EssayJudge (Su et al., 2025) has been introduced as a multimodal benchmark that includes both textual and visual inputs, covering 125 topics and annotated across ten fine-grained scoring traits. This enables more comprehensive evaluation of AES systems, especially those leveraging MLLMs.

### A.2 Multimodal Large Language Models

MLLMs have experienced rapid development in recent years and have been widely adopted across various domains (Yuan et al., 2025; Song et al., 2025b; Yan et al., 2024a). Their core advantage lies in the ability to jointly process visual and textual inputs to handle a range of complex tasks (Xi et al., 2023; Huo et al., 2024; Yan et al., 2024c; Yan and Lee, 2024; Zou et al., 2025; Dang et al., 2024; Huo et al., 2025; Chen et al., 2025a). On the proprietary side, MLLMs such as GPT-4o (Hurst et al., 2024) and Gemini-1.5 (DeepMind, 2025) have demonstrated state-of-the-art performance in multimodal reasoning, instruction following, and question answering tasks (Chang et al., 2024; Yan et al., 2024b,a; Zheng et al., 2024; Yan et al., 2025a). Meanwhile, open-source MLLMs have made notable advances in terms of accessibility and modularity. LLaVA-NEXT (Liu et al., 2024a) employs pretrained encoders and adapters to align vision and language representations efficiently. Other representative MLLMs—such as Qwen2.5-VL (Chen et al., 2025c), DeepSeek-VL (Lu et al., 2024), InternVL (Chen et al., 2025b), Yi-VL (Young et al., 2024), LLaMA3-VL (Dubey et al., 2024), and MiniCPM-V (Hu et al., 2024)—have introduced a variety of fusion mechanisms, including visual projection heads, mixture-of-experts architectures, and image-grounded token masking. These MLLMs have been applied to a wide range of domains, including education and medical diagnostics (Qu et al., 2025; Zou et al., 2024; Zhou et al., 2024; Huang et al., 2024), showcasing the expanding scope and depth of MLLM capabilities.

Building upon these diverse MLLMs, our proposed CAFES multi-agent framework flexibly incorporates different MLLMs as the backbone for each agent module, enabling collaborative interaction between student and teacher MLLMs to enhance the accuracy and robustness of AES.

## B Additional Experimental Details

### B.1 Trait-Specific Rubrics

In this section, we introduce the rubrics of the 10 traits which is similar to EssayJudge. The rubrics are detailed in Table 4 to Table 13. Each trait is assessed using a numerical score ranging from 0 to 5. A score of 5 represents high-quality performance with respect to the trait being evaluated, while a score of 0 represents low-quality performance in the same regard.

### B.2 Prompt for CAFES framework

Our agent-based CAFES framework consists of three modules—Initial Scorer, Feedback Pool Generator, and Reflective Scorer—each employing customized prompt designs for their functional roles. While all agents operate under a unified trait-based rubric schema, the input structure and expected output vary to support multi-stage evaluation. The details are shown in Figure 12 to 14.

### B.3 Model Sources

Table 14 details specific sources for the various student MLLMs. The hyperparameters for the experiments are set to their default values unless specified otherwise.

### B.4 Average Trait-Specific Score Comparison

Closed-source MLLMs tend to adopt a more rigorous scoring strategy compared to open-source MLLMs. This trend is supported by both quantitative and distributional evidence. First, as shown in the Figure 15, closed-source models consistently assign lower average scores than open-source models across most traits, regardless of whether the scores are initial or final (Su et al., 2025; Kundu and Barbosa, 2024). Second, Figure 16 reveals that closed-source models exhibit slightly higher score variance (0.81 vs. 0.79), indicating a broader and possibly more cautious distribution of judgments. Together, these findings suggest that closed-source MLLMs are more aligned with rigorous rubric interpretation, both when directly scoring and when acting as student models within the CAFES framework.

### B.5 More examples of Findings

Beyond the examples of Claude-3.5-Sonnet and InternVL2.5-8B presented in the main paper, additional cases support this observation, which is shown in Figure 17.

