This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICCVDM71812.2026.11709134, 2026 7th International Conference on Computer Vision and Data Mining (ICCVDM)

Automated Essay Marking with Data-Efficient Augmentation and Score-Token Weighted Loss

Yamama Shakaa*, Nadine Aburumman, Anastasia Anagnostou, Simon Taylor, Yongmin Li
Department of Computer Science, Brunel University London, London, UK
{yamama.shakaa, nadine.aburumman, anastasia.anagnostou, simon.taylor, yongmin.li}@brunel.ac.uk
*Corresponding author

***Abstract*—Automated Essay Scoring (AES) research has predominantly focused on holistic scoring, yet rubric-based trait scoring provides richer diagnostic feedback for both students and instructors. While encoder-based models such as BERT have achieved competitive trait scoring performance, they are limited to score prediction and cannot generate the explanatory feedback that makes rubric-based assessment educationally useful. Large Language Models (LLMs), by contrast, naturally support both scoring and feedback generation within a single unified framework. However, most existing work on fine-tuned LLM-based rubric scoring relies on either large annotated corpora or synthetic data augmentation to achieve competitive performance, limiting practical adoption in settings where neither is readily available. This paper investigates whether a small, carefully balanced dataset of real essays can enable a fine-tuned LLM to achieve competitive rubric-based trait scoring. We fine-tune Qwen2.5-7B-Instruct using QLoRA on progressively augmented subsets of DREsSNew (371, 478, and 736 essays), targeting underrepresented score bands at each stage, and propose a score-token weighted loss function that concentrates the training signal on score-bearing tokens. Data augmentation drives the majority of the performance improvement: expanding from 371 to 736 essays yields gains of +0.042 QWK on Content, +0.021 on Organization, +0.041 on Language, 0.024 on Total. Score-token weighting provides a further modest but consistent improvement, most notably on Content (+0.025). Our best configuration achieves per-trait QWK of 0.631 (Content), 0.612 (Organization), 0.625 (Language), and 0.669 (Total) on the DREsSNew test set using only 736 real essays — remaining competitive with models trained on datasets over 65 times larger while preserving the generative capability needed for downstream feedback generation.**

***Index Terms*—Text Mining, Data Augmentation, Educational Data Mining, Automated Essay Scoring, QLoRA**

I. Introduction

Automated Essay Scoring (AES) has been an active area of natural language processing research for decades, with applications ranging from large-scale standardised testing to classroom writing instruction. Traditional AES systems based on handcrafted features or encoder-only models such as BERT have achieved competitive scoring performance on both holistic and trait-based tasks [1], [2], but they are fundamentally constrained to score prediction and cannot generate the diagnostic, rubric-aligned feedback that makes writing assessment educationally meaningful. This limitation has motivated a shift toward Large Language Models (LLMs), which can naturally unify scoring and feedback generation within a single framework by predicting structured scores and producing explanatory text in a single generation.

Recent work has demonstrated the potential of LLM-based AES across several directions. Rubric-aware chain-of-thought (CoT) prompting has improved scoring transparency by decomposing assessment into explicit reasoning traces before aggregating a final score [3]. Fine-tuning open-weight LLMs with multi-task LoRA on trait-specific rubrics has shown strong results on IELTS writing assessment. For instance, supervised fine-tuning of LLaMA-3-8B on 5,088 expert-annotated essays across four IELTS traits (Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range) achieves per-trait QWK scores ranging from 0.705 to 0.748 [4]. Comparative analyses of prompting versus fine-tuning have further shown that even a compact fine-tuned model substantially outperforms zero-shot or few-shot prompting of much larger models, including GPT-4 on per-trait scoring tasks [1], [5].

However, across these studies, strong performance has consistently been tied to dataset scale. The IELTS fine-tuning work draws on a private dataset of over 5,000 expert-annotated essays, of which 3,000 are used for supervised fine-tuning [4], with additional stages for reward modelling and Reinforcement Learning from Human Feedback (RLHF), a setup that presupposes access to large human preference datasets and multi-stage training infrastructure. The state-of-the-art rubric-based scoring system on DREsSNew combines 2,279 real classroom essays with over 40,000 synthetically generated samples [1], and even data-efficiency-focused comparative studies use full benchmark datasets [5]. None of these works directly investigate how far a compact model can be pushed with an order of magnitude less data through training-data curation and loss design alone.

This paper addresses that gap directly. We fine-tune Qwen2.5-7B-Instruct [6] using Quantized Low-Rank Adaptation (QLoRA) [7] on the DREsSNew benchmark [1], a real-classroom dataset of English as a Foreign Language(EFL) argumentative essays scored by expert instructors on three rubric dimensions. Rather than relying on scale or synthetic augmentation, we investigate two targeted interventions: a score-distribution-aware data augmentation strategy that progressively expands the training set from 371 to 736 real essays by targeting underrepresented score bands, and a score-token weighted loss function that concentrates the training signal on the three numeric score tokens embedded within each structured JSON output.

Copyright 'For the purpose of open access, the author has applied a 'Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.'

The main contributions of this paper are:

1) A **score-distribution-aware augmentation strategy** that targets underrepresented score bands using each trait's own median bucket count as a data-driven threshold.
2) A **score-token weighted loss function** for generative LLM fine-tuning that upweights the loss contribution of score-bearing tokens relative to surrounding output structure.

II. Related Work

A. Evolution of Automated Essay Scoring

Automated Essay Scoring (AES) has progressed from early statistical systems such as Project Essay Grade (PEG) [8] to neural architectures based on LSTMs and CNNs that capture richer semantic and contextual signal [2], [9]. Encoder-based transformer models such as BERT subsequently achieved competitive holistic and trait-level scoring performance [2], but remain limited to score prediction and cannot generate the diagnostic feedback that makes rubric-based assessment educationally useful. This limitation, together with the emergence of instruction-tuned Large Language Models (LLMs) such as the LLaMA and Qwen families, has driven a shift toward LLM-based AES, which unifies scoring and feedback generation within a single generative framework.

