ADBIAS: Adjusting Multi-LLM Biases via Trait-based Quantitative Analysis for Automated Essay Scoring

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have significantly enhanced the performance of automated essay scoring (AES). However, relying on a single LLM often results in inconsistent evaluations due to its inherent biases. To overcome this challenge, We introduce ADBIAS, a novel multi-agent AES framework designed to systematically identify and mitigate model-specific biases across multiple LLMs—namely, GPT-4o, Claude 3.5 Sonnet, LLaMA 4 Maverick, and Gemini 2.5 Flash. ADBIAS follows a three-stage process: (1) generating trait-level scores and rationales from each LLM, (2) quantifying scoring tendencies using the Many-Facet Rasch Model (MFRM), and (3) producing final scores via a bias-aware Meta-LLM that integrates metadata including bias information. Empirical results on the ASAP and ASAP++ datasets show that ADBIAS improves scoring accuracy (+6.4% QWK) and substantially reduces bias variance (–57.9%) compared to both single-model and ensemble baseline models. By incorporating explicit bias modeling and calibrated aggregation, ADBIAS advances the reliability, fairness, and interpretability of LLM-based essay evaluation.

1 Introduction

The integration of automated essay scoring (AES) into educational contexts has been increasingly recognized, especially for its potential to alleviate grading workloads and provide timely and effective feedback.

Early AES systems were designed to predict a single holistic score. This approach gained popularity due to its ability to reduce human raters' workload and simplify the evaluation of overall writing performance (Page, 1966; Page et al., 1997). However, relying on a single score to represent the multiple dimensions of writing ability has raised concerns about its inadequacy in capturing the multifaceted nature of writing.

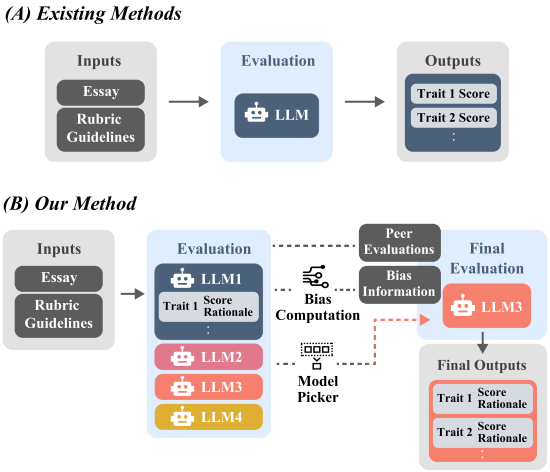

Figure 1: Unlike existing methods (A) that rely on a single LLM, our ADBIAS framework (B) first generates independent evaluations from multiple LLMs. It then quantitatively computes the scoring bias of each model and integrates this bias information into a final Meta-LLM to produce a fairer and more consistent evaluation.

To overcome these challenges, research has shifted toward multi-trait scoring, aiming to capture the diverse dimensions of writing more accurately. Building on this, recent approaches have been developed to enhance the explainability and reliability of scoring systems by leveraging large language models (LLMs) to generate rationales—either before scoring, as supporting evidence for generating score predictions, or after scoring, to explain the reasoning behind their scoring decisions (Chu et al., 2025; Do et al., 2025b).

Despite these advancements, the reliability of current AES systems remains questionable, as the underlying LLMs may carry inherent biases.

One of the key challenges in applying LLMs to AES lies in the variability of scoring outcomes across models. Even when evaluating the same essay, different LLMs often produce divergent scores due to inherent differences in their evaluative tendencies. For example, GPT-3.5 tends to regress

