

Think Before You Score: Pedagogical Chain-of-Thought Annotation and Reinforcement Learning for Low-Resource Multilingual Essay Evaluation

Wei Tian^{1,*} Yuhao Zhou^{1,*} Shenwei Li¹ Xiayu Sun¹ Man Lan^{1,2,†}

¹School of Computer Science and Technology, East China Normal University

²Shanghai Institute of Artificial Intelligence for Education, East China Normal University

tianwei@stu.ecnu.edu.cn, mlan@cs.ecnu.edu.cn

*Equal contribution. †Corresponding author.

Abstract

Automated Essay Scoring (AES) in low-resource, non-English settings faces a compounded challenge: existing corpora record only a final holistic score, discarding the intermediate reasoning human examiners rely on. This **annotation gap** drives **Rubric Drift**—the tendency of LLMs to apply Anglocentric grading norms onto localized curricula. We introduce **MPEAB-2430**, a multilingual benchmark spanning Malay, Hindi, and Arabic curricula, annotated with **Pedagogical Chain-of-Thought (P-CoT)** trajectories capturing sub-dimensional analyses, explicit error attribution, and curriculum-conformant feedback from language-specific experts. Leveraging this modest-scale supervision, we propose **RL-CRAF++**, combining CoT-SFT with Constrained GRPO whose rewards—including a novel **ordinal ranking consistency** signal targeting the QWK objective—are derived entirely from human annotations, bypassing external LLM judges. A 4B RL-CRAF++ model consistently surpasses 3-shot frontier models (GPT-5, DeepSeek-V3.2), with P-CoT supervision accounting for $\sim 66\%$ of alignment gains and the redesigned RL stage contributing $\sim 34\%$, yielding near-zero calibration bias and actionable feedback confirmed by a downstream revision experiment. Code and data: <https://github.com/TW-NLP/MPEAB>.

1 Introduction

Automated Essay Scoring (AES) models trained on English-centric corpora often fail to generalize to low-resource, non-Western curricula (Singh et al., 2023). A likely contributing factor is a fundamental **Annotation Gap** in the standard dataset paradigm. As illustrated in Figure 1, while human evaluation is a sequential, multi-dimensional cognitive process, traditional AES corpora collapse this workflow into a single holistic scalar. This loss of

intermediate pedagogical reasoning may encourage models toward shortcut learning, potentially leading to **Rubric Drift**—the tendency to apply Anglocentric grading norms onto localized educational content (Li and Liu, 2024).

To address this gap, supervision should capture the *process* of evaluation, not just the *outcome*. We introduce **MPEAB-2430**, a multilingual AES benchmark spanning Malay, Hindi, and Arabic national curricula. MPEAB-2430 formalizes the examiner’s workflow through **Pedagogical Chain-of-Thought (P-CoT)** annotations: sub-dimensional rationales, curriculum-specific error taxonomies, and pedagogical feedback. By enforcing *ontological isomorphism* across languages, MPEAB-2430 aims to prevent cross-cultural rubric contamination while preserving localized pedagogical standards.

Exploiting such high-fidelity but modest-scale data (~ 550 training essays per language) requires extreme sample efficiency. We propose **RL-CRAF++**, a two-stage alignment pipeline. First, **CoT-SFT** employs behavioral cloning on P-CoT trajectories to encourage the model to articulate reasoning *before* scoring. Second, **Constrained GRPO** performs multi-objective reinforcement learning using rewards derived *exclusively* from human annotations. Crucially, we introduce an **ordinal ranking consistency reward** (R_{rank}) that directly aligns the optimization objective with QWK’s rank-agreement semantics, substantially amplifying the RL stage’s contribution beyond calibration alone.

Our experiments demonstrate that a compact 4B parameter RL-CRAF++ model consistently outperforms zero-shot and 3-shot frontier models, including GPT-5 and DeepSeek-V3.2, when trained in-domain on ~ 550 essays per language. Ablations reveal that P-CoT drives $\sim 66\%$ of the total alignment gain over plain SFT, with the RL stage contributing a complementary $\sim 34\%$ —substantially higher than achieved by prior reward designs. We

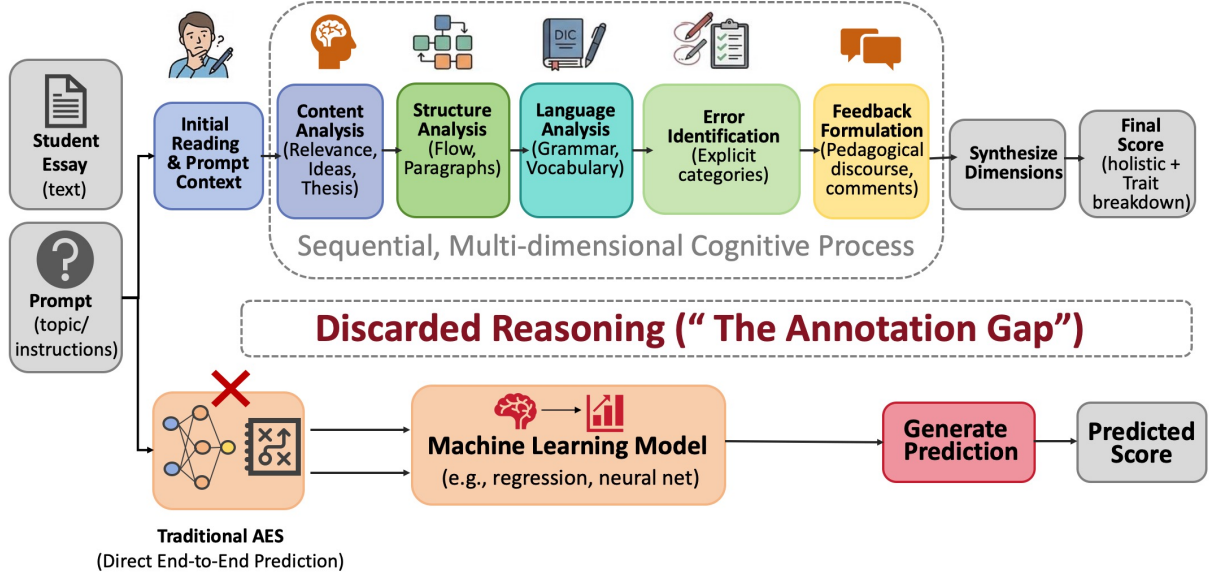


Figure 1: The **Annotation Gap** in Automated Essay Scoring. Traditional AES models (bottom) collapse the sequential, multi-dimensional cognitive workflow of human examiners (top) into a direct scalar prediction, discarding intermediate pedagogical reasoning.

further demonstrate pedagogical value via a downstream revision experiment: essays revised using RL-CRAF++ feedback improve by +6.5 points on average, versus +4.6 for CoT-SFT, confirming that the RL stage instills actionable, curriculum-conformant guidance that process-SFT alone cannot provide.

Our primary contributions are:

- **MPEAB-2430:** A typologically diverse AES benchmark with full P-CoT trajectories for Malay, Hindi, and Arabic curricula, designed to expose the annotation gap in low-resource settings.
- **Process-Oriented Supervision:** A formal framework showing that unrolling the examiner’s reasoning trace substantially reduces shortcut learning, accounting for ~66% of total alignment gains over scalar-only SFT.
- **RL-CRAF++ Algorithm with Ordinal Ranking Reward:** A sample-efficient, judge-free RL pipeline whose novel R_{rank} signal directly optimizes QWK-aligned rank consistency, amplifying the RL contribution to ~34% of total gains while enabling curriculum-conformant feedback.
- **Downstream Revision Evidence:** An “LLM-as-a-Student” experiment demonstrating statistically significant score improvements from RL-CRAF++ feedback over CoT-SFT alone

($p < 0.001$, all languages).

2 The MPEAB-2430 Benchmark

2.1 Design Philosophy: From Outcome to Process

Traditional AES corpora record only the terminal node of an examiner’s cognitive process, discarding intermediate pedagogical reasoning. We hypothesize that this **annotation gap** encourages models toward Rubric Drift, as they may learn spurious surface correlates rather than curriculum-aligned evaluation mechanisms. MPEAB-2430 therefore operationalises AES not as scalar prediction, but as a structured cognitive trajectory. We formalise the Pedagogical Chain-of-Thought (P-CoT) as:

$$\mathcal{T}_{\text{P-CoT}} : x \rightarrow (\mathbf{c}, \mathbf{s}, \mathbf{l}, E, \mathbf{f}, y) \quad (1)$$

where $\mathbf{c}, \mathbf{s}, \mathbf{l}$ are rationale-score pairs for content, structure, and language; E is a curriculum-defined error set; $\mathbf{f}$ is pedagogical feedback; and y is the holistic scalar. This formulation enables dense CoT-SFT targets, deterministic judge-free rewards, and multi-faceted evaluation that decouples holistic accuracy from rationale alignment.

2.2 Corpus Construction and Typological Diversity

We deliberately selected three typologically and orthographically distinct languages—Malay (Austronesian), Hindi (Indo-Aryan), and Arabic

Lang.	Curriculum	Total	Train	Dev	Test
Malay	SPM/UEC	820	574	82	164
Hindi	CBSE	780	546	78	156
Arabic	MOE	830	581	83	166
Total	—	2,430	1,701	243	486

Table 1: MPEAB-2430 split sizes and curriculum provenance. Each language constitutes an isolated evaluation environment.

(Semitic)—to test evaluation alignment across diverse morphosyntactic landscapes. Essays were sourced from national examination centres with the consent of the participating schools and students and their guardians, and were fully anonymised. The three curricula (SPM/UEC, CBSE, MOE) are treated as strictly independent evaluation environments; we do not project cross-lingual standards.

Table 1 details the corpus splits. Full per-split score statistics and sub-dimension distributions are provided in Appendix F; cross-split grade distributions confirm minimal systematic leakage. All prompts are strictly disjoint across splits to evaluate prompt generalization within the target curriculum.

2.3 The P-CoT Annotation Protocol

Reliability through structural grounding. To ensure high-fidelity supervision, two native curriculum experts annotated each essay via a Language-Specific Annotation Protocol (LSAP). Discrepancies exceeding 10 holistic points, or any error-attribution conflicts, triggered formal adjudication by a senior examiner. Because raters were required to ground their scores in intermediate structural steps, MPEAB-2430 achieves high inter-rater reliability: holistic QWK reaches 0.86 (Malay), 0.83 (Hindi), and 0.81 (Arabic), with robust error-list agreement (Cohen’s $\kappa = 0.82$ –0.87).

