EssayJudge: A Multi-Granular Benchmark for Assessing Automated Essay Scoring Capabilities of Multimodal Large Language Models

Jiamin Su[1,2,*], **Yibo Yan**[1,3,*], **Fangteng Fu**[1], **Han Zhang**[1], **Jingheng Ye**[4], **Xiang Liu**[1,3], **Jiahao Huo**[1], **Huiyu Zhou**[5], **Xuming Hu**[1,2,3,†]

[1]The Hong Kong University of Science and Technology (Guangzhou)

[2]Beijing Future Brain Education Technology Co., Ltd.

[3]The Hong Kong University of Science and Technology, [4]Tsinghua University

[5]Guangxi Zhuang Autonomous Region Big Data Research Institute

jsu360@connect.hkust-gz.edu.cn, yanyibo70@gmail.com, xuminghu@hkust-gz.edu.cn

Abstract

Automated Essay Scoring (AES) plays a crucial role in educational assessment by providing scalable and consistent evaluations of writing tasks. However, traditional AES systems face three major challenges: ❶ reliance on handcrafted features that limit generalizability, ❷ difficulty in capturing fine-grained traits like coherence and argumentation, and ❸ inability to handle multimodal contexts. In the era of Multimodal Large Language Models (MLLMs), we propose ESSAYJUDGE, the **first multimodal benchmark to evaluate AES capabilities across lexical-, sentence-, and discourse-level traits**. By leveraging MLLMs' strengths in trait-specific scoring and multimodal context understanding, ESSAYJUDGE aims to offer precise, context-rich evaluations without manual feature engineering, addressing longstanding AES limitations. Our experiments with 18 representative MLLMs reveal gaps in AES performance compared to human evaluation, particularly in discourse-level traits, highlighting the need for further advancements in MLLM-based AES research. The dataset and code are available at this URL.

1 Introduction

Automated Essay Scoring (AES) has become an essential tool in educational assessment, providing efficient and consistent scoring for large-scale writing tasks (Ye et al., 2025; Ramesh and Sanampudi, 2022; Li and Liu, 2024; Wu et al., 2024; Xia et al., 2024). While AES systems have significantly reduced the workload of human graders, they still *face challenges in delivering accurate and detailed evaluations, particularly for trait-specific scoring*, which assesses individual aspects of writing quality, such as coherence, creativity, and argumentation (Song et al., 2024; Pack et al., 2024; Ruseti et al., 2024). Such detailed feedback is critical for guiding students in improving their writing skills, but remains difficult to achieve with existing methods.

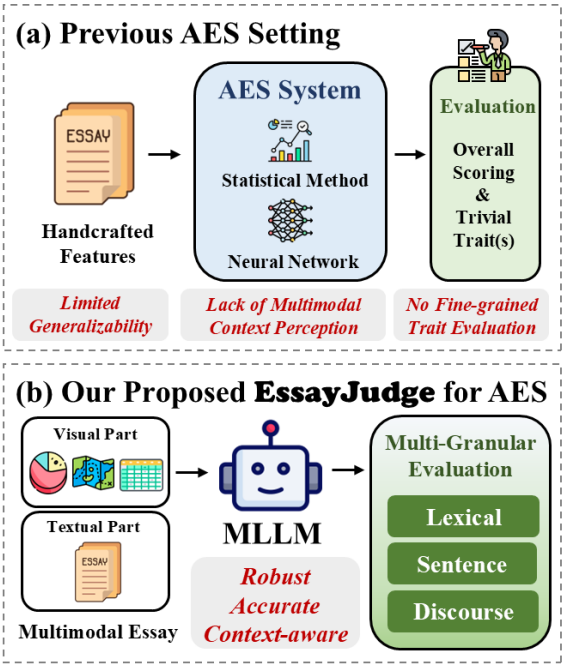


Figure 1: Comparison of task settings between the previous evaluation paradigm (a) and our proposed ESSAYJUDGE benchmark (b) on automated essay scoring task.

Traditional AES approaches, including statistical models such as Support Vector Machines, rely heavily on handcrafted features such as word frequency and essay length (Yang et al., 2024; Jansen et al., 2024; Uto et al., 2020). As illustrated in Figure 1 (a), they often suffer from ❶ relying on manually engineered features, thus limiting the generalizability across diverse data; ❷ failing to model fine-grained traits such as logical structure and argument persuasiveness; ❸ inability to handle multimodal context, thus struggling to deliver comprehensive and context-aware evaluations (Lim et al., 2021; Uto, 2021; Wang et al., 2022).

The emergence of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) offers a promising solution to these chal-

*Co-first authors with equal contribution.

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 6363–6389

lenges (Xiao et al., 2024b; Mansour et al., 2024; Ding and Zou, 2024; Luo et al., 2025). Unlike traditional models, LLMs can capture rich semantic representations directly from text, allowing them to evaluate essays holistically and provide detailed feedback on specific traits (Gao et al., 2024; Maity and Deroy, 2024). Furthermore, MLLMs extend this capability by integrating text and image inputs, enabling a deeper understanding of multimodal essay context (Lee et al., 2023; Xu et al., 2024). This not only improves scoring accuracy but also addresses the need for nuanced evaluations in more complex writing scenarios.

Therefore, we propose ESSAYJUDGE, the **first multimodal benchmark for assessing the multi-granular AES capabilities of MLLMs**. As shown in Figure 1 (b), ESSAYJUDGE aims to address the aforementioned gaps by ❶ refraining from manually engineered features, using MLLMs' inherent capabilities to automatically capture complex linguistic patterns and contextual cues, which allows for better generalization across diverse datasets; ❷ leveraging MLLMs' ability to model multi-granular traits (*i.e.,* lexical-, sentence-, and discourse-levels), enabling more precise and nuanced trait-specific evaluations; ❸ incorporating multimodal inputs including text and image components, enabling MLLM to handle complex context, which is crucial for essays on intricate topics.

Through extensive experiments, we evaluated 18 representative open-source and closed-source MLLMs, yielding the following key insights: (i) open-source MLLMs generally perform poorly in AES compared to closed-source MLLMs, particularly GPT-4o; (ii) closed-source MLLMs tend to assign lower scores across multiple traits compared to human evaluators, reflecting their stricter scoring criteria; (iii) closed-source MLLMs perform better in evaluating essays based on single-image setting compared to multi-image one. In general, our findings highlight that there is still a noticeable gap in AES performance compared to human evaluators, particularly in discourse-level traits, underscoring the necessity for further LLM research.

Our contributions can be summarized as follows:

- We introduce the **first multimodal AES dataset** ESSAYJUDGE, comprising over 1,000 high-quality multimodal English essays, each of which has undergone rigorous multi-round human annotation and verification.
- We propose a **trait-specific scoring framework** that enables comprehensive evaluation with ten multi-granular metrics, covering three dimensions: lexical, sentence, and discourse levels.
- We conduct an **in-depth evaluation of 18 state-of-the-art MLLMs and human evaluation** on ESSAYJUDGE, using Quadratic Weighted Kappa (QWK) as the primary metric to assess the trait-specific scoring performance.

Benchmarks	Venue	Size	#Topics	Modality	#Traits
$ASAP_{AES}$ (Cozma et al., 2018)	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., 2024)	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 (Ours)	-	1054	125	T, I	10

Table 1: Comparison between previous AES benchmarks and our proposed ESSAYJUDGE. The cells highlighted in red indicate the highest number for *#Topics* and *#Traits* columns, and the unique modality for *Modality* column.

By bridging the gaps in existing AES benchmarks, ESSAYJUDGE contributes to the development of more accurate, robust, and context-aware MLLM-based essay scoring systems in the era of AGI.

2 Related Work

2.1 AES Datasets

Existing AES datasets have advanced the field but remain some limitations (shown in Table 1) (Ke and Ng, 2019; Li and Ng, 2024b,a). For example, $ASAP_{AES}$ is notable for its size, enabling high-performance prompt-specific systems (Cozma et al., 2018). However, differing score ranges across prompts and heavy preprocessing (*e.g.*, removal of paragraph structures and named entities) reduce its utility. ASAP++ is an extension of ASAP that introduces trait-specific scores (Mathias and Bhattacharyya, 2018; Li and Ng, 2024a). However, its traits are coarse-grained, with all content-based traits (*e.g.*, coherence, persuasiveness, and thesis clarity) grouped into a single "CONTENT" category. The CLC-FCE dataset includes holistic scores and linguistic error annotations, supporting grammatical error detection alongside scoring tasks, but the small number of essays per prompt hinders the development of prompt-specific systems (Yannakoudakis et al., 2011; Li and Ng, 2024b). TOEFL11 dataset focuses on native language identification and provides only coarse-grained proficiency labels (low, medium & high), which do not fully capture essay quality. ICLE

(Granger et al., 2009) and ICLE++ (Li and Ng, 2024c) datasets provide some of the most detailed trait-specific annotations, with ICLE++ scoring essays on 10 dimensions of writing quality. Nevertheless, these datasets are still constrained by limited topic diversity. Similarly, The AAE corpus includes 102 persuasive essays and only focuses on argument structure (Stab and Gurevych, 2014). To address the aforementioned limitations, we propose the ESSAYJUDGE benchmark, which features multimodal context, 125 unique essay topics, and comprehensive scoring across 10 distinct traits.