B. Rubric-Aligned Fine-Tuning for Explainable Scoring

Recent work has combined structured reasoning with fine-tuning to make LLM-based essay scoring more transparent. Xu et al. [3] proposed QwenScore+, a framework that integrates rubric-aware chain-of-thought (CoT) reasoning with a two-stage optimization pipeline built on the Qwen model family. Rather than relying solely on zero-shot prompt engineering, their approach applies supervised fine-tuning (SFT) on a specialized dataset of 5,088 expert-annotated essays to decompose evaluations into trait-specific dimensions, before leveraging Proximal Policy Optimization (PPO) to align the generated textual feedback with human expert preferences. By structuring the pipeline into explicit multi-trait pathways, QwenScore+ achieves strong human-machine agreement, with QWK scores across localized trait dimensions reaching 0.78–0.84. While this architecture substantially improves scoring traceability and feedback quality over holistic prediction, it highlights a distinct data-efficiency gap: it depends on thousands of fully annotated expert examples and a resource-intensive multi-stage training pipeline to achieve robust calibration.

C. Multi-Adapter Fine-Tuning of LLMs for Trait-Level Scoring and Feedback

Xu et al. [4] fine-tuned LLaMA-3-8B using a multi-task LoRA (M-LoRA) architecture on a private dataset of 5,088 expert-annotated IELTS essays. A separate lightweight adapter is trained for each of four core IELTS traits (Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range) while the underlying LLaMA-3 backbone remains frozen, mitigating catastrophic forgetting across traits. A subsequent PPO-based RLHF stage refines the personalization and quality of the generated textual feedback. While this work demonstrates robust multi-dimensional scoring and feedback generation, it does not examine how performance scales under a constrained training budget, nor the interaction between dataset size and loss-weighting strength across trait-specific adapters.

D. Structured Reasoning and Multi-Agent Frameworks

A parallel line of work improves interpretability and robustness through multi-agent decomposition rather than single-model fine-tuning. AutoSCORE [10] separates rubric-relevant component extraction from final scoring across two dedicated agents, forcing the reasoning path through human-aligned intermediate states and particularly benefiting smaller language models with limited reasoning capacity. Multi-Agent Rubric-guided Essay Scorer (MARES) [11] and the Roundtable Essay Scoring (RES) framework [12] instead simulate a "society of agents" that engage in dialectical reasoning or debate to reach a consensus score. Comparative analysis of single- versus multi-agent architectures [13] suggests that multi-agent systems are particularly effective at flagging critically weak essays through specialized veto logic, while single-agent models remain a more cost-effective choice for mid-range assessment. Complementary work has focused on specialized evaluator models rather than general-purpose LLMs: the Prometheus family [14], [15] induces fine-grained evaluation capability through large synthetic instruction-response corpora, and LCES [16] formulates AES as a zero-shot pairwise comparison task to provide an accurate open-source alternative to proprietary APIs. These approaches improve interpretability and robustness but, like the fine-tuning frameworks above, are developed and evaluated at a scale far beyond the low-resource setting this paper targets.

E. Large Language Models as Alternatives to Human Evaluators

A critical question underlying all of the above is whether LLMs can serve as viable alternatives to human experts for text evaluation at all. Chiang and Lee [17] were among the first to propose "LLM evaluation" as a reproducible, faster, and cheaper alternative to human assessment, showing that models such as InstructGPT and ChatGPT can achieve high consistency with expert human ratings on tasks like open-ended story generation. This potential is further supported by G-EVAL [18], which uses chain-of-thought prompting together with token probabilities to achieve stronger alignment with human judgments on Natural Language Generation (NLG) evaluation tasks, and by Tang et al. [19], who found that well-crafted prompt engineering allows GPT-4 to reach scoring accuracies that approach human-rater baselines on multi-dimensional writing assessment.

However, other studies highlight significant divergences between model behavior and human grading practice. Mathew et al. [20] argue that LLMs do not grade essays like humans:

models tend to overrate short, underdeveloped essays while over-penalizing minor grammatical errors in longer, more sophisticated writing. Yancey et al. [21] similarly found that a traditional machine learning baseline (XGBoost) outperformed GPT-3.5 and GPT-4 in rating short L2 essays on the Common European Framework of Reference (CEFR) scale, suggesting that LLMs lack the precision of feature-driven approaches for high-stakes quality prediction. A further concern is a "generative bias," in which LLM-based evaluators may prefer outputs generated by other LLMs over high-quality human-written text. To bridge this gap, specialized open-source evaluators such as Prometheus and Prometheus 2 [14], [15] have been developed to mirror human and GPT-4 judgments more closely across both direct assessment and pairwise ranking formats by training on massive collections of fine-grained rubrics and reference materials, rather than relying on a general-purpose LLM's evaluation ability out of the box.

