Article

Not peer-reviewed version

Explainable AI for Education: Enhancing Essay Scoring via Rubric-Aligned Chain-of-Thought Prompting

Wenbo Xu * , Muhammad Shahreeza , Wai Lam Hoo , Wudao Yang

Posted Date: 28 April 2025

doi: 10.20944/preprints202504.2338.v1

Keywords: automated essay scoring (AES); large language models (LLMs); chain-of-thought prompting; rubric alignment; explainability

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2025 doi:10.20944/preprints202504.2338.v1

Article

Explainable AI for Education: Enhancing Essay Scoring via Rubric-Aligned Chain-of-Thought Prompting

Wenbo Xu [1,2,*], Muhammad Shahreeza [1], Wai Lam Hoo [1] and Wudao Yang [3]

[1] Universiti Malaya, Faculty of Computer Science and Information Technology, Kuala Lumpur 50603, Malaysia
[2] Jiangxi Arts and Ceramics Technology Institute, Faculty of Digital Arts, Jingdezhen 333001, China
[3] Yunnan Minzu University, School of Mathematics and Computer Science, Kunming 650504, China
[*] Correspondence: s2125850@siswa.um.edu.my or xuwenboxue_xwb@163.com; Tel.: +60-18-324-9182

Abstract: Automated Essay Scoring (AES) systems increasingly leverage fine-tuned large language models (LLMs) to enhance scoring accuracy and feedback generation. However, current LLM-based AES approaches often lack interpretability and consistent alignment with human scoring rubrics, limiting their practical adoption in educational settings. This study proposes QwenScore+, a novel framework that integrates rubric-aware Chain-of-Thought (CoT) prompting with reinforcement learning from human feedback (RLHF) to improve the transparency, quality, and educational alignment of automated feedback. QwenScore+ is evaluated on a proprietary IELTS writing dataset comprising over 5,000 essays annotated with trait-level scores and expert-written feedback. Experimental results demonstrate that QwenScore+ significantly outperforms strong baselines such as BERT, GPT-3.5, and GPT-4 in feedback generation, achieving higher BLEU, ROUGE-L, and cosine similarity scores, alongside improvements in trait-level scoring measured by quadratic weighted kappa (QWK). Furthermore, rubric-aligned CoT prompting enables the generation of feedback that better mirrors human reasoning patterns, as confirmed through automatic metrics and human evaluations. These findings highlight the potential of combining explainable reasoning strategies with human-aligned reward optimization to develop more transparent, reliable, and pedagogically valuable AES systems for real-world applications.

Keywords: automated essay scoring (AES); large language models (LLMs); chain-of-thought prompting; rubric alignment; explainability

1. Introduction

The emergence of Explainable AI (XAI) in education has underscored the critical need for transparency and interpretability in intelligent systems, particularly in high-stakes assessment scenarios where human trust and pedagogical value are paramount [1–4]. Automated Essay Scoring (AES), a central task of educational natural language processing (NLP), has evolved from traditional feature-based models to deep neural networks such as BERT [5], and more recently to fine-tuned Large Language Models (LLMs) like GPT-3.5 and GPT-4 [6–8]. These models demonstrate superior contextual understanding and scoring performance, typically measured by metrics such as Quadratic Weighted Kappa (QWK) [9,10]. Yet, they often operate as black boxes, providing minimal insight into the rationale behind their scoring and feedback outputs [10–12].

Recent advances in intelligent classroom environments have enabled the seamless integration of assessment tools within digital learning platforms. As depicted in Figure 1, student essays can be collected in real time and automatically evaluated using AI-powered AES systems [7,13]. This shift toward scalable, rubric-aligned essay evaluation underscores the need for models that are not only accurate but also interpretable and pedagogically grounded [4,9,14].

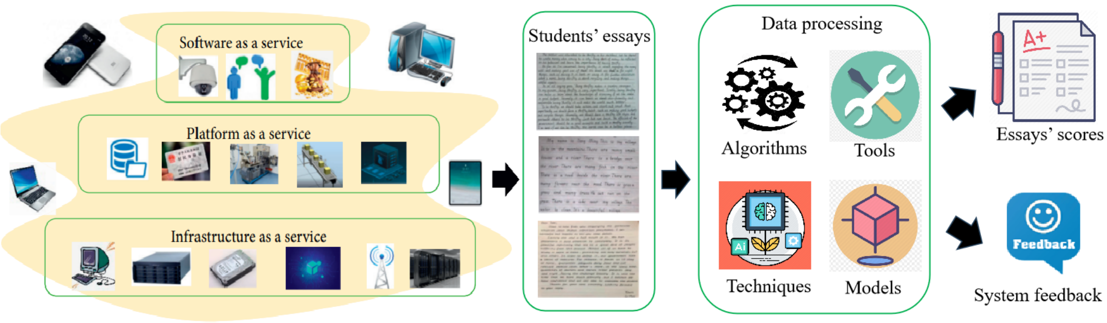


Figure 1. Automated essay scoring workflow in digital classrooms.

This lack of interpretability, combined with limited access to high-quality annotated data, presents significant challenges for practical deployment. Most publicly available AES datasets offer only coarse-grained scores without detailed expert feedback, limiting the efficacy of supervised fine-tuning [9,15]. Worse still, most existing AES datasets either lack fine-grained feedback or are inconsistently annotated, creating a bottleneck for model transparency and generalization [7,16].

To address these limitations, this study constructs a high-quality private IELTS [1] essay dataset, annotated by certified IELTS writing experts using the globally standardized Task 2 Band Descriptors [2]. Each essay is paired with detailed, trait-level feedback aligned with four core scoring dimensions: Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. The dataset not only provides robust training signals for LLM fine-tuning but also supports interpretability research by anchoring model behavior in well-defined educational constructs [11,17,18].

Built upon this foundation, we introduce QwenScore+, a novel rubric-aware Chain-of-Thought (CoT) prompting framework for Automated IELTS essay evaluation, implemented on Qwen[3]—an advanced Chinese-English bilingual LLM. To the best of our knowledge, this is the first study to apply rubric-aligned CoT prompting on Qwen for trait-level essay scoring and personalized feedback generation. While CoT prompting has demonstrated promising results in domains such as arithmetic reasoning and question answering [17,19,20], its integration into educational assessment—particularly within rubric-guided writing evaluation—remains underexplored [21–23].

Addressing this gap, our approach decomposes the scoring process into intermediate reasoning steps that are explicitly aligned with rubric criteria, enabling the model to generate interpretable feedback and trait-level scores that reflect human evaluative reasoning [23,24]. To further enhance the quality and alignment of model outputs, we incorporate Reinforcement Learning from Human Feedback (RLHF), enabling reward-based optimization that guides the model toward human-preferred scoring and feedback behaviors [25,26].

In summary, the key contributions of this work are:

1. Construction of a rubric-aligned IELTS dataset with fine-grained expert feedback: We present a new AES benchmark composed of over 5,000 essays with expert-annotated trait scores and paragraph-level feedback, facilitating high-fidelity model training and evaluation.
2. Introduction of rubric-aware Chain-of-Thought prompting in AES, implemented on Qwen: Our framework explicitly integrates scoring rubrics into multi-step reasoning prompts, significantly improving interpretability and scoring alignment with educational standards.
3. Improved scoring transparency via reasoning traceability and RLHF on Qwen: By coupling rubric traits with intermediate reasoning steps and fine-tuning via Reinforcement Learning from Human Feedback (RLHF), the proposed system—developed on the Qwen language model—generates

[1] The International English Language Testing System (IELTS) is a widely accepted standardized test designed to evaluate the English language proficiency of non-native speakers, commonly used in academic, immigration, and professional contexts.

[2] Official IELTS Task 2 Band Descriptors available from the British Council: https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf

[3] Qwen is a family of open-source large language models developed by Alibaba, available at https://github.com/QwenLM/Qwen.

coherent, criterion-referenced feedback, advancing the development of trustworthy and pedagogically useful AES models.