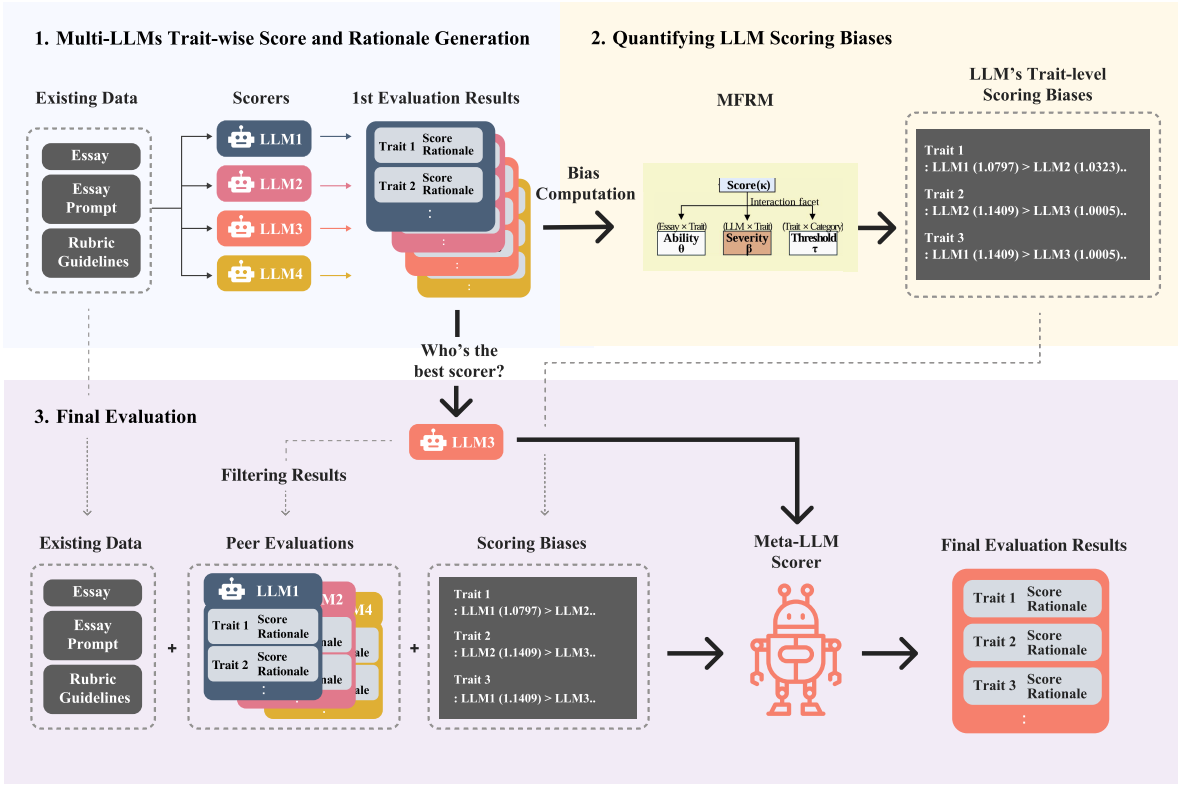

Figure 2: The overall pipeline of our proposed ADBIAS framework, consisting of three main stages: (1) Multi-LLMs Trait-wise Score and Rationale Generation, (2) Quantifying LLM Scoring Biases using MFRM, and (3) Final Evaluation with a Meta-LLM.

toward mid-range scores, GPT-4 yields a broader score distribution, and Claude 2 exhibits a more conservative scoring pattern, frequently assigning lower scores (Tang et al., 2024). These model-specific biases raise concerns about the reliability, consistency, and fairness of AES systems. A recent study empirically confirmed model-specific scoring biases by analyzing inter-model variance and F1 score differences on identical prompts (Ito and Ma, 2025).

To address these issues, the use of multi-LLM collaboration in AES systems has recently emerged as a promising alternative.

A notable example is the GET framework, which quantifies the bias of each LLM, labels its evaluative tendency in textual form, and then has each model independently assess the essay. Based on both labeled tendencies and individual evaluations, the framework adopts a Tree-of-Thought (ToT) structure to derive a final judgment through a structured debate process (Ito and Ma, 2025).

While this approach provides a structured mechanism for incorporating model-specific tendencies, it leaves open the question of how these biases are explicitly captured and systematically leveraged to improve scoring reliability. In GET, such bias signals primarily serve as qualitative references rather than being formally structured or calibrated, making it difficult to trace or adjust for bias in a consistent and measurable way. As a result, the extent to which bias influences the final evaluation remains unclear, highlighting the need for a more explicit and accountable framework for bias-aware scoring.

To address these challenges, we propose **ADBIAS** (Adjusting Multi-LLM Biases via Trait-Based Quantitative Analysis for Automated Essay Scoring), a framework that explicitly quantifies model-specific scoring biases and incorporates them into the evaluation process. This approach aims to mitigate LLM-inherent bias and **lay the groundwork for a more reliable and consistent AES system**.

Fig. 2 illustrates the proposed framework, which is structured around three core stages of operation.:

(1) We employ four state-of-the-art LLMs (GPT-4o, Claude 3.5 Sonnet, Llama 4 Maverick, and Gemini 2.5 Flash) to independently generate

trait-specific scores and evaluation rationales for each essay.

(2) We apply the Many-Facet Rasch Measurement (MFRM) model to systematically quantify model-specific scoring biases based on the trait-specific scores produced by each LLM.

(3) We provide the outputs of multiple LLMs, including trait-level scores, evaluation rationales, and quantified bias information. These outputs are passed to a final evaluator model, the Meta-LLM (Llama 4 Maverick), which references the peer evaluations and their associated bias tendencies to generate the final score and rationale for each essay.

2 Related Work

2.1 Holistic Scoring

Early AES research focused on predicting a single holistic score for essays. However, this approach faced limitations in capturing the multi-dimensional nature of writing and providing trait-specific feedback essential for subjective assessment interpretability (Taghipour and Ng, 2016). To address this, AES research shifted toward trait-based scoring, enabling more fine-grained assessment and improving pedagogical utility.(Hussein et al., 2020).

2.2 Advancements in Trait-Based AES: Cross-Prompt Generalization

As trait-based scoring evolved, the need for cross-prompt generalization—aiming to maintain scoring consistency across unseen prompts—has emerged as a key objective. Recent approaches have leveraged trait-attention mechanisms and multi-task learning architectures to enhance scoring performance while jointly predicting both holistic and trait-level scores (Ridley et al., 2021).

Notably, the ProTACT model adopted a structure that explicitly captures the interaction between prompts and traits, which helps maintain consistent trait-level scoring across different prompts. (Do et al., 2023). Furthermore, Grammar-aware Cross-Prompt Trait Scoring (GAPS), which incorporates grammatical error correction, has been proposed to improve trait-specific evaluation by focusing on general grammatical features regardless of the prompt (Do et al., 2025a).

However, studies have shown that cross-prompt AES can be effectively achieved using simple neural network architectures, indicating that model complexity alone is not essential for improving generalization performance (Li and Ng, 2024). This underscores the importance of efficient feature engineering and model design in achieving robust cross-prompt generalization.

2.3 The Emergence of LLMs and Innovation in AES

Numerous studies have demonstrated the effectiveness of LLM-based AES systems in both scoring accuracy and feedback quality. For example, GPT-based models have achieved QWK scores up to 0.68 and high agreement with human feedback (Liew and Tan, 2024), and show strong performance in both holistic and trait-level scoring (Mansour et al., 2024; Seßler et al., 2024). Prompt engineering techniques—such as few-shot learning and temperature calibration—have been shown to improve alignment with human judgment, particularly in dimensions like Ideas and Organization (Tang et al., 2024).