2.2 AES Systems

AES research focuses on three main categories: heuristic approaches, machine learning approaches, and deep learning approaches (Li and Ng, 2024a). Heuristic AES approaches focus on holistic scoring by combining trait scores such as Organization, Coherence, and Grammar into a weighted sum. Trait-specific scores are computed using rules, like assessing Organization based on a five-paragraph format (Attali and Burstein, 2006). Machine learning approaches (*e.g.*, Logistic Regression and Support Vector Machine) rely on handcrafted features, such as lexical (Chen and He, 2013), length-based (Vajjala, 2016; Yannakoudakis and Briscoe, 2012), and discourse features (Yannakoudakis and Briscoe, 2012), and perform well in within-prompt scoring but struggle with generalization to new prompts. Deep learning approaches, particularly those using Transformer architectures like BERT (Wang et al., 2022), have advanced AES by learning essay representations directly from text, enabling multi-trait and cross-prompt scoring. Among these, LLM-based approaches stand out for their ability to leverage commonsense knowledge and understand complex instructions (Mizumoto and Eguchi, 2023). By using prompts, LLMs can perform AES in zero-shot settings with rubrics alone (Lee et al., 2024) or in few-shot settings with minimal labeled data (Mansour et al., 2024; Xiao et al., 2024a). These methods enhance flexibility, scalability, and performance, especially in low-resource scenarios.

2.3 Multimodal Large Language Models

MLLMs have brought significant advancements to diverse tasks and applications (Xi et al., 2023; Huo et al., 2024; Yan et al., 2024d; Yan and Lee, 2024; Zou et al., 2025; Dang et al., 2024). Proprietary MLLMs such as GPT-4o (Hurst et al., 2024) and Gemini-1.5 (DeepMind, 2024b) have shown remarkable capabilities in multimodal challenges, excelling in areas such as multimodal reasoning and QA (Chang et al., 2024; Yan et al., 2024c,b; Zheng et al., 2024; Yan et al., 2025a). At the same time, Open-source MLLMs have made considerable strides. For instance, LLaVA-NEXT (Liu et al., 2024) utilizes a pretrained vision encoder to generate visual embeddings, which are then aligned with text embeddings through a lightweight adapter, enabling effective multimodal understanding. Similarly, MLLMs such as Qwen2-VL (Wang et al., 2024), DeepSeek-VL (Lu et al., 2024a), InternVL (Chen et al., 2024, 2025), MiniCPM (Hu et al., 2024), Ovis (Lu et al., 2024b), LLaMA3 (Dubey et al., 2024) and Yi-VL (Young et al., 2024) implement innovative projection techniques to combine visual and textual features effectively, enabling many multimodal applications. These models showcase the growing potential of MLLMs in advancing both research and practical applications that rely on multimodal data (Qu et al., 2025; Zou et al., 2024; Zhou et al., 2024; Huang et al., 2024). Therefore, we introduce ESSAYJUDGE, a novel benchmark designed to evaluate MLLMs' capability to score essays with multimodal context, paving the way for AGI systems (Xiao et al., 2024b; Tate et al., 2024; Yan et al., 2024a, 2025b).

3 Dataset

3.1 Data Collection

This section describes the process of constructing our dataset to ensure high-quality data for analysis, as illustrated in Figure 2. Unlike traditional datasets that often rely on publicly available sources or textbook modifications, the data for this study originates from a K-12 Education Organization. This organization provides a repository of essays graded by experienced educators, guaranteeing the credibility and reliability.

From the original dataset, four primary fields were retained: (i) *Image*, which contains the image of the writing Topics; (ii) *Question*, which contains the text of the writing prompt; (iii) *Essay*, representing the student's written work; (iv) *Overall Score*, reflecting the final assessment provided by professional educators. To enhance the dataset's quality, a series of processing and cleaning steps were applied. These steps included removing essays with incomplete or low-quality responses and selecting topics that met the criteria for reliability and diversity. Through this rigorous process, we curated the ESSAYJUDGE dataset consisting of 1,054

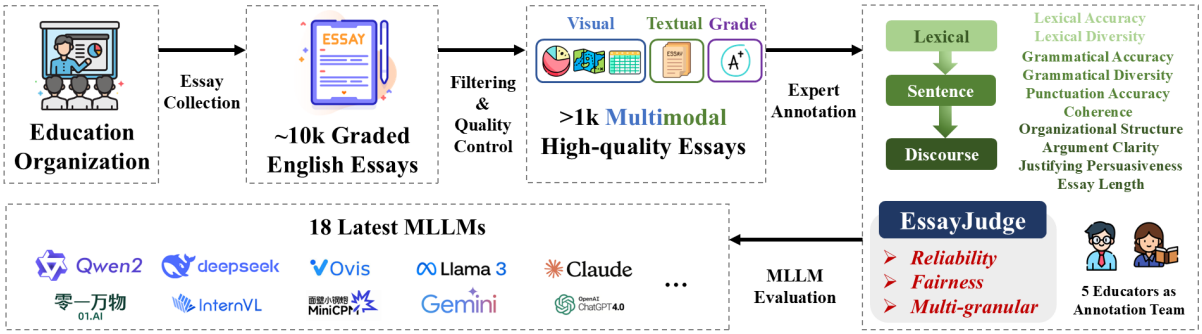


Figure 2: Roadmap illustration of ESSAYJUDGE dataset collection, construction and annotation.

multimodal essays with 125 topics. These essays, covering a broad spectrum of writing abilities, form a solid foundation for facilitating AES research.

3.2 Data Annotation Scheme

Through discussion with English teachers and linguistic experts, we identified ten traits of essays and categorized them into three levels of granularity, progressing from fine-grained lexical features to broader sentence-level structures and, finally, to discourse-level characteristics. This hierarchical structure reflects a natural flow from the smallest units of language to the overall coherence and persuasiveness of an essay, providing a comprehensive framework for evaluation. The rubrics (with a scale of 0 to 5) can be seen in Appendix A. Higher scores indicate a stronger performance.

At the **lexical level**, the focus is on the precision and diversity of word usage, which forms the foundation of effective expression. Traits including ❶ *lexical accuracy* and ❷ *lexical diversity* assess how well the writer uses vocabulary to convey meaning, including the correctness of word choice, spelling, and semantic appropriateness, as well as the variety and richness of vocabulary demonstrated in the essay. These fine-grained features ensure that the language is both accurate and expressive.

Moving to the **sentence level**, the evaluation shifts to the internal quality of sentences and the connections between them. Traits including ❸ *grammatical accuracy* and ❹ *grammatical diversity* examine the correctness and variety of grammatical structures, reflecting the writer's ability to construct well-formed and diverse sentences. ❺ *Punctuation accuracy* ensures that punctuation marks are used appropriately to enhance clarity and readability. Additionally, ❻ *coherence* is assessed at this level, focusing on how smoothly sentences connect through effective transitions, logical relationships, and appropriate use of conjunctions. This intermediate granularity highlights the writer's ability to build logical and linguistically sound sentence structures that support the flow of ideas.

At the **discourse level**, the evaluation considers the overall structure, argumentation, and coherence of the essay as a whole. Traits including ❼ *organizational structure* assess how well the essay is organized across its introduction, body, and conclusion, ensuring ideas are logically and clearly presented. ❽ *Argument clarity* evaluates the explicitness and focus of the central argument, while ❾ *justifying persuasiveness* measures the strength of evidence and reasoning provided to support the argument. Finally, ❿ *essay length* ensures that the essay meets the required length while maintaining depth and focus. This broader granularity captures the writer's ability to integrate all elements into a cohesive and persuasive whole.