## C More Essay Scoring examples

To further illustrate the effectiveness of our multi-agent AES framework, we include several additional essay cases (as shown in Figure 18 to 20). Each example consists of the essay topic, student's essays, initial scores, feedback pool, and final revised scores after reflection. These examples highlight different error types, model reasoning behaviors, and improvement patterns across traits.

| Benchmarks | Venue | Size | #Topics | Modality | #Traits |
|---|---|---|---|---|---|
| ASAP$_{\text{AES}}$ (Cozma et al., 2018a) | ACL | 17,450 | 8 | T | 0 |
| ASAP++ (Mathias and Bhattacharyya, 2018) | ACL | 10,696 | 6 | T | 8 |
| CLC-FCE (Yannakoudakis et al., 2011) | ACL | 1,244 | 10 | T | 0 |
| TOEFL11 (Lee et al., 2024a) | EMNLP | 1,100 | 8 | T | 0 |
| ICLE (Granger et al., 2009) | COLING | 3,663 | 48 | T | 4 |
| AAE (Stab and Gurevych, 2014) | COLING | 102 | 101 | T | 1 |
| ICLE++ (Li and Ng, 2024c) | NAACL | 1,008 | 10 | T | 10 |
| CREE (Bailey and Meurers, 2008) | BEA | 566 | 75 | T | 1 |
| EssayJudge (Su et al., 2025) | - | 1054 | 125 | T, I | 10 |

Table 3: Comparison between previous AES datasets.

| Score | Scoring Criteria |
|---|---|
| 5 | The central argument is clear, and the first paragraph clearly outlines the topic of the image and question, providing guidance with no ambiguity. |
| 4 | The central argument is clear, and the first paragraph mentions the topic of the image and question, but the guidance is slightly lacking or the expression is somewhat vague. |
| 3 | The argument is generally clear, but the expression is vague, and it doesn't adequately guide the rest of the essay. |
| 2 | The argument is unclear, the description is vague or incomplete, and it doesn't guide the essay. |
| 1 | The argument is vague, and the first paragraph fails to effectively summarize the topic of the image or question. |
| 0 | No central argument is presented, or the essay completely deviates from the topic and image. |

Table 4: Rubrics for evaluating the argument clarity of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Transitions between sentences are natural, and logical connections flow smoothly; appropriate use of linking words and transitional phrases. |
| 4 | Sentences are generally coherent, with some transitions slightly awkward; linking words are used sparingly but are generally appropriate. |
| 3 | The logical connection between sentences is not smooth, with some sentences jumping or lacking flow; linking words are used insufficiently or inappropriately. |
| 2 | Logical connections are weak, sentence connections are awkward, and linking words are either used too little or excessively. |
| 1 | There is almost no logical connection between sentences, transitions are unnatural, and linking words are very limited or incorrect. |
| 0 | No coherence at all, with logical confusion between sentences. |

Table 5: Rubrics for evaluating the coherence of the essays.

**Task Definition:** You are an experienced English writing examiner. Please evaluate the student's essay by assigning a score (0-5) for each of the ten traits and a confidence level (1–10) that reflects how certain you are about each score, where 1 is least certain and 10 is completely certain.

**Rubrics:** {Trait-specific corresponding rubrics}

**Below is the reference content:**
Image: "{image}"
Essay Topic: "{question}"
Student's Essay: "{essay}"

**Instruction:** Please provide your answer in the same style and format as the example. Use the exact trait names as shown (with proper capitalization) Return your response strictly in JSON format without any additional text, explanations, or code block delimiters (no triple backticks).