2. Related Work

This section reviews the relevant literature across three key dimensions: (1) the integration of XAI into AES, (2) the development of CoT prompting mechanisms, and (3) the design of rubric-aligned scoring strategies. While each area has witnessed notable progress in recent years, existing research still exhibits critical gaps in applying these methods to educational assessment contexts. The following subsections analyze the advancements and limitations in these domains, and identify the unexplored intersections that this study aims to address.

2.1. Research and Challenges of XAI in AES

In recent years, XAI has been increasingly integrated into educational assessment to enhance the transparency and reliability of automated scoring systems. Several studies have proposed methods such as score visualization, feature attribution, and attention-based interpretability to uncover the decision-making processes of scoring models. However, most of these approaches are grounded in conventional feature engineering or shallow neural networks, making them less effective when applied to large-scale, black-box systems like LLMs [10].

More recent efforts have focused on enabling rationale-based AES. For instance, Chu et al. [24] proposed a framework combining scoring-LLMs (S-LLMs) with generative explanations to support multi-trait scoring. Similarly, Cohn et al. [23] introduced the CoTAL framework, which integrates human scoring logic and prompt engineering to build a more interpretable and generalizable AES system. While such studies have advanced the application of XAI in AES, they primarily focus on output-level interpretability and often lack the systematic integration of rubric criteria, feedback strategies, and internal reasoning paths.

2.2. Progress and Transfer Gaps in CoT Reasoning Mechanisms

CoT prompting has been widely applied in domains such as mathematical reasoning, logical QA, and commonsense inference. Its core strength lies in explicitly exposing the step-by-step reasoning process, thereby improving the accuracy and interpretability of model outputs [19,20]. For example, Cao et al. [17] proposed the Tree-of-Thought framework, which supports branching decision pathways and excels in complex reasoning tasks. Diao et al. [25] further optimized reasoning consistency through active prompting strategies.

Despite these advancements, the application of CoT in educational assessment—particularly in AES—remains largely unexplored. No existing study has systematically employed CoT as a scoring inference mechanism to guide feedback generation aligned with rubric-based evaluation. This work pioneers the integration of CoT into the essay scoring pipeline, further enhanced by RLHF, to construct human-preference-aligned reasoning paths for scoring.

2.3. Exploration and Limitations of Rubric-Aligned Scoring Mechanisms

Rubrics serve as foundational tools in writing assessment, providing a standardized and interpretable basis for scoring. Early systems like e-rater employed hand-crafted features to partially align scoring with rubric criteria, leveraging rule-based and linguistic feature extraction [27]. More recent research has aimed to embed rubric dimensions into deep learning architectures or prompt templates to improve alignment, consistency, and control.

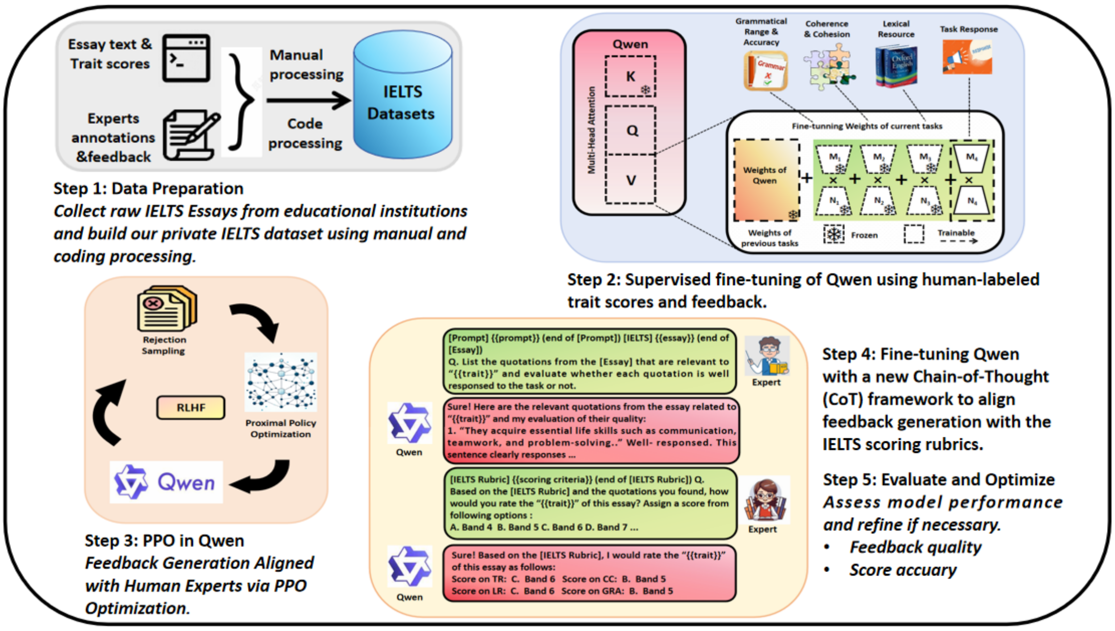
For example, Hashemi et al. [12] introduced LLM-R UBRIC, a calibrated framework for aligning language model scoring with multidimensional rubrics. Henkel et al. [18] incorporated rubric-aware CoT prompting in the AMMORE dataset to enhance assessment granularity, particularly in handling edge cases. While these studies show promise in rubric integration, most fail to tightly couple rubric logic with interpretable inference mechanisms, thus limiting their capacity to produce cognitively aligned feedback and scores.

doi:10.20944/preprints202504.2338.v1

3. Datasets and Methods

This section outlines the datasets used in this study and the methodological framework adopted to integrate CoT reasoning into AES. We first introduce the publicly available datasets and our proprietary IELTS corpus, which provides fine-grained, expert-annotated data for both scoring and feedback generation. We then present the CoT-enhanced scoring and feedback framework, detailing how reasoning steps, scoring criteria, and feedback strategies are systematically embedded into the model's prompting and generation process.

The following figure provides an overview of the entire methodological pipeline adopted in this study. Figure 2 illustrates the main components, including dataset input, rubric-aware prompting, structured CoT reasoning, and final score and feedback generation stages.



Step 1: Data Preparation
Collect raw IELTS Essays from educational institutions and build our private IELTS dataset using manual and coding processing.

Step 2: Supervised fine-tuning of Qwen using human-labeled trait scores and feedback.

Step 3: PPO in Qwen
Feedback Generation Aligned with Human Experts via PPO Optimization.

[Prompt] {{prompt}} (end of [Prompt]) [IELTS] {{essay}} (end of [Essay])
Q. List the quotations from the [Essay] that are relevant to "{{trait}}" and evaluate whether each quotation is well responsed to the task or not.

Sure! Here are the relevant quotations from the essay related to "{{trait}}" and my evaluation of their quality:
1. "They acquire essential life skills such as communication, teamwork, and problem-solving.." Well- responsed. This sentence clearly responses ...

[IELTS Rubric] {{scoring criteria}} (end of [IELTS Rubric]) Q. Based on the [IELTS Rubric] and the quotations you found, how would you rate the "{{trait}}" of this essay? Assign a score from following options :
A. Band 4 B. Band 5 C. Band 6 D. Band 7 ...

Sure! Based on the [IELTS Rubric], I would rate the "{{trait}}" of this essay as follows:
Score on TR: C. Band 6 Score on CC: B. Band 5
Score on LR: C. Band 6 Score on GRA: B. Band 5

Step 4: Fine-tuning Qwen with a new Chain-of-Thought (CoT) framework to align feedback generation with the IELTS scoring rubrics.

Step 5: Evaluate and Optimize
Assess model performance and refine if necessary.
- ***Feedback quality***
- ***Score accuary***

Figure 2. Overview of the proposed methodology integrating CoT reasoning into AES for structured scoring and feedback generation.

3.1. Public Datasets and IELTS Private Dataset

Finding suitable datasets is critical for advancing AES effectiveness, as they provide essential data for model evaluation. For an AES system to accurately evaluate an essay, it must be trained to recognise patterns in the data and to understand the characteristics of good and bad essays. The datasets used to train and evaluate an AES system must include various essays, including essays of different lengths, different topics, and written by different authors. This will ensure that the system can properly identify and evaluate the various types of essays that it will be asked to assess. To the best of our knowledge, seven English composition datasets are widely accepted and used in the AES field. Their details are demonstrated in Table 1. In some studies, researchers refer to the datasets they use as corpora. For convenience, we will refer to this concept uniformly as datasets in this paper.