3.3 Datasets Annotation Procedure

To ensure a thorough and objective evaluation of the traits, we enlisted two experienced experts in English education, who independently assessed all 10 traits for each essay. After scoring, we compared the results and calculated the differences between the two sets of scores. For traits where the score difference was less than or equal to 1, we took the average of the two scores to establish the ground-truth score. In cases where the score difference exceeded 1, we asked another independent team consisting of three senior annotators to review the essays and discuss the traits with the team. They finally reached a consensus on the final ground-truth score, ensuring a fair and reliable outcome.

3.4 Data Details

ESSAYJUDGE dataset comprises a substantial collection of 1,054 multimodal essays designed for AES (See details in Appendix B). The dataset is categorized based on the *number of images per question*, with 66.7% being single-image questions and the remaining 33.3% multi-image questions.

4 Experiment and Analysis

4.1 Experimental Setup

Evaluation Groups. We meticulously categorized diverse MLLMs into distinct groups to assess their capabilities in trait-specific AES. (i) The **Open-Source MLLMs** category encompassed models such as Yi-VL (Young et al., 2024), Qwen2-VL (Wang et al., 2024), DeepSeek-VL (Lu et al., 2024a), LLaVA-NEXT (Liu et al., 2024), InternVL2 (Chen et al., 2024), InternVL2.5 (Chen et al., 2025), MiniCPM-V2.6 (Hu et al., 2024), MiniCPM-LLaMA3-V2.5 (Hu et al., 2024), Ovis1.6-Gemma2 (Lu et al., 2024b), and LLaMA-3.2-Vision (Dubey et al., 2024), each demonstrating their unique strengths and capabilities in multi-granular essay scoring. (ii) The **Closed-Source MLLMs** featured proprietary models like Qwen-Max (Team, 2024), Step-1V (StepFun, 2024), Gemini-1.5-Pro (DeepMind, 2024b), Gemini-1.5-Flash (DeepMind, 2024a), Claude-3.5-Haiku (Anthropic, 2024a), Claude-3.5-Sonnet (Anthropic, 2024b), GPT-4o-mini (OpenAI, 2024), and GPT-4o (Hurst et al., 2024), providing a comparison point for the performance of models that are not publicly accessible. (iii) Lastly, the **Human Performance** category served as a benchmark for human-level intelligence, enabling us to assess how closely MLLMs emulate human cognitive abilities (More details in Appendix C.3). The detailed prompts for MLLMs and sources of MLLMs are provided in Appendix C.4 and C.5.

Evaluation Metric. We employ Quadratic Weighted Kappa (QWK) (Ke and Ng, 2019; Li and Ng, 2024b,a) as our metric for scoring the similarity, which is widely used to evaluate the agreement between model scores and the ground truth. 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 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

Closed-source MLLMs demonstrate significant superiority over open-source MLLMs in essay scoring tasks, with GPT-4o achieving the strongest overall performance. Table 2 illustrates that closed-source MLLMs consistently outperform open-source MLLMs across ten traits. This advantage is likely due to the high-quality proprietary datasets and advanced training techniques leveraged by closed-source models (Yu et al., 2024; Wang et al., 2023), enabling them to achieve more balanced and robust performance. GPT-4o stands out as the best-performing model, achieving the highest QWK scores across multiple traits except for Argument Clarity, highlighting its exceptional capability. Among open-source MLLMs, InternVL2 emerges as the best performer. However, its overall performance remains behind closed-source ones, underscoring the gap between open-source and closed-source models in terms of capturing complex and evaluative subtleties.

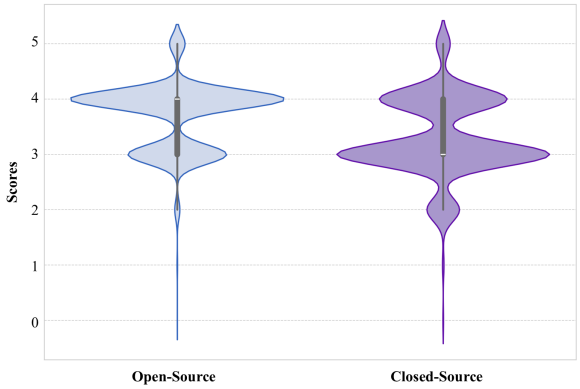


Figure 3: The open-source and closed-source MLLMs' distribution of average scores among ten traits.

Closed-source MLLMs exhibit distinct scoring patterns compared to open-source counterparts, characterized by greater score variability and stricter adherence to rubrics. As revealed in Figure 3, closed-source models demonstrate significantly higher score variance (0.49 vs. 0.34), suggesting enhanced capacity to differentiate essay quality through broader score distribution across performance levels. This contrasts with open-source models' tendency to cluster scores in the mid-range (3-4 points), reflecting limited discriminative capacity for quality extremes. As shown in Figure 5, the scoring rigor of closed-source systems is further evidenced by their consistent assignment of lower scores across critical traits including Argument Clarity, Coherence, and Linguistic Features, with human ratings typically intermediate between the two model types. This systematic conservatism stems from closed-source models' strict adherence to quantitative rubtics (Kundu and Barbosa, 2024),

MLLMs	#Para.	Lexical Level		Sentence Level				Discourse Level			
		LA	LD	CH	GA	GD	PA	AC	JP	OS	EL
Open-Source MLLMs											
Yi-VL (Young et al., 2024)	6B	0.07	0.05	0.08	0.09	0.05	0.09	0.05	0.13	0.05	0.07
Qwen2-VL (Wang et al., 2024)	7B	0.20	0.26	0.13	0.21	0.16	0.12	0.17	0.10	0.14	0.15
DeepSeek-VL (Lu et al., 2024a)	7B	0.09	0.12	0.12	0.13	0.35	0.06	0.18	0.21	0.08	0.09
LLaVA-NEXT (Liu et al., 2024)	8B	0.02	0.04	0.03	0.11	0.10	0.02	0.12	0.15	0.02	0.10
InternVL2 (Chen et al., 2024)	8B	0.28	0.27	0.34	0.36	0.31	0.33	0.25	0.29	0.31	0.29
InternVL2.5 (Chen et al., 2025)	8B	0.14	0.29	0.29	0.29	0.31	0.26	0.15	0.21	0.25	0.22
MiniCPM-V2.6 (Hu et al., 2024)	8B	0.18	0.07	0.08	0.16	0.09	0.04	0.12	0.35	0.06	0.24
MiniCPM-LLaMA3-V2.5 (Hu et al., 2024)	8B	0.37	0.27	0.36	0.29	0.34	0.29	0.09	0.18	0.21	0.09
Ovis1.6-Gemma2 (Lu et al., 2024b)	9B	0.15	0.11	0.13	0.39	0.27	0.36	0.11	0.13	0.14	0.21
LLaMA-3.2-Vision (Dubey et al., 2024)	11B	0.20	0.16	0.17	0.14	0.11	0.12	0.09	0.17	0.17	0.16
Closed-Source MLLMs											
Qwen-Max (Team, 2024)	-	0.57	0.51	0.52	0.56	0.48	0.40	0.34	0.54	0.45	0.41
Step-1V (StepFun, 2024)	-	0.52	0.40	0.49	0.50	0.46	0.37	0.26	0.39	0.31	0.25
Gemini-1.5-Pro (DeepMind, 2024b)	-	0.52	0.46	0.57	0.56	0.51	0.35	0.29	0.46	0.54	0.28
Gemini-1.5-Flash (DeepMind, 2024a)	-	0.46	0.40	0.48	0.53	0.41	0.33	0.33	0.42	0.47	0.28
Claude-3.5-Haiku (Anthropic, 2024a)	-	0.59	0.54	0.53	0.50	0.57	0.40	0.35	0.39	0.48	0.33
Claude-3.5-Sonnet (Anthropic, 2024b)	-	0.66	0.60	0.58	0.66	0.60	0.57	0.33	0.46	0.39	0.35
GPT-4o-mini (OpenAI, 2024)	-	0.64	0.56	0.54	0.58	0.54	0.45	0.33	0.57	0.45	0.46
GPT-4o (Hurst et al., 2024)	-	0.89	0.89	0.87	0.85	0.61	0.65	0.30	0.80	0.79	0.70
Human Performance											
Human performance	-	0.91	0.91	0.89	0.93	0.56	0.86	0.72	0.86	0.88	0.77

Table 2: Comparison of open-source and closed-source MLLM performance (QWK). The highest and second highest scores among MLLMs in each column are highlighted in red and blue, respectively.

prioritizing error penalization (grammatical inaccuracies, logical inconsistencies) and standardization over nuanced language appreciation. While ensuring transparency and reproducibility, this approach may undervalue creative language use, contrasting with human raters' contextual flexibility and open-source models' reduced capacity to assess linguistic complexity effectively.