Figure 12: Prompt for Initial Scorer.

| Score | Scoring Criteria |
|---|---|
| 5 | Word count is 150 words or more, with the content being substantial and without obvious excess or brevity. |
| 4 | Word count is around 150 words, but slightly off (within 10 words), and the content is complete. |
| 3 | Word count is noticeably too short or too long, and the content is not sufficiently substantial or is somewhat lengthy. |
| 2 | Word count deviates significantly, failing to fully cover the requirements of the prompt. |
| 1 | Word count is far below the requirement, and the content is incomplete. |
| 0 | Word count is severely insufficient or excessive, making it impossible to meet the requirements of the prompt. |

Table 6: Rubrics for evaluating the essay length of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Sentence structure is accurate with no grammatical errors; both simple and complex sentences are error-free. |
| 4 | Sentence structure is generally accurate, with occasional minor errors that do not affect understanding; some errors in complex sentence structures. |
| 3 | Few grammatical errors, but more noticeable errors that affect understanding; simple sentences are accurate, but complex sentences frequently contain errors. |
| 2 | Numerous grammatical errors, with sentence structure affecting understanding; simple sentences are occasionally correct, but complex sentences have frequent errors. |
| 1 | A large number of grammatical errors, with sentence structure severely affecting understanding; sentence structure is unstable, and even simple sentences contain mistakes. |
| 0 | Sentence structure is completely incorrect, nonsensical, and difficult to understand. |

Table 7: Rubrics for evaluating the grammatical accuracy of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Uses a variety of sentence structures, including both simple and complex sentences, with flexible use of clauses and compound sentences, demonstrating rich sentence variation. |
| 4 | Generally uses a variety of sentence structures, with appropriate use of common clauses and compound sentences. Sentence structures vary, though some sentence types lack flexibility. |
| 3 | Uses a variety of sentence structures, but with limited use of complex sentences, which often contain errors. Sentence variation is somewhat restricted. |
| 2 | Sentence structures are simple, primarily relying on simple sentences, with occasional attempts at complex sentences, though errors occur frequently. |
| 1 | Sentence structures are very basic, with almost no complex sentences, and even simple sentences contain errors. |
| 0 | Only uses simple, repetitive sentences with no complex sentences, resulting in rigid sentence structures. |

Table 8: Rubrics for evaluating the grammatical diversity of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Fully addresses and accurately analyzes all important information in the image and prompt (*e.g.*, data turning points, trends); argumentation is in-depth and logically sound. |
| 4 | Addresses most of the important information in the image and prompt, with reasonable analysis but slight shortcomings; argumentation is generally logical. |
| 3 | Addresses some important information in the image and prompt, but analysis is insufficient; argumentation is somewhat weak. |
| 2 | Mentions a small amount of important information in the image and prompt, with simple or incorrect analysis; there are significant logical issues in the argumentation. |
| 1 | Only briefly mentions important information in the image and prompt or makes clear analytical errors, lacking reasonable reasoning. |
| 0 | Fails to mention key information from the image and prompt, lacks any argumentation, and is logically incoherent. |

Table 9: Rubrics for evaluating the justifying persuasiveness of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Vocabulary is accurately chosen, with correct meanings and spelling, and minimal errors; words are used precisely to convey the intended meaning. |
| 4 | Vocabulary is generally accurate, with occasional slight meaning errors or minor spelling mistakes, but they do not affect overall understanding; words are fairly precise. |
| 3 | Vocabulary is mostly correct, but frequent minor errors or spelling mistakes affect some expressions; word choice is not fully precise. |
| 2 | Vocabulary is inaccurate, with significant meaning errors and frequent spelling mistakes, affecting understanding. |
| 1 | Vocabulary is severely incorrect, with unclear meanings and noticeable spelling errors, making comprehension difficult. |
| 0 | Vocabulary choice and spelling are completely incorrect, and the intended meaning is unclear or impossible to understand. |

Table 10: Rubrics for evaluating the lexical accuracy of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Vocabulary is rich and diverse, with a wide range of words used flexibly, avoiding repetition. |
| 4 | Vocabulary diversity is good, with a broad range of word choices, occasional repetition, but overall flexible expression. |
| 3 | Vocabulary diversity is average, with some variety in word choice but limited, with frequent repetition. |
| 2 | Vocabulary is fairly limited, with a lot of repetition and restricted word choice. |
| 1 | Vocabulary is very limited, with frequent repetition and an extremely narrow range of words. |
| 0 | Vocabulary is monotonous, with almost no variation, failing to demonstrate vocabulary diversity. |