F. Prompting vs. Fine-Tuning vs. Frozen Embeddings

Hallaç and Oğul [5] systematically compared prompting, fine-tuning, and frozen representation learning on the Feedback Prize–ELL dataset (3,911 essays scored across six analytic traits). A fine-tuned LLaMA-3.2-1B model achieved the highest per-trait QWK (0.655), while frozen embeddings from the same model used to train a traditional Support Vector Regression (SVR) model yielded a comparable QWK of 0.639, suggesting that scoring-relevant information is already well encoded in intermediate representations without task-specific weight updates. In contrast, zero- and few-shot prompting of substantially larger LLMs, including Phi-4:14B and GPT-4 Turbo, trailed well behind (QWK ≤ 0.41), frequently suffering from structured-output sensitivity and parser failures. The strong performance of frozen LLM embeddings paired with SVR in prior work [5] — achieving QWK 0.639 without any weight updates — suggests that scoring-relevant information is already encoded in LLM representations without task-specific training. The relative failure of zero- and few-shot prompting despite the same underlying model knowledge suggests that the bottleneck is not understanding essay quality, but expressing that understanding as precise, consistently-formatted rubric scores. Fine-tuning addresses exactly this gap, rather than teaching the model what good writing is, it teaches the model to reliably translate its implicit quality assessment into the structured output format required for rubric-based scoring.

G. Standardized Benchmarks and Generative Data Bottlenecks

Yoo et al. [1] introduced the DREsS benchmark to directly address training-set scale and score imbalance in AES. DREsS totals 48.9K samples across three sub-datasets: $\text{DREsS}_{\text{New}}$ (2,279 real-classroom EFL essays), $\text{DREsS}_{\text{Std}}$ (6,515 samples unified from existing sources), and $\text{DREsS}_{\text{CASE}}$ (40,185 synthetic essays generated via their Corruption-based Augmentation strategy for Essays (CASE), which introduces sentence-level errors into well-written essays to mimic authentic learner mistakes). Training purely on $\text{DREsS}_{\text{New}}$ compresses performance to a total QWK of 0.471 (BERT) and 0.466 (Llama 3.1 8B); CASE augmentation raises these to 0.685 and 0.691, respectively. Crucially, even when GPT-4o is given explicit EFL student personas, it fails to generate essays that match their intended rubric scores, yielding only 0.225 average QWK for generative augmentation [1], compared to 0.661 for corpus-specific corruption-based synthesis (CASE). This highlights that generating effective synthetic training data for EFL essay scoring is not as straightforward as prompting a general-purpose model, it requires corpus-specific strategies that deliberately replicate learner error patterns, underscoring a persistent data-efficiency gap that our work directly addresses.

H. Advanced Training Objectives for Multi-Trait Optimization

Beyond data scale, recent work has targeted the training objective itself. Scoring-aware Multi-reward Reinforcement Learning (SaMRL) [22] integrates Quadratic Weighted Kappa (QWK) directly into training through a bidirectional reward signal combined with an Mean Squared Error (MSE) penalty, better capturing the ordinal nature of human ratings than standard cross-entropy or unweighted MSE losses. Trait-Aware Policy Optimization (TAPO) [23] extends this further with trait-level credit assignment, providing localized feedback to specific tokens within a generated score sequence and overcoming the purely sequence-level scalar rewards of standard Group Relative Policy Optimization (GRPO). These advances in fine-grained, token-level supervision motivate our investigation of a score-token weighted loss as a lightweight, data-efficient alternative to full RL-based objectives.

I. Gap

Across these studies, performance gains are consistently tied to dataset scale, requiring over 5,000 expert-annotated essays for multi-trait IELTS architectures [4] or a 40,000+ synthetic expansion to stabilize baseline performance on $\text{DREsS}_{\text{New}}$ [1]. Structured multi-agent and specialized-evaluator approaches [10], [12], [14], [16] improve interpretability but do not target data efficiency, while advanced RL-based objectives such as SaMRL and TAPO [22], [23] operate primarily at sequence- or token-level reward design in large-scale settings. None of these works directly investigates how far a compact model can be pushed on an order of magnitude less data (fewer than 1,000 primary samples) through targeted training-data curation and localized score-token weighted loss design alone. This leaves a distinct operational gap regarding data efficiency and objective optimization, which our score-band-aware median augmentation strategy and score-token weighted loss are explicitly intended to address.

III. Methodology

A. Task Formulation

Given a student essay e and its associated writing prompt p, the goal is to predict three rubric-based trait scores $\hat{y}_c, \hat{y}_o, \hat{y}_l \in \{1.0, 1.5, \ldots, 5.0\}$ for Content, Organization, and Language

respectively, along with a derived total score $\hat{y}_t = \hat{y}_c + \hat{y}_o + \hat{y}_l \in [3.0, 15.0]$.

B. Evaluation Metric: Quadratic Weighted Kappa

We use Quadratic Weighted Kappa (QWK) as the primary evaluation metric, following [1]. QWK measures the agreement between predicted scores $\hat{y}$ and human-assigned scores y, accounting for the ordinal nature of the scoring scale. Unlike simple accuracy, QWK penalises disagreements proportionally to their distance — a prediction that is two score levels away from the true score incurs a much larger penalty than one that is off by only half a point. This makes QWK particularly appropriate for essay scoring, where partial credit matters and a prediction of 1.0 for a true score of 5.0 is far worse than a prediction of 4.5.

Formally, QWK is defined as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{1}$$

where O_{ij} is the observed count of essays where the true score is i and the predicted score is j, E_{ij} is the expected count under the assumption of independent ratings, and $w_{ij} = \frac{(i-j)^2}{(N-1)^2}$ is the quadratic weight penalising disagreements by their squared distance. A QWK of 1.0 indicates perfect agreement; 0.0 indicates agreement no better than chance; negative values indicate systematic disagreement. Values above 0.6 are conventionally interpreted as substantial agreement [2], [24]. Since the scoring scale uses 0.5-point increments, QWK is computed by scaling all scores by a factor of 2 (mapping the range $[1.0, 5.0]$ to integer values $[2, 10]$) before applying the standard Cohen's kappa with quadratic weights [25]. This preserves the quadratic penalisation structure while ensuring compatibility with integer-based QWK implementations. All QWK scores reported in this paper are computed using this scaling approach via scikit-learn's `cohen_kappa_score` function [26].