Despite their widespread use, these datasets present several limitations. Most notably, they lack detailed, rubric-aligned trait annotations, personalized feedback, and expert-level error type categorization—components increasingly recognized as critical for explainable AES development. Moreover, few of them incorporate comprehensive manual quality control or record inter-rater consistency, limiting their reliability for fine-grained evaluation tasks.

To address these gaps, a proprietary IELTS writing dataset was constructed and annotated as part of this study. The dataset comprises 5,088 essays collected from IELTS candidates, with each essay independently scored by two certified IELTS examiners across four official traits: Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA).

In cases where trait-level score discrepancies exceeded one full band, senior adjudicators were engaged to arbitrate and ensure scoring validity.

Table 1. Commonly used public datasets in automated essay scoring.

Dataset	Writer	Samples	Prompts	Scoring	Rubric	Annotations
CLC-FCE[1]	Non-native	1,244	10	Holistic (1–40)	ESOL	80 linguistic error types
EFCAMDAT[2]	Non-native	1,180,310	1,447	Holistic (A1–D4)	CEFR	Grammar errors & POS tags
TOEFL11[3]	Non-native	12,100	8	Holistic (Low/Med/High)	TOEFL	None
SciEntsBank[4]	Native	15,357	15	Holistic (Assumed–Unaddressed)	FOSS	145,911 facet entailment
ASAP	–	12,980	8	Holistic (1–6)	OWSG	–
ASAP-SAS	Native	27,588	10	Holistic (0–2/3)	SOEI	None
ASAP++[5]	–	10,688	6	Attribute-based (4/5)	OWSG	–
ICLEv3[6]	Non-native	9,529	258	Holistic (B2/C1/C2)	CEFR	POS tagging
DREsS[7]	Non-native	1,782	22	Holistic (1–5) & Traits	EFL	None

[1] Cambridge Learner Corpus (FCE): https://www.ilexir.co.uk/datasets/index.html. [2] EF-Cambridge Open Language Database: https://ef-lab.mmll.cam.ac.uk/EFCAMDAT.html. [3] TOEFL11 dataset: https://catalog.ldc.upenn.edu/LDC2014T06. [4] Student Response Analysis corpus (Beetle & SciEntsBank): https://huggingface.co/datasets/nkazi/SciEntsBank. [5] ASAP: https://www.kaggle.com/c/asap-aes; SAS: https://www.kaggle.com/competitions/asap-sas/data; ASAP++: https://cfilt.iitb.ac.in/~egdata/. [6] International Corpus of Learner English (ICLEv3): https://corpora.uclouvain.be/cecl/icle/trial/. [7] DREsS: Rubric-based Essay Scoring on EFL Writing: https://arxiv.org/pdf/2402.16733.pdf.

Beyond trait scores, the dataset includes fine-grained feedback annotations aligned with the IELTS public band descriptors. Annotators were trained through a rubric calibration process, and their inter-rater reliability was evaluated using Fleiss' Kappa ($\kappa \geq 0.80$) across all traits, demonstrating strong scoring consistency. A data cleaning pipeline combining human validation and GPT-4-based prompt-engineered feedback normalization was employed to enhance the dataset's structural integrity and reduce annotation variance.

A sample JSON format of the annotated dataset is illustrated in Appendix B, providing insight into its structural design, scoring fields, and feedback schema. This IELTS dataset serves as a foundation for the model's supervised fine-tuning (SFT), reward model training, and reinforcement learning with human feedback (RLHF), contributing to both scoring accuracy and interpretability in AES. Building upon this curated resource, the following section introduces the methodological framework that integrates Chain-of-Thought reasoning into the automated scoring and feedback generation process.

3.2. *CoT-Based Integration of Feedback Generation and Scoring Mechanism*

Even with impressive progress, AES systems have always struggled to provide structured and explainable feedback against human grading criteria. CoT framework decomposes difficult evaluation tasks into systematic, step-by-step sub-tasks so that scoring and generation of feedback are transparent as well as consistent with pre-defined IELTS writing criteria. With the application of CoT, the model assesses essays along four main dimensions: TR, CC, LR, and GRA. This decomposition makes it possible to have a fine-grained assessment of essay quality, enhancing the accuracy as well as explainability of automatic scoring models.

As compared to conventional AES models returning a holistic score with no explicit reasoning procedure, the current research presents a new CoT-based method that breaks down essay assessment into systematic steps. Such breakdown improves transparency as it makes it easier for automatic models to provide explanations of scores and provide more interpretable feedback.

3.2.1. Structured Reasoning for Feedback and Scoring

Traditional AES models typically provide a single holistic score even without assessing the essay's strengths and weaknesses. AES models are opaque, as test-takers cannot see why they received a particular score. Unlike AES models, the CoT framework allows for structured, step-wise reasoning with a view to improving scoring precision as well as quality of feedback. CoT systematically breaks down the evaluation into a succession of reasoning steps, with scores as well as quality of feedback aligned with set IELTS assessment criteria.

Following the recent progress of Chain-of-Thought prompting [19], we employ a hierarchical evaluation scheme where the scores for each of the four traits—TR, CC, LR, and GRA—are assessed separately and combined to form a final rating. This decomposition avoids interference between dimensions as well as improves the overall robustness and interpretability of the assessment:

$$\text{Score}_{final} = f(\text{Score}_{TR}, \text{Score}_{CC}, \text{Score}_{LR}, \text{Score}_{GRA}) \tag{1}$$

where $f(\cdot)$ denotes a weighted aggregation function that integrates the individual scores for each dimension.

The multiple-step reasoning approach facilitates more specific feedback generation through:

- Feature Extraction: Determining linguistic as well as content-based features pertinent to each of the scoring dimensions.
- Rubric mapping: Matching extracted features with pre-established IELTS band descriptors for rubric consistency.
- Hierarchical Justification: Produces steps of structured justification that account for the given scores.
- Actionable Feedback Generation: Aligning with scoring criteria so that test-takers have valid, interpretable, and actionable guidance for improvement.

Unlike traditional AES models, which depend upon a one-pass decision-making mechanism, the CoT framework progressively refines its judgment through step-wise inference. This systematic approach aids scoring consistency as well as improves pedagogical value of the scores by directly associating writing weaknesses with respective scoring criteria.

To supplement rubric alignment, attention-based alignment is used. This allows the model to locate salient essay attributes and determine their alignment with dimension-specific rubric descriptors, for example, as described for Task Response or Lexical Resource. Through the application of context-aware attention throughout the essay, the model selectively attends to linguistically and structurally pertinent items, modifying trait scores according to human rater expectations. This alignment method guarantees that each scoring generated is data-driven as well as explicitly rooted in pre-established evaluation rubrics, making subsequent rationale for scores interpretable as well as pedagogically relevant.

By incorporating CoT-based structured reasoning with attention-based scoring, this methodology greatly improves the reliability and interpretability of IELTS essay scoring by automating the structured evaluation. This structured assessment provides a basis for incorporating logical reasoning into the scoring pipeline, as discussed in detail next.

3.2.2. Implementation of CoT in the Scoring Model

A multi-step logical reasoning framework is utilized to deploy CoT within an AES scheme that incorporates a transformer-based CoT decoder alongside QWEN. It adopts a pipeline architecture with a number of crucial elements:

- Feature Extraction Layer: Fine-tuned to extract linguistic features from essays, highlighting key elements within the four IELTS scoring dimensions.
- Rubric Alignment Layer: This layer applies attention-based matching between extracted features and IELTS rubrics, consistently applying evaluation criteria.

- Logical Reasoning Layer: The CoT reasoning framework is used to infer step-by-step justifications for each assigned score, making the scoring process transparent and interpretable.
- Feedback Generation Module: The system generates personalized, structured feedback using fine-tuned reinforcement learning strategies.

Mathematically, the rubric alignment score S_d for dimension d is computed as:

$$S_d = \sigma(W_d \cdot \text{Attention}(Q_d, K_d, V_d)) \tag{2}$$

Where W_d represents task-specific learnable parameters, and Q_d, K_d, V_d are query, key, and value representations used in the attention mechanism for rubric alignment.