Table 11: Rubrics for evaluating the lexical diversity of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | The essay has a well-organized structure, with clear paragraph divisions, each focused on a single theme. There are clear topic sentences and concluding sentences, and transitions between paragraphs are natural. |
| 4 | The structure is generally reasonable, with fairly clear paragraph divisions, though transitions may be somewhat awkward and some paragraphs may lack clear topic sentences. |
| 3 | The structure is somewhat disorganized, with unclear paragraph divisions, a lack of topic sentences, or weak logical flow. |
| 2 | The structure is unclear, with improper paragraph divisions and poor logical coherence. |
| 1 | The paragraph structure is chaotic, with most paragraphs lacking clear topic sentences and disorganized content. |
| 0 | No paragraph structure, content is jumbled, and there is a complete lack of logical connections. |

Table 12: Rubrics for evaluating the organizational structure of the essays.

| Score | Scoring Criteria |
|---|---|
| 5 | Punctuation is used correctly throughout, adhering to standard rules with no errors. |
| 4 | Punctuation is mostly correct, with occasional minor errors that do not affect understanding. |
| 3 | Punctuation is generally correct, but there are some noticeable errors that slightly affect understanding. |
| 2 | There are frequent punctuation errors, some of which affect understanding. |
| 1 | Punctuation errors are severe, significantly affecting comprehension. |
| 0 | Punctuation is completely incorrect or barely used, severely hindering understanding. |

Table 13: Rubrics for evaluating the punctuation accuracy of the essays.

**Task Definition:** You are an experienced English writing examiner. Your task is to provide detailed positive feedback on a student essay across ten traits: Argument Clarity, Justifying Persuasiveness, Organizational Structure, Coherence, Essay Length, Grammatical Accuracy, Grammatical Diversity, Lexical Accuracy, Lexical Diversity, Punctuation Accuracy.

**Rubrics:** {Trait-specific corresponding rubrics}

**Below is the reference content:**
Image: "{image}"
Essay Topic: "{question}"
Student's Essay: "{essay}"

**Instruction:** please generate your feedback dimension by dimension.Your output must be in natural language paragraphs. Do not use JSON, code blocks, or bullet points. Start each dimension with the tag in square brackets, for example: [Argument Clarity]
Sample Format: [Argument Clarity] The opening paragraph clearly introduces the topic of the image and outlines the overall trend, effectively setting up the structure for later analysis.

Figure 13: Prompt for Feedback Pool Manager.

**Task Definition:** You are evaluating a set of essay scores originally provided by another assistant reviewer. A detailed feedback report—including both positive and negative comments across 10 traits—is available for reference, but should not be treated as an absolute judgment. Your task is to serve as a more careful and critical second-round reviewer. Do not assume the original scores are correct — examine each trait carefully and revise any score that appears inaccurate or unsupported by the essay.

**Rubrics:** {Trait-specific corresponding rubrics}

**Below is the reference content:**
Image: "{image}"
Essay Topic: "{question}"
Student's Essay: "{essay}"
Original Scores :"{score}"
Feedback Report :"{feedback}"

**Instruction:** If revision is needed, return only the affected dimensions with new scores and brief reasoning. Otherwise, confirm the original scores. Return your response strictly in JSON format without any additional text, explanations, or code block delimiters (no triple backticks). Only raw JSON is accepted.

