

A Unified Framework to Elicit Structured Feedback for Interpretable Multi-Trait Essay Scoring

Shihang Yang^{1,2} Sanwoo Lee^{1,2} Ningning Zhao³ Yunfang Wu^{1,2*}

¹National Key Laboratory for Multimedia Information Processing, Peking University

²School of Computer Science, Peking University

³School of Chinese Language and Literature, Beijing Normal University

sh_yang@stu.pku.edu.cn, wuyf@pku.edu.cn

Abstract

Multi-trait Automated Essay Scoring (AES) requires rubric-grounded reasoning across interdependent traits, rather than isolated score prediction. Existing feedback-enhanced methods often decouple feedback from scoring or assess traits independently, weakening score-feedback consistency and rubric alignment. We propose **HiFTS**, a unified autoregressive framework that generates hierarchical CoT feedback before predicting trait-level and holistic scores. HiFTS distills rubric-grounded hierarchical CoT feedback from a teacher LLM and trains student models to jointly generate feedback and scores. HiFTS further applies Group Relative Policy Optimization with a composite reward balancing score agreement, calibration, feedback quality, and structural validity. At inference, a lightweight global prior provides holistic guidance to reduce drift during long-form reasoning. We also introduce **CFMS-34**, a Chinese multi-trait AES dataset with 951 essays annotated with holistic scores and 34 rubric-based traits. Experiments on CFMS-34 and ASAP++ show that HiFTS achieves strong holistic and trait-level scoring while producing coherent, rubric-aligned feedback. We release our data and code at <https://github.com/Atiyahsama/HiFTS>.

1 Introduction

Grounding writing assessment in multi-faceted feedback is essential for providing learners with constructive guidance and ensuring transparency of automated scoring. Traditionally, however, automated essay scoring (AES) models have focused on score-only predictions, pursuing high agreement with human raters at the cost of leaving the underlying decision process black-box to the end users (Taghipour and Ng, 2016; Yang et al., 2020; Wang and Liu, 2025). While human-crafted features add partial interpretability via linear models (Yupei and

Renfen, 2021) or post-hoc activation alignment (Fiacco et al., 2023), such model-centric explanations remain challenging for humans to interpret.

Driven by breakthroughs in large language models (LLMs) (Ouyang et al., 2022; OpenAI, 2023), recent AES studies have sought to mirror human-like evaluation by predicting scores alongside natural language rationales. While early attempts harnessed LLMs’ inherent reasoning potential via in-context learning (Yancey et al., 2023; Stahl et al., 2024) and sophisticated prompting (Wang et al., 2026a), they found a substantial gap in scoring accuracy compared to traditional supervised models (Lee et al., 2024). Consequently, studies have employed supervised fine-tuning (SFT) or reinforcement learning (RL), which demonstrate a narrowed accuracy gap while preserving LLMs’ strengths in generating rationales (Chu et al., 2025; Li and Pan, 2025; Do et al., 2026). Notably, these methods are increasingly evaluated on *multi-trait* AES tasks (Chu et al., 2025; Do et al., 2024b) to test their capacity for fine-grained evaluation across traits with intricate dependencies.

Despite the improvements, current supervised approaches lack a unified framework that generates multi-trait feedback and scores jointly in a single autoregressive process. For instance, feedback is generated from an external LLM to augment inputs for training score-only model (Chu et al., 2025). Methods that natively generate feedback exhibit either a lack of inter-trait dependencies due to independent trait-by-trait generation (Do et al., 2026), or a disconnect between the score and the rationale stemming from a separate regression head (Li and Pan, 2025).