The reasoning trajectory. The P-CoT annotation maps directly to an autoregressive sequence, encapsulated in `<think>...</think>` tags:

1. **Content Rationale & Score (C):** Discourse on relevance, elaboration, and maturity ([0, 40]).
2. **Structure Rationale & Score (S):** Discourse on paragraphing, cohesion, and logical flow ([0, 30]).
3. **Language Rationale & Score (L):** Morphosyntactic evaluation culminating in a definitive **[ERROR LIST]** bounded by the LSAP ontology ([0, 30]).

Layer	Content (abbreviated; <i>italics</i> = English gloss)
Essay excerpt	<i>“Hari jadi saya datang... Tapi mak bagi buku tulis... Moral: jangan sombong.”</i> [My birthday came... But mum gave me a notebook... Moral: don’t be arrogant.]
Content (C)	Isi hampir tiada; hanya ayat ringkas tanpa huraian. [Almost no content; only short sentences, no elaboration.] Score: 0/40
Structure (S)	9 ayat pendek; tiada perenggan. [9 short sentences; no paragraphing.] Score: 3/30
Language (L) + Error List	Bahasa basahan (<i>nak, je</i>); imbuhan meN- hilang. [Colloquial register; missing meN- prefix.] [ERROR LIST]: Kesilapan Imbuhan, Kesilapan Kata Sendi Nama, Kesilapan Bahasa Basahan, Kesilapan Subjek-Predikat. Score: 14/30
Teacher Comment	<i>Karangan ini terlalu pendek... [This essay is too short... Advice: start with sentence-writing practice.]</i>
Final Score	17/100

Table 2: Condensed P-CoT trajectory (Malay SPM/UEC, with English gloss). The model is trained to generate this entire cognitive sequence *before* emitting the final scalar.

4. **Teacher Comment:** Curriculum-conformant feedback in the specific pedagogical register.

5. **Holistic Aggregation:** $y = C + S + L$.

Table 2 demonstrates this cognitive unrolling. The **[ERROR LIST]** is bounded by the LSAP, which follows *ontological isomorphism*: nine abstract error slots shared across languages, each instantiated independently by native experts (full taxonomies in Appendix H).

3 Method: RL-CRAF++

As illustrated in Figure 2, RL-CRAF++ operationalises the P-CoT annotation layers through a two-stage training paradigm. The pipeline is designed to be sample-efficient, achieving convergence with ~550 essay budgets per language.

3.1 Stage 1: Pedagogical Rationale Elicitation (CoT-SFT)

Standard SFT on holistic scores is prone to *shortcut learning*. Stage 1 casts AES as a behavioral cloning task over the examiner’s cognitive sequence, formatting the P-CoT annotations as a structured autoregressive trajectory enclosed in `<think>` tags:

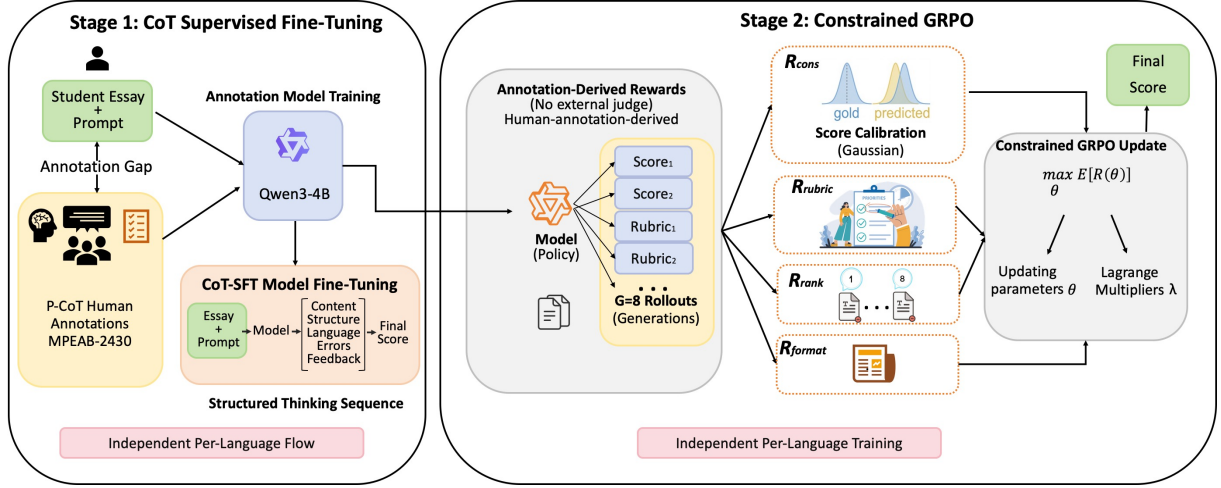


Figure 2: The RL-CRAF++ framework. **Stage 1 (CoT-SFT)** performs behavioral cloning on pedagogical reasoning traces. **Stage 2 (Constrained GRPO)** optimizes curriculum alignment via human-derived rewards including ordinal ranking consistency, eliminating the need for external LLM judges.

Autoregressive P-CoT Target

```
<think>
[Content Analysis] → C score
[Structure Analysis] → S score
[Language Analysis] → [ERROR LIST] →
L score
[Teacher Comment]
</think>
final_score (C + S + L)
```

By requiring the model to generate the full intermediate rationale before committing to a score, CoT-SFT provides dense, structured supervision that decomposes the complex scoring manifold into tractable sub-tasks, acting as an implicit regularizer that helps a 4B parameter model generalize from scarce data (Doi et al., 2024; Cohn et al., 2026).

3.2 Stage 2: Judge-Free Constrained Policy Optimization (GRPO)

While CoT-SFT instills the *form* of the evaluation procedure, cross-entropy loss is agnostic to the *magnitude* of scoring errors and cannot directly optimize rank agreement—the quantity that QWK measures. We address this via Constrained GRPO (Shao et al., 2024) with a redesigned reward suite.

Constrained Optimisation Objective. We maximize holistic score consistency while satisfying lower bounds on rubric attribution (τ_r) and ordinal ranking consistency (τ_k):

$$\begin{aligned} \max_{\theta} \quad & \mathbb{E}_{(x,r) \sim \mathcal{D}} [R_{\text{cons}}(y; x, r)] \\ \text{s.t.} \quad & \mathbb{E}[R_{\text{rubric-w}}(y)] \geq \tau_r = 0.73, \\ & \mathbb{E}[R_{\text{rank}}(y)] \geq \tau_k = 0.68 \end{aligned} \quad (2)$$

We solve this via Lagrangian relaxation with dual variables $\mu_r, \mu_k \geq 0$ updated via projected gradient ascent (step 0.01). For GRPO rollouts, the composite reward is:

$$R = \lambda_1 R_{\text{cons}} + \lambda_2 R_{\text{rubric-w}} + \lambda_3 R_{\text{rank}} + \lambda_4 R_{\text{format}} \quad (3)$$

with $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.40, 0.25, 0.25, 0.10)$, selected by grid search on MPEAB-2430 dev splits (Appendix J).

Human-Annotation-Derived Rewards. All rewards are computed *deterministically* against MPEAB-2430 P-CoT annotations, bypassing proprietary LLM judges. We note that while this design avoids English-centric LLM biases, human annotations inherently encode localized curriculum biases; our rewards are better understood as *curriculum-grounded* rather than bias-free.

1. Score Consistency (R_{cons}): A sharpened Gaussian ($\sigma = 2.0$) over the holistic residual, supplemented by sub-score alignment:

$$R_{\text{cons}} = 2.0 e^{-\frac{(s-s^*)^2}{8}} + 0.5 \sum_{d \in \{C, S, L\}} e^{-\frac{(\hat{s}_d - s_d^*)^2}{8}} \quad (4)$$

2. Severity-Weighted Rubric Attribution ($R_{\text{rubric-w}}$): An upgrade over binary Set-F1 that weights each LSAP error category by its curriculum-grounded severity tier $w_c \in \{1.0, 0.75, 0.50\}$ (morphosyntactic > lexical/semantic > orthographic):

$$R_{\text{rubric-w}} = \frac{\sum_c w_c \cdot \frac{2|E_c^{\text{pred}} \cap E_c^{\text{gold}}|}{|E_c^{\text{pred}}| + |E_c^{\text{gold}}|}}{\sum_c w_c} \quad (5)$$

Severity tiers are defined per LSAP ontology (Appendix H) and require no additional annotation.

3. Ordinal Ranking Consistency (R_{rank}): [New] QWK measures rank agreement, yet prior AES reward designs optimize pointwise error, creating an objective mismatch. We introduce a batch-level pairwise ranking reward that directly targets QWK’s ordinal semantics. For a training batch $\mathcal{B}$ and essay x_i with predicted score $\hat{s}_i$ and gold s_i^* :

$$R_{\text{rank}}(i) = \frac{1}{|\mathcal{B}| - 1} \sum_{j \neq i} \mathbf{1} \left[\begin{matrix} \text{sgn}(\hat{s}_i - \hat{s}_j) = \\ \text{sgn}(s_i^* - s_j^*) \end{matrix} \right] \quad (6)$$

This pairwise ranking accuracy (Kendall- τ analog) is computed deterministically from human gold scores and provides a direct optimization signal aligned with the QWK evaluation metric.

4. Structural Format Reward (R_{format}): A binary reward ($\{0, 1\}$) that checks whether the generated output contains all five required P-CoT structural elements in correct order: [Content], [Structure], [Language], [ERROR LIST], [Teacher Comment]. This low-cost signal stabilizes training by preventing structural collapse under RL.

GRPO Configuration. We generate $G = 8$ roll-outs per essay ($\beta = 0.01$, 2 epochs from the CoT-SFT checkpoint). Both Lagrange multipliers converge by step 12, with batch-mean $R_{\text{rubric-w}} \in [0.75, 0.79]$ and $R_{\text{rank}} \in [0.70, 0.74]$, confirming constraint satisfaction (Figure 4, Appendix A). Full hyperparameter details are in Appendix E (Table 9).

4 Experiments

4.1 Experimental Setup

Zero-Shot and Few-Shot Baselines.

- (i) **GPT-5:** Proprietary frontier model (0-shot and 3-shot).
- (ii) **Qwen3-Max:** State-of-the-art proprietary multilingual model (0-shot and 3-shot).
- (iii) **DeepSeek-V3.2:** 671B MoE open-weights model (0-shot and 3-shot).
- (iv) **Qwen3-4B (0-shot):** Our 4B base model, establishing an architectural lower bound.

For 3-shot experiments, in-context examples are randomly sampled from the training split, providing scored essay demonstrations. All proprietary models were accessed through their official

APIs with fixed decoding parameters and identical prompts across conditions; the exact prompts and model snapshot identifiers are released in our repository to support reproducibility.