To further refine feedback quality, this study incorporates PPO, a reinforcement learning strategy, to enhance feedback alignment with human expert evaluations. Given a set of expert-annotated reference feedback samples F^*, the reward function maximizes alignment between model-generated feedback F and human references:

$$R(F, F^*) = \text{cosine}(\text{Embed}(F), \text{Embed}(F^*)) \tag{3}$$

By iteratively fine-tuning the model with PPO updates, the system refines its ability to generate concise, informative, and personalized feedback aligned with expert assessments.

Together, these components establish a foundation for reliable scoring and personalized feedback generation in AES systems.

3.2.3. Algorithmic Implementation of the CoT Framework

The CoT framework logically structures the essay feedback and scoring process, ensuring systematic evaluation aligned with IELTS scoring standards. Algorithm 1 outlines the logical steps for CoT-based structured scoring and feedback generation.

Algorithm 1 Enhanced CoT-Based Essay Feedback and Scoring using QWen LLM

```
Input: IELTS essay text
Output: Feedback, Scores, and Confidence Levels for TR, CC, LR, and GRA
Initialize QWen LLM model: QWen_LLM
for each essay do
    Step 1: Preprocess essay text (tokenization, normalization)
    Step 2: Generate detailed feedback for each dimension (TR, CC, LR, GRA)
    for each dimension in {TR, CC, LR, GRA} do
        Extract feedback from generated text for the specific dimension
    end for
    Step 3: Compare feedback with Band Descriptors
    for each dimension in {TR, CC, LR, GRA} do
        Compute weighted similarity with band descriptors
        Assign score based on highest similarity
        Compute confidence score
        if confidence score < threshold then
            Re-evaluate with an alternative prompt
        end if
    end for
    Step 4: Output final scores and store reasoning steps
    Store final scores, confidence levels, and reasoning logs
end for
```

This algorithmic methodology provides a systematic and open evaluation procedure. By breaking down scoring and feed-forward generation into explicit steps, the CoT approach improves the reliability and explainability of AES models. Empirical results show that the CoT-augmented scoring system surpasses that of conventional AES models with regards to scoring reliability and feed-forward quality by achieving superior QWK and BLEU scores as well as superior human assessment ratings. Merging logical reasoning, rubric alignment, as well as reinforcement learning methods assures that feed-forward from automatic scoring resembles that of human expert assessments, finally enhancing the educational value of automatic essay scoring methods.

3.3. Evaluation Metrics

To evaluate the relevance and quality of the provided feedback, the current work utilized a blend of lexical overlap as well as semantic similarity measures. It chose these measures due to their effectiveness as tested for natural language generation tasks as well as their applicability to measuring the quality of educational feedback.

Specifically, three metrics were adopted:

- BLEU [28]: Measures n-gram precision between generated and reference feedback, capturing lexical fluency and local accuracy.
- ROUGE-L [29]: Evaluates the longest common subsequence between texts, reflecting content recall and structural consistency.
- Cosine Similarity: Computes the semantic similarity between feedback embeddings.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{4}$$

where **A** and **B** denote the embedding vectors of the model-generated feedback and the human expert feedback, respectively. The numerator $\mathbf{A} \cdot \mathbf{B}$ represents the dot product between the two vectors, while the denominator $\|\mathbf{A}\| \|\mathbf{B}\|$ is the product of their Euclidean norms.

The incorporation of these measures facilitates the holistic assessment that captures correspondence as well as semantic alignment at a deeper level between model-provided feedback and human references. Such strategies of evaluation align perfectly with the goals of rubric-based and pedagogically rich feedback generation as pursued through this research.

Aside from automatic scoring, a human evaluation procedure was performed. IELTS specialists rated the feedback according to a five-point Likert scale of relevance, specificity, and clarity as common practices for subjective quality assessment [30]. Additionally, pairwise preference ratings were gathered for identification of which versions of the feedback were rated as more educationally beneficial, using comparative evaluation techniques as found in standard human alignment tasks for feedback [31]. This qualitative assessment serves as a supplement to automatic measures, offering a holistic quality estimation of the feedback.

4. Experimental Settings and Results

This section presents the experimental configuration and key findings related to the evaluation of the proposed AES framework. It includes implementation details, model comparison results, and both automatic and human evaluation outcomes. The experiments are designed to assess trait-level scoring accuracy, feedback quality, and the impact of key components such as CoT reasoning, PPO, and model scaling. Each subsection corresponds to a distinct experimental focus, contributing to a comprehensive understanding of the system's effectiveness.

4.1. Experimental Environment and Parameter Settings

The QWEN-based models used in this study were implemented using the PyTorch 2.0.1 framework, with HuggingFace Transformers and PEFT libraries to support parameter-efficient tuning. All experiments were conducted on Nvidia A800 GPUs, each equipped with 80GB of VRAM. For all fine-tuning and inference tasks, we adopted a batch size of 16 due to the model size constraints, and used the AdamW optimizer with an initial learning rate of 2×10^{-5}, decaying by a factor of 0.1 upon encountering a plateau in validation loss. An early stopping strategy based on validation loss was adopted to retain optimal model weights, with a maximum training limit of 50 epochs.

doi:10.20944/preprints202504.2338.v1

The QWEN model backbone employed for scoring and feedback generation consists of 32 transformer layers with a hidden size $H = 4096$, 32 attention heads, and a feed-forward network with an intermediate size of 11008. The maximum input token length was set to 512 tokens for scoring tasks and 1024 tokens for feedback generation, while the output sequences were truncated to 150 tokens for concise feedback outputs.

For reinforcement learning fine-tuning, the PPO algorithm was adopted. The clipping parameter was set to $\epsilon = 0.2$, the advantage estimation parameter to $\lambda = 0.95$, and the discount factor to $\gamma = 0.99$. Separate learning rates were used for the policy and value networks, set to 1×10^{-5} and 5×10^{-5}, respectively. Reward signals were computed using a cosine similarity-based reward model aligned with rubric-based feedback quality.

All experimental conditions were executed over 10 independent runs, with average performance metrics reported to ensure robustness. For baseline models, including LLaMA-3-based scoring systems, hyperparameters were selected according to their official open-source configurations, with necessary adjustments made to ensure consistency in input length and evaluation criteria across all settings.

4.2. Trait-Level Scoring Performance with the CoT Framework

For the trait-level scoring task incorporating the CoT framework, the model outputs dimension-specific predictions across the four IELTS traits: TR, CC, LR and GRA. Quadratic Weighted Kappa (QWK) was used as the primary metric to assess alignment with human annotations, while Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) quantified the trait-level prediction deviation. These complementary metrics enable fine-grained evaluation of scoring consistency and accuracy.

As visualized in Figure 3, the original CoT-based model demonstrated superior performance, particularly in CC (QWK = 0.6814, MAE = 0.2600, RMSE = 0.5099) and GRA (QWK = 0.5640). These results underscore the advantage of incorporating explicit reasoning steps, which appear especially beneficial for traits requiring discourse-level understanding and syntactic complexity.