C. Base Model and Fine-Tuning Framework

We fine-tune Qwen2.5-7B-Instruct [6] using Quantized Low-Rank Adaptation (QLoRA) [7]. The model is loaded in 4-bit NF4 precision with bfloat16 computation, reducing GPU memory from approximately 14 GB to 5 GB. Low-rank adapters (LoRA) [27] are injected into all attention and feed-forward projection matrices with rank r=8 and scaling factor α=16. We chose a low rank for two reasons. First, prior work shows that weight updates during downstream adaptation exhibit a low intrinsic rank, so small r values are often sufficient to match full fine-tuning performance [27]. Second, Wang et al. [28] find that for LLaMA 2-7B, a similarly-sized model, r=8 already captures a mean of 92.9% of gradient information across layers (min. 85.1%), with further increases to r=128 yielding only marginal gains (99.3%), suggesting higher ranks offer diminishing returns at this scale. This choice also suited our hardware, as all experiments run on a single NVIDIA RTX 5090 (32 GB VRAM), where a low-rank configuration keeps memory and storage overhead manageable.

The model is trained in a one-pass configuration: given a system prompt containing the full scoring rubric and a user message containing the essay and its associated writing prompt, the model generates a structured JSON response containing three trait scores, a derived total score, and holistic feedback in a single generation. This unified one-pass format is chosen over a decomposed three-adapter approach based on preliminary experiments on DS-371, in which the one-pass model achieved substantially higher total QWK (0.721 vs. 0.614 on a separate 130-essay balanced test set), with the most notable advantage on Language (0.634 vs. 0.449). These results should be interpreted with caution given the small test size and its balanced score distribution, which differs from the naturally left-skewed DREsSNew test set used throughout this paper [1]. Nonetheless, the consistent Language advantage motivates our architectural choice. A representative prompt and output example are provided in Appendix A. Training hyperparameters: learning rate 2×10^{-4} with cosine decay, effective batch size 8 (per-device batch 1, gradient accumulation 8), 3 epochs, maximum sequence length 2,048 tokens. Gradient checkpointing is enabled to reduce VRAM usage. The best checkpoint is selected based on minimum validation loss. Qwen2.5-7B-Instruct was selected for several reasons. First, it is a fully open-weight model that runs on a single consumer GPU without API dependency, consistent with our practical deployment motivation. Second, at the 7B parameter scale, Qwen2.5 demonstrated strong instruction-following capability at the time of our experiments [6], which is critical for producing well-formed structured JSON output reliably across all test essays. Third, the Qwen model family has been independently validated for rubric-based essay scoring tasks [3], providing a degree of prior evidence for its suitability for this domain. Finally, Qwen2.5 is fully compatible with 4-bit QLoRA fine-tuning without custom modifications, enabling straightforward reproducibility of our pipeline.

D. Score-Token Weighted Loss

Standard cross-entropy loss assigns equal weight to every output token, including JSON structural tokens (braces, key names, punctuation) that the base model already produces reliably without fine-tuning. Since QWK depends exclusively on three numeric score values embedded within the JSON output, we propose concentrating the training signal on those tokens.

We define a per-token weighted cross-entropy loss:

$$\mathcal{L} = \frac{\sum_i w_i \cdot \ell_i}{\sum_i w_i} \tag{2}$$

where ℓ_i is the cross-entropy loss at token position i, $w_i = \lambda$ for score-bearing tokens and $w_i = 1$ otherwise, and prompt tokens are masked with $w_i = 0$ following standard instruction fine-tuning practice.

Score-bearing positions are identified by anchoring the search to each score's specific JSON key (e.g. `"content_score"`) before locating the value token immediately following, rather than searching for the bare numeric

value. This key-anchored approach correctly handles the common case where two or more traits share the same score value. We test multiple weight multipliers: $\lambda \in \{4, 6, 8\}$ on DS-478 and $\lambda \in \{3, 6\}$ on DS-736, with λ=0 (unweighted) as the baseline for both.

IV. Experiments and Results

A. Score-Distribution-Aware Data Augmentation

All fine-tuned models use essays drawn from **DREsS-New** [1], a real-classroom dataset of argumentative essays written by EFL undergraduate students at a South Korean university during timed 40-minute in-class writing tests. Each essay is scored by a single expert instructor on three rubric dimensions — Content, Organization, and Language — on a scale of 1.0 to 5.0 in 0.5-point increments. A fixed held-out test set of 456 essays is used for all evaluations. According to [1] real-classroom essay datasets are naturally left-skewed because instructors are often hesitant to give very low scores leading to a systematic underrepresentation of the lowest score range. A model trained on such imbalanced data learns a biased scoring function that performs well on common mid-range scores but poorly at the extremes. Three progressively augmented training sets are constructed as described below; Figure 1 illustrates their score distributions per trait.

DS-371 is the initial training set of 371 essays drawn from DREsSNew, manually curated to achieve an approximately bell-shaped score distribution across all three traits, with most essays concentrated in the mid-range scores (2.5–4.0) and fewer at the extremes. Score distributions show near-zero skewness (Content: -0.09, Organization: -0.22, Language: $+0.02$) with means of 3.25, 3.30, and 3.16 respectively. Despite this overall balance, extreme score bands remain severely thin: Content and Organization each have fewer than 10 essays at score 1.0, and Language has no essays at score 1.0 and only 6 at score 5.0. The distribution shows that the model has insufficient signal to learn reliable scoring behaviour.