To address the black-box nature of LLMs, recent work has focused on enhancing scoring transparency. RMTS incorporates LLM-generated rationales into smaller models (Chu et al., 2025), while RaDME combines LLM reasoning with lightweight scoring to improve explainability (Do et al., 2025b). In addition, hybrid approaches integrating linguistic features with LLMs have shown promise in improving both in-domain and cross-domain performance (Hou et al., 2025).

2.4 Multi-LLM Collaboration and Reasoning Ability Enhancement

To address the inherent limitations of single-LLM systems—such as model-specific biases and inconsistency—the paradigm of **multi-LLM collaboration** has been introduced (Feng et al., 2025). An Ensemble Tree-of-Thought (ToT) framework, which employs simulated debate among multiple LLMs, has also been proposed to enhance evaluation accuracy and the quality of generated explanations (Ito and Ma, 2025). In addition, research on LLM-based creativity assessment has shown that multi-LLM collaboration can improve the originality of generated content (Zhao et al., 2025).

3 Dataset

To enable trait-level modeling and systematic bias analysis, we utilize the original ASAP corpus and

its extended version, ASAP++. These datasets provide reliable annotations across diverse prompts and scoring traits. According to their providers, both datasets were collected with consent for academic use and are publicly available for non-commercial research purposes.

3.1 ASAP

The Automated Student Assessment Prize (ASAP) dataset comprises 12,978 English essays from U.S. students (grades 7–10) across eight writing prompts. Each essay received holistic scores from two to three expert raters. Prompts 7 and 8 also include trait-level scores for Organization, Conventions, Style, and Sentence Fluency.

This dataset, publicly available via Kaggle, currently lacks explicitly stated official licenses or terms of use. We exclusively use it for academic research, refraining from commercial use or redistribution.

3.2 ASAP++

ASAP++ extends the original dataset by adding trait-level annotations to prompts 1–6, which previously lacked them. In this study, we use ASAP++ annotations for prompts 1–6 and retain the original ASAP trait scores for prompts 7–8, forming a unified corpus that supports prompt-trait level modeling of both performance and scoring bias.

The ASAP++ dataset is available under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) License. This resource can be accessed online at https://cfilt.iitb.ac.in/~egdata/.

Derivatives from data accessed for research purposes in this study, including our framework, should not be used outside of research contexts.

Pr	Grade	Es	Score Range	Traits
1	8	1783	1–6	Cont, Org, WC, SF, Conv
2	10	1800	1–6	Cont, Org, WC, SF, Conv
3	10	1726	0–3	Cont, PA, Nar, Lang
4	10	1772	0–3	Cont, PA, Nar, Lang
5	8	1805	0–4	Cont, PA, Nar, Lang
6	10	1800	0–4	Cont, PA, Nar, Lang
7	7	1569	0–3	Cont, Org, Style, Conv
8	10	723	1–6	Cont, Org, WC, SF, Conv, Voice

Table 1: Trait abbreviations: Cont = *Content*, Org = *Organization*, WC = *Word Choice*, SF = *Sentence Fluency*, Conv = *Conventions*, PA = *Prompt Adherence*, Nar = *Narrativity*, Lang = *Language*, Style = *Style*, Voice = *Voice*

4 Methodology

We introduce ADBIAS, a three-stage framework designed to improve both the accuracy and fairness of automated essay scoring by explicitly quantifying and correcting for model-specific bias. The system leverages multiple LLMs to independently evaluate essays, applies a Rasch-based model to estimate trait-level bias, and then synthesizes all information through a bias-aware Meta-LLM.

4.1 Overview of ADBIAS

ADBIAS consists of the following three stages:

1. **Independent Scoring by Multiple LLMs:** Four diverse LLMs (GPT-4o, Claude 3.5, Llama 4, and Gemini 2.5) independently evaluate each essay based on a shared prompt and rubric. All models follow a standardized prompt structure, operate under deterministic conditions (temperature = 0), and produce both trait-level scores and rationales.

2. **Bias Quantification via MFRM:** Using trait scores from the LLMs, we apply the Many-Facet Rasch Model (MFRM) to estimate model-specific severity parameters (β). These values capture systematic scoring tendencies and are converted into odds ratios that are embedded into the Meta-LLM's input.

3. **Final Scoring with Meta-LLM:** The Meta-LLM (LLaMA 4) aggregates peer evaluations and associated bias signals to produce final trait scores. Instead of naïvely averaging peer outputs, the model performs structured reasoning that accounts for each peer's scoring tendencies, enabling a more fair and consistent evaluation.

4.2 Model Selection

We intentionally selected four LLMs with diverse architectures and behaviors to capture a wide spectrum of scoring tendencies:

- **GPT-4o (OpenAI):** Strong general-purpose model with top-tier performance
- **Claude 3.5 (Anthropic):** Known for its conservative and bias-sensitive scoring style
- **LLaMA 4 (Meta):** Stable open-source model, chosen as the Meta-LLM

- **Gemini 2.5 (Google):** Lightweight and efficient, used for examining speed–bias tradeoffs

All models were run under identical conditions and prompts to ensure fairness and comparability.

For all models, **temperature = 0** was fixed to ensure deterministic evaluation results.

4.3 Prompt Design

Each model received a two-part prompt:

- A **system message** defining the model's role as an evaluator and providing the full rubric.
- A **user message** containing the prompt, student essay, and instructions to produce trait-wise rationales and scores.