Motivated by these insights, we propose **HiFTS** (Hierarchical Feedback-to-Score Reasoning for Multi-Trait Essay Scoring), a unified framework that reinforces LLMs to elicit structured multi-trait feedback and scores. Specifically, our hierarchical input template prompts the model to capture

*Corresponding author.

inter-trait relations by generating global-to-local rationales followed by rubric-aligned scores. After a warm-up SFT with CoT traces prompted by the template, we use group-relative policy optimization (GRPO) (Shao et al., 2024) to align the feedback traces with ground-truth trait scores as verifiable rewards. At inference time, the generation process is further conditioned on a prior score predicted by a lightweight BERT model (Devlin et al., 2019), which improves scoring stability and accuracy.

Furthermore, we present **CFMS-34**, a Chinese multi-trait AES dataset with 34 fine-grained trait scores annotated by human experts. This dataset addresses a major gap where recent multi-trait AES methods are predominantly validated and optimized on English benchmarks (Ridley et al., 2021; Do et al., 2024a; Wang and Liu, 2025). CFMS-34 serves as a rigorous testbed for cross-lingual evaluation, and we find that recent high-performing methods on the standard English benchmark, ASAP++ (Mathias and Bhattacharyya, 2018), do not necessarily succeed on CFMS-34. In contrast, experiments on both CFMS-34 and ASAP++ demonstrate that our proposed HiFITS improves both LLMs’ scoring accuracy and the quality of feedback. Fine-grained and ablation analyses further show that these gains come from hierarchical CoT supervision, GRPO-based alignment, and essay-specific prior guidance at inference.

Our contributions are summarized as follows:

- We construct and release **CFMS-34**, a fine-grained multi-trait Chinese AES dataset curated by expert annotators, promoting cross-lingual validation of multi-trait AES methods.
- We propose **HiFITS**, a unified LLM post-training framework that elicits structured feedback and multi-trait knowledge.
- During inference, we coordinate the LLM with a lightweight BERT model to enhance scoring accuracy and stability via prior-guided generation.

2 Related Work

Multi-trait essay scoring. Automated essay scoring (AES) has progressed from feature- and semantics-based scoring systems to neural and pre-trained language model based approaches (Foltz et al., 1999; Attali and Burstein, 2006; Taghipour and Ng, 2016; Yang et al., 2020; Wang et al.,

2022; Das et al., 2024). Multi-trait AES extends holistic scoring by predicting scores for rubric-defined dimensions, enabling more fine-grained assessment of writing quality (Mathias and Bhattacharyya, 2018, 2020; Ridley et al., 2021; Kumar et al., 2022). Recent studies improve multi-trait scoring with cross-prompt modeling, trait-aware representations, rubric-assisted features, mixture-of-experts, and graph-based trait interactions (Do et al., 2023; Wang and Liu, 2025; Eltanbouly et al., 2025; Li and Ng, 2025; Zhao et al., 2026). Another line of work casts multi-trait AES as autoregressive score generation, where language models generate structured trait scores and can be further optimized with score-related objectives (Do et al., 2024a,b). However, most approaches model trait dependencies implicitly, with shallow relation modules, or through order-sensitive score generation (Ridley et al., 2021; Do et al., 2023, 2024a; Li and Ng, 2025). Concurrent work studies trait-aware post-training for score-only autoregressive generation (Wang et al., 2026b); in contrast, HiFITS integrates hierarchical CoT feedback with score generation, so that trait-level scores are produced after global-to-local rubric-grounded reasoning.

Feedback-based scoring and alignment. Recent AES work explores LLMs for rubric-aware scoring and feedback generation through prompting, supervised training, and distillation (Yancey et al., 2023; Stahl et al., 2024; Lee et al., 2024; Wang et al., 2026a; Harada et al., 2026). To improve reliability, subsequent methods incorporate LLM-generated rationales or feedback into supervised scoring pipelines, jointly train scoring and feedback modules, or distill score-guided feedback from stronger teacher models (Chu et al., 2025; Li and Pan, 2025; Do et al., 2026). Despite their effectiveness, these methods often separate reasoning from scoring: feedback may be produced externally, numerical scores may be predicted by separate heads, or trait-level feedback may be generated independently. Post-training methods such as RLHF, preference optimization, and GRPO align generative models with task-specific, verifiable, or score-related rewards (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024; Do et al., 2024b). Our work follows this direction but aligns a single autoregressive process that produces structured multi-trait feedback then generates calibrated trait and holistic scores.