Figure 14: Prompt for Reflective Scorer.

| MLLMs | Source | URL |
|---|---|---|
| InternVL2.5-2B | local checkpoint | https://huggingface.co/OpenGVLab/InternVL2-2B |
| InternVL2.5-4B | local checkpoint | https://huggingface.co/OpenGVLab/InternVL2-4B |
| InternVL2.5-8B | local checkpoint | https://huggingface.co/OpenGVLab/InternVL2-8B |
| InternVL2.5-26B | local checkpoint | https://huggingface.co/OpenGVLab/InternVL2-26B |
| Qwen2.5-VL-3B | local checkpoint | https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct |
| Qwen2.5-VL-32B | local checkpoint | https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct |
| LLaMA-3.2-Vision-11B | local checkpoint | https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct |
| LLaMA-3.2-Vision-90B | local checkpoint | https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct |
| Claude-3.5-Sonnet | `claude-3.5-sonnet-20241022` | https://www.anthropic.com/claude/sonnet |
| Gemini-2.5-Flash | `gemini-2.5-flash-preview-04-17` | https://deepmind.google/technologies/gemini/flash |
| GPT-4o-mini | `gpt-4o-mini-2024-07-18` | https://platform.openai.com/docs/models/gpt-4o-mini |

Table 14: Sources of our evaluated MLLMs.

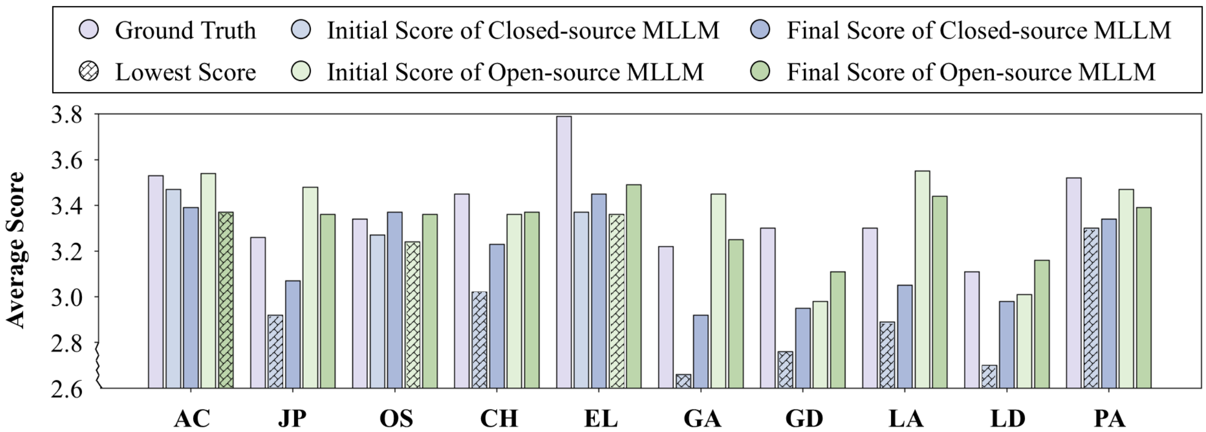



Figure 15: Average trait-specific scores assigned by closed-source and open-source MLLMs at both the initial stage and after revision through the CAFES framework.

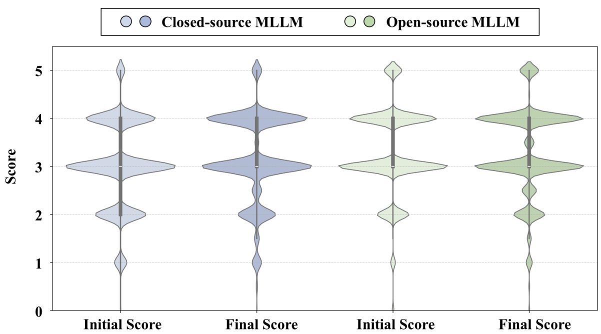
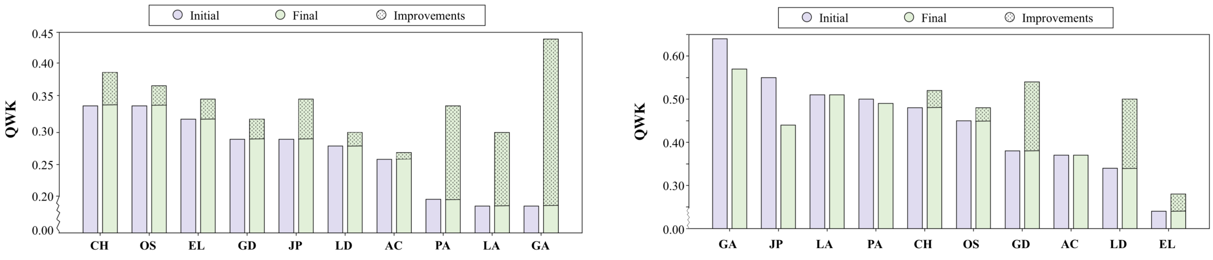