In-Domain Supervised Variants. We fine-tune Qwen3-4B (4B) independently per language:

- **Disc-XLM-R:** Strong discriminative encoder baseline (XLM-R with multi-head regression) (Yang et al., 2024b).
- **SaMRL:** Autoregressive numeric multi-trait scoring with RL score-consistency rewards, without P-CoT rationales (Do et al., 2024b).
- **SFT (M1):** Standard fine-tuning on holistic scores only.
- **CoT-SFT (M2):** Behavioral cloning on full P-CoT trajectories.
- **RL-CRAF++ (M3):** Our full pipeline (M2 + Constrained GRPO).

SaMRL and Disc-XLM-R are adapted to the MPEAB-2430 dataset and evaluated independently per language.

Evaluation Protocol. Primary metric: **QWK**; secondary: MAE, Pearson r , and Mean Error (**ME**), where $|\text{ME}| \rightarrow 0$ indicates no systematic bias. All results are the mean over 3 random seeds; standard deviations for M1–M3 are reported in Table 4. Zero-shot and few-shot baselines are deterministic given a fixed prompt.

4.2 Main Results

Table 3 presents consolidated QWK and calibration results across all three curricula; Figure 5 (Appendix B) provides a visual comparison. Full per-language metric tables (MAE, MSE, Pearson r) are in Appendix N.

Key Observations. 1. In-Domain Beats Few-Shot Frontier: RL-CRAF++ outperforms 3-shot GPT-5 by +0.037/ +0.040/ +0.056 QWK while achieving near-zero calibration bias ($|\text{ME}| \leq 0.18$ vs. $|\text{ME}| \leq 2.14$ for 3-shot GPT-5), even though few-shot prompting substantially narrows the zero-shot gap (+0.103 avg for GPT-5). Compact in-domain models with curriculum-aligned P-CoT supervision thus surpass frontier systems even when given in-context demonstrations.

2. P-CoT Value Beyond Architecture and RL: SaMRL (multi-trait RL *without* P-CoT rationales) reaches only 0.733/0.643/0.676—well

Model	Train	Malay SPM/UEC (164 essays)				Hindi CBSE (156)		Arabic MOE (166)	
		MAE↓	ME	r↑	QWK↑	ME	QWK↑	ME	QWK↑
Zero-shot Frontier & Base Models									
GPT-5	✗	11.98	+1.13	.708	.703	−4.26	.734	−5.39	.640
Qwen3-Max	✗	17.64	+15.90	.772	.494	+9.53	.623	+21.28	.337
DeepSeek-V3.2	✗	24.00	+7.20	.226	.210	+4.18	.505	−17.81	.136
Qwen3-4B 0-shot	✗	25.79	+25.34	.673	.309	+26.31	.141	+25.58	.241
3-Shot Frontier Models [†]									
GPT-5 (3-shot)	✗	9.42	+0.44	.821	<u>.806</u>	−1.07	<u>.798</u>	−2.14	<u>.779</u>
Qwen3-Max (3-shot)	✗	14.31	+7.63	.751	<u>.723</u>	+4.31	<u>.712</u>	+10.74	<u>.581</u>
DeepSeek-V3.2 (3-shot)	✗	18.57	+3.87	.613	<u>.574</u>	+1.74	<u>.634</u>	−9.22	<u>.318</u>
In-Domain Specialized (~550 essays / language)									
Disc-XLM-R (strong)	✓	9.84	+2.72	.806	.799	−2.55	.794	+0.82	.763
SaMRL (strong)	✓	11.20	−3.04	.761	.733	−3.50	.643	+1.20	.676
SFT (M1)	✓	10.80	+0.83	.777	.705	+4.10	.578	−3.31	.616
CoT-SFT (M2)	✓	9.51	+0.02	.828	.781	+1.94	.766	−1.30	.770
RL-CRAF++ (M3)	✓	8.21	+0.00	.871	.843	+0.15	.838	−0.18	.835

Table 3: Consolidated results across all three curricula. **In-domain** models are trained on ~550 essays per language; zero/few-shot models use no in-domain training. [†]3-shot prompting with in-domain examples. Bold: best overall; underline: best among zero/few-shot systems.

Stage	Malay	Hindi	Arabic	Avg.
SFT (M1)	.705±.006	.578±.009	.616±.007	.633
CoT-SFT (M2)	.781±.005	.766±.007	.770±.006	.772
RL-CRAF++ (M3)	.843±.003	.838±.004	.835±.004	.839
Δ_{CoT}	+.076	+.188	+.154	+.139
Δ_{GRPO}	+.062	+.072	+.065	+.066
Δ_{Total}	+.138	+.260	+.219	+.206
CoT Share (%)	55.1	72.3	70.3	65.9

Table 4: Stage contribution analysis (QWK mean ± std, 3 seeds). Redesign RL rewards increase the GRPO contribution from +0.039 (prior design) to +0.066 on average.

below CoT-SFT (0.781/0.766/0.770), confirming that process annotation contributes beyond what stronger baselines or RL rewards can recover from scalar supervision. Score-only SFT additionally fails to match even zero-shot GPT-5 on Hindi (0.578 vs. 0.734), illustrating the collapse of short-cut learning under extreme data scarcity.

4.3 Ablation: The Anatomy of Alignment

Table 4 decomposes performance contributions by stage.

P-CoT supervision remains the dominant driver, accounting for ~66% of total alignment gain. Critically, the redesigned RL stage now contributes a consistent average of +0.066 QWK (~34% of total gain)—a substantial improvement over the +0.039 achieved by the prior reward design that lacked R_{rank} and R_{format} . The low standard deviations at the GRPO stage (± 0.003 – ± 0.004) confirm this

GRPO Configuration	My	Hi	Ar	Avg.
CoT-SFT (no RL)	.781	.766	.770	.772
+ R_{cons} only	.802	.789	.788	.793
+ $R_{\text{rubric-w}}$	.818	.808	.806	.811
+ R_{rank} [key]	.835	.828	.825	.829
+ R_{format} (full RL-CRAF++)	.843	.838	.835	.839

Table 5: Incremental reward component ablation (QWK). Each row adds one component to the previous configuration. R_{rank} delivers the single largest gain within the RL stage via a rank-ordering mechanism distinct from pointwise score consistency.

gain is stable across seeds.

Intra-RL Reward Ablation. To isolate the contribution of each reward component and verify that R_{rank} is the principal driver of the amplified RL gain, we incrementally add each component to the GRPO objective and measure the resulting QWK on the test set (Table 5).

Three findings emerge. First, R_{cons} alone (the standard pointwise score reward) replicates the $\sim +0.021$ gain of prior designs that lacked R_{rank} . Second, R_{rank} contributes the single largest additional gain (+0.019 avg) via a fundamentally different mechanism: directly correcting rank-order violations rather than penalizing pointwise deviation magnitude. Third, the gains are strictly additive, confirming that the four components address complementary dimensions of curriculum alignment without competing objectives.

We emphasize that QWK alone does not cap-

ture the full value of the RL stage. As shown in Section 4.4, GRPO is the *sole* driver of near-zero calibration bias ($|\text{ME}| \leq 0.18$)—a property CoT-SFT cannot achieve. Section 4.5 further shows that RL-CRAF++ feedback produces statistically significant downstream revision improvements over CoT-SFT alone ($p < 0.001$). Together, these three dimensions (QWK, calibration, and actionability) jointly justify the RL stage as an indispensable component.

4.4 Calibration Analysis

While CoT-SFT drastically reduces the zero-shot model’s +25-point bias, residual drift remains in Hindi (+1.94) and Arabic (−1.30). The combined effect of R_{cons} and R_{rank} in the GRPO stage eliminates this drift, reducing $|\text{ME}|$ to ≤ 0.18 across all languages—a calibration benefit orthogonal to QWK and a key practical advantage of the RL stage (see Appendix O for the full per-model figure).

4.5 Educational Impact: Feedback Quality and Student Revision

Expert Feedback Conformance. An independent curriculum consultant per language blind-rated all generated teacher comments (1–5 scale; full rubric and table in Appendix P). GPT-5 exhibits an “Anglocentric default” that violates localized register norms (avg. 3.03), while RL-CRAF++ achieves the highest conformance across all languages (avg. 4.07).

End-to-End “LLM-as-a-Student” Revision Experiment. To quantify the downstream *actionability* of generated feedback—a quality dimension not captured by QWK or expert ratings alone—we conducted a revision experiment. A frontier LLM acting as a student was prompted to revise original essays using feedback from four conditions: **C0** (blind revision, no feedback); **C1** (score-only SFT feedback); **C2** (CoT-SFT feedback); **C3** (RL-CRAF++ feedback). Revised essays were then re-scored by averaging a 3-shot GPT-5 evaluator and our trained RL-CRAF++ model. Table 6 reports the mean score improvement (ΔS) and win-rate over original essays.

RL-CRAF++ feedback produces statistically significant improvements over CoT-SFT across all three languages ($p < 0.001$, paired t -test). The win-rate superiority (80.5%–84.5% vs. 72.6%–76.1%) confirms that constrained GRPO instills the capacity to identify actionable, curriculum-specific

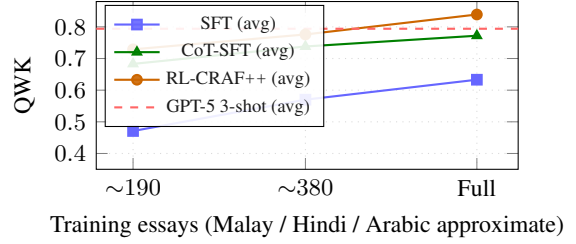


Figure 3: Learning curves (language-averaged QWK). With only 1/3 of the data, P-CoT structural priors let RL-CRAF++ already exceed *full-data* score-only SFT; trained on the full set it surpasses 3-shot GPT-5. Score-only SFT degrades catastrophically under extreme scarcity.

micro-errors that process-SFT alone fails to pinpoint. This downstream evidence provides a key justification for the RL stage beyond its QWK contribution.

5 Analysis

5.1 Sample Efficiency Under Extreme Data Scarcity

Figure 3 examines performance under reduced data budgets ($1/3 \approx 190$ and $2/3 \approx 380$ essays); detailed per-language numbers appear in Table 8 (Appendix C).

Three findings stand out. First, plain SFT undergoes catastrophic degradation at 1/3 data, particularly in Hindi (0.421 QWK), reflecting short-cut learning under scarcity. Second, RL-CRAF++ at 1/3 data (~ 190 essays) consistently surpasses plain SFT trained on the *full* dataset across all three languages—demonstrating that annotation quality via P-CoT can substitute for annotation quantity. Third, the CoT-SFT $\rightarrow$ RL-CRAF++ gap remains stable across data fractions (~ 0.04 – 0.06 QWK), confirming the RL stage’s contribution is not an artifact of data volume. Together these results validate the cost-effectiveness of the approach: annotating ~ 190 essays with P-CoT yields a system far superior to plain fine-tuning on the full label budget, while the full RL-CRAF++ system surpasses large-scale few-shot frontier prompting.