Trait evaluations followed a fixed sequence to reduce interference across traits, and all outputs were returned in a structured dictionary format for easy parsing and analysis.

4.4 Bias Estimation with MFRM

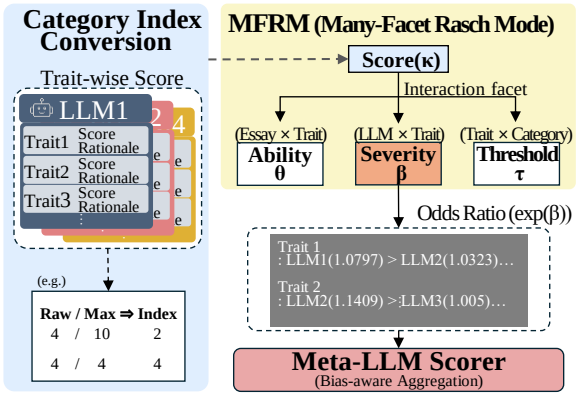

Figure 3: Overview of MFRM-based bias estimation and aggregation.Each LLM generates trait-level scores and rationales. The Many-Facet Rasch Model (MFRM) estimates scorer severity (β), trait thresholds (τ), and essay ability (θ), which inform the Meta-LLM to produce bias-aware final scores.

MFRM was used to estimate three key parameters:

$$L_{n,j,s,t} := \theta_{n,t} - \beta_{j,s,t} - \tau_{s,t} \quad (1)$$

$$\Pr(X_{n,j,t} = k) = \frac{\exp\left(\sum_{s=0}^{k-1} L_{n,j,s,t}\right)}{\sum_{c=0}^{m_t-1} \exp\left(\sum_{s=0}^{c-1} L_{n,j,s,t}\right)} \quad (2)$$

- **Ability (θ):** The student's latent proficiency for each trait
- **Severity (β):** Trait-specific bias of each model across prompts
- **Threshold (τ):** Difficulty of each trait's scoring categories

All parameters were estimated using a PyTorch implementation with warm-up scaling and online connection to the Meta-LLM, allowing real-time bias adjustment.

- θ: `nn.Embedding(`$N_p N_t$`,1)`
- β: `nn.Embedding(`$N_r N_s$`,1)`
- τ: `cum.softplus(Embedding)`
- ID constraints (batch): center θ, β; `fix` τ_1`; scalings` $1.0, 0.5, 0.2 \rightarrow 0.5$

4.5 Severity Fit Verification (Infit · Outfit Analysis)

To verify the trait-specific evaluation fit of the models, Infit and Outfit Mean-Square (MSQ) values were calculated. For a total of 144 model $\times$ set $\times$ trait combinations observed from 8 essay sets $\times$ 4 models, the proportion of MSQ values falling within the recommended range ($0.5 \leq$ MSQ ≤ 1.5) was measured to confirm evaluation fit.

4.6 Meta-LLM Selection and Meta-Evaluation Method

LLaMA 4 was chosen as the Meta-LLM due to its lenient scoring pattern and minimal bias variance, making it a stable anchor for integration. During inference, the Meta-LLM receives all peer scores, rationales, and β metadata, and generates final scores through reasoning that balances strict and lenient perspectives.

This bias-aware aggregation process enables ADBIAS to outperform traditional ensembles by producing more consistent and human-aligned scoring decisions.

4.7 Meta-Prompting via ADBIAS

Meta-LLM (LLaMA4 Maverick) receives:

- Raw rationales from all LLMs
- Trait-specific severity, OR, threshold
- Consistency warnings when rationale contradicts score (Zhao et al., 2025)
- Self-consistency probing (Wang et al., 2023) ensures coherence in meta-LLM output.

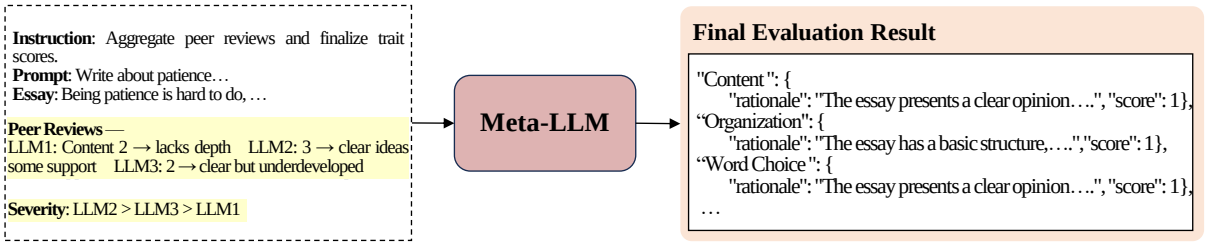

Figure 5: Final evaluation process via Meta-LLM. It generates bias-aware trait scores by aggregating peer reviews(scores and rationales) and incorporating severity(β) metadata, enabling fairer and more consistent evaluations.

5 Experiments & Results

This section provides an empirical assessment of the proposed ADBIAS framework. We assess both scoring accuracy, using Quadratic Weighted Kappa (QWK), and fairness, measured by rater bias variance (β), through statistical analysis and ablation studies conducted on the ASAP and ASAP++ datasets.

5.1 Dataset and Experimental Design

We conducted experiments using the full ASAP and ASAP++ datasets for two main reasons:

First, to enable robust estimation of scoring bias. Since MFRM estimates severity (β) separately for each prompt–trait pair, evaluating across all sets helps isolate rater bias from prompt-specific difficulty, preventing potential confounding that can arise in single-prompt setups.

Second, to examine the generalizability of our framework in realistic educational settings, where essay prompts vary in grade level, genre, rubric design, and scoring range. Including all sets allows us to evaluate the universality of model-specific biases and the flexibility of our system.