Multimodal inputs play a critical role in improving the performance of LLMs in AES tasks. Our ablation study compared the performance of models in two settings, including multimodal inputs (text and images) versus text-only inputs. As shown in Figure 4, when image information was removed, GPT-4o experienced a decrease in QWK scores across all ten traits. These findings underscore that visual features provide essential evaluative dimensions that are inaccessible to text-only approaches, especially when images contain critical information supporting the argument. Additional evaluation results for other models are provided in the Appendix C.2.

4.3 Trait-Specific Analysis

Closed-source MLLMs perform poorly in evaluating Argument Clarity and Essay Length while excelling at lexical-level assessment. As shown in

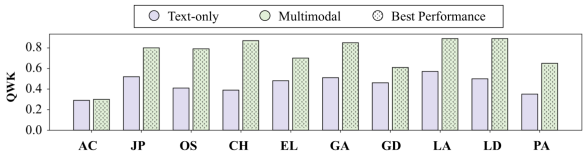


Figure 4: GPT-4o's QWK values across traits for text-only and multimodal settings.

Figure 6, their low QWK values for Argument Clarity and Essay Length are because closed-source LLMs rely heavily on surface-level features like grammar and vocabulary, making them ineffective in handling complex arguments that require contextual understanding and reasoning. For Essay Length, they often hallucinate word counts (Rawte et al., 2023) and over-rely on quantitative measures, leading to an overvaluation of verbosity and an undervaluation of concise but effective writing (Jeon and Strube, 2021). In contrast, closed-source MLLMs' strong lexical performance is due to exposure to a extensive, high-quality datasets (Shi et al., 2023), enabling superior precision and variety in lexical choice.

MLLMs demonstrate outstanding performance in assessing Coherence when grading essays related to line charts. For example, as shown in Figure 7, GPT-4o achieves a high QWK value of 0.89 for the coherence trait when evaluat-

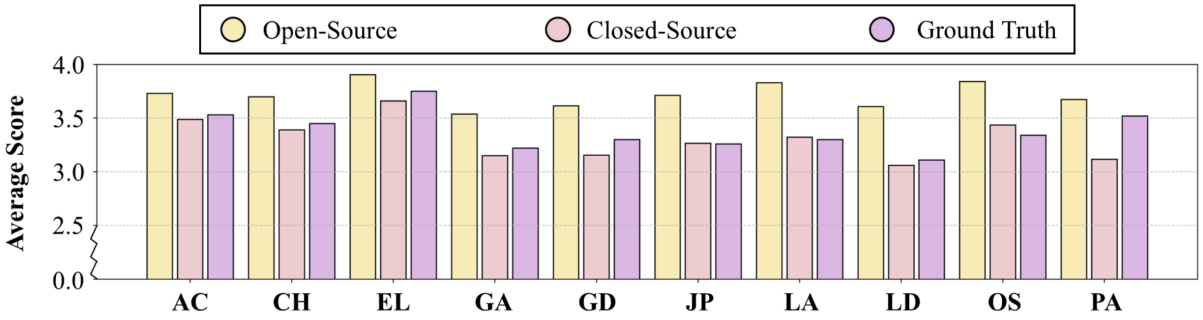


Figure 5: The average scores of open-source MLLMs, closed-source MLLMs and ground-truth across ten traits.

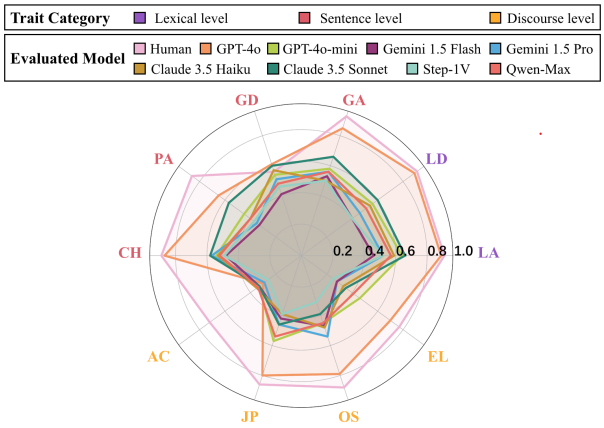


Figure 6: The average QWK values of representative models across three granularities of ten traits.

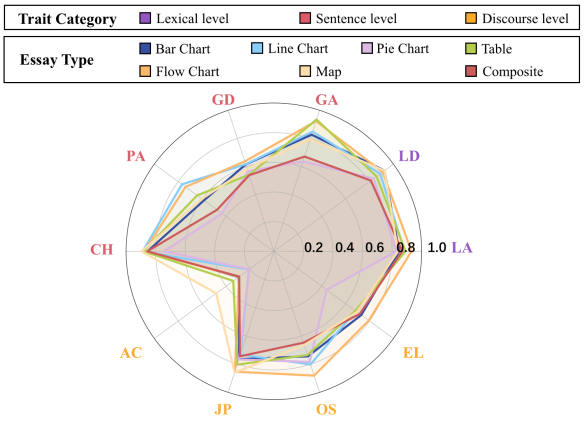


Figure 7: GPT-4o's mean QWK values across different essay types across three granularities of ten traits.

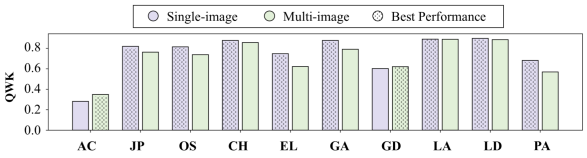


Figure 8: Comparison of GPT-4o's QWK values across ten traits between single-image and multi-image settings.

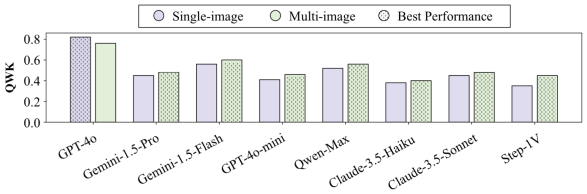


Figure 9: The **closed-source** MLLMs' QWK values of JP trait among single-image & multi-image settings.

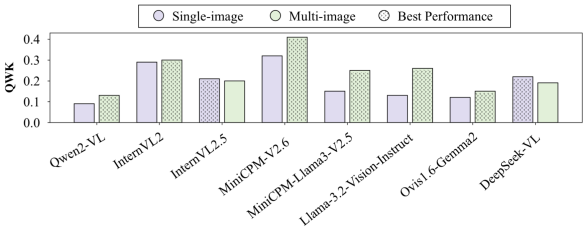


Figure 10: The **open-source** MLLMs' QWK values of JP trait among single-image & multi-image settings.

ing line chart essays. The results of the top three open-source and top three closed-source MLLMs are provided in the appendix D. This is primarily because line charts have a highly linear information structure that emphasizes trends and changes, making the logical relationships between data points clear and easy to follow (Islam et al., 2024). Additionally, the core information in line chart descriptions typically focuses on key points such as turning points, peaks, and troughs, which simplifies the logical chain and allows models to effectively capture and evaluate the coherence of the text.

4.4 Analysis of #image Setting

Most closed-source MLLMs perform better in evaluating essays with single-image setting. As shown in Figure 8, GPT-4o performed better on single-image tasks except for JP and GD traits. Appendix E shows that among all evaluated closed-source MLLMs, only three models do not follow this pattern. Single-image tasks are simpler and more focused, typically requiring students to describe one logical theme. This clear structure makes it easier for models to capture key information and provide accurate evaluations, without

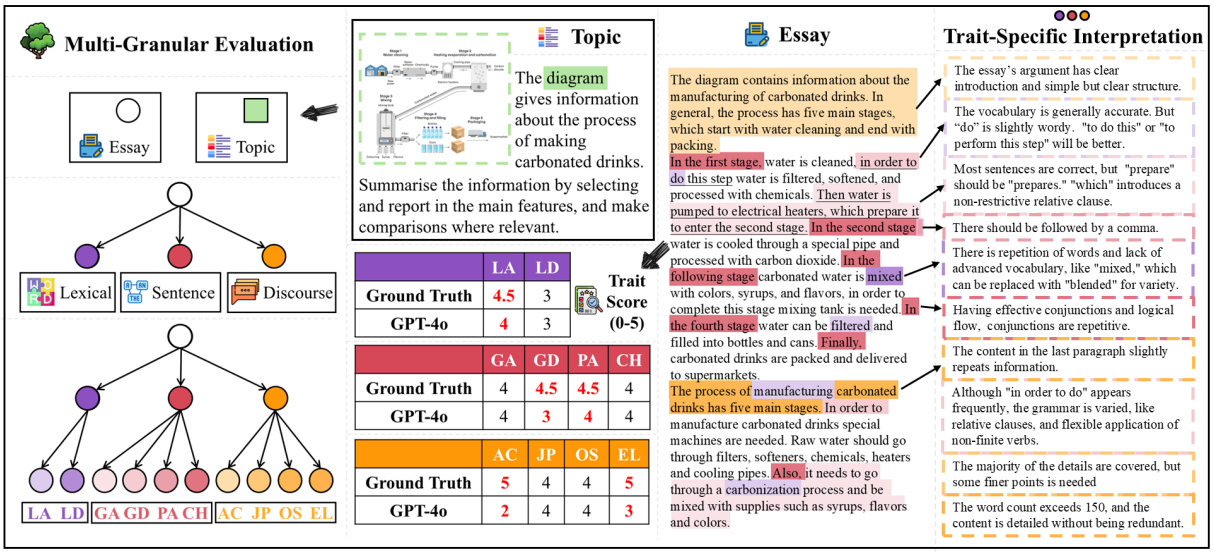


Figure 11: The example case of multi-granular evaluation for the essay.

the need for complex comparisons or logical integration across multiple images. In contrast, multi-image tasks involve comparing, relating, and integrating data from multiple sources, which increases task complexity and the likelihood of errors in both student responses and model evaluations.