Figure 16: Score distributions of closed-source and open-source MLLMs at both the initial scoring stage and after revision through the CAFES framework.




(a) Qwen2.5-VL-3B (b) GPT-4o-mini

Figure 17: Improvements of QWK score across all traits based on different student MLLM.

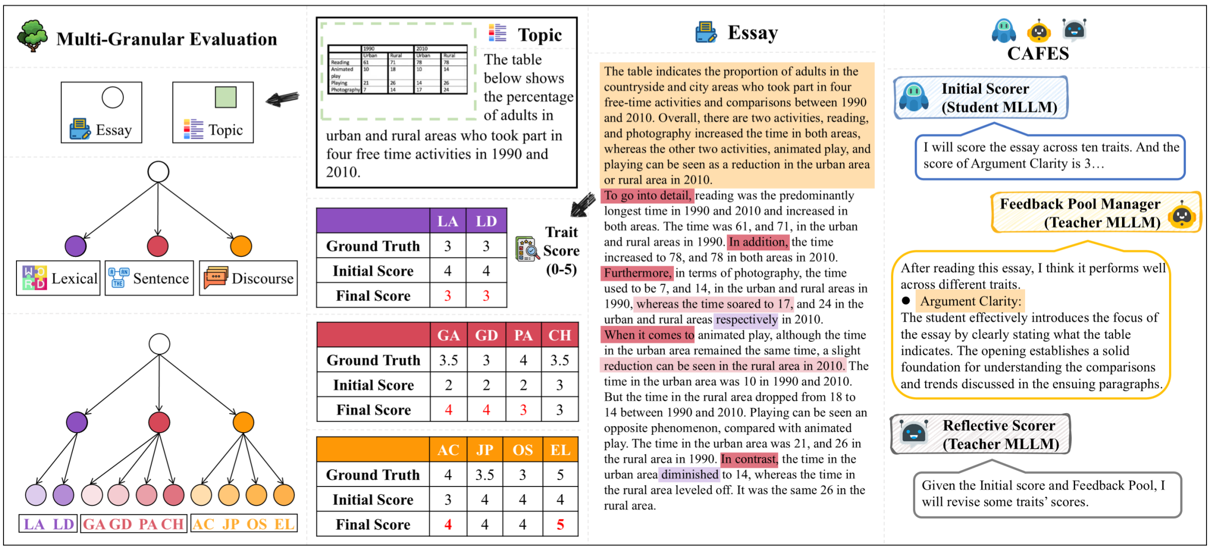



Figure 18: A case study illustrating CAFES 's score revision process. And the student MLLM is Claude-3.5-Sonnet, and the teacher MLLM is GPT-4o.