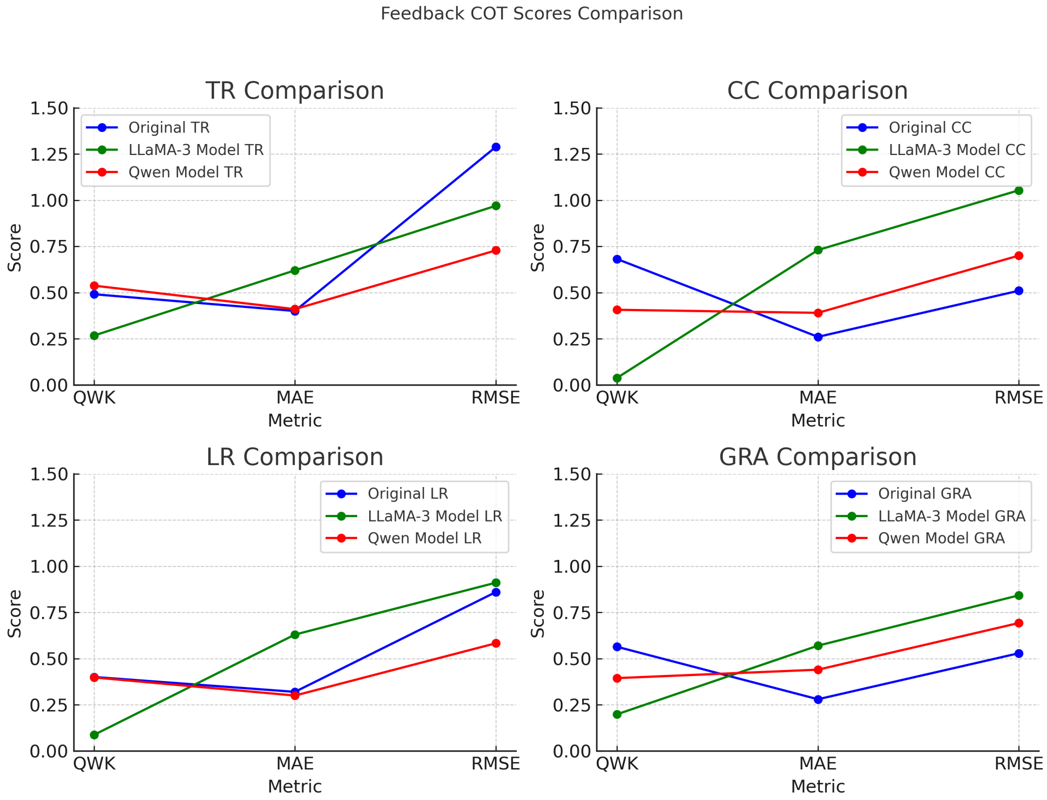


Figure 3. Comparison of Model Performance Across Scoring Dimensions Using CoT Framework.

In contrast, the LLaMA-3 baseline without CoT reasoning yielded the weakest performance across all traits. Notably, the QWK scores for CC and LR dropped to 0.0378 and 0.0871, respectively, and all trait-level MAE and RMSE values increased significantly. These results demonstrate the limitations of end-to-end holistic prediction without structured trait-specific inference.

The Qwen variant incorporating both CoT and rubric-alignment attention exhibited trait-specific improvements. It achieved the highest QWK in TR (0.5372) and the lowest MAE and RMSE in LR (0.3000 and 0.5831, respectively), while CC and GRA showed only moderate gains. This suggests that rubric-guided attention mechanisms may be more effective for certain traits, such as lexical quality, than for others.

To examine whether larger models yield better trait-level predictions, Qwen-7B, 14B, and 72B were compared under identical CoT and rubric-alignment settings. As shown in Table 2, Qwen-72B achieved slightly higher scores, but improvements were marginal. This suggests that in data-constrained scenarios, model scale alone does not guarantee significant gains, highlighting the importance of annotated data quality and task-specific alignment.

Table 2. Comparison of Different Qwen Model Versions in Trait-Level Scoring Performance

Model Variant	Avg. QWK	Avg. MAE	Avg. RMSE
Qwen-7B (CoT only)	0.5253	0.3750	0.7851
Qwen-14B (CoT + Rubric Attention)	0.5372	0.3602	0.7726
Qwen-72B (CoT + Rubric Attention)	0.5410	0.3561	0.7685

These findings collectively validate the effectiveness of CoT reasoning in improving trait-level scoring performance. In particular, CoT-enhanced models showed strong alignment with human ratings in coherence and grammar-related traits. While rubric alignment and model scaling provide incremental benefits, the most substantial improvements stem from incorporating explicit, structured reasoning aligned with scoring rubrics.

4.3. Feedback Generation Performance: Impact of PPO and CoT

To assess the contributions of PPO and CoT reasoning in feedback generation, ablation experiments were conducted. The baseline system lacked both PPO and CoT mechanisms, while other configurations introduced either or both techniques. As shown in Table 3, the combination of PPO and CoT achieved the highest scores across all automatic metrics: BLEU, ROUGE-L, and embedding-based cosine similarity. PPO alone contributed to alignment with human preferences, while CoT enhanced the linguistic richness and trait-specific clarity of the feedback—particularly in coherence and grammar-related aspects.

Table 3. Impact of PPO and CoT on Feedback Quality

Model Variant	BLEU	ROUGE-L	Cosine Similarity
Baseline (no PPO, no CoT)	0.2371	0.4012	0.6125
+ PPO only	0.2698	0.4265	0.6480
+ CoT only	0.2535	0.4321	0.6601
+ PPO + CoT	**0.2873**	**0.4588**	**0.6856**

Beyond automatic evaluation, a human assessment was conducted to evaluate the educational usefulness and clarity of the generated feedback. IELTS instructors rated feedback using a five-point Likert scale and provided pairwise preference judgments between model outputs. As visualized in Figure 4, both Likert scores and preference rates increased consistently from the baseline to the PPO+CoT configuration, highlighting the pedagogical value of combining structured reasoning with reinforcement optimization.

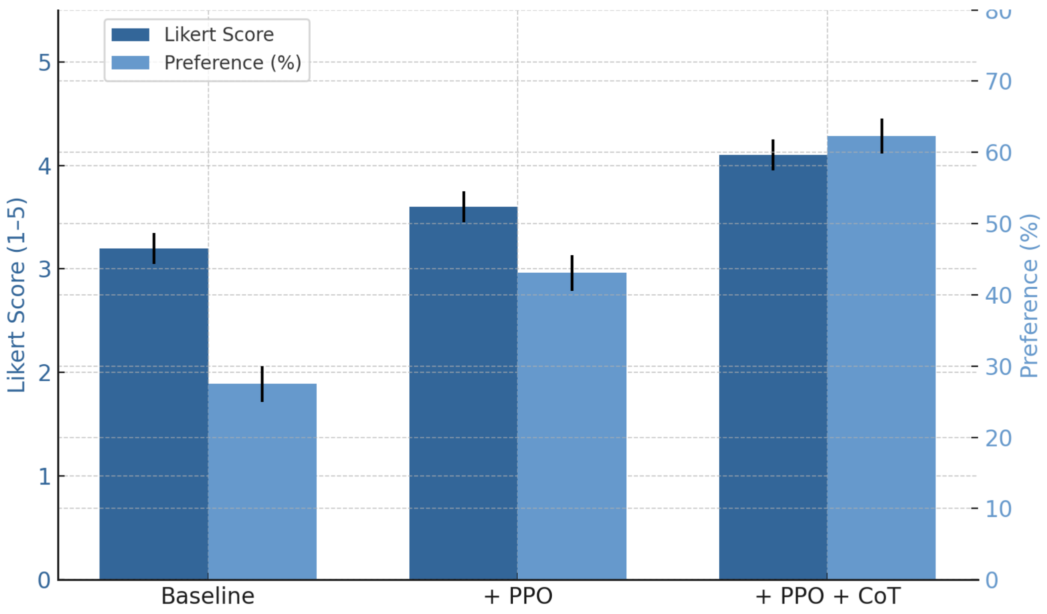


Figure 4. Human evaluation of feedback quality across model variants. The left axis shows the average Likert score (1–5), and the right axis shows the preference percentage. Error bars represent 95% confidence intervals.

These results collectively suggest that integrating reinforcement learning and CoT reasoning significantly enhances both the objective quality and subjective helpfulness of generated feedback, contributing to more interpretable and learner-aligned AES outputs.

4.4. Effect of Qwen Model Size on Feedback Generation

To further investigate the impact of model capacity on feedback generation quality, additional experiments were conducted using three versions of the Qwen model family: Qwen-7B, Qwen-14B, and Qwen-72B. All models were trained using the same CoT prompting and PPO fine-tuning framework, under identical data and parameter settings, to isolate the effect of model size.

As shown in Table 4, increasing model size resulted in only marginal improvements across BLEU, ROUGE-L, and cosine similarity scores. The largest model, Qwen-72B, achieved only a 2.5% improvement in BLEU and a 0.6% improvement in cosine similarity compared to Qwen-7B. This performance plateau suggests that under limited data conditions, the benefits of scaling up model size are constrained.