For Justifying Persuasiveness, most MLLMs perform better when evaluating essays with multi-image setting. We selected the top-performing eight models from both open-source and closed-source MLLMs, as shown in Figure 9 and Figure 10. Unlike most traits, where MLLMs tend to perform better on single-image tasks, the evaluation of Justifying Persuasiveness shows a distinct advantage in multi-image settings. This may be because multi-image tasks inherently provide richer and more diverse data points, enabling students to construct more evidence-based arguments.

4.5 Case Study

Figure 11 shows the detailed multi-granular evaluation for one of the essays. Additional examples are shown in Appendix F. Specifically, we can find that argument clarity is the most discrepant from the ground truth. *Argument clarity* is crucial in AES, as it directly reflects whether the author's central ideas are in alignment with the essay requirement (Falk and Lapesa, 2023). Leading models like GPT-4o show relatively poor performance in assessing argument clarity, which is illustrated in Figure 2. However, argument clarity serves as a key indicator of a model's reasoning abilities and its capacity to integrate and process complex information, including visual elements. The ability to clearly present and logically connect ideas is essential for both multimodal understanding and reasoning, making it a critical benchmark for evaluating the AES performance of MLLMs. Addressing these challenges in future MLLMs could significantly improve their ability to assess essays with complex reasoning and enhance their multimodal integration capabilities.

5 Conclusion

In this work, we presented ESSAYJUDGE, the first multimodal benchmark designed to evaluate the AES capabilities of MLLMs across lexical, sentence, and discourse-level traits. Addressing longstanding limitations in traditional AES approaches, ESSAYJUDGE leverages MLLMs' inherent strengths in contextual understanding and multimodal analysis, enabling more precise, trait-specific evaluations without reliance on handcrafted features. Our comprehensive evaluation of 18 representative MLLMs highlights current limitations of MLLM-based AES systems. Notably, closed-source MLLMs such as GPT-4 demonstrate superior performance compared to open-source counterparts, yet a significant gap remains in achieving human-level accuracy.

We envision that ESSAYJUDGE will not only drive innovation in AES but also serve as a stepping stone toward broader applications of MLLMs in educational assessment and beyond. The research community can address the challenges identified and foster the development of more accurate and interpretable AES systems towards AGI.

Limitations

Despite the findings we demonstrate in our work, there still exist minor limitations:

1. The datasets used in this study primarily consist of essays written by non-native speakers of English, making it unclear whether our conclusions apply to essays written by native speakers, such as those in the ASAP dataset. However, since our rubrics are broadly applicable and not designed specifically for non-native speakers, we believe the conclusions can be generalized to essays written by native speakers as well.

2. Although our study covers diverse topics, including healthcare, biology, demographics, environment, education and so on, there is still a demand for a more generalized benchmark. Further expansion is needed to address a wider variety of writing contexts and disciplines, ensuring its broader applicability across different writing tasks.

Acknowledgements

This work was supported by Open Project Program of Guangxi Key Laboratory of Digital Infrastructure (Grant Number: GXDIOP2024015); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

Anthropic. 2024a. Claude 3.5 haiku.

Anthropic. 2024b. Claude 3.5 sonnet.

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 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.

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, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. 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.

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

Google DeepMind. 2024a. Gemini 1.5 flash.

Google DeepMind. 2024b. Gemini 1.5 pro.

Linqian Ding and Di Zou. 2024. Automated writing evaluation systems: A systematic review of grammarly, pigai, and criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, pages 1–53.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Rujun Gao, Hillary E Merzdorf, Saira Anwar, M Cynthia Hipwell, and Arun Srinivasa. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, page 100206.

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, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 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*.

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

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.

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.

Sungho Jeon and Michael Strube. 2021. Countering the influence of essay length in neural essay scoring. In *Proceedings of the second workshop on simple and efficient natural language processing*, pages 32–38.

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? *Preprint*, arXiv:2409.13120.

Gyeong-Geon Lee, Lehong Shi, Ehsan Latif, Yizhu Gao, Arne Bewersdorff, Matthew Nyaaba, Shuchen Guo, Zihao Wu, Zhengliang Liu, Hui Wang, et al. 2023. Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*.

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

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.

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. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024b. Ovis: Structural embedding alignment for multimodal large language model. *Preprint*, arXiv:2405.20797.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Subhankar Maity and Aniket Deroy. 2024. The future of learning in the age of generative ai: Automated question generation and assessment with large language models. *arXiv preprint arXiv:2410.09576*.

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.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

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.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Stefan Ruseti, Ionut Paraschiv, Mihai Dascalu, and Danielle S McNamara. 2024. Automated pipeline for multi-lingual automated essay scoring with readerbench. *International Journal of Artificial Intelligence in Education*, pages 1–22.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

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.

StepFun. 2024. Introducing step1v.

Tamara P Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can ai provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

Qwen Team. 2024. Introducing qwen1.5.

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 nonnative learner essays: Investigating the role of linguistic features. *CoRR*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482.

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.

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, et al. 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. 2024a. Human-ai collaborative essay scoring: A dual-process framework with llms. *Preprint*, arXiv:2401.06431.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024b. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

Tianlong Xu, Richard Tong, Jing Liang, Xing Fan, Haoyang Li, and Qingsong Wen. 2024. Foundation models for education: Promises and prospects. *arXiv preprint arXiv:2405.10959*.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024a. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

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. 2024b. 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, et al. 2024c. 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. 2024d. 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, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

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, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606.

A Trait-Specific Rubrics

In this section, we introduce the rubrics used to annotate the 10 traits for each essay in our dataset. The rubrics are detailed in Table 5 to Table 14. 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 ESSAYJUDGE Dataset Details

B.1 Dataset Scope

In the ESSAYJUDGE benchmark, the essay requirements are all set at university-level difficulty. The essays *highly depend on visual information*. Therefore, when writing the essays, students need to use the information provided by the pictures as evidence to support their arguments. This creates a unique multimodal challenge and provides a basis for evaluating the ability of MLLMs to handle diverse and complex information in the multimodal context of AES.

B.2 Dataset Categorization

The Figure 3 above provides a detailed breakdown of the ESSAYJUDGE dataset, which consists of 1054 multimodal essays. The dataset is categorized based on the type of image it includes, with 66.7% (703 essays) containing a single image and 33.3% (351 essays) containing multiple images. Further classification is made based on the type of visual content within these essays, with the most common type being flow charts (28.9%), followed by bar charts (20.0%), and tables (14.5%). Other image types include line charts (13.8%), pie charts (6.7%), maps (5.9%), and composite charts (10.2%). This dataset provides valuable insights into the diversity of multimodal elements incorporated in essay content.

B.3 Dataset Topic

Our dataset includes 125 distinct essay topics, which span a wide range of themes such as population, environment, education, production, evolution, and so on. The topics represent a diverse array of subjects, offering a broad scope for analysis. In the Table 15, we highlight the top five most frequent topics in the dataset, providing an overview of the predominant themes present in the essays.

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 3: Key statistics of ESSAYJUDGE dataset.

B.4 Annotation Details

During the annotation process, we found that when two experts independently scored for the first time, the proportion of the score difference less than or equal to 1 was 94.8%. This data fully indicates that the scoring consistency of the two experts is very high, reflecting that the scoring of the two experts is relatively accurate. Table 4 shows the proportions of the score difference less than or equal to 1 based on specific traits.