DS-478 augments DS-371 with 107 additional essays drawn from the remaining DREsSNew pool not used in training or testing. Augmentation targets score bands falling below each trait's own median bucket count — computed directly from DS-371 — . Content-1.0 increases from 9 to 18 essays, Organization-1.0 from 8 to 16, and Language-1.0 from 0 to 5. Despite targeting the median threshold, Language extremes remain severely constrained by pool availability — Language-5.0 reaches only 11 essays after exhausting the available DREsSNew pool at that score level, falling short of the median target. The resulting DS-478 distribution retains a roughly bell-shaped form but is more spread out than DS-371, with improved coverage at the low extremes.

DS-736 further augments DS-478 with 258 real essays sourced from the broader DREsS dataset [1], which incorporates essays from ASAP [29], ASAP++ [30], and ICNALE-EE [31] with scores standardised to the same 1.0–5.0 rubric scale by the original dataset authors. Essays sourced from ASAP and ASAP++ were cleaned to remove corpus-specific anonymisation placeholders (e.g. `@PERSON1`, `@CAPS1`) and assigned unique ID prefixes to avoid collision with DREsS-New identifiers. The resulting DS-736 distribution exhibits a two-tailed shape with elevated counts at both low and high extremes. Extreme-score coverage improves substantially over DS-478: Content-1.0 increases to 151 essays, Language-1.0 to 81, and Language-5.0 to 51. Table I summarises the three training sets.

TABLE I
Training Dataset Summary.

Dataset	n	C Mean	O Mean	L Mean	Min Bucket
DS-371	371	3.25	3.30	3.16	6
DS-478	478	3.25	3.30	3.18	5
DS-736	736	2.89	2.98	2.88	51

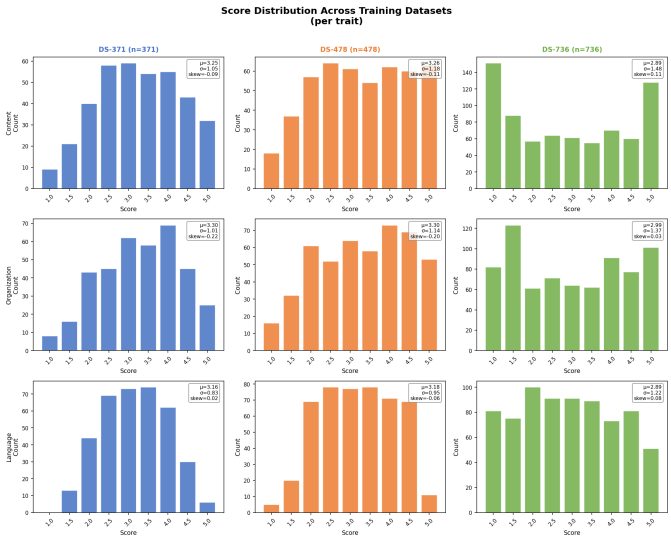


Fig. 1. Score distributions per trait across the three training sets. DS-736 shows substantially improved extreme-score coverage compared to DS-371 and DS-478.

B. Stratified Train/Validation Splitting

To prevent the split from removing rare extreme-score examples from training, we use a stratified split based on total score (rounded to the nearest 0.5). Score bands with fewer than 14 examples (derived as $\lceil 2/\min(\text{val_size}, 1-\text{val_size})\rceil$ with val_size=0.15) are retained entirely in training. This yields approximately 626 training and 110 validation essays for DS-736.

C. Main Results

Table II presents per-trait and total QWK for all configurations on the 456-essay DREsSNew test set.

Zero-shot prompting of frontier models, tested without any task-specific training data, yields poor trait-level performance despite full rubric provision: Qwen3-32B achieves 0.225 total QWK and DeepSeek achieves 0.221. Five-shot prompting improves results — DeepSeek improves to 0.310 — but all frontier configurations remain far below fine-tuned

performance, consistent with prior findings that task-specific fine-tuning substantially outperforms zero-shot and few-shot prompting for rubric-based trait scoring, regardless of frontier model size [1], [5].

Our baseline (DS-371, unweighted) already substantially outperforms all frontier baselines with 0.641 total QWK using only 371 real essays. Our best configuration (DS-736, λ=3) achieves 0.631 (Content), 0.612 (Organization), and 0.625 (Language), with a total QWK of 0.669, representing improvements of +0.067, +0.024, +0.034, and +0.028 over the DS-371 baseline on Content, Organization, Language, and Total respectively. Notably, our best configuration matches or exceeds both CASE-augmented baselines on Language and Content reported in Yoo et al. [1] — achieving Language 0.625 compared to 0.607 (BERT + CASE) and 0.589 (Llama 3.1 8B + CASE), and Content 0.631 matching Llama 3.1 8B + CASE (0.631) and approaching BERT + CASE (0.642) — despite using real essays only and no synthetic augmentation. The result suggests that targeted real-data augmentation at extreme score bands, combined with mild score-token weighting, can match or exceed the scoring performance of systems trained on much more extensive synthetic corpora.

D. Effect of Dataset Size

Expanding from DS-371 to DS-736 without score-token weighting yields consistent gains: +0.042 on Content, +0.021 on Organization, and +0.041 on Language. Language benefits most, consistent with its having the most severe extreme-score underrepresentation in DS-371 — Language=1.0 had zero training examples in DS-371, increasing to 81 in DS-736.

These results confirm that targeted score-band augmentation is the primary driver of improvement. Notably, the unweighted DS-736 model (0.665 total) already outperforms all DS-478 configurations regardless of weighting multiplier.

E. Effect of Score-Token Weighted Loss

On DS-478, increasing λ from 0 to 8 improves total QWK from 0.633 to 0.656 (+0.023). On DS-736, λ=3 outperforms both the unweighted baseline and more aggressive weighting. Increasing to λ=6 degrades performance across all traits relative to λ=3, suggesting that once data scarcity is addressed, only mild loss emphasis is beneficial. This interaction between data quality and loss weighting is a notable observation: the two interventions are complementary rather than substitutable.