Table 4. Feedback Generation Quality Across Qwen Model Versions

Model Variant	BLEU	ROUGE-L	Cosine Similarity
Qwen-7B (CoT + PPO)	0.2873	0.4588	0.6856
Qwen-14B (CoT + PPO)	0.2931	0.4622	0.6880
Qwen-72B (CoT + PPO)	0.2945	0.4630	0.6891

This observation aligns with prior studies on model scaling laws, which emphasize the interdependence between model size, data availability, and compute resources [32,33]. In the context of automated feedback generation for AES, where high-quality annotated feedback is scarce and costly to obtain, increasing model capacity alone does not yield substantial quality gains.

These findings reinforce the importance of investing in high-quality, diverse, and representative training datasets for effective model scaling in educational NLP applications. Particularly in specialized tasks such as IELTS feedback generation, data-centric approaches may offer more practical and interpretable improvements than scaling parameters alone.

5. Discussion

This section synthesizes key findings from the experimental evaluation of QWENScore+, focusing on its interpretability, feedback quality, and architectural contributions. The discourse presents how the application of CoT reasoning, parameter-efficient fine-tuning, and reinforcement learning improves the performance as well as the understandability of automatic essay scoring techniques. This section does not confine experimental conditions but relates theoretical insights to educational application implications.

5.1. CoT-Based Reasoning and Interpretability

One of the fundamental design features of QWENScore+ is its application of CoT reasoning. In contrast with direct regression models, CoT architecture informs the model to proceed with structured steps: essay content analysis, generating rubric-aligned feedback, and logical inference of trait-specific scores. This modular design improves scoring accuracy as well as interpretability [19,20].

By using CoT, the model can produce middle-out rationales tracing the decision-making progress across dimensions. Such rationales are open explanations, mapping the model's output to people's assessment logic for enhancing explainability, usability, as well as faith with pedagogical applications.

Interestingly, CoT models performed the best in CC and TR—dimensions most dependent on discourse structure as well as content relevance—indicating that interpretability as well as accuracy can be attained together. Additionally, rubric-sensitive mechanisms within the Qwen variant even anchored trait scoring more firmly to semantically pertinent cues.

5.2. Feedback Structuring via CoT and Rubric Alignment

Beyond quantitative gains, qualitative gains were also seen within the structure of the feedback with the inclusion of the CoT framework [34]. By making the model reason step-by-step, from essay strength and weakness identification to mapping against trait-specific criteria, CoT provided a more logical, pedagogically consistent structure to the generated feedback.

This structured rationale method greatly enhanced the interpretability of the feedback, especially for educational purposes where precision and rubric consistency are paramount. Not only was the resulting feedback more organized but also more consistent with the IELTS assessment framework, as indicated by enhanced ROUGE-L and Cosine Similarity scores.

Additionally, CoT and PPO had complementary effects: CoT organized the internal reasoning flow structure, while PPO fine-tuned the output for alignment with human preferences. Combined, these elements improved the objective quality as well as subjective acceptability of the generated feedback, as attested by human judgment.

5.3. Interpretability through Structured Reasoning

The QWENScore+ architecture uses two complementary training approaches of SFT and RLHF supplemented with a CoT reasoning mechanism. It is designed to improve the explainability and pedagogical calibration of the LLMs when they are applied for automatic scoring of essays as well as generating feedback.

SFT operates as a localized strategy of optimization targeting the four IELTS core scoring attributes: TR, CC, LR and GRA. It learns from exemplar data annotated by experts and gains task-specific knowledge. It provides feedback that is compliant with the professional rater's expectations.

The addition of CoT further improves explainability by breaking up the scoring into intermediary steps of reasoning. Instead of generating a final overall score directly, the model is directed to initially produce dimension-level feedback and subsequently project these explanations onto rubric-mapped scores. This explicit inference path benefits traceability as well as alignment with human assessors' thought processes.

Furthermore, the application of RLHF, specifically through PPO, fine-tunes these steps of reasoning through reward-based feedback [35,36]. Human preferences are utilized to reinforce desirable output patterns, allowing the model to generate feedback that is not only accurate but also contextually appropriate, clear, and useful to learners.

The combined effect of CoT, RLHF, and SFT demonstrates that performance and interpretability are by no means incompatible. Instead, step-wise reasoning mechanisms are able to make scores more reliable while at the same time making it even more transparent, which is a crucial factor for building faith in computer-based learning systems.

Together, the results at the level of traits, generating feedback, and generalization show that the architecture of QWENScore+ gains from a holistic optimization approach. Instead of independent improvements, CoT reasoning, fine-tuning with LoRA, and RLHF reinforcement learning all work together synergistically at the level of modules to promote consistency, interpretability, as well as robustness.

5.4. Theoretical and Practical Implications

The emergence of QWENScore+ holds substantial theoretical as well as practical implications for the AES community. Architecturally, the model is a hybrid model that combines discriminative and generative modeling. Integrating encoder-type modules (as BERT for rubric alignment) with generative LLMs (such as Qwen) enables a single scoring-feedback pipeline capable of executing both numeric evaluation as well as text-based explanation tasks.

From a training perspective, the two-stage optimization approach—SFT followed by RLHF—provides a blueprint for aligning LLM behavior with human instructional goals [37,38]. SFT ensures rubric fidelity, while RLHF encourages adaptability and refinement through human-in-the-loop supervision. This approach aligns with new directions of human-aligned language model research and creates new opportunities for educational AI application.

Practically, the efficacy of the QWENScore+ model demonstrates the feasibility of rubric-sensitive, scalable scoring for IELTS-level high-stakes testing. It provides real-time trait-level ratings, reduces the weight of grading on teachers, as well as gives students individualized feedback based on formal criteria. Its ability to mirror the weighing of learner as well as expert expectations, as shown from scores from human evaluation, makes it more effective for classroom learning as well as online learning.

Nevertheless, challenges exist. They include input length constraints of scoring rubrics, consistency of training data, as well as difficulty in fully representing advanced rubric logic using general-purpose LLMs. Resolving these will involve creating more diverse training data sets, adaptive rubric encoding methods, as well as potentially hybrid models integrating symbolic and neural reasoning.

In summary, the QWENScore+ methodology advances the theoretical foundations of interpretable scoring and feedback modeling while providing a sound basis for practical application within educational contexts. Subsequent revisions must consider fairness, reducing bias, and equal access issues as well, particularly when applied to diverse student populations within real-world deployments.

6. Limitations and Future Work

While the proposed rubric-aligned CoT prompting framework demonstrates strong performance and interpretability, several limitations remain that warrant further research.

First, the generalization capability of the model across diverse prompts and rhetorical structures is still limited. CoT prompting improved rubric alignment but showed occasional hallucinations when handling underrepresented or atypical prompts [19,39]. Future work could expand prompt diversity and integrate discourse-aware training mechanisms to enhance robustness.

Second, despite using trained annotators, minor inconsistencies in human scoring introduced noise into the training data. Refining annotation protocols and incorporating automated adjudication strategies, such as GPT-assisted scoring review, may improve label reliability and model stability.

Third, the computational cost of RLHF remains high, posing challenges for deployment in resource-limited settings [40–42]. Research into efficient alternatives like QLoRA, AdapterFusion, or lightweight RL algorithms could make training more scalable and accessible.

Additionally, input length limitations in current architectures hinder the integration of full-length essays and rubrics within a single inference step. Hierarchical encoding or long-context transformer models may offer practical solutions to this constraint.

Finally, the potential for embedded biases in pretraining data necessitates ongoing fairness evaluation. Future developments should incorporate bias auditing pipelines, counterfactual data augmentation, and interpretability mechanisms to ensure equitable outcomes across learner populations [43,44].

7. Conclusions

This study presents a comparative evaluation of rubric-aligned CoT prompting across multiple open-source LLMs for AES. Using a high-quality, expert-annotated IELTS dataset, the study demonstrates that CoT prompting significantly improves scoring interpretability and feedback quality. Among the evaluated models, Qwen showed the best alignment with human scoring criteria and pedagogical feedback standards.