Dimension	Code	Definition / Trait Description
Content	C01–C03	Authenticity, sufficiency, and novelty of selected materials.
	C04–C05	Alignment with the central theme; depth of thought and originality.
	C06	Multi-perspective thinking and appropriateness of material presentation.
Structure	S01	Logical paragraph division and structural clarity.
	S02–S03	Use of foreshadowing, transitions, and cohesive devices.
	S04–S05	Structural coherence (e.g., “General-Specific-General”) and thread development.
	S06–S08	Relevance among paragraphs, sentences, and the central theme.
	S09–S10	Balance between detailed and concise writing (emphasis on key points).
	S11–S12	Quality of the opening (engagement) and ending (summary/insight).
Expression	E01–E02	Sentence fluency, grammatical correctness, and sophistication of word choice.
	E03–E04	Degree of descriptive detail and logical narrative order.
	E05–E07	Use of rhetorical devices, varied expression modes, and writing techniques.
	E08–E10	Observational perspective (holistic vs. focused) and narrative methods.
	E11–E12	Precision in capturing key characteristics of characters/events with clear prioritization.
Conventions	Cv01–Cv02	Standard formatting and accurate punctuation usage.
	Cv03–Cv04	Orthographic correctness and aesthetic handwriting.

Table 1: Definition of scoring traits in the CFMS-34 dataset.

3 CFMS-34 Dataset Construction

3.1 Data Collection and Curation

In this paper, we present **CFMS-34**, a Chinese multi-trait essay scoring dataset featuring high-level dimensions that encompass fine-grained sub-traits, constructed through a rigorous pedagogical assessment process. The dataset consists of 951 essays written by primary school students during in-class examinations. To ensure consistent writing conditions and reflect authentic classroom settings, all essays were completed in response to unit-level writing prompts from textbooks within a 50-minute time limit.

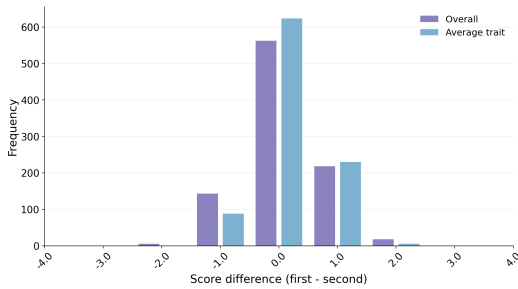


Figure 1: Distribution of score differences between the two raters.

Two trained experts in Chinese language education independently scored each essay based on strict guidelines. These guidelines define 34 fine-grained sub-traits categorized into four high-level dimensions: Content, Structure, Expression, and Conventions, as summarized in Table 1. Ultimately,

the annotation process provides 34 sub-trait scores and an overall score, all on a scale of 0 to 5. For 97.4% of the samples, the experts agreed on the overall score within one-point difference, and the human–human QWK on the overall score is 0.57, indicating a solid inter-rater consistency, as shown in Figure 1. Across the 34 sub-traits, the average agreement is 0.48 exact match, 0.90 within one point, and 0.41 QWK (see Appendix B for the full per-trait table).

Statistic	Value
Total Essays	951
Language	Chinese (Simplified)
Raters per Essay	2 (Human Experts)
Avg. Essay Length	492.4 Chinese characters
Score Dimensions	Overall + 34 Sub-traits
Score Range	0.0 – 5.0
Overall Score (mean $\pm$ std)	3.64 $\pm$ 0.74
Human–Human QWK (overall)	0.57

Table 2: Statistics of the CFMS-34 dataset.

We split CFMS-34 into training, development, and test sets using an 8:1:1 ratio. To ensure a reliable evaluation benchmark, the test set explicitly reserves samples with perfect inter-rater agreement on the overall score