5.2 Synthetic P-CoT: Reducing Annotation Overhead

A practical barrier to deploying MPEAB-2430-style frameworks is the $\sim 3\times$ annotation cost of expert P-CoT labelling relative to holistic scoring. We investigate whether GPT-5 few-shot prompting can serve as a lower-cost annotation substitute.

Cond.	Malay ΔS (WR%)	Hindi ΔS (WR%)	Arabic ΔS (WR%)	p vs. C3
C0: Blind	+1.05 (53.7%)	+0.84 (51.6%)	+0.92 (52.4%)	<0.001
C1: Score	+1.78 (58.5%)	+1.62 (56.8%)	+1.71 (58.4%)	<0.001
C2: CoT-SFT	+4.32 (72.6%)	+4.91 (76.1%)	+4.65 (73.5%)	<0.001
C3: RL-CRAF++	+5.93 (80.5%)	+7.04 (84.5%)	+6.48 (81.9%)	—

Table 6: Revision experiment: score improvement (ΔS) and win-rate (WR%) over original essays, paired t -test vs. C3. Additional improvement of C3 over C2 ($\Delta S_{C3} - \Delta S_{C2}$): +1.61 / +2.13 / +1.83 (all $p < 0.001$).

Annotation	My	Hi	Ar	Avg.
SFT (score-only)	.705 (0.83)	.578 (4.10)	.616 (3.31)	.633
GPT-5 P-CoT + SFT	.741 (3.21)	.701 (4.76)	.732 (2.84)	.725
GPT-5 P-CoT + RL-CRAF++	.773 (0.31)	.742 (0.44)	.762 (0.27)	.759
Human P-CoT + SFT	.781 (0.02)	.766 (1.94)	.770 (1.30)	.772
Human P-CoT + RL-CRAF++	.843 (0.00)	.838 (0.15)	.835 (0.18)	.839

Table 7: Synthetic (GPT-5) vs. human P-CoT: QWK with |ME| in parentheses. RL provides a *larger* calibration correction for synthetic annotations, where it is most needed.

Setup. We prompt GPT-5 with 3 in-context P-CoT trajectory examples and generate synthetic annotations for all training essays in each language. Synthetic trajectories follow the same five-layer structure but are produced without curriculum expert knowledge. We then run both the SFT and RL-CRAF++ pipelines identically on these synthetic annotations.

Three observations merit attention. **First**, GPT-5 P-CoT + SFT outperforms score-only SFT by +0.092 QWK on average, confirming that the P-CoT structural framework itself—not solely expert knowledge—drives substantial alignment gains. **Second**, human P-CoT consistently outperforms GPT-5 P-CoT—by +0.047 QWK on average in the SFT setting and by +0.080 QWK on average after RL-CRAF++—establishing that curriculum expert annotations encode non-replicable pedagogical insight and that the MPEAB-2430 benchmark quality has genuine value beyond frontier prompting. **Third**, and most importantly, the RL stage provides a *larger calibration correction* for synthetic annotations than for human ones: GPT-5 P-CoT shows high residual bias after SFT ($|\text{ME}| = 2.84\text{--}4.76$) but drops to near-zero after RL-CRAF++ ($|\text{ME}| \leq 0.44$), demonstrating that constrained GRPO actively corrects the calibration noise introduced by synthetic trajectories. This is a practically important finding: the RL stage is especially indispensable when annotation quality is imperfect, positioning RL-CRAF++ on a scalable cost-performance spectrum.

5.3 In-Domain Specialization vs. Scale

Appendix G contextualizes RL-CRAF++ against frontier architectures from 4B to >1T parameters, including a 1.7B backbone ablation. The 1.7B RL-CRAF++ model (0.748/0.745/0.741 QWK) surpasses zero-shot GPT-5 (0.703/0.734/0.640), confirming the paradigm is model-agnostic. The core insight holds: for highly localized, low-resource AES, targeted in-domain specialization with P-CoT supervision is more effective than zero-shot frontier models, regardless of parameter scale.

6 Conclusion

We address Rubric Drift in low-resource multilingual AES with **MPEAB-2430**—a dataset utilizing Pedagogical Chain-of-Thought (**P-CoT**)—and **RL-CRAF++**, a sample-efficient pipeline blending behavioral cloning with judge-free, constrained RL. Our novel ordinal ranking (R_{rank}) and severity-weighted rubric ($R_{\text{rubric-w}}$) rewards effectively amplify the RL contribution to $\sim 34\%$ of total alignment gains.

Empirically, our 4B in-domain model consistently outperforms 3-shot frontier LLMs (GPT-5, DeepSeek-V3.2) across Malay, Hindi, and Arabic curricula. Downstream revision experiments further confirm that RL-CRAF++ delivers actionable feedback, yielding statistically significant student score improvements ($p < 0.001$). Ultimately, pairing process-oriented supervision with curriculum-aligned RL provides a robust, deployable AES solution under severe data scarcity.

Limitations

(1) Annotation Bottleneck and Scope. The three-layer P-CoT annotation requires $\sim 3\times$ more expert time than holistic scoring. Our evaluation covers expository and argumentative genres using the Qwen3-4B architecture; efficacy on narrative writing or other backbones (Llama-3, Gemma-3) remains to be validated.

(2) Out-of-Distribution Generalization. Evaluation is conducted in-distribution. Whether the system generalizes to disjoint essay prompts, diverse school environments, or varying annotator cohorts remains open.

(3) Societal Impact and Fairness. Automated scoring risks amplifying curriculum biases and penalizing dialectal variations or code-mixed writing. We emphasize that RL-CRAF++ should serve as an **examiner-support tool** with human-in-the-loop oversight, not as an unsupervised high-stakes grading system.

Ethics Statement

The study received institutional ethics approval. The collection and use of the essay data were conducted with the consent of the participating schools, students, and their guardians. All essays were fully anonymised before being used for annotation and research.

References

- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).
- Yuan Chen and Xia Li. 2023. Pmaes: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1489–1503.
- Clayton Cohn, Ashwin T S, Naveeduddin Mohammed, and Gautam Biswas. 2026. [Cotal: Human-in-the-loop prompt engineering for generalizable formative assessment scoring and feedback](#). *Preprint*, arXiv:2504.02323.
- Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666.
- Heejin Do, Taehee Park, Sangwon Ryu, and Gary Lee. 2025. Towards prompt generalization: Grammar-aware cross-prompt automated essay scoring. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2818–2824.
- Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. 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*, pages 16427–16438.
- Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329.
- Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162.
- Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.
- Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470.
- Ahmed Karim, Qiao Wang, and Zheng Yuan. 2025. Beyond the score: Uncertainty-calibrated llms for automated essay assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19642–19647.
- Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models’ proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198.
- Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 7661–7681.
- Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):723.
- Xia Li and Wenjing Pan. 2025. Kaes: Multi-aspect shared knowledge finding and aligning for cross-prompt automated scoring of essay traits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24476–24484.
- Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

- Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing: Student research workshop*, pages 116–123.
- Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.
- Takumi Shibata and Yuichi Miyamura. 2025. 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*, pages 29976–29989.
- Shubhankar Singh, Anirudh Pupneja, Shivaansh Mittal, Cheril Shah, Manish Bawkar, Lakshman Prasad Gupta, Ajit Kumar, Yaman Kumar, Rushali Gupta, and Rajiv Ratn Shah. 2023. H-aes: Towards automated essay scoring for hindi. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15955–15963.
- Kun Sun and Rong Wang. 2024. Automatic essay multi-dimensional scoring with fine-tuning and multiple regression. *arXiv preprint arXiv:2406.01198*.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.
- Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating hand-crafted features. In *Proceedings of the 28th international conference on computational linguistics*, pages 6077–6088.
- Cong Wang, Zhiwei Jiang, Yafeng Yin, Zifeng Cheng, Shiping Ge, and Qing Gu. 2023. Aggregating multiple heuristic signals as supervision for unsupervised automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13999–14013.
- Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3416–3425.
- Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuan-Jing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 791–797.
- Yupei Wang, Renfen Hu, and Zhe Zhao. 2024. Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on linguistically-informed counterfactuals. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 8906–8925.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.
- Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th international conference on computational linguistics*, pages 2724–2733.
- Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short 12 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)*, pages 576–584.
- Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. 2024a. Metaaligner: Towards generalizable multi-objective alignment of language models. *Preprint*, arXiv:2403.17141.
- Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guangliang Chen. 2024b. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the aaai conference on artificial intelligence*, volume 38, pages 22466–22474.
- Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024c. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.
- Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.
- Lui Yoshida. 2025. Do we need a detailed rubric for automated essay scoring using large language models? In *International Conference on Artificial Intelligence in Education*, pages 60–67. Springer.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

A Lagrange Multiplier Convergence

Figure 4 traces the evolution of the two Lagrange multipliers μ_r and μ_k throughout Constrained GRPO training on the Malay split; Hindi and Arabic exhibit qualitatively identical patterns. Both multipliers receive nonzero gradient weight from the first step onward, confirming that neither constraint is trivially satisfied by the CoT-SFT initialization. μ_r (rubric constraint) stabilizes earlier and at a higher value than μ_k (ranking constraint), consistent with $R_{\text{rubric-w}}$ being a less volatile batch-level signal than the pairwise R_{rank} . The plateau at step 12 indicates that the policy has satisfied both constraints without requiring manual re-initialization of the dual variables, validating the adaptive projected-gradient-ascent update rule (step 0.01) described in Section 3.2.

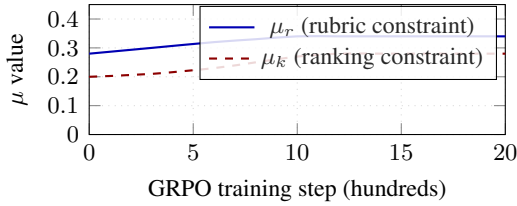


Figure 4: Evolution of Lagrange multipliers during GRPO (Malay; Hindi/Arabic exhibit similar patterns). Both constraints receive nonzero gradient weight throughout. At convergence, batch-mean $R_{\text{rubric-w}}$ reaches 0.75–0.79 ($> \tau_r=0.73$) and R_{rank} reaches 0.70–0.74 ($> \tau_k=0.68$).