The key contributions of this research are threefold: (1) proposing a rubric-aware prompting strategy that enhances reasoning transparency in AES; (2) systematically comparing multiple LLMs to assess their suitability for CoT integration; and (3) validating the approach through both automatic metrics and human evaluation.

Overall, this framework offers a promising pathway toward developing more explainable, trustworthy, and pedagogically meaningful AES systems. Future research should focus on extending the method to multilingual and multimodal contexts, enabling adaptive feedback delivery, and improving fairness and scalability in real-world educational environments.

Author Contributions: Conceptualization, Wenbo Xu; methodology, Wenbo Xu; software, Wenbo Xu; validation, Wenbo Xu, Muhammad Shahreeza and Wai Lam Hoo; formal analysis, Wenbo Xu; investigation, Wenbo Xu; resources, Wenbo Xu; data curation, Wenbo Xu and Wudao Yang; writing—original draft preparation, Wenbo Xu; writing—review and editing, Muhammad Shahreeza; visualization, Wudao Yang; supervision, Muhammad Shahreeza; project administration, Wenbo Xu; funding acquisition, Wai Lam Hoo. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the UM Research Maintenance Fee under Grant RMF1489-2021 and UM Research University Faculty Grant (GPF009D-2019).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Universiti Malaya Research Ethics Committee (UMREC) of the University of Malaya (protocol code UM.TNC2/UMREC_3267, approved on 13 March 2024). This approval is valid from March 2024 to March 2027.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The source code, datasets, and supplementary materials are available at: https://github.com/Owenxu0409/LLMs-AES-IELTS.

Acknowledgments: We would like to express our gratitude to Mengjing Cai, chairman of Qingdao Northern Education Consulting Co., LTD, for providing 5,088 IELTS datasets with detailed expert annotations, which were instrumental in facilitating the progress of this project.

Conflicts of Interest: The authors declare no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

AES	Automated Essay Scoring
AI	Artificial Intelligence
ASAP	Automated Student Assessment Prize
CC	Coherence and Cohesion

CoT	Chain of Thought
GRA	Grammatical Range and Accuracy
IELTS	International English Language Testing System
LLMs	Large Language Models
LR	Lexical Resource
ML	Machine Learning
MAE	Mean Absolute Error
NLP	Natural Language Processing
PPO	Proximal Policy Optimization
QWK	Quadratic Weighted Kappa
RLHF	Reinforcement Learning from Human Feedback
RMSE	Root Mean Squared Error
SFT	Supervised Fine-tuning
TOEFL	Test of English as a Foreign Language
TR	Task Response
XAI	Explainable Artificial Intelligence

Appendix A. IELTS Writing Task 2 Rubrics—Band 5 Example

This appendix presents the detailed descriptors for Band 5 performance on the IELTS Writing Task 2 rubric, organized by scoring dimension.

Task Response (TR)

- Addresses the task only partially; the format may be inappropriate in places.
- Expresses a position but the development is not always clear and there may be no conclusions drawn.
- Presents some main ideas but these are limited and not sufficiently developed; there may be irrelevant detail.

Coherence and Cohesion (CC)

- Presents information with some organisation but there may be a lack of overall progression.
- Makes inadequate, inaccurate or over-use of cohesive devices.
- May be repetitive because of lack of referencing and substitution.
- May not always use paragraphs logically.

Lexical Resource (LR)

- Uses a limited range of vocabulary, but this is minimally adequate for the task.
- May make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader.

Grammatical Range and Accuracy (GRA)

- Uses only a limited range of structures.
- Attempts complex sentences but these tend to be less accurate than simple sentences.
- May make frequent grammatical errors and punctuation may be faulty; errors can cause some difficulty for the reader.

The complete band descriptors are available at the official IELTS website: https://ielts.org/organisations/ielts-for-organisations/ielts-scoring-in-detail

Appendix B. IELTS Dataset JSON Visualization

Table A1 presents an example of a data entry from the annotated IELTS dataset used in this study.

Table A1. IELTS Dataset Example

Essay ID	0001
Title	Some people claim that many things that children are taught at school are a waste of time. Other people argue that everything they study at school is useful at some time. Discuss both views and give your own opinion.
Essay	Consumerism seems impossible to escape for people living in our day and age. Now, more than ever, advertisers attempt to bombard all of our waking moments with persuasive sales pitches... In general, the more and more pervasive advertisements are depriving consumers of their free will and abducting their power of choice...
Scores	**TR**: 4, **CC**: 4, **LR**: 5, **GRA**: 5, **Overall Score**: 4.5
Annotations	**TR Positives**: Addressing the Prompt: The essay discusses both views on school education... **TR Negatives**: Depth of Analysis: The essay does not fully develop the arguments for both sides... **CC Positives**: Logical Flow: The essay attempts to maintain a logical flow with some connecting phrases... **CC Negatives**: Transitions between sentences are often awkward or missing... **LR Positives**: Range of Vocabulary: The essay demonstrates an attempt to use a range of vocabulary... **LR Negatives**: Word Choice: Some word choices are inaccurate and awkward... **GRA Positives**: Sentence Structure Variety: The essay attempts a variety of sentence structures... **GRA Negatives**: Frequent Grammatical Errors: The essay contains numerous grammatical errors...
Suggestions	**Introduction**: The introduction should be clearer and more concise... **Body Paragraph 1**: Improve the logical structure and clarity of sentences... **Body Paragraph 2**: Use clear examples and logical transitions to support your points... **Body Paragraph 3**: Discuss the counterargument with clear examples... **Conclusion**: Strengthen the conclusion by summarizing key points more effectively...

Appendix C. Prompting Instructions for LLMs

GPT-4 Zero-Shot Scoring Instructions for IELTS Writing

This Excel file contains the following columns:

- **essay_id**: A unique identifier for each essay.
- **essay_prompt**: The prompt to which the essay responds.
- **essay**: The actual content of the essay.

The task is to randomly select 1,000 records from this file for scoring. Please adhere to the following guidelines:

- TR, CC, LR, and GRA scores must be integers ranging from 0 to 9.
- The Overall score should be the average of these four scores, rounded to the nearest 0.5 (for example, 6.25 should be rounded to 6.5).
- The scoring should consider both the **essay_prompt** and the **essay** content.
- Please save the **essay_id** and the original scores for the randomly selected essays, as they will be needed for comparative analysis later.

GPT-4 Few-Shot Scoring Instructions for IELTS Writing

In a few-shot learning setting, please follow these additional guidelines along with the zero-shot scoring instructions:

- Provide exemplar essays with corresponding band scores for TR, CC, LR, and GRA. These exemplars should include:
  - **Low-Band (0-4)**: Essays that have significant issues in addressing the task, coherence, vocabulary, and grammar, with frequent errors.
  - **Mid-Band (5-6)**: Essays that meet basic task requirements but have some issues with idea development, coherence, and occasional errors in vocabulary and grammar.
  - **High-Band (7-9)**: Essays that fully address the task, are well-organized, use a varied vocabulary, and have minimal grammatical errors.
- Ensure exemplars cover a range of prompts and include concise explanations for the assigned scores.

Task: Grade the following essay using the *IELTS Writing Task 2 Band Descriptors* and the full band range (0-9).

- **Provide an explanation for your score**: [Insert Score and Explanation]
- **Essay to be scored**: [Insert Essay Text]

LLaMA-3 Zero-Shot Scoring Instructions for IELTS Writing

This instruction set is intended for a fine-tuned LLaMA-3 model to perform automated scoring on IELTS Writing Task 2 essays in a zero-shot setting. The input data is structured with the following fields:

- **essay_id**: A unique identifier assigned to each essay.
- **essay_prompt**: The writing task prompt provided to the candidate.
- **essay**: The complete written response from the candidate.

Please follow the guidelines below when assigning scores:

- Assign integer band scores from 0 to 9 for the four IELTS scoring criteria: Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA).
- Compute the **Overall score** as the arithmetic mean of the four trait scores and round to the nearest 0.5 increment (e.g., 6.25 → 6.5).
- All scores must be inferred based on both the prompt and the essay content.
- Store the predicted scores along with the corresponding **essay_id** for further evaluation and comparison.