F. Comparison with Published Baselines

Without synthetic augmentation, our model matches or exceeds DREsSNew-only baselines reported in Yoo et al. [1] on all three traits: Content 0.631 vs. 0.414 (BERT) and 0.413 (Llama 3.1 8B); Organization 0.612 vs. 0.311 (BERT) and 0.375 (Llama 3.1 8B); Language 0.625 vs. 0.487 (BERT) and 0.426 (Llama 3.1 8B). Against full DREsS + CASE models trained on 48,900+ essays, our model matches Content (0.631 vs. 0.631 for Llama + CASE) and exceeds Language (0.625 vs. 0.589), with the remaining gap on Organization (0.612 vs. 0.771).

V. Discussion

A. Data Augmentation as the Primary Lever

Our results consistently show that targeted score-band augmentation drives the majority of performance improvement, with score-token weighted loss providing a complementary but secondary refinement. The gains from DS-371 to DS-736 reflect the model learning from score ranges it previously had no exposure to — particularly Language=1.0, absent from DS-371 entirely. These results reveal that data augmentation and score-token weighting are complementary rather than substitutable interventions: weighting provides meaningful gains when data is thin (DS-478, λ=8: +0.023 total), but its benefit diminishes and eventually reverses as data coverage improves (DS-736, λ=3 is optimal, λ=6 actively hurts). The practical implication : prioritising real essay collection at underrepresented score bands is more valuable than tuning the loss function, though mild weighting remains a useful refinement once data coverage is adequate.

B. Why Mild Weighting Outperforms Aggressive Weighting on Richer Data

A high weight multiplier forces the model to focus so heavily on the score tokens that it neglects the quality of the surrounding text. On a small or highly imbalanced dataset, this focus is helpful because it boosts a weak signal. However, on a richer dataset, this same concentration distorts the model's learning process and hurts its ability to generalize. This is consistent with broader observations in the imbalanced learning literature, where oversampling and loss reweighting are most effective when imbalance is severe and can become harmful beyond the point where imbalance has been substantially corrected [32].

C. The Organisation Gap

The most notable limitation relative to the DREsS state of the art is the gap on Organization (0.612 vs. 0.771 for Llama 3.1 8B + CASE). This is consistent with the CASE augmentation strategy, which generates structurally diverse synthetic essays across a range of quality levels, providing richer supervision for Organization-specific patterns than is achievable from a small set of real essays alone. Several directions may help close this gap without synthetic generation: targeted collection of real essays specifically selected to cover diverse organizational structures across the full score range; augmentation strategies that treat Organization score as the primary selection criterion independently of other traits; or rubric-aware prompt engineering that provides stronger explicit supervision on structural features during fine-tuning. Investigating these directions, alongside the potential of targeted synthetic augmentation for Organization patterns specifically, remains an avenue for future work.

D. Post-Hoc Calibration

Table III reports the effect of three post-hoc calibration methods — Isotonic Regression, Quantile Mapping, and Beta Calibration — fitted and evaluated directly on the 456-essay

TABLE II
PER-TRAIT QWK ON THE DREsSNEW TEST SET (456 ESSAYS).

Model	Training Data	Content	Org	Lang	Total
DREsS Paper Baselines [1]					
BERT	DREsSNew (∼1,368)	0.414	0.311	0.487	0.471
Llama 3.1 8B	DREsSNew (∼1,368)	0.413	0.375	0.426	0.466
BERT + CASE	Full DREsS (48,900+)	0.642	0.750	0.607	0.685
Llama 3.1 8B + CASE	Full DREsS (48,900+)	0.631	0.771	0.589	0.691
GPT-4o zero-shot	None	0.310	0.322	0.231	0.304
GPT-4o 5-shot	None	—	—	—	0.428
Frontier Model Baselines (This Work)					
Qwen3-32B zero-shot + rubric	None	0.155	0.324	0.189	0.225
Qwen3-32B 5-shot	None	0.252	0.342	0.298	0.225
DeepSeek zero-shot + rubric	None	0.193	0.267	0.192	0.221
DeepSeek 5-shot	None	0.249	0.361	0.275	0.310
Fine-Tuned Models — This Work					
Qwen2.5-7B, DS-371, λ=0	371 essays	0.564	0.588	0.591	0.641
Qwen2.5-7B, DS-478, λ=0	478 essays	0.611	0.641	0.532	0.633
Qwen2.5-7B, DS-478, λ=4	478 essays	0.617	0.603	0.567	0.639
Qwen2.5-7B, DS-478, λ=6	478 essays	0.600	0.612	0.594	0.638
Qwen2.5-7B, DS-478, λ=8	478 essays	0.627	0.619	0.595	0.656
Qwen2.5-7B, DS-736, λ=0	736 essays	0.606	0.609	0.632	0.665
Qwen2.5-7B, DS-736, λ=6	736 essays	0.622	0.595	0.601	0.649
Qwen2.5-7B, DS-736, λ=3	**736 essays**	**0.631**	**0.612**	**0.625**	**0.669**

test set, representing an optimistic upper bound on achievable calibration gain. Even under this best-case scenario, all three methods degrade performance across all traits relative to the uncalibrated baseline, with Beta Calibration producing the largest drops (Content: −0.234, Organization: −0.126). Since calibration fails to improve results even when fitted on the test set itself, it is unlikely to provide any benefit under realistic cross-validated conditions. The failure of post-hoc calibration is not indicative of a model deficiency, but rather suggests that score-band augmentation has already addressed the systematic prediction bias that calibration methods are designed to correct. The Wasserstein distance, which measures the average displacement in score points needed to transform the predicted score distribution into the human score distribution, ranges from 0.138 (Content) to 0.177 (Organization) across all three traits, all well below half a scoring increment (0.25 points). This confirms that by ensuring adequate representation of all score levels during training, the augmented model's predicted score distribution already aligns closely with the true distribution, leaving little systematic bias for post-hoc methods to correct, and attempting to apply them actively degrades performance.