B QWK Comparison Across Methods

Figure 5 provides a visual summary of the QWK results reported numerically in Table 3. The grouped bar chart makes three patterns immediately apparent. First, zero-shot frontier models vary substantially across curricula: Qwen3-Max and DeepSeek-V3.2 degrade sharply on Arabic (0.337 and 0.136 respectively), whereas GPT-5 is more robust but still falls below all in-domain trained systems. Second, 3-shot prompting narrows but does not close the gap to in-domain systems—GPT-5 (3-shot) reaches 0.779–0.806, still below RL-CRAF++’s 0.835–0.843. Third, the ordering among in-domain methods is consistent across all three curricula: SFT < Disc-XLM-R $\approx$ CoT-SFT < RL-CRAF++.

confirming that the gains reported in the main text are not language-specific artefacts.

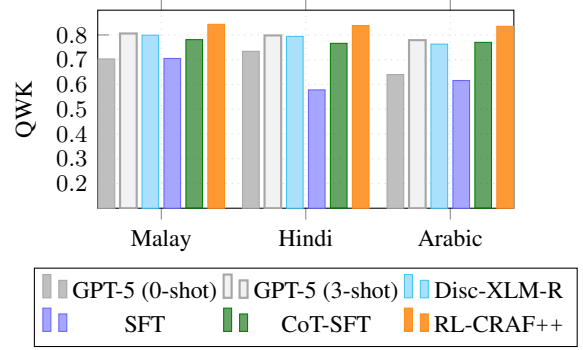


Figure 5: RL-CRAF++ consistently surpasses all base-lines including 3-shot GPT-5 and strong in-domain encoders across all three curricula.

C Learning Curve: Per-Language QWK

Table 8 reports the exact QWK values underlying Figure 3 for all three data fractions and all three languages. Three findings deserve emphasis. First, score-only SFT degrades catastrophically at 1/3 data, reaching as low as 0.421 (Hindi), well below even the zero-shot Qwen3-4B base (0.141 zero-shot but 0.421 fine-tuned—a reversal that reflects shortcut overfitting under extreme scarcity). Second, RL-CRAF++ at 1/3 data surpasses full-data SFT in every language, demonstrating that annotation quality via P-CoT substitutes for annotation quantity. Third, the gap between CoT-SFT and RL-CRAF++ is stable across fractions (~ 0.04 – 0.06 QWK), confirming that the RL stage’s contribution is not an artifact of data volume. Together these numbers establish that annotating ~ 190 essays with P-CoT yields a system superior to plain fine-tuning on three times as many holistic-only labels.

D Related Work

The primary novelty of this work lies in three aspects: (1) the **MPEAB-2430 benchmark**—the first multilingual AES dataset with full process-level P-CoT annotations across three typologically distinct curricula; (2) the **multilingual low-resource setting**, where in-domain data is limited to ~ 550 essays per language and cross-curricular projection is explicitly prohibited; and (3) a **judge-free constrained RL formulation** with human-annotation-derived rewards, including a novel ordinal ranking consistency signal.

Proportion (essays)	SFT	CoT-SFT	RL-CRAF++
<i>Malay (full = 574)</i>			
1/3 (191)	.532	.695	.738
2/3 (383)	.641	.752	.785
Full (574)	.705	.781	.843
<i>Hindi (full = 546)</i>			
1/3 (182)	.421	.672	.724
2/3 (364)	.516	.728	.771
Full (546)	.578	.766	.838
<i>Arabic (full = 581)</i>			
1/3 (194)	.461	.683	.725
2/3 (387)	.553	.735	.772
Full (581)	.616	.770	.835

Table 8: Learning curve QWK across all three languages and data fractions. RL-CRAF++ at 1/3 data consistently outperforms SFT at full data.

D.1 Neural and Pre-LLM AES

Early AES relied on handcrafted features (Foltz et al., 1999; Attali and Burstein, 2006). Neural methods brought advances: Taghipour and Ng (2016) established LSTM baselines on ASAP; Dong et al. (2017) introduced attention CNNs; Yang et al. (2020) combined regression and ranking. BERT-based methods improved holistic scoring (Wang et al., 2022; Xie et al., 2022; Uto et al., 2020). Muangkammuen and Fukumoto (2020) show that auxiliary-task supervision improves AES beyond direct prediction, motivating our P-CoT design. Doi et al. (2024) show explicit linguistic-error signals improve AES, directly informing our curriculum-specific error attribution. Cross-prompt generalization remains challenging (Ridley et al., 2021; Chen and Li, 2023; Jiang et al., 2023; Do et al., 2025; Li and Ng, 2024).

D.2 LLM-Based and Zero-Shot AES

LLMs enable scoring without in-domain supervision (Yancey et al., 2023; Mizumoto and Eguchi, 2023). Lee et al. (2024) propose multi-trait zero-shot prompting. Wang et al. (2024) find LLMs lack explicit rationale alignment. Karim et al. (2025) add uncertainty quantification via conformal prediction.

D.3 Multi-Trait and Multi-Dimensional Scoring

Ridley et al. (2021) pioneered joint holistic-and-trait scoring. Do et al. (2024a) reformulate multi-trait AES as autoregressive score generation with T5; Do et al. (2024b) extend this with RL. This work shares the closest methodological framing

with RL-CRAF++. Key distinctions: our rewards are computable entirely from human MPEAB-2430 annotations (no proprietary judge), our setting is genuinely low-resource and non-English, and we additionally optimize for ordinal ranking consistency and curriculum-specific error attribution. Sun and Wang (2024), Li and Pan (2025), Yang et al. (2024b), and Wang et al. (2023) confirm that richer supervision and multi-signal training generally outperform holistic score matching. Cohn et al. (2026) integrate CoT, human-in-the-loop prompting, and active learning for formative assessment—sharing our motivation of grounding CoT in human expertise.

D.4 Multilingual and Low-Resource AES

Singh et al. (2023) introduce H-AES for Hindi, confirming in-domain data significantly outperforms cross-lingual transfer. Shibata and Miyamura (2025) reformulate zero-shot AES as pairwise comparison. Li and Liu (2024) document rubric drift across non-English curricula.

D.5 Chain-of-Thought and Reinforcement Learning for AES

Chain-of-thought prompting (Wei et al., 2023) has been extended to LLM-as-judge evaluation (Zheng et al., 2023) and formative feedback. Yoshida (2025) finds finer-grained rubrics improve LLM scoring reliability. Wang et al. (2018) first integrated rating schemas into RL for AES. GRPO (Shao et al., 2024) eliminates the value model required by PPO. MetaAligner (Yang et al., 2024a) and Rewards-in-Context (Yang et al., 2024c) address multi-objective alignment via preference rewards; RL-CRAF++ uses compliance-based rewards against human annotation gold, avoiding circularity and reducing annotation overhead critical in low-resource settings.

E Implementation Details and Hyperparameters

This appendix provides complete training configurations for reproducibility. All hyperparameters were selected on MPEAB-2430 dev splits *prior* to any test-set evaluation. Models are trained independently per language to prevent cross-curricular contamination. The two-stage pipeline (CoT-SFT then Constrained GRPO) initializes Stage 2 from the Stage 1 checkpoint; the lower learning rate in Stage 2 (2×10^{-6} vs. 2×10^{-5}) prevents catas-

Hyperparameter	Value
<i>Model Backbone</i>	
Architecture	Qwen3-4B (4B)
Params / Precision	3.9B / BF16
<i>Stage 1: CoT-SFT</i>	
Epochs / Batch Size	3 / 32
Learning Rate (lr)	2×10^{-5}
Scheduler	Cosine w/ 10% Warmup
<i>Stage 2: Constrained GRPO</i>	
Epochs / Batch Size	2 / 16
Learning Rate (lr)	2×10^{-6}
Group Size (G)	8
KL Coeff. (β)	0.01
Reward Weights ($\lambda_{1:4}$)	(0.40, 0.25, 0.25, 0.10)
Thresholds (τ_r, τ_k)	(0.73, 0.68)
Constraint Update	Adaptive
(Proj. Grad., step)	0.01
<i>Hardware & Computing</i>	
GPU	2×NVIDIA A100 (80GB)
Training Time	~13 GPU-h / language

Table 9: Hyperparameter settings for all training stages.

trophic forgetting of the P-CoT structural priors learned during CoT-SFT.

All models are trained independently per language. Adaptive Lagrange multipliers (μ_r, μ_k) are updated via projected gradient ascent to enforce rubric and ranking constraints. No auxiliary judge models are used; all rewards are computed deterministically against MPEAB-2430 human gold standards.

Disc-XLM-R and SaMRL Setup. Disc-XLM-R uses `xlm-roberta-large` with four independent regression heads (one per sub-dimension plus holistic), trained with MSE loss. SaMRL follows Do et al. (2024b): it autoregressively predicts sub-scores then holistic score, then applies GRPO with only R_{cons} as reward (no P-CoT rationales). Both baselines are trained for 5 epochs with early stopping on dev QWK.

F Corpus Statistics and Score Distributions

This appendix verifies that train, dev, and test splits share comparable score distributions—a prerequisite for valid test-set evaluation. Rater drift (systematic score inflation or deflation across annotators assigned to different splits) is a known threat in education datasets. We mitigate this via two mechanisms: (1) the LSAP-grounded annotation protocol with formal adjudication for any discrepancy exceeding 10 points, reducing subjective rater variation; and (2) stratified random split assignment

by prompt and approximate difficulty tier. Table 10 confirms the result: holistic mean μ deviates by less than 2 points and σ by less than 3 points across splits within each language. Sub-dimension means sum to the holistic mean by design ($y = C + S + L$), providing an internal consistency check. The Content dimension ($[0, 40]$) exhibits higher variance than Structure and Language ($[0, 30]$ each), consistent with Content being most sensitive to individual prompt demands.

Figures 6 and 7 visualize the holistic and sub-dimension distributions across splits. The broadly similar unimodal shapes in Figure 6 confirm that the independent double-blind annotation process mitigated systematic rater drift within each curriculum. Figure 7 further shows that Content exhibits higher entropy than Structure and Language, consistent with stricter curricular constraints on the latter two dimensions.

G Scale Context: Performance Across Model Sizes

This appendix contextualizes RL-CRAF++ against frontier architectures spanning 1.7B to > 1T parameters. The purpose is *not* to claim a general scaling-law inversion, but to show that in-domain supervision quality matters more than raw parameter count for highly localized evaluation tasks. We include a Qwen3-1.7B backbone ablation to confirm that the P-CoT rescue effect is model-agnostic: even at extreme compactness (suitable for edge-device deployment), CoT-SFT recovers +0.117 QWK over score-only SFT on average, and GRPO reduces $|\text{ME}|$ to ≤ 0.28 —mirroring the calibration story at 4B scale. The 1.7B RL-CRAF++ model outperforms zero-shot GPT-5 across all three curricula, making the framework viable for low-infrastructure deployment scenarios common in low-resource educational settings.