Traits	#Essays	Proportion
AC	900	85.4%
JP	1,035	98.2%
OS	1,016	96.4%
CH	1,038	98.5%
EL	920	87.3%
GA	1,025	97.2%
GD	1,044	99.1%
LA	1,034	98.1%
LD	1,044	99.1%
PA	936	88.8%
Total	9,992	94.8%

Table 4: The proportions of the score difference less than or equal to 1 based on specific traits.

C Additional Experimental Details

C.1 Analysis on Scaling Law

We conducted additional experiments using a larger open-source MLLM—Qwen-VL (72B). Compared to Qwen-VL (7B), this model shows a significant performance improvement, consistent with the expected scaling law (as shown in the Table 16). However, even with increased model size, these models still underperform compared to most closed-source MLLMs, which is in line with our conclusions.

C.2 Multimodal Ablation Study

We also evaluated GPT-4o-mini. Figure 12 shows that GPT-4o-mini exhibited a decline in nine out of ten traits when image information was removed, with the exception of lexical diversity. Except for GPT-4o and GPT-4o-mini, we also evaluated text-only LLM - GPT-3.5, which also showed noticeable performance drops (as shown in Table 17 below). This finding further underscores the importance of multimodal inputs, as visual features provide essential evaluative dimensions that text-only approaches cannot capture, especially when images contain critical supporting information.

C.3 Human Performance Evaluation

In the Human Performance section, four postgraduate students with excellent English backgrounds independently evaluated the essays, scoring them across ten distinct traits. To ensure the reliability of their assessments, each student was assigned an approximately equal number of the 1,054 essays, ensuring a balanced workload.

The evaluations were conducted independently, with no discussions permitted between the evaluators to preserve the integrity of the scoring process. Each postgraduate student provided their assessments based solely on their expertise and understanding of the scoring criteria.

This analysis underscores the ability of human assessments to capture the distinct traits of essays, highlighting their role in reflecting the nuances of natural intelligence. These evaluations also reveal the gap between large language models and human cognitive capabilities, serving as a benchmark for the advancements that machine intelligence strives to achieve.

C.4 Prompt for MLLM Evaluation

For the evaluation of MLLMs, we designed prompts that consist of four distinct parts: Task Definition, Rubrics, Reference Content, and Instruction. The details are shown in Figure 13.

The input to the model includes the question text, the accompanying image(s), the student's essay, as well as the specific trait to be evaluated and its corresponding rubrics.

The output should be the only a numerical score, in line with the requirements set forth in the Task Definition. However, given that some models, especially the open-source MLLMs, tend to deviate from the only score task and produce outputs beyond what is expected, we explicitly reinforce the requirement for a numerical score in the final Instruction section of the prompt. This redundancy aims to ensure adherence to the evaluation task and improve the reliability of the scoring process.

C.5 Model Sources

Table 18 details specific sources for the various MLLMs we evaluated. The hyperparameters for the experiments are set to their default values unless specified otherwise.

D More on Trait-Specific Analysis

We present the performance results of the top three open-source and closed-source MLLMs in grading essays related to line charts. Aside from the best performer, GPT-4o, other models evaluated include Claude-3.5-Sonnet, GPT-4o-mini, InternVL2, MiniCPM-LLaMA3-V2.5, and InternVL2.5. As shown in Figure 14 to Figure 18 These models also demonstrated outstanding performance in assessing coherence when grading essays related to line charts.

E More on Analysis of #image

This section presents the evaluation results of all assessed closed-source MLLMs. As shown in Figure 19 to Figure 25, most closed-source MLLMs, except for the Gemini series and Qwen-Max, perform better when grading essays based on single-image tasks compared to multi-image tasks.

F More Multimodal Essay Scoring Examples

This section provides additional examples of multimodal essay scoring based on Multi-Granular rubrics, which is shown in Figure 26 to Figure 28, showcasing the application of our proposed framework to a diverse set of essays that incorporate both textual and visual elements.

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 5: 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 6: Rubrics for evaluating the coherence of the essays.

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 7: 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 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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 14: Rubrics for evaluating the punctuation accuracy of the essays.

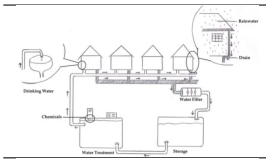
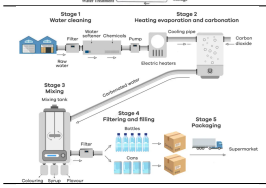
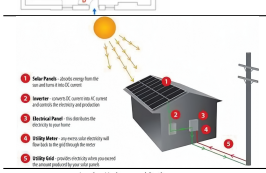
Image	Topic	Frequency
	The diagram below shows how rain water is collected and then treated to be used as drinking water in an Australian town. Summarise the information by selecting and reporting the main features and make comparisons where relevant. You should write at least 150 words.	23
	The diagram gives information about the process of making carbonated drinks. Summarise the information by selecting and report in the main features, and make comparisons where relevant. You should write at least 150 words.	23
	The diagrams below show the existing ground floor plan of a house and a proposed plan for some building work.Summarise the information by selecting and reporting the main features and make comparisons where relevant. You should write at least 150 words.	16
	The diagram below shows how solar panels can be used to provide electricity for domestic use.Write a report for a university, lecturer describing the information shown below. You should write at least 150 words.	16
	The graph shows Underground Station passenger numbers in London.Summarise the information by selecting and reporting the main features, and make comparisons where relevant. You should write at least 150 words.	16

Table 15: The top five most frequent topics in the dataset with images, topics, and frequencies.

Trait	Qwen2-VL (7B)	Qwen2-VL (72B)
Argument Clarity	0.17	0.34
Justifying Persuasiveness	0.10	0.52
Organizational Structure	0.14	0.48
Coherence	0.13	0.55
Essay Length	0.15	0.40
Grammatical Accuracy	0.21	0.51
Grammatical Diversity	0.16	0.55
Lexical Accuracy	0.20	0.52
Lexical Diversity	0.26	0.38
Punctuation Accuracy	0.12	0.42

Table 16: QWK values of Qwen2-VL (7B) and Qwen2-VL (72B).

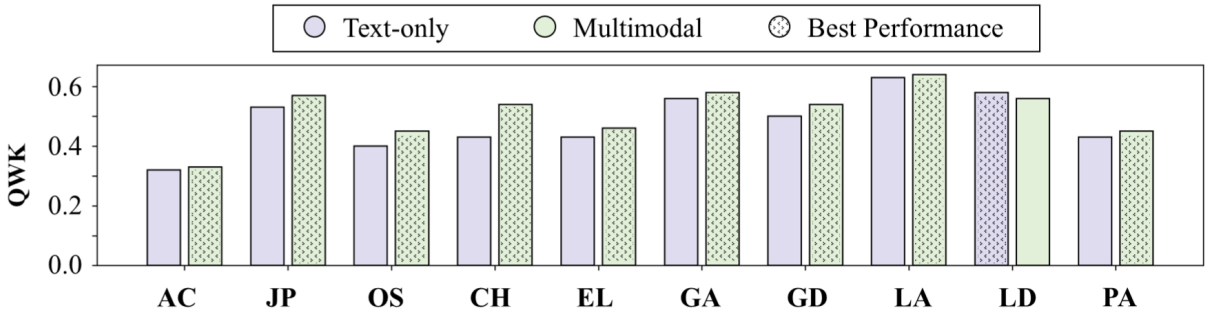


Figure 12: GPT-4o-mini's QWK values across traits for text-only and multimodal settings.

Trait	GPT-4o	GPT-4o-mini	GPT-3.5
Argument Clarity	0.30	0.33	0.23
Justifying Persuasiveness	0.8	0.57	0.24
Organizational Structure	0.79	0.45	0.18
Coherence	0.87	0.54	0.37
Essay Length	0.70	0.46	0.35
Grammatical Accuracy	0.85	0.58	0.42
Grammatical Diversity	0.61	0.54	0.43
Lexical Accuracy	0.89	0.64	0.21
Lexical Diversity	0.89	0.56	0.24
Punctuation Accuracy	0.65	0.45	0.28

Table 17: QWK values of GPT-4o, GPT-4o-mini and GPT-3.5 without image inputs.

Task Definition: Assume you are a professional English Educator. You need to score the {Trait} in the student's essay. Based on the essay topic and image prompt, as well as the student's essay, please assign a score (0-5) according to the criteria in the rubric. The output should be only the score.

Rubrics: {Trait-specific corresponding rubrics}

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

Instruction: Please output only the number of the score (e.g. 5)

Figure 13: Prompt for trait-specific AES task.