Accordingly, each of the four LLMs independently scored the essays, producing trait-specific scores and rationales. We recorded their raw scores, rationale texts, MFRM-estimated β values, and Meta-LLM outputs, and compared them against human scores. QWK and β variance were used as the primary evaluation metrics.

5.2 Results

5.2.1 RQ-A: Do different LLMs produce significantly different trait-specific score distributions?

To assess whether different LLMs show distinct scoring tendencies, we conducted statistical tests across 36 prompt–trait combinations. To analyze the data, a one-way ANOVA assessed mean differences, while Levene's test evaluated variance equality. Multiple comparisons were adjusted using the Bonferroni correction.

All 36 combinations showed statistically significant differences in both means and variances ($p < 0.05$). The average F-statistic across conditions was 137.1 for means and 108.2 for variances.

These results confirm that different LLMs exhibit systematic and model-specific scoring biases—not random variation—when evaluating the same prompt and rubric. This underscores the importance of bias estimation and correction as adopted in the ADBIAS framework.

5.2.2 RQ-B: Is the MFRM-derived severity parameter (β) a reliable indicator of rater bias?

We estimated 144 severity parameters (β) across all LLMs $\times$ prompt $\times$ trait combinations. To verify their reliability, we examined both model fit and predictive power.

For model fit, we used Infit and Outfit Mean Square (MSQ) statistics. 69.4% (100 out of 144) of β estimates fell within the accepted Rasch range (0.5 to 1.5), significantly higher than chance ($p \approx 1.7 \times 10^{-6}$)). Median Infit and Outfit values were 1.12 and 1.02, respectively. The average standard error of β was 0.044 logits (maximum: 0.076), indicating high estimation precision.

To assess predictive power, we computed the Pearson correlation between each model's β value and the deviation of its average trait scores from the overall mean. A strong negative correlation (r = –0.91, $p < 10^{-36}$) confirmed that higher β values correspond to lower average scores—demonstrating that β accurately captures scoring strictness.

Taken together, these results show that the β parameter is both statistically sound and a meaningful signal of model-specific rater bias.

5.2.3 RQ-C: Does injecting β metadata into the Meta-LLM improve accuracy and reduce bias?

We compared two conditions: a β-Aware Meta-LLM that incorporates rater bias metadata, and a β-Blind baseline using LLaMA-4 without such information.

Metric	β-Blind (LLaMA-4)	β-Aware (Meta-LLM)	Δ (Abs / %)	Test Stat.	p-value	95% CI
Accuracy (QWK)	0.4326	0.4600	+0.0275 (+6.4%)	$t(45) = 2.11$	0.041	[+0.0021, +0.0527]
Mean $\lvert\beta\rvert$	0.0775	0.0380	–0.0395 (–50.9%)	–	–	–
Var β	0.00192	0.00081	–0.00111 (–57.9%)	Levene $F = 4.17$	0.045	–

Table 2: Comparison between β-Blind and β-Aware (Meta-LLM) settings across QWK accuracy, mean absolute β, and β variance.

In terms of accuracy, the β-Aware condition achieved an average QWK of 0.4600, significantly higher than the β-Blind baseline (0.4326), with a gain of +6.4% ($t(45) = 2.11, p = 0.041$).

In terms of fairness, the average absolute β decreased from 0.0775 to 0.0380 (–50.9%), and β variance dropped from 0.00192 to 0.00081 (–57.9%), with Levene's test confirming the significance ($p = 0.045$).

At the essay set level, QWK improved in 6 out of 8 sets, and β variance decreased across all sets—most notably in Set 4 (–86%) and Set 6 (–72%).

ES	QWK β-Aware	QWK β-Blind	Δ	Var β-Aware	Var β-Blind	β Var.
1	**0.396**	0.388	**+0.007**	0.00224	0.00175	-28%
2	0.544	**0.550**	-0.005	0.00026	0.00180	**-85%**
3	**0.555**	0.490	**+0.065**	0.00053	0.00064	-18%
4	**0.570**	0.491	**+0.080**	0.00029	0.00201	**-86%**
5	**0.579**	0.514	**+0.065**	0.00022	0.00036	-38%
6	**0.431**	0.386	**+0.045**	0.00031	0.00111	**-72%**
7	0.298	**0.350**	-0.053	0.00101	0.00266	**-62%**
8	**0.462**	0.407	**0.055**	0.00103	0.00084	-23%

Table 3: Comparison of QWK and variance metrics across scoring conditions.

When grouped by prompt difficulty (based on MFRM threshold medians), easier prompts showed larger QWK gains, while harder prompts saw greater reductions in β variance. A Pearson correlation ($r \approx$ `0.41) confirmed that accuracy gains decline as difficulty increases, whereas bias mitigation remains robust.

5.2.4 RQ-D: How does the β-Aware Meta-LLM compare to single LLMs and simple ensembles?

We compared the β-Aware Meta-LLM with four single models (GPT-4o, Claude 3.5, Gemini 2.5, LLaMA-4) and an unweighted ensemble.

Trait	QWK β-Aware	QWK β-Blind	Δ	Var β-Aware	Var β-Blind	β Var.
Cont	**0.516**	0.448	**+0.068**	0.00105	0.00175	**–40%**
Conv	0.272	**0.357**	–0.085	0.00096	0.00119	–19%
Lang	**0.535**	0.467	+0.068	0.00038	0.00353	–89%
Narr	0.592	0.523	+0.069	0.00041	0.00187	–78%
Org	**0.505**	0.423	+0.082	0.00039	0.00158	–75%

Table 4: Trait-wise QWK scores and β variance comparisons between β-aware and β-blind settings.