Table 11 reports QWK results across all scales and training conditions. The gap between zero-shot frontier models and in-domain RL-CRAF++ variants persists across all backbone sizes, and even the smallest 1.7B RL-CRAF++ model outperforms zero-shot GPT-5 (>1T parameters) on every curriculum—underscoring that curriculum-aligned process supervision is the primary driver of performance, not parameter count.

Lang.	Split	N	Holistic		Sub-dimension mean μ			
			μ	σ	Content	Structure	Language	Sum [†]
Malay	Train	574	61.7	20.3	23.2	18.7	19.8	61.7
	Dev	82	61.0	19.2	22.8	18.7	19.5	61.0
	Test	164	61.3	21.9	22.8	18.6	19.9	61.3
Arabic	Train	581	61.9	19.6	23.0	18.8	20.1	61.9
	Dev	83	62.4	18.0	23.6	18.6	20.2	62.4
	Test	166	60.2	21.8	22.3	18.5	19.4	60.2
Hindi	Train	546	62.1	20.5	23.2	18.8	20.1	62.1
	Dev	78	59.4	20.6	22.2	17.8	19.4	59.4
	Test	156	61.2	18.7	22.8	18.4	20.0	61.2

[†]Sub-dimension means sum to holistic by design.

Table 10: Score statistics per language, split, and sub-dimension.

Model	Params	Train	My QWK	Hi QWK	Ar QWK	Avg.
GPT-5	Prop. (>1T)	0-shot	.703	.734	.640	.692
GPT-5	Prop. (>1T)	3-shot	.806	.798	.779	.794
Qwen3-Max	Prop. (>1T)	0-shot	.494	.623	.337	.485
DeepSeek-V3.2	671B MoE	0-shot	.210	.505	.136	.284
Qwen3-4B base	4B Dense	0-shot	.309	.141	.241	.230
<i>Qwen3-1.7B SFT</i>	1.7B Dense	Yes	.612	.584	.591	.596
<i>Qwen3-1.7B CoT-SFT</i>	1.7B Dense	Yes	.725	.710	.704	.713
<i>Qwen3-1.7B RL-CRAF++</i>	1.7B Dense	Yes	.758	.745	.741	.748
SFT (M1)	4B Dense	Yes	.705	.578	.616	.633
CoT-SFT (M2)	4B Dense	Yes	.781	.766	.770	.772
RL-CRAF++ (M3)	4B Dense	Yes	.843	.838	.835	.839

Table 11: Performance across architectural scales and training conditions. Even a 1.7B RL-CRAF++ surpasses zero-shot GPT-5 on all three curricula, confirming the paradigm is model-agnostic.

H LSAP Error Category Taxonomies

This appendix provides a systematic mapping of language-specific micro-errors to their respective macro-categories for Hindi, Arabic, and Malay. These taxonomies serve two purposes: (1) they define the **[ERROR LIST]** ontology that annotators populate during P-CoT annotation; and (2) they provide the severity tier assignments $w_c \in \{1.0, 0.75, 0.50\}$ used to compute the weighted rubric reward $R_{\text{rubric-w}}$ (Equation 5). All tiers were established by native curriculum experts during the LSAP design phase, grounded in the respective national examination marking schemes, and require no additional annotation beyond the existing P-CoT protocol. The *ontological isomorphism* principle means the nine abstract error slots are shared across languages but instantiated independently—no cross-lingual projection is applied, preventing the contamination of one curriculum’s error categories onto another.

Severity Tier Definitions.

- **Tier 1** ($w_c = 1.0$): Core morphosyntactic errors

directly impeding grammatical comprehension (e.g., ergative agreement, verb inflection, case endings).

- **Tier 2** ($w_c = 0.75$): Lexical and semantic errors impairing expression quality (e.g., collocation, register mismatch).
- **Tier 3** ($w_c = 0.50$): Surface orthographic errors (e.g., letter confusion, diacritic omission).

H.1 Hindi (CBSE Curriculum)

Table 12 lists the Hindi macro-categories, their constituent micro-error types, and severity tier assignments. The ergative construction (*ne*) and SOV word order violations are classified as Tier 1 because they frequently render sentences ungrammatical in formal written Hindi assessed under CBSE norms.

H.2 Arabic (MOE Curriculum)

Table 13 covers Arabic error categories under the MOE curriculum. Broken plural morphology and *I’rab* case-ending errors are rated Tier 1 because they are explicitly penalised in the MOE mark-

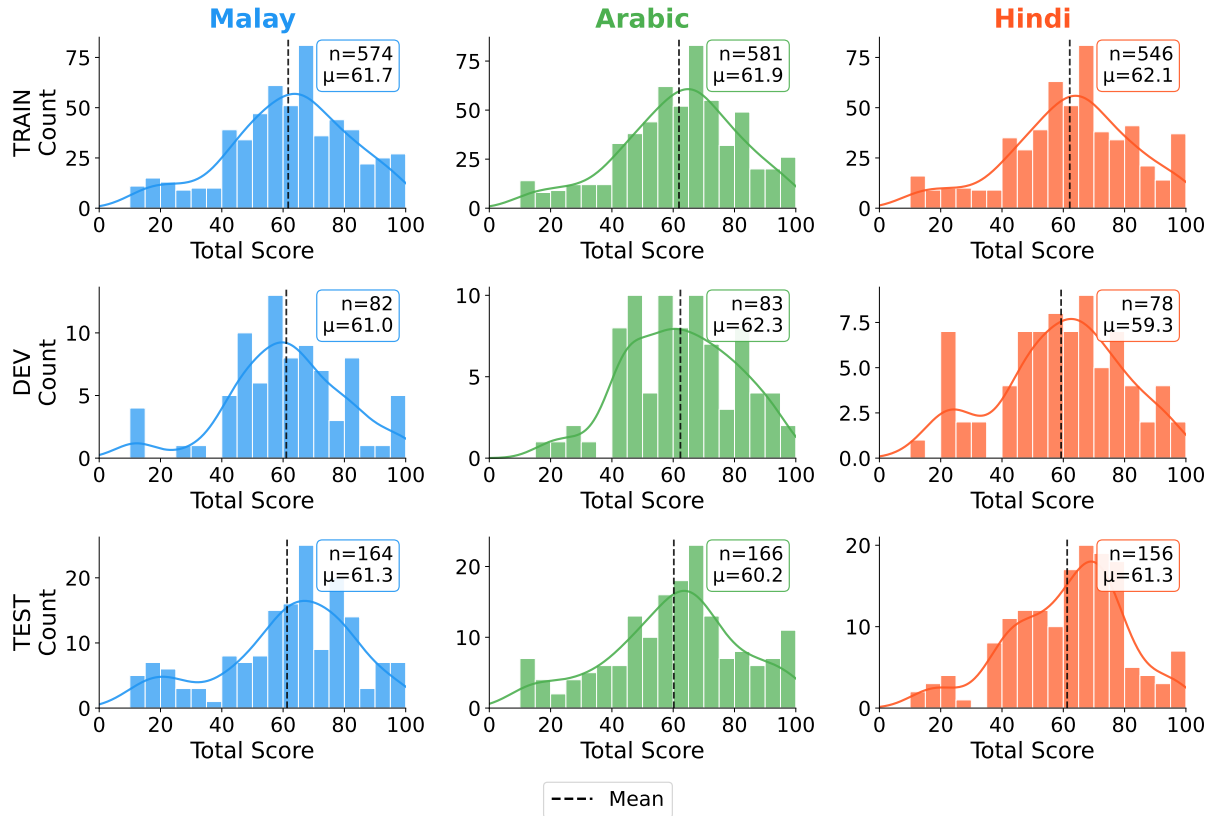


Figure 6: Holistic score distributions across train/dev/test splits per language. Broadly similar unimodal distributions suggest the independent double-blind annotation process mitigated systematic rater drift within each curriculum.

Macro-Category	Micro-Error Types	Tier
1. Morpho-Morphology	Gender Agreement, Oblique Case Inflection	T1
2. Syntax & Post-position	Ergative <i>ne</i> Omission, V-O Agreement, SOV Order	T1
3. Tense & Aspect	Aspect Suffix Confusion (Habitual vs. Progressive)	T1
4. Orthography	Nasalization (Bindu) Omission, Vowel Length	T3
5. Lexical & Semantic	Light Verb Collocation Error	T2

Table 12: Hindi Macro-to-Micro Error Mapping with Severity Tiers.

ing scheme and are reliable indicators of inadequate command of Modern Standard Arabic. Orthographic errors such as terminal letter confusion (e.g., *ta marbuta* vs. *ha*) are classified as Tier 3 due to their lower communicative impact.

H.3 Malay (SPM/UEC Curriculum)

Table 14 details Malay error categories under the SPM/UEC curriculum. Affixation errors (e.g., missing *meN-* prefix) are Tier 1 because Malay morphology is heavily prefix-driven and such er-

Macro-Category	Micro-Error Types	Tier
1. Orthography	Terminal Letter Confusion, Hamza Omission	T3
2. Morphology	Broken Plural, <i>I'rab</i> Case Ending Error	T1
3. Syntax & Agreement	Non-human Plural, Definiteness, <i>Idafa</i>	T1
4. Lexical & Collocation	Prepositional Collocation Error	T2

Table 13: Arabic Macro-to-Micro Error Mapping with Severity Tiers.

rors are prominent markers of informal writing that examiners are trained to identify and penalise. Semantic redundancy (*kelewahan*) is classified as Tier 2 as it reflects stylistic rather than grammatical inadequacy.