MLLMs	Source	URL
Yi-VL-6B	local checkpoint	https://huggingface.co/01-ai/Yi-VL-6B
Qwen2-VL-7B	local checkpoint	https://huggingface.co/Qwen/Qwen2-VL-7B
DeepSeek-VL-7B	local checkpoint	https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat
InternVL2-8B	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2-8B
InternVL2.5-8B	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2_5-8B
MiniCPM-V 2.6-8B	local checkpoint	https://huggingface.co/openbmb/MiniCPM-V-2_6
MiniCPM-Llama3-V 2.5-8B	local checkpoint	https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5
LLaMA-3.2-Vision-Instruct-11B	local checkpoint	https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
Qwen-Max	qwen-vl-max-0809	https://modelscope.cn/studios/qwen/Qwen-VL-Max
Step-1V	step-1v-32k	https://platform.stepfun.com/docs/llm/vision
Gemini 1.5 Pro	gemini-1.5-pro-latest	https://deepmind.google/technologies/gemini/pro/
Gemini 1.5 Flash	gemini-1.5-flash-latest	https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash
Claude 3.5 Haiku	claude-3.5-haiku-20241022	https://www.anthropic.com/claude/haiku
Claude 3.5 Sonnet	claude-3.5-sonnet-20241022	https://www.anthropic.com/claude/sonnet
GPT-4o-mini	gpt-4o-mini-2024-07-18	https://platform.openai.com/docs/models/gpt-4o-mini
GPT-4o	gpt-4o-2024-08-06	https://platform.openai.com/docs/models/gpt-4o

Table 18: Sources of our evaluated MLLMs.

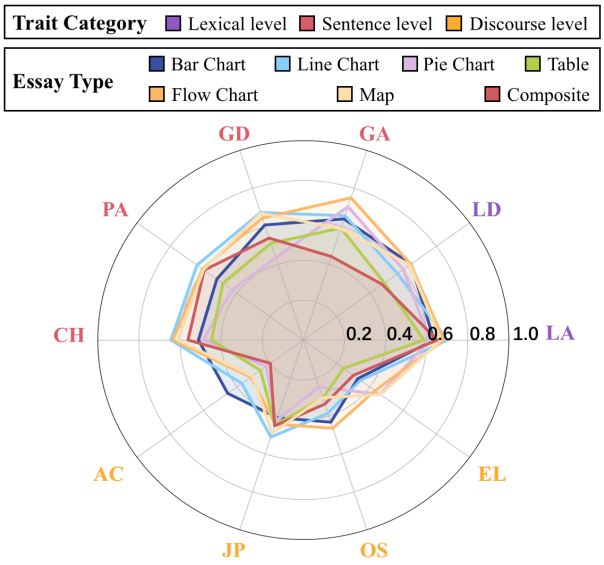


Figure 14: The relation between essay type and Claude-3.5-Sonnet's QWK.

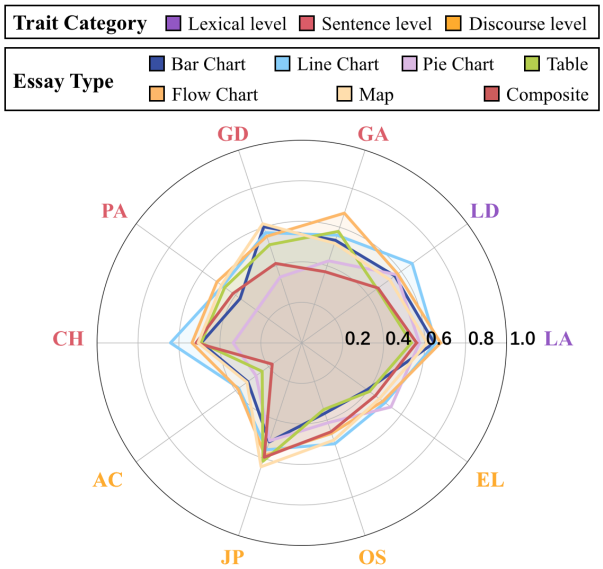


Figure 15: The relation between essay type and GPT-4o-mini's QWK.

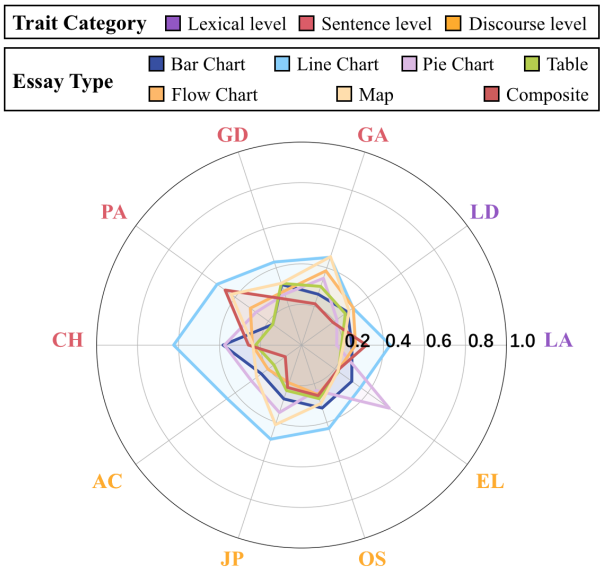


Figure 16: The relation between essay type and InternVL2's QWK.

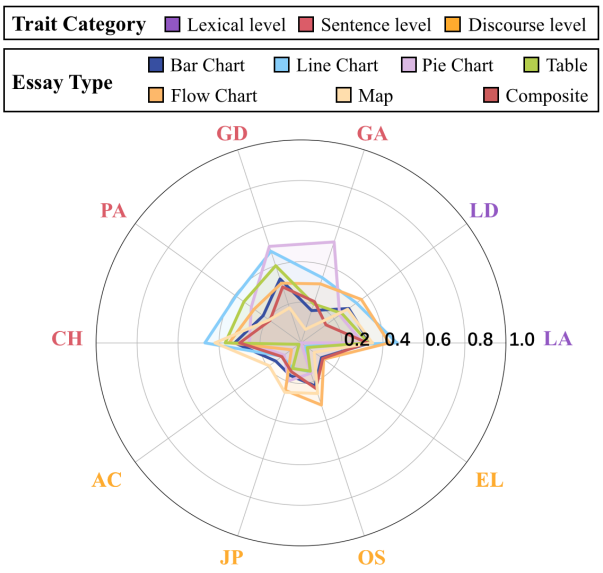


Figure 17: The relation between essay type and MiniCPM-LLaMA3-V2.5's QWK.

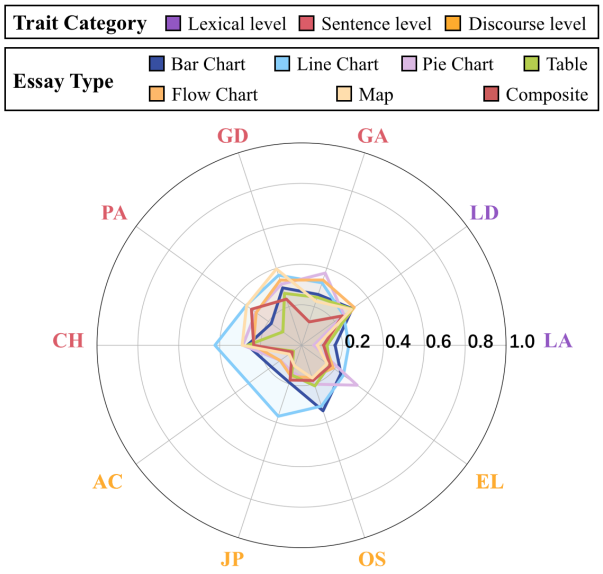


Figure 18: The relation between essay type and InternVL2.5's QWK.

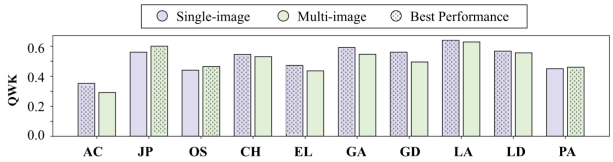


Figure 19: GPT-4o-mini's QWK values across traits for single-image and multi-image settings.

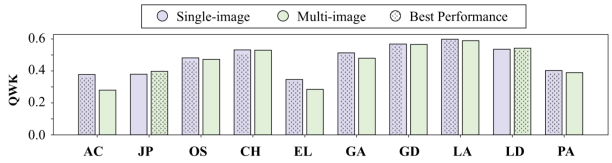


Figure 20: Claude 3.5 Haiku's QWK values across traits for single-image and multi-image settings.