E. Implications for Practice

Our findings demonstrate that a 7B parameter model fine-tuned on fewer than 750 real essays can achieve rubric-based trait scoring competitive with systems trained on tens of thousands of examples, running entirely on a single consumer GPU without API dependency. Furthermore, the use of an LLM preserves the generative capability needed for downstream feedback generation — distinguishing this approach from encoder-only AES systems that achieve competitive scoring but cannot produce explanatory text.

TABLE III
POST-HOC CALIBRATION RESULTS ON BEST MODEL (DS-736, λ=3). CALIBRATION FITTED AND EVALUATED ON THE 456-ESSAY TEST SET, REPRESENTING AN UPPER BOUND ON ACHIEVABLE CALIBRATION GAIN.

Calibration	Content	Org	Lang	Total
None (baseline)	**0.631**	**0.612**	**0.625**	**0.669**
Isotonic Regression	0.542	0.560	0.575	0.610
Quantile Mapping	0.604	0.609	0.581	0.647
Beta Calibration	0.397	0.486	0.597	0.556

*Calibration was fitted and evaluated on the test set, providing an upper bound on calibration gain. Degradation under this optimistic setting suggests calibration is unlikely to help under realistic cross-validated conditions.

F. Limitations

All experiments use a single model family (Qwen2.5) and a single dataset (DREsSNew); whether findings generalise to other architectures, languages, or domains requires further investigation. The optimal weight multiplier (λ=3) was determined empirically and may require retuning for other datasets. The LoRA rank (r=8) was likewise fixed based on theoretical and empirical evidence from prior work rather than an ablation on our own dataset [27], [28]; whether a larger rank would improve scoring consistency on such a small training set remains an open question we did not test directly. We also did not investigate how variations in system prompt wording affect the stability of the model's structured JSON score generation; prompt sensitivity is a known factor in LLM output reliability and represents an important direction for future work. Finally, each configuration was trained and evaluated once. Given the stochastic nature of both LLM training and inference, results should ideally be averaged over multiple runs, and small

differences between configurations should be interpreted with caution.

VI. Conclusion

This paper investigated whether targeted data curation and loss design can substantially reduce annotation requirements for rubric-based automated essay scoring with large language models. Fine-tuning Qwen2.5-7B-Instruct using QLoRA on progressively augmented subsets of DREsSNew — expanding from 371 to 736 real essays through score-distribution-aware augmentation — we demonstrated that careful coverage of underrepresented score bands is the primary driver of per-trait scoring improvement. A complementary score-token weighted loss function provided further modest refinement, with λ=3 on DS-736 achieving the best overall configuration: per-trait QWK of 0.631 (Content), 0.612 (Organization), and 0.625 (Language).

These results, achieved using 736 real essays and no synthetic data, match or exceed models trained on datasets over 65 times larger on two of three traits, and substantially outperform frontier models under both zero-shot and five-shot prompting. A key finding is that score-token weighting and data augmentation are complementary: weighting provides meaningful gains when data is thin but becomes less necessary as data coverage improves.

Future work should investigate strategies to close the Organization gap, including targeted real-data collection for structurally diverse essays and rubric-aware prompt engineering, explore reinforcement learning from human feedback (RLHF) as a complementary alignment stage for all three traits, extend evaluation to additional model families and domains, and activate the generative capability of the fine-tuned model for full formative feedback generation.

Acknowledgment

The authors used Claude (Anthropic, version 4.6 Sonnet) as an AI writing assistant to help draft and refine portions of this manuscript, and NotebookLM (Google, default web version) for literature organization and synthesis. All experimental work, results, analysis, and final editorial decisions were made by the authors.

References

[1] H. Yoo, J. Han, S.-Y. Ahn, and A. Oh, "DREsS: Dataset for rubric-based essay scoring on EFL writing," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 13 439–13 454.

[2] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022, doi:10.1007/s10462-021-10068-2.

[3] W. Xu, M. Shahreeza, W. L. Hoo, and W. Yang, "Explainable AI for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting," *International Journal of Modern Physics C*, pp. 1–26, 2025.

[4] W. Xu, M. S. S. Kassim, and R. Mahmud, "Enhancing IELTS writing automated scoring with M-LoRA fine-tuned LLAMA-3 and human feedback-driven PPO reinforcement learning," *Scientific Reports*, vol. 16, p. 10865, 2026, doi:https://doi.org/10.1038/s41598-026-43318-w.

[5] İ. R. Hallaç and H. Oğul, "From prompting to fine-tuning: A comprehensive evaluation of llm strategies for automatic essay assessment," *SSRN preprint*, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5699710

[6] Qwen Team, "Qwen2.5 technical report," Alibaba Group, Tech. Rep., 2024. [Online]. Available: https://arxiv.org/abs/2412.15115

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 10 088–10 115. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html

[8] E. B. Page, "The imminence of grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.

[9] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882–1891, doi:10.18653/v1/D16-1193.

[10] Y. Wang, Z. Ding, X. Wu, S. Sun, N. Liu, and X. Zhai, "AutoSCORE: Enhancing automated scoring with multi-agent large language models via structured component recognition," vol. 40, no. 48, pp. 40 898–40 906, 2026.