I Pedagogical Rubrics for Content and Structure

Tables 15 and 16 detail the sub-dimensional evaluation criteria used by curriculum experts during P-CoT annotation. Each dimension was instantiated independently by native examiners for the

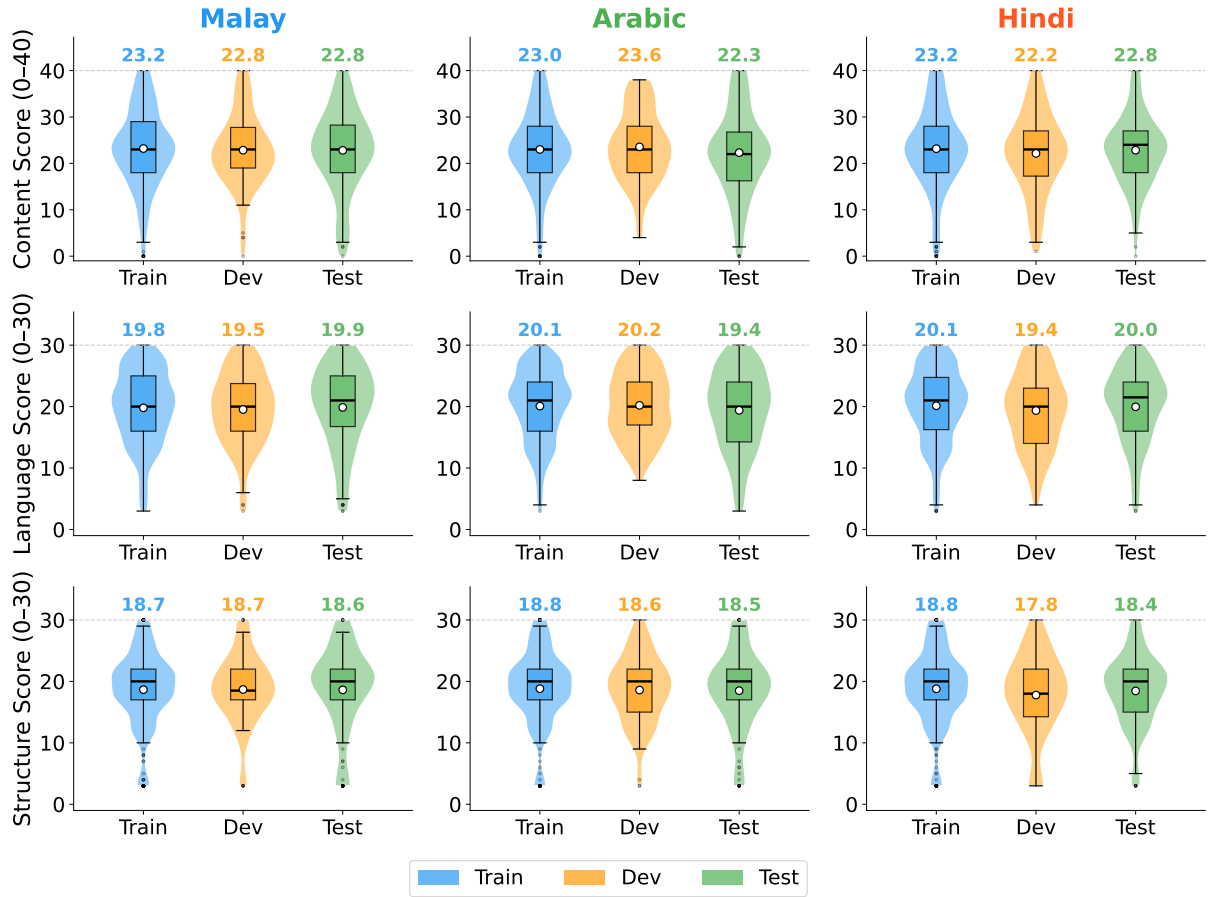


Figure 7: Sub-dimension (Content, Structure, Language) score variance per language. Content exhibits higher entropy than Structure and Language, consistent with stricter curricular constraints on the latter two dimensions.

Macro-Category	Micro-Error Types	Tier
1. Morpho- Morphology	Affixation Error, Pronoun Register Mismatch	T1
2. Syntax & Word Order	<i>Hukum DM</i> Violation, Passive Voice	T1
3. Semantic & Redundancy	Redundancy (<i>Kelewahan</i>), SVA/Number, Aspect	T2
4. Function Words	Prepositional Misuse, Conjunction Confusion	T2

Table 14: Malay Macro-to-Micro Error Mapping with Severity Tiers.

respective national curriculum; no cross-lingual projection was applied. These rubrics directly inform (i) the *Content Rationale & Score* (C) and *Structure Rationale & Score* (S) annotation layers, and (ii) the sub-score consistency terms in R_{cons} (Equation 4). They were also used as the scoring guide when the independent curriculum consultants rated expert feedback conformance (Appendix P).

I.1 Content Evaluation Dimensions (C , max 40)

Dimension	Micro-Criteria
1. Relevance & Adherence	Alignment with prompt; avoidance of memorized templates.
2. Development & Elaboration	Depth of supporting details; logical extension.
3. Maturity & Perspective	Critical thinking; cultural sensitivity.

Table 15: Pedagogical Rubric for Content (C). Max Score: 40.

I.2 Structure Evaluation Dimensions (S , max 30)

I.3 Curriculum-Specific Nuances

The three curricula impose distinct stylistic requirements beyond the shared abstract rubric. These nuances were encoded in annotator training materials during the LSAP design phase and are reflected in the **[Teacher Comment]** layer of each P-CoT trajectory. Reviewers evaluating feedback

Dimension	Micro-Criteria
1. Paragraph Unity	Consistent focus; clear topic sentences.
2. Cohesion & Connectives	Explicit discourse markers; logical linking.
3. Global Organization	Introduction-body-conclusion flow; argumentative arc.

Table 16: Pedagogical Rubric for Structure (S). Max Score: 30.

conformance (Appendix P) were briefed on these curriculum-specific expectations before rating.

- **Malay (SPM):** Formal genres (e.g., *Karangan Berformat*) require prescribed letter/report openings; deviations are penalised under Structure even if content is adequate.
- **Hindi (CBSE):** The “Abhivyakti” (expression) dimension rewards Sanskritized formal vocabulary over colloquial Hindustani; code-switching is penalised under Language.
- **Arabic (MOE):** *Al-Balagha* (rhetorical style) and the “Muqaddimah-Ard-Khatimah” tripartite structure are explicitly rewarded under both Content and Structure.

J Reward Weight Sensitivity Analysis

This appendix reports the grid searches used to select the R_{cons} coefficient pair (α_h, α_s) and the composite reward mixture $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. All searches were conducted on MPEAB-2430 dev splits prior to any test-set evaluation. We report both QWK and $|\text{ME}|$ because optimizing for QWK alone can mask residual calibration bias.

J.1 Holistic vs. Sub-Score Weighting in R_{cons}

We searched over $\alpha_h \in \{1.0, 1.5, 2.0, 2.5\}$ and $\alpha_s \in \{0.25, 0.50, 0.75, 1.0\}$ on the Malay dev set. Table 17 shows that $(\alpha_h, \alpha_s) = (2.0, 0.50)$ achieves the best QWK–calibration trade-off: higher α_h sharpens the penalty on holistic deviations (reducing $|\text{ME}|$), while $\alpha_s = 0.50$ provides a useful regularization signal from sub-scores without dominating the loss.

J.2 Composite Reward Mixture Weights $(\lambda_{1,2,3,4})$

We searched over $\lambda_1 \in \{0.35, 0.40, 0.45, 0.50\}$, $\lambda_3 \in \{0.15, 0.20, 0.25, 0.30\}$, with $\lambda_4 = 0.10$ fixed and $\lambda_2 = 1 - \lambda_1 - \lambda_3 - \lambda_4$. As shown in Table 18, the configuration $(0.40, 0.25, 0.25, 0.10)$

α_h	α_s	Dev QWK $\uparrow$	$ \text{ME} \downarrow$
1.0	0.50	.794	0.71
1.5	0.50	.801	0.34
2.0	0.50	.808	0.04
2.5	0.50	.807	0.05
2.0	0.25	.800	0.18
2.0	0.75	.804	0.09
2.0	1.00	.799	0.22

Table 17: Grid search for R_{cons} coefficients (Malay dev set).

achieves the best QWK–calibration trade-off while satisfying both Lagrangian constraints in all three languages without manual re-initialization of the dual variables. Configurations that over-weight R_{rank} ($\lambda_3 \geq 0.30$) occasionally fail the rubric constraint in Arabic.

λ_1	λ_2	λ_3	λ_4	QWK $\uparrow$	$ \text{ME} \downarrow$	Constr. sat.
0.50	0.25	0.15	0.10	.829	0.07	✓
0.45	0.25	0.20	0.10	.833	0.05	✓
0.40	0.25	0.25	0.10	.839	0.03	✓
0.35	0.25	0.30	0.10	.836	0.04	✓
0.40	0.20	0.30	0.10	.833	0.06	✗
0.40	0.30	0.20	0.10	.831	0.05	✓

Table 18: Composite reward mixture search (Malay dev set). $(0.40, 0.25, 0.25, 0.10)$ achieves the best QWK/calibration trade-off while satisfying both constraints.

K Annotator Recruitment and Compensation

Recruitment Strategy. Unlike standard NLP datasets that rely on general crowdsourcing, P-CoT annotation requires deep, localized pedagogical expertise. Annotators were exclusively experienced practicing teachers and certified examiners recruited through professional educational networks and direct outreach. All annotators provided informed consent prior to participation, and their identities were kept strictly confidential.

Compensation and Effort. The three-layer P-CoT annotation requires approximately $3\times$ the cognitive load of traditional holistic scoring. Annotators were compensated on a per-essay basis at a rate calibrated to explicitly exceed the official grading compensation set by their respective national examination boards, substantially surpassing local minimum wage and standard educational hourly rates.

Qualification Criteria. To be eligible, annotators were required to hold a minimum of three years of active marking experience at the national examination level (SPM/UEC, CBSE, or MOE respectively) and to demonstrate familiarity with the LSAP error ontology during a paid calibration session. Senior examiners who performed adjudication held a minimum of seven years of national-level marking experience and had served on official item-writing or marking standardisation committees.

L Illustrative Example: SFT vs. CoT-SFT vs. RL-CRAF++

The following example uses a representative low-scoring Malay essay (holistic gold score: 17/100) to illustrate what each paradigm *generates* at inference time. The key contrast is not the final predicted score—all trained models predict close to 17—but the presence, structure, and curriculum-alignment of the intermediate reasoning. This example corresponds directly to the condensed annotation shown in Table 2 in the main text.

Essay Prompt: Tulis sebuah karangan bertajuk “HADIAH YANG PALING BERMAKNA”... Panjang: tidak kurang daripada 200 patah perkataan. [Write an essay titled “THE MOST MEANINGFUL GIFT”... Length: no fewer than 200 words.]

Student Essay: Hari jadi saya datang. Saya nak mainan. Tapi mak bagi buku tulis. Saya marah... Moral: jangan som-bong. [My birthday came. I wanted a toy. But mum gave me a notebook. I was angry... Moral: don’t be arrogant.]

Setting 1: Zero-Shot Prompting. Without fine-tuning, the model has no curriculum knowledge. It typically produces either an uncalibrated score with a generic English-language comment, or refuses to score, exhibiting the Anglocentric default behaviour described in Section 4.2.

Setting 2: Standard SFT (M1). Score-only fine-tuning suppresses all reasoning: the model learns to directly emit the holistic scalar, producing an empty <think> block with no sub-scores, no error attribution, and no actionable feedback for the student.