Method	Avg. QWK	Δ QWK*	Var β	Δ Var β
Meta-LLM (β-Aware)	0.460	—	0.0008	—
Claude 3.5	0.471	+0.011	0.0017	+101%
GPT-4o	0.398	–0.062	0.0041	+395%
Gemini 2.5	0.420	–0.040	0.0033	+292%
LLaMA-4	0.433	–0.027	0.0020	+137%
Majority Voting	0.430	–0.030	—	—

Table 5: Comparison of average QWK, variance in β, and relative changes across LLM-based methods.

In terms of QWK, the Meta-LLM achieved 0.4600, outperforming GPT-4o (0.398), Gemini (0.420), LLaMA (0.433), and the ensemble (0.430). Only Claude slightly exceeded it (0.471), but the difference (+0.011) is marginal and not practically significant.

In terms of bias, the Meta-LLM recorded the lowest β variance (0.0008), reducing rater bias by 44.6% compared to Claude and by over 70% relative to GPT-4o and Gemini.

These findings support the claim that ADBIAS achieves a balanced trade-off between accuracy and fairness that neither single LLMs nor simple ensemble methods can offer. The framework's structural integration of bias signals enables both improved performance and more equitable scoring.

5.3 Case Studies

5.3.1 Prompt 1. Essay 525

Trait	Peer Mean	Meta Score	Human Score	Claude	Gemini	GPT	LLaMA
Cont	4.5	5.0	5.0	4.0	5.0	5.0	4.0
Conv	3.5	4.0	4.0	3.0	4.0	4.0	3.0
Org	4.5	5.0	5.0	4.0	5.0	5.0	4.0
SF	3.5	4.0	4.0	3.0	4.0	4.0	3.0
WC	3.5	4.0	4.0	3.0	4.0	4.0	3.0

Table 6: Trait-level scores for essay #525 across peer models, Meta-LLM, human rating, and individual LLMs.

In this essay, for traits such as content and organization, Claude and LLaMA assigned a conservative score of 4, while GPT and Gemini assigned a 5. The Human rater gave a 5 for this item, and the Meta model also assigned the same score as Human. Particularly for conventions, sentence_fluency, and

word_choice, Claude and LLaMA repeatedly gave a score of 3, showing a somewhat lower evaluation, whereas Gemini and GPT gave a 4. In this situation where Claude and LLaMA showed a consistent downward trend, and GPT and Gemini showed a somewhat lenient trend, Meta compromised these extremes to ultimately derive a score that precisely matched the Human evaluation. This is significant in that the Meta rater makes reliable judgments without simply averaging the biases of multiple LLMs.

5.3.2 Prompt 5. Essay 12389

Trait	Peer Mean	Meta Score	Human Score	Claude	Gemini	GPT	LLaMA
Cont	1.75	1.0	1.0	2.0	1.0	2.0	2.0
Lang	2.00	2.0	2.0	2.0	1.0	3.0	2.0
Nar	1.75	1.0	1.0	2.0	1.0	2.0	2.0
PA	1.75	1.0	1.0	2.0	1.0	2.0	2.0

Table 7: Trait-level scores for essay #12389 from peer models, Meta-LLM, human raters, and individual LLMs.

This essay generally received low scores. For the traits of content, narrativity, and prompt_adherence, Human assigned a score of 1 to all. Claude and LLaMA also mostly gave low scores around 2, and Gemini's scores were generally concentrated between 1 and 2. However, GPT uniquely assigned a 3 for the language trait, giving a relatively lenient evaluation. Meta predicted a 2 for this item, same as Human, suppressing overestimation, and consistently presented scores identical to Human scores for the remaining traits. This case shows that when GPT attempted to overestimate in some items, Meta did not uncritically reflect this but made adjustments based on the overall LLM tendencies and Human criteria.

5.3.3 Prompt 8. Essay 21385

Trait	Peer Mean	Meta Score	Human Score	Claude	Gemini	GPT	LLaMA
Cont	5.00	5.0	5.0	5.0	4.0	6.0	5.0
Conv	5.00	5.0	5.0	5.0	5.0	5.0	5.0
Org	5.00	5.0	5.0	5.0	4.0	6.0	5.0
SF	5.00	5.0	5.0	5.0	4.0	6.0	5.0
Voice	5.25	5.0	5.0	6.0	4.0	6.0	5.0
WC	5.00	5.0	5.0	5.0	4.0	6.0	5.0

Table 8: Trait-level scores for essay #21385 from peer models, Meta-LLM, human raters, and individual LLMs.

This essay was a high-scoring essay where the Human evaluation score was a perfect 5 for all traits. Claude assigned a 5 for all items, consistent with Human, and LLaMA and Meta also recorded the same scores. In contrast, Gemini gave a 4 for all traits, consistently evaluating lower than Human, while GPT showed overestimation by assigning a 6 for traits like voice, word_choice, and organization. Even in this situation where Gemini underestimated and GPT overestimated, Meta remained steadfast and consistently assigned a 5, the same score as Human. This shows that through the bias signal inserted into the Meta model, extreme biases among LLMs were not directly reflected, and the existing judgment capability of LLaMA was maintained.

These three cases demonstrate that ADBIAS does not simply average the scores of multiple LLMs but performs intelligent calibration based on consistency, rubric fidelity, and alignment with human evaluation. In particular, by suppressing extreme scores and placing greater weight on the judgments of reliable models, ADBIAS consistently produces **more stable and human-like evaluation results** than individual LLM raters.