[11] W. Cheng and Z. Li, "A multiagent scorer for improved AES of international students' essays," *International Journal of High Speed Electronics and Systems*, vol. 34, p. 2540759, 2025. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0129156425407594

[12] J. Jang, A. Moon, M. Jung, Y. Kim, and S. J. Lee, "LLM-agents at the roundtable: A multi-perspective and dialectical reasoning framework for essay scoring," *arXiv preprint arXiv:2509.14834*, 2025. [Online]. Available: https://arxiv.org/abs/2509.14834

[13] J. A. Idowu and A. Almasoud, "Specialists or generalists? multi-agent and single-agent llms for essay grading," *arXiv preprint arXiv:2601.22386*, 2026. [Online]. Available: https://arxiv.org/abs/2601.22386

[14] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo, "Prometheus: Inducing fine-grained evaluation capability in language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=8euJaTveKw

[15] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, "Prometheus 2: An open source language model specialized in evaluating other language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 4334–4353.

[16] T. Shibata and Y. Miyamura, "LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 29 988–30 001.

[17] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?" *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 15 607–15 631, 2023.

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-EVAL: NLG evaluation using GPT-4 with better human alignment," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

[19] X. Tang, H. Chen, D. Lin, and K. Li, "Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments," *Heliyon*, vol. 10, no. 14, p. e34262, 2024.

[20] J. G. Mathew, S. Taher, A. Kundu, and D. Barbosa, "LLMs do not grade essays like humans," *Preprint*, 2026. [Online]. Available: https://arxiv.org/abs/2603.23714

[21] K. P. Yancey, G. LaFlair, A. Verardi, and J. Burstein, "Rating short l2 essays on the CEFR scale with GPT-4," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 576–584.

[22] H. Do, S. Ryu, and G. G. Lee, "Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 16 427–16 438.

[23] Z. Wang, S. Lee, J. Wang, C. Miao, W. Li, and Y. Wu, "Trait-aware policy optimization for autoregressive multi-trait essay scoring," *arXiv preprint arXiv:2605.25731*, 2026. [Online]. Available: https://arxiv.org/abs/2605.25731
[24] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[25] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[26] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html
[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
[28] S. Wang, L. Yu, and J. Li, "LoRA-GA: Low-rank adaptation with gradient approximation," in *Advances in Neural Information Processing Systems*, vol. 37. NeurIPS, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/62c4718cc334f6a0a62fb81c4a2095a1-Paper-Conference.pdf
[29] B. Hamner *et al.*, "The Hewlett Foundation: Automated essay scoring (ASAP)," Kaggle Competition, 2012. [Online]. Available: https://www.kaggle.com/c/asap-aes
[30] S. Mathias and P. Bhattacharyya, "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)*. European Language Resources Association, 2018. [Online]. Available: https://aclanthology.org/L18-1187/
[31] S. Ishikawa, "The ICNALE and sophisticated contrastive interlanguage analysis of asian learners of english," *Learner Corpus Studies in Asia and the World*, vol. 1, pp. 91–118, 2013. [Online]. Available: https://www.researchgate.net/publication/285651859_The_ICNALE_and_sophisticated_contrastive_interlanguage_analysis_of_Asian_learners_of_English
[32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5128907

APPENDIX

A. System Prompt

The system prompt comprises four components: a persona definition, full rubric with behavioural anchors at all nine score levels (1.0–5.0 in 0.5 increments), critical scoring rules, and step-by-step task instructions.

Listing 1. System prompt structure

```
You are an expert University Writing Examiner.
Your task is to evaluate a student's essay across
three dimensions: Content, Organization, and
    Language.
```

```
[SCORING RUBRIC - BEHAVIORAL ANCHORS]
CONTENT:
5.0 - Fully developed argument, multiple
      well-supported reasons, insightful analysis
4.5 - Strong argument with well-supported reasons;
      minor gaps in depth or insight
...
1.0 - No clear argument, irrelevant or missing
      content

ORGANIZATION:
5.0 - Clear intro/body/conclusion, smooth
      transitions, logical flow throughout
...
1.0 - No recognisable structure

LANGUAGE:
5.0 - Accurate grammar, wide vocabulary, varied
      sentence structures, academic tone
...
1.0 - Severe errors throughout, meaning is often
      unclear
```

```
[CRITICAL SCORING RULES]
- Use 0.5 increments only.
- Never default to round numbers.
CONTENT: Must directly address the university
  prompt; clear stance with two developed
  reasons -> 4.0 or above; vague stance -> 2.0
  or below.
ORGANIZATION: Clear intro/body/conclusion with
  smooth transitions -> 4.0 or above; missing
  structure -> 2.0 or below.
LANGUAGE: Grammar errors never impede
  understanding -> 4.0 or above; errors
  frequently impede understanding -> 2.0 or
  below.

[TASK]
1. Read the essay against all three rubric
   dimensions.
2. Reason through each dimension step by step.
3. Assign a score using 0.5 increments.
4. Sum three scores for final_mark (3.0-15.0).
5. Write holistic feedback referencing all
   three dimensions.
```

B. Model Output Example

Listing 2. Model JSON output (student ID 420)

```
{
  "student_id": "420",
  "section_scores": {
    "content_score": 2.5,
    "organization_score": 2.5,
    "language_score": 3.0
  },
  "final_mark": 8.0,
  "general_feedback": {
    "holistic_feedback": "The essay presents a
    clear stance that attendance should be
    optional, but the argument is ......."
  }
}
```

The output demonstrates trait-specific scoring at 0.5-point granularity, correct total derivation (2.5 + 2.5 + 3.0 = 8.0), and rubric-aligned feedback referencing specific textual evidence — all in a single generation pass.