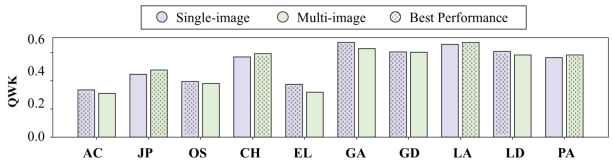


Figure 21: Claude 3.5 Sonnet's QWK values across traits for single-image and multi-image settings.

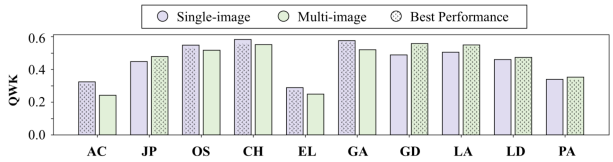


Figure 22: Gemini 1.5 Pro's QWK values across traits for single-image and multi-image settings.

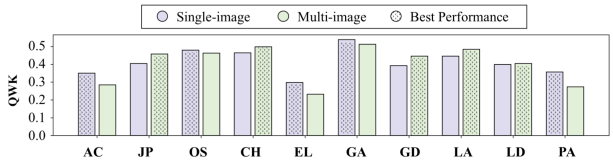


Figure 23: Gemini 1.5 Flash's QWK values across traits for single-image and multi-image settings.

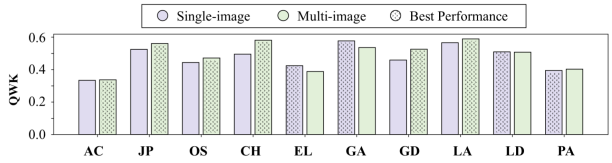


Figure 24: Qwen-Max's QWK values across traits for single-image and multi-image settings.

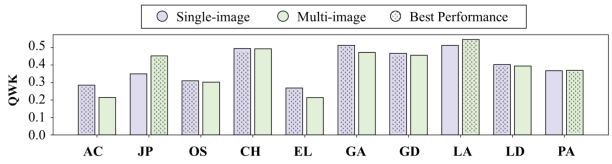


Figure 25: Step-1V's QWK values across traits for single-image and multi-image settings.

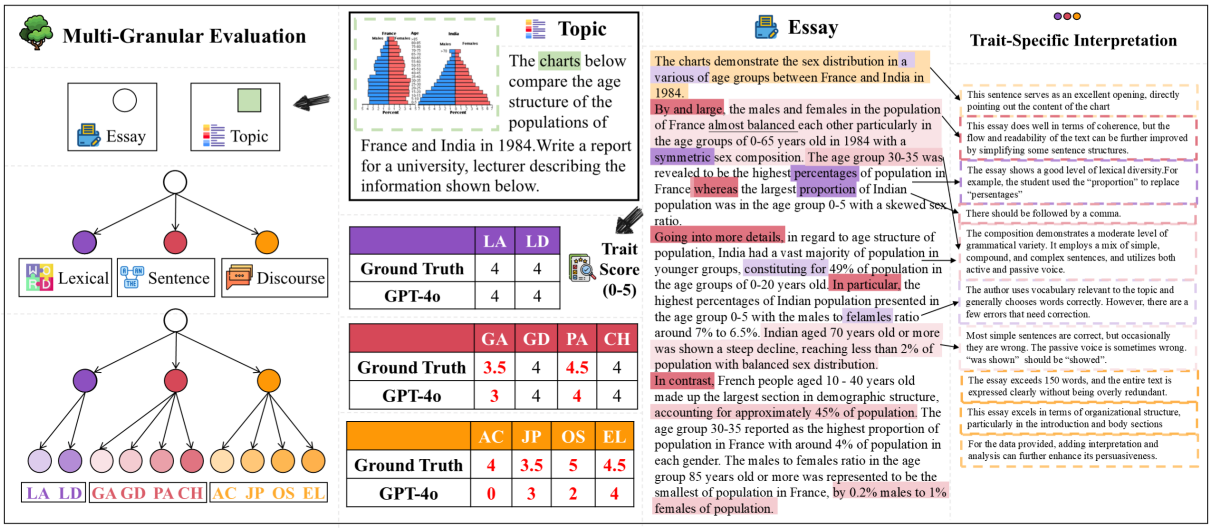


Figure 26: The example one of multi-granular evaluation for the essay.

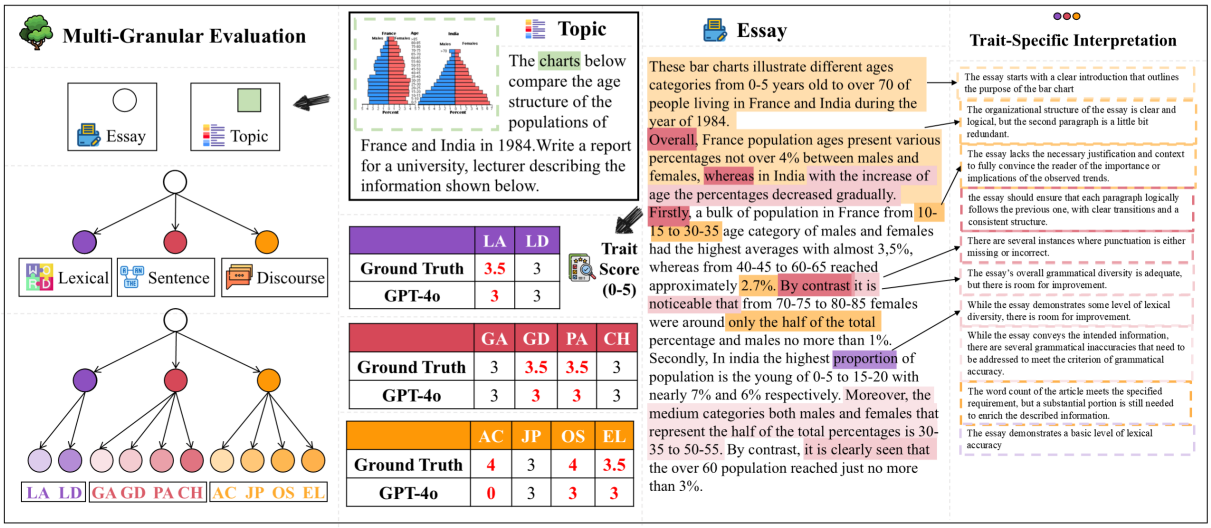


Figure 27: The example two of multi-granular evaluation for the essay.

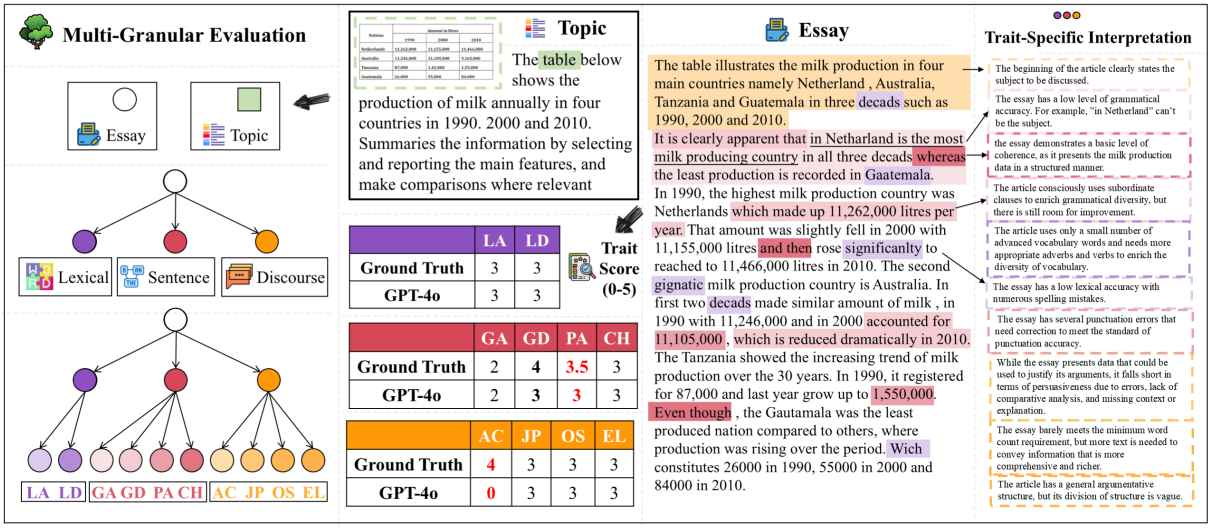


Figure 28: The example three of multi-granular evaluation for the essay.