Through the experiments above, it was comprehensively verified that the proposed ADBIAS framework is effective in improving evaluation accuracy and substantially mitigating inter-model scoring bias in automated essay evaluation.

6 Conclusion

We introduced ADBIAS, a modular framework for bias-aware essay scoring that integrates multiple LLMs through trait-wise evaluation and Rasch-based calibration. By explicitly modeling rater bias and incorporating severity metadata into a Meta-LLM, our approach enables more consistent and fair scoring across diverse prompts. Empirical results demonstrate that ADBIAS improves both agreement with human ratings and inter-model consistency, particularly under conditions of high trait-level variance. While our framework addresses a key gap in LLM-based AES, several open challenges remain, which we outline below.

Limitation

While ADBIAS represents a step toward more accountable and transparent essay scoring, several limitations warrant further investigation.

Model set dependency

Our framework relies on a predefined set of four LLMs. Although chosen for diversity, the system's performance is inevitably shaped by the characteristics of these specific models. In future work, integrating adaptive model selection or confidence-weighted aggregation strategies could make the framework more flexible and robust to model variability.

Unmodeled rationale quality

Although each peer model provides rationales alongside trait scores, the Meta-LLM does not assess their quality directly. This leaves open the risk that misleading or low-quality rationales could influence final decisions. Incorporating rationale evaluation modules or rationale-weighted scoring could strengthen reliability and interpretability.

Simplified bias modeling

We adopt a trait-specific, additive formulation of bias via MFRM. While effective, this approach may overlook more complex interactions—such as non-linear dependencies between traits, prompt difficulty, or topic familiarity. Future extensions might explore neural or hybrid approaches to bias modeling that go beyond additive assumptions.

Limited domain and language scope

Our evaluation focuses exclusively on English-language essays from the ASAP dataset. The generalizability of ADBIAS to other languages, genres (e.g., argumentative, narrative), or domains (e.g., scientific writing) remains an open question. Extending this framework to cross-lingual or cross-domain settings could reveal new dimensions of model bias and calibration.

Alignment with human scoring standards

While ADBIAS improves inter-model agreement, it is not explicitly optimized for alignment with human raters. Closing this gap may require additional supervision signals—such as human-provided rationales or calibration through reference scores—or training the Meta-LLM with contrastive examples to reflect human expectations more closely.

We view ADBIAS as a foundation rather than a finished solution—one that offers a principled and extensible framework for multi-agent AES. By explicitly adjusting bias, employing a modular architecture, and supporting a wide range of LLMs, the framework offers a scalable, fair, and interpretable solution for automated writing assessment.

References

SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. 2025. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. *Preprint*, arXiv:2410.14202.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1545. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2024. Autoregressive score generation for multi-trait essay scoring. *Preprint*, arXiv:2403.08332.

Heejin Do, Taehee Park, Sangwon Ryu, and Gary Geunbae Lee. 2025a. Towards prompt generalization: Grammar-aware cross-prompt automated essay scoring. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2818–2824, Albuquerque, New Mexico. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. 2025b. Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring. arXiv preprint. ArXiv:2502.20748 [cs.CL].

Thomas Eckes. 2023. Rater bias in writing assessment: A many-facet rasch modeling approach. *Language Testing*, 40(1):123–145.

Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Shannon Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi, and Yulia Tsvetkov. 2025. When one llm drools, multi-llm collaboration rules. arXiv preprint. ArXiv:2502.04506 [cs.CL].

Zhaoyi Hou, Haotian Zhang, and Simin Liu. 2025. Improve llm-based automatic essay scoring with linguistic features. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume PMLR 273, pages 41–65.

William T. Hoyt and Jamie L. Kern. 1999. Rater effects in performance assessments. *Journal of Educational and Behavioral Statistics*, 24(3):253–277.

Mohamed Hussein, Nobal B. Niraula, and Vasile Rus. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. In *International Conference on Artificial Intelligence in Education*.

Yuki Ito and Qiang Ma. 2025. Ensemble tot of llms and its application to automatic grading system for supporting self-learning. arXiv preprint. ArXiv:2502.16399 [cs.IR].

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Prompting large language models for zero-shot essay scoring via multi-trait specialization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ting-Yun Lei, Ming-Hsiu Chen, and Shuo-Huan Ma. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. In *Advances in Neural Information Processing Systems*.

Yuxuan Li and Hwee Tou Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jia Kang Liew and Chew Lim Tan. 2024. On automated essay grading using large language models. arXiv preprint. ArXiv:2402.00001.

John M. Linacre. 1989. *Many-Facet Rasch Measurement*. MESA Press.

Sarah Mansour, Dhairya Thakkar, and Swapnil Raghuvanshi. 2024. Can large language models automatically score proficiency of written essays? arXiv preprint. ArXiv:2403.00002.

Michael Matta, Sterett H. Mercer, and Milena A. Keller-Margulis. 2023. Implications of bias in automated writing quality scores for fair and equitable assessment decisions. *School Psychology*, 38(3):173–181.

Carol M. Myford and Edward W. Wolfe. 2004. Investigating rater effects in the writing section of a large-scale language test. *Journal of Educational Measurement*, 41(2):129–161.

Ellis B. Page. 1966. The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5):238–243.

Ellis B. Page, John P. Poggio, and Timothy Z. Keith. 1997. Computer analysis of student essays: Finding trait differences in student profile. In *Proceedings of the Annual Meeting of the American Educational Research Association*, Chicago, IL, USA. ERIC Document Reproduction Service No. ED411316.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. Refiner: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Andrew Ridley, Sidney D'Mello, and Shawn Kelly. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*.