Qwen Fine-Tune Scoring Instructions for IELTS Writing

For fine-tuning the latest version of the Qwen model on IELTS Writing Task 2 scoring, the following instructions should be strictly followed:

- **Task Separation for Scoring and Feedback:** Fine-tune the model to treat scoring and feedback generation as distinct subtasks. This modular training structure enables Qwen to optimize for accurate evaluation and meaningful feedback independently.
- **Trait-Level Scoring and Feedback:** Configure the model to process each of the four IELTS scoring criteria—Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA)—independently. Each criterion should receive both a score and detailed feedback aligned with the IELTS rubric.
- **Scoring Format (Discrete Band Selection):** When training for scoring prediction, adopt a classification-style format with predefined options for each trait. For example:

  - **TR:** Select the most appropriate band score for the essay's task response:
    - A. Band 5
    - B. Band 6
    - C. Band 7
    - D. Band 8
  - Apply a similar discrete choice format for CC, LR, and GRA.
- **Feedback Prompting Strategy:** After predicting the score for each dimension, the model should generate feedback explicitly aligned with that score level. For instance, if Band 6 is predicted for TR, the feedback must explain what aspects justify Band 6 and what improvements are necessary to reach Band 7 or above.

Task Objective: Once the model is fully fine-tuned, perform the following on each input essay:

- **Scoring:** Assign a band score (0–9) for TR, CC, LR, and GRA based on the discrete band options.
- **Feedback Generation:** Produce tailored feedback for each scoring dimension, grounded in the rubric descriptors and the essay's performance.
- **Essay to be scored:** [Insert Essay Text]
- **Provide scores, explanations, and feedback:**
  - **TR:** [Insert Score], [Insert Explanation], [Insert Feedback]
  - **CC:** [Insert Score], [Insert Explanation], [Insert Feedback]
  - **LR:** [Insert Score], [Insert Explanation], [Insert Feedback]
  - **GRA:** [Insert Score], [Insert Explanation], [Insert Feedback]

References

1. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
2. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x.
3. Shute, V.J. Focus on Formative Feedback. *Review of Educational Research* **2008**, *78*, 153–189. https://doi.org/10.3102/0034654307313795.
4. Andrade, H.G. Using Rubrics to Promote Thinking and Learning. *Educational Leadership* **2000**, *57*, 13–18.
5. Uto, M. A review of deep–neural automated essay scoring models. *Behaviormetrika* **2021**, *48*. https://doi.org/10.1007/s41237-021-00142-y.
6. Krumsvik, R.J. GPT-4's capabilities in handling essay-based exams in Norwegian: an intrinsic case study from the early phase of intervention. *Frontiers in Education* **2025**, *10*. https://doi.org/10.3389/feduc.2025.1444544.
7. Li, W.; Liu, H. Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications* **2024**, *11*. https://doi.org/10.1057/s41599-024-03209-9.
8. Chiang, C.H.; Chen, W.C.; Kuan, C.Y.; Yang, C.; Lee, H.y. Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. arXiv preprint arXiv:2407.05216, 2024.
9. Shermis, M.D.; Burstein, J. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*; Routledge: New York, 2013.
10. Kumar, V.S.; Boulanger, D. Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education* **2020**, *31*. https://doi.org/10.1007/s40593-020-00211-5.
11. Hong, S.; Cai, C.; Du, S.; Feng, H.; Liu, S.; Fan, X. "My Grade is Wrong!" A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays. arXiv preprint arXiv:2409.07453, 2024.
12. Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; Kedzie, C. LLM-R UBRIC: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 13806–13834.
13. Fagbohun, O.; Iduwe, N.P.; Abdullahi, M.; Ifaturoti, A.; Nwanna, O.M. Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *J Artif Intell Mach Learn & Data Sci* **2024**, *2*. https://doi.org/10.51219/JAIMLD/oluwole-fagbohun/19.

14. Lipnevich, A.A.; McCallen, L.N.; Miles, K.P.; Smith, J.K. Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science* **2014**, *42*. https://doi.org/10.1007/s11251-013-9299-9.
15. Zhang, H.; Litman, D. Co-Attention Based Neural Network for Source-Dependent Essay Scoring. *arXiv preprint arXiv:1908.01993* **2019**.
16. Alves da Silva, W.; Coelho de Araujo, C. Automated ENEM Essay Scoring and Feedbacks: A Prompt-Driven LLM Approach. arXiv preprint, 2023.
17. Cao, Y.; Yao, S.; Zhao, J.; Yu, D.; Shafran, I.; Narasimhan, K.; Griffiths, T.L. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *NeurIPS* **2024**.
18. Henkel, O.; Horne-Robinson, H.; Dyshel, M.; Ch, N.; Moreau-Pernet, B.; Abood, R. Learning to Love Edge Cases in Formative Math Assessment: Using the AMMORE Dataset and Chain-of-Thought Prompting to Improve Grading Accuracy. arXiv preprint arXiv:2409.17904, 2023.
19. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* **2022**.
20. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916* **2022**.
21. Zhang, Z.; Zhang, A.; Li, M.; Smola, A. Automatic Chain of Thought Prompting in Large Language Models. arXiv preprint arXiv:2210.03493, 2022.
22. Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.; et al. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
23. Cohn, C.; Hutchins, N.; T, A.; Biswas, G. CoTAL: Human-in-the-Loop Prompt Engineering, Chain-of-Thought Reasoning, and Active Learning for Generalizable Formative Assessment Scoring. *IEEE Transactions on Learning Technologies* **2025**.
24. Chu, S.Y.; Kim, J.W.; Wong, B.; Yi, M.Y. Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs. arXiv preprint arXiv:2410.14202, 2025.
25. Diao, S.; Wang, P.; Lin, Y.; Pan, R.; Liu, X.; Zhang, T. Active Prompting with Chain-of-Thought for Large Language Models. arXiv preprint arXiv:2302.12246, 2024.
26. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. REACT: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
27. Attali, Y.; Burstein, J. Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning, and Assessment* **2006**, *4*.
28. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318.
29. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, 2004, pp. 74–81.
30. Likert, R. A technique for the measurement of attitudes. *Archives of Psychology* **1932**, *22*, 5–55.
31. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P. Learning to summarize with human feedback. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 33, pp. 3008–3021.
32. Kaplan, J.; McCandlish, S.; Henighan, T.; et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* **2020**.
33. Hoffmann, J.; et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* **2022**.
34. Ustalov, D.; Panchenko, A.; Biemann, C. Chain of Thought Prompting Improves Educational Feedback Generation. *arXiv preprint arXiv:2303.12112* **2023**.
35. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Joseph, N.; Chen, A.; Goldie, A.; Mirhoseini, A.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* **2023**.
36. Zhou, Z.; Xie, T.; Liu, Z.; Zhang, W.; Tang, J. A Survey of Alignment Techniques for Large Language Models. *arXiv preprint arXiv:2307.11057* **2023**.
37. OpenAI. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774, 2024. arXiv preprint arXiv:2303.08774.
38. Pan, Y.; Ma, J.; Tang, S.; Gao, Y.; Liu, Y. Recent Advances in Human-Aligned Language Models: From Pre-training to Reinforcement. *arXiv preprint arXiv:2312.00531* **2023**.

39. Ji, Z.; Lee, N.; Fries, J.; Li, T.; Tan, Z.; Zhang, B.; Jaggi, M.; Anandkumar, A.; Goldwasser, D.; Wang, Y. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)* **2023**, *55*, 1–38.
40. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* **2022**.
41. Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2022.
42. Yang, Y.; Tao, C.; Fan, X. LoRA-LiteE: A Computationally Efficient Framework for Chatbot Preference-Tuning. *arXiv preprint arXiv:2411.09947* **2024**.
43. Chinta, S.V.; Wang, Z.; Yin, Z.; Hoang, N.; Gonzalez, M.; Le Quy, T.; Zhang, W. FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications. *arXiv preprint arXiv:2407.18745* **2024**.
44. Lacmanovic, S.; Skare, M. Artificial intelligence bias auditing – current approaches, challenges and lessons from practice. *Review of Accounting and Finance* **2025**, *24*, 63–89.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.