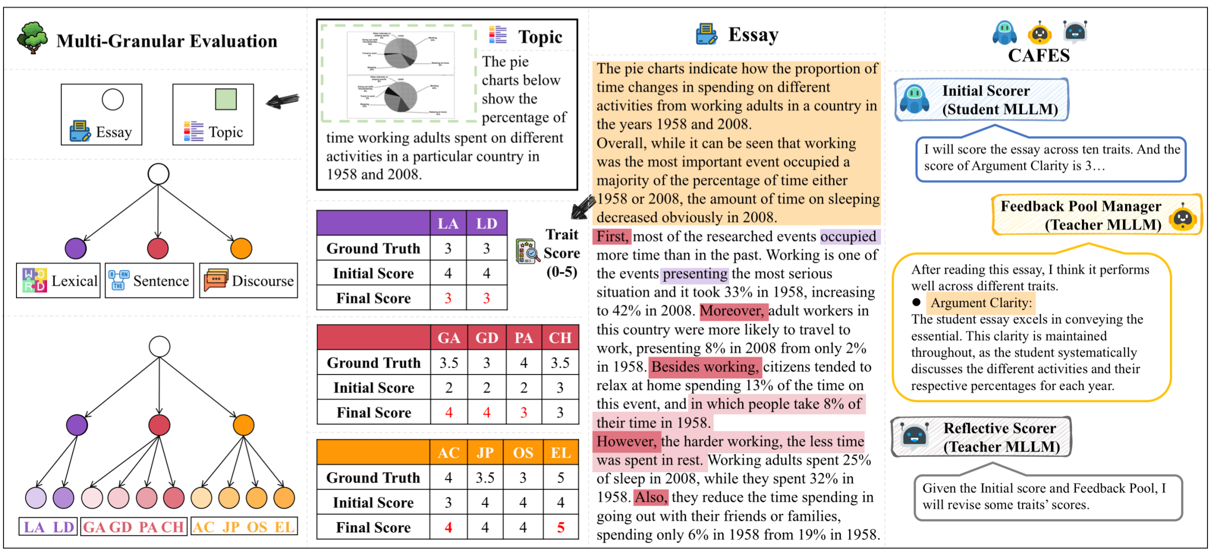



Figure 19: A case study illustrating CAFES 's score revision process. And the student MLLM is GPT-4o-mini, and the teacher MLLM is GPT-4o.

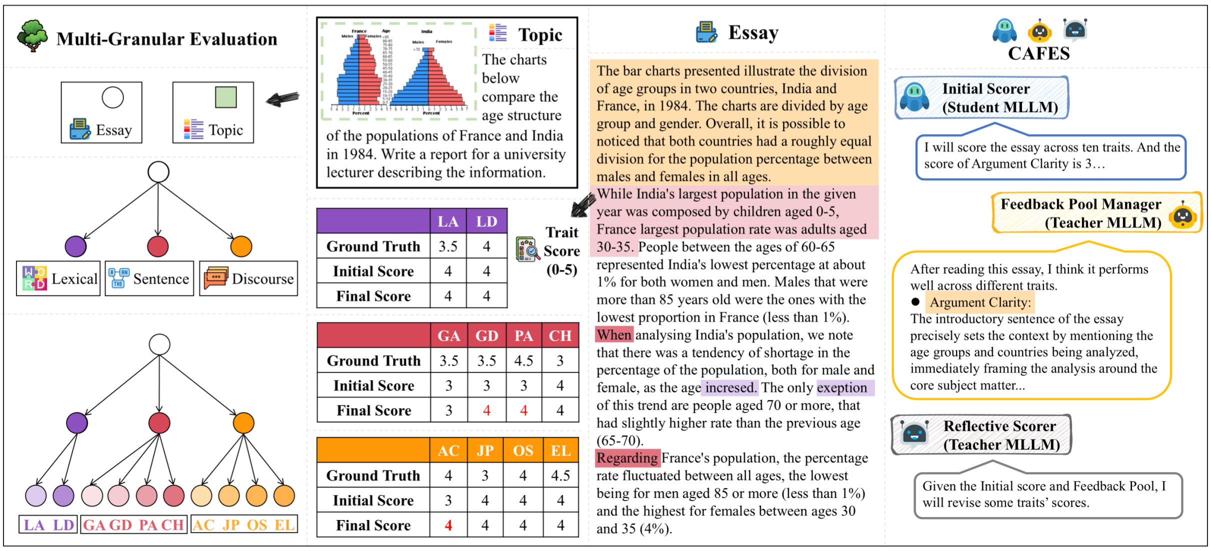



Figure 20: A case study illustrating CAFES 's score revision process. And the student MLLM is Qwen2.5-VL-32B, and the teacher MLLM is GPT-4o.