K. Rubei and 1 others. 2025. Bias-aware multi-trait scoring for automated essay evaluation. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. To appear.

Kathrin Seßler, Melanie Fürstenberg, Barbara Deml, and Enkelejda Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. arXiv preprint. ArXiv:2411.16337 [cs.CL].

John Stamper, Kenneth Koedinger, and Steven Ritter. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, USA. Association for Computational Linguistics.

Shuyao Tang, Robert Knewton, and Xuan Hu. 2024. Harnessing llms for multi-dimensional writing assessment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Masaki Uto and Maomi Ueno. 2020. Generalized multifacet rasch model for evaluating rater performance. *Psychometrika*, 85(3):630–654.

Tim vor der Brück and Marc Pouly. 2024. Estimating text similarity based on semantic concept embeddings. arXiv preprint. ArXiv:2401.04422.

Xuezhi Wang, Jason Wei, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Advances in Neural Information Processing Systems*.

Tomohiro Yamashita. 2024. An application of many-facet rasch measurement to evaluate automated essay scoring: A case of chatgpt-4.0. Language Testing (submitted).

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research*.

Appendix A Supplementary Tables and Figures

Table 9: Summary of AES-related Research Features: A = *Multi-Trait*, B = *LLM*, C = *Multi-LLM*, D = *Rationale*, E = *MFRM*

Reference	Model	A	B	C	D	E
Hussein et al. (2020)	LSTM-AES	✓	✗	✗	✗	✗
Ridley et al. (2021)	Trait-attn.	✓	✗	✗	✗	✗
Do et al. (2023)	ProTACT	✓	✗	✗	✗	✗
Li & Ng (2024)	ShallowNN	✓	✗	✗	✗	✗
Liew & Tan (2024)	4LLMs	✓	✓	✗	✗	✗
Paul et al. (2024)	REFINER	✓	✓	✗	✓	✗
Tang et al. (2024)	GPT, LLaMA	✓	✓	✗	✓	✗
Yamashita et al. (2024)	GPT	✓	✓	✗	✗	✓
Chu et al. (2025)	RMTS	✓	✓	✗	✓	✗
Do et al. (2025a)	RaDME	✓	✓	✗	✓	✗
Do et al. (2025b)	GAPS	✓	✓	✗	✓	✗
Hou et al. (2025)	LLM+LF	✓	✓	✗	✗	✗
Ito & Ma (2025)	GET	✓	✓	✓	✓	✗
Anonymous (2025)	ADBIAS	✓	✓	✓	✓	✓

Table 10: QWK scores by model across different traits.

Trait	Claude	Gemini	GPT	LLaMA	**ADBIAS**
Content	0.564	0.484	0.448	0.454	0.488
Conv.	0.306	0.265	0.296	0.390	0.314
Lang.	0.529	0.541	0.401	0.467	0.485
Narr.	0.608	0.559	0.521	0.523	0.553
Org.	0.552	0.477	0.427	0.428	0.471
PA	0.504	0.390	0.389	0.415	0.425
SF	0.427	0.401	0.402	0.498	0.432
Style	0.241	0.250	0.215	0.267	0.255
Voice	0.379	0.471	0.466	0.393	0.486
WC	0.443	0.385	0.382	0.455	0.428
AVG	0.455	0.417	0.395	0.429	0.454

Table 11: QWK scores by model across different prompts (essay sets).

Prompt	Claude	Gemini	GPT	LLaMA	**ADBIAS**
S1	0.419	0.360	0.338	0.388	0.396
S2	0.567	0.528	0.480	0.550	0.544
S3	0.542	0.512	0.480	0.490	0.555
S4	0.592	0.561	0.483	0.491	0.570
S5	0.595	0.557	0.447	0.514	0.579
S6	0.492	0.346	0.367	0.386	0.431
S7	0.306	0.296	0.229	0.350	0.298
S8	0.450	0.383	0.439	0.407	0.462
AVG	0.495	0.443	0.408	0.447	0.479

LLM Prompt	**Generation Output**

System Prompt:
You are a rater for essays written by students...
Use the scoring criteria
`{rubric_guide}`
User Prompt:
– Write a justification for each trait...
– Assign a score from 1–6...
Return as Python dict:...
	`{"essay_id":"1",`
	`"essay_set":"1",`
	`"scores":{. . .}}`

Figure 6: Prompt structure and expected generation format. All LLMs followed a unified prompt design consisting of system and user instructions. The prompt guides the model to generate trait-wise rationales and scores in a consistent Python dictionary format, facilitating reliable evaluation and parsing.

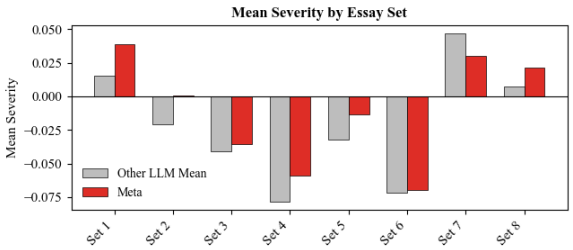

Figure 7: (a) Mean Severity for Meta and other LLMs

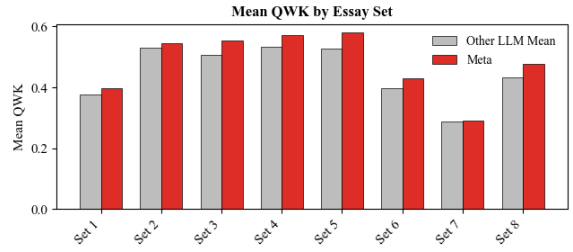

Figure 8: (b) Mean QWK scores for Meta and other LLMs