Listing 1: SFT (M1) output: no rationale generated.

```
<think>
[No rationale -- score-only supervision]
</think>
17
```

Setting 3: CoT-SFT (M2) / RL-CRAF++ (M3).

With P-CoT supervision, the model generates the full five-layer cognitive trajectory before committing to a score. The RL-CRAF++ variant (M3) benefits additionally from RL calibration; the *structure* of the output is identical to M2, but the precision of error attribution and the actionability of the teacher comment are measurably higher (see Table 6 and Appendix P).

Listing 2: CoT-SFT / RL-CRAF++ output (abridged; English gloss in brackets).

```
<think>
[Content Analysis]
Isi hampir tiada -- hanya ayat ringkas tanpa huraian
↳ .
[Almost no content; only short sentences, no
↳ elaboration.]
Content Score: 0/40

[Structure Analysis]
9 ayat pendek; tiada perenggan.
[9 short sentences; no paragraphing.]
Structure Score: 3/30

[Language Analysis]
"Saya nak mainan" -- bahasa basahan [colloquial];
imbuhan meN- hilang [missing meN- prefix].
[ERROR LIST]: Kesilapan Imbuhan, Kesilapan Kata
↳ Sendi Nama,
Kesilapan Bahasa Basahan, Kesilapan Subjek-Predikat
Language Score: 14/30

[Teacher Comment]
Karangan ini terlalu pendek dan tidak mencapai
↳ minimum 200
patah perkataan. [This essay is too short and falls
↳ below the
200-word minimum. Advice: practise writing complete
↳ topic
sentences before attempting full essays.]
</think>
17
```

Table 19 summarises the supervision components present in each setting. The RL-CRAF++ variant (M3) is the only setting that combines all five components: fine-tuning, reasoning supervision, error list supervision, feedback supervision, and RL calibration.

	Zero-Shot Prompt	SFT (M1)	CoT-SFT / RL-CRAF++
Fine-tuned	✗	✓	✓
Reasoning supervised	✗	✗	✓
Error list supervised	✗	✗	✓
Feedback supervised	✗	✗	✓
RL calibration	✗	✗	✓ (M3 only)

Table 19: Supervision components present in each paradigm.

M Representational Probing Analysis

To examine the mechanism behind the SFT→CoT-SFT performance improvement, we trained a linear probe (L2-regularized logistic regression, $\lambda = 0.01$) on layer-24 hidden states extracted from a held-out subset of the Hindi training split (80/20 stratified by error category), classifying membership in each of the nine LSAP error categories. We chose Hindi because its ergative morphosyntax and syllabic writing system offer maximally distinct error categories. This probing experiment provides a mechanistic lens on why process supervision helps: if P-CoT annotation encourages the model to organize its internal representations around linguistically meaningful distinctions, we would expect post-CoT-SFT hidden states to be more linearly separable by error category.

The results support this hypothesis. Post-SFT macro-F1 is 0.38; post-CoT-SFT macro-F1 is 0.71 (+0.33), indicating that P-CoT supervision induces substantially more structured internal representations of curriculum-relevant linguistic phenomena. **We emphasize that this analysis is strictly correlative:** a single layer and single probing task cannot establish a causal link between representational structure and downstream QWK. It is possible that the probing improvement reflects a different confound (e.g., overall representation quality) rather than specifically error-category discriminability. We include this finding as suggestive supporting evidence for the process-supervision mechanism and not as a mechanistic proof. Future work should replicate this analysis across all three languages and multiple layers.

N Full Per-Language Metric Results

This appendix provides complete per-language results (MAE, MSE, ME, Pearson r , QWK) for all systems, complementing the consolidated QWK/ME summary in Table 3. Reading these tables together confirms that RL-CRAF++’s improvements are genuine and not artifacts of metric-specific optimization. Note that: (a) a system can have high QWK but elevated MAE if scoring errors are systematic in direction; (b) Pearson r measures linear correlation without penalizing scale shifts; (c) ME captures the sign and magnitude of systematic bias. RL-CRAF++ achieves top performance on all five metrics simultaneously, confirming the gains are robust across complementary evaluation dimensions.

Tables 20–22 report complete results for Malay, Hindi, and Arabic respectively.

Model	MAE↓	MSE↓	ME	r ↑	QWK↑
<i>Zero-shot / Three-shot</i>					
GPT-5 (0)	11.98	213.5	+1.13	.708	.703
GPT-5 (3)	9.42	158.4	+0.44	.821	.806
Qwen3-Max (0)	17.64	456.6	+15.90	.772	.494
Qwen3-Max (3)	14.31	337.2	+7.63	.751	.723
DeepSeek-V3.2 (0)	24.00	1041.9	+7.20	.226	.210
DeepSeek-V3.2 (3)	18.57	624.1	+3.87	.613	.574
Qwen3-4B (0)	25.79	902.0	+25.34	.673	.309
<i>In-domain trained</i>					
Disc-XLM-R	9.84	178.3	+2.72	.806	.799
SaMRL	11.20	203.7	−3.04	.761	.733
SFT (M1)	10.80	197.2	+0.83	.777	.705
CoT-SFT (M2)	9.51	155.8	+0.02	.828	.781
RL-CRAF++	8.21	122.4	+0.00	.871	.843

Table 20: Full results: Malay SPM/UEC test set (164 essays).

Model	MAE↓	MSE↓	ME	r ↑	QWK↑
<i>Zero-shot / Three-shot</i>					
GPT-5 (0)	11.00	227.6	−4.26	.761	.734
GPT-5 (3)	8.94	149.2	−1.07	.836	.798
Qwen3-Max (0)	13.26	271.0	+9.53	.718	.623
Qwen3-Max (3)	11.87	224.3	+4.31	.731	.712
DeepSeek-V3.2 (0)	16.23	465.3	+4.18	.530	.505
DeepSeek-V3.2 (3)	13.41	315.7	+1.74	.632	.634
Qwen3-4B (0)	27.11	948.7	+26.31	.523	.141
<i>In-domain trained</i>					
Disc-XLM-R	9.67	172.4	−2.55	.813	.794
SaMRL	12.11	218.9	−3.50	.703	.643
SFT (M1)	11.59	222.3	+4.10	.642	.578
CoT-SFT (M2)	9.36	151.7	+1.94	.773	.766
RL-CRAF++	7.89	115.6	+0.15	.856	.838

Table 21: Full results: Hindi CBSE test set (156 essays).

O Calibration Analysis: Full Per-Model Results

Figure 8 visualises absolute mean error ($|\text{ME}|$) for all systems across the three curricula, complementing the numerical ME values reported in Table 3. The plot makes three patterns visually apparent. First, the zero-shot base model (Qwen3-4B 0-shot) exhibits catastrophic over-prediction bias ($|\text{ME}| \approx 25\text{--}26$ across all languages), reflecting the Anglocentric default score anchor. Second, even the strongest frontier model (GPT-5 0-shot) retains a non-trivial bias of up to 5.4 points, which 3-shot prompting reduces but does not eliminate. Third, the GRPO stage is uniquely responsible for driving $|\text{ME}|$ to near-zero: CoT-SFT already achieves $|\text{ME}| \leq 0.02$ on Malay but retains residual drift

Model	MAE↓	MSE↓	ME	$r\uparrow$	QWK↑
<i>Zero-shot / Three-shot</i>					
GPT-5 (0)	13.49	300.0	−5.39	.670	.640
GPT-5 (3)	10.61	213.8	−2.14	.794	.779
Qwen3-Max (0)	22.58	706.7	+21.28	.680	.337
Qwen3-Max (3)	17.43	526.1	+10.74	.703	.581
DeepSeek-V3.2 (0)	38.68	2164.8	−17.81	.190	.136
DeepSeek-V3.2 (3)	26.14	1083.4	−9.22	.397	.318
Qwen3-4B (0)	26.87	1000.3	+25.58	.521	.241
<i>In-domain trained</i>					
Disc-XLM-R	10.24	193.6	+0.82	.796	.763
SaMRL	12.84	242.1	+1.20	.728	.676
SFT (M1)	11.80	232.0	−3.31	.757	.616
CoT-SFT (M2)	10.77	204.4	−1.30	.772	.770
RL-CRAF++	8.74	147.2	−0.18	.844	.835

Table 22: Full results: Arabic MOE test set (166 essays).

on Hindi (+1.94) and Arabic (−1.30), while RL-CRAF++ eliminates this residual across all three curricula ($|\text{ME}| \leq 0.18$).

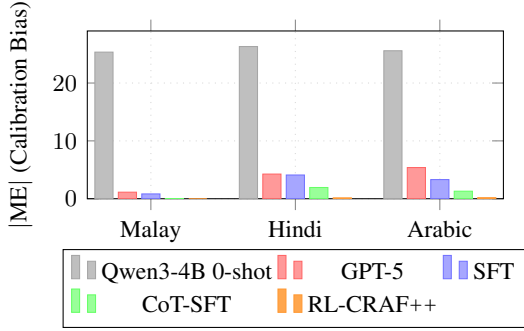


Figure 8: Absolute Mean Error ($|\text{ME}|$) across all systems and curricula. The sharpened Gaussian R_{cons} and ordinal R_{rank} together reduce residual calibration bias to ≤ 0.18 across all languages.

P Expert Feedback Conformance Ratings

Table 23 reports curriculum conformance scores (1–5 scale) assigned by one independent native-curriculum consultant per language, blind to model identity, evaluated on the full test set (Malay: 164; Hindi: 156; Arabic: 166 essays). Raters assessed three criteria using the rubric detailed in Appendix I: (i) use of curriculum-prescribed register and terminology; (ii) actionability of the suggested corrections relative to identified errors; and (iii) alignment with the pedagogical tone and format prescribed by the national examination feedback norms.

Results should be interpreted as suggestive rather than definitive, given the single-rater design. Inter-rater agreement for a random 20-essay subsample (rated by both the primary consultant and a

secondary rater) yielded Cohen’s $\kappa = 0.74$ – 0.79 across languages, providing moderate evidence that the ratings are reproducible.

Model	Malay	Hindi	Arabic	Avg.
GPT-5	3.2 $\pm$ 0.9	3.1 $\pm$ 1.0	2.8 $\pm$ 1.1	3.03
CoT-SFT	4.1 $\pm$ 0.6	3.8 $\pm$ 0.7	3.7 $\pm$ 0.8	3.87
RL-CRAF++	4.3$\pm$0.5	4.0$\pm$0.6	3.9$\pm$0.7	4.07

Table 23: Expert conformance ratings (1–5 scale, full test set per language). $\pm$ denotes std across individual essays. GPT-5’s lower scores reflect Anglocentric register defaults that violate localized feedback norms; RL-CRAF++ achieves the highest conformance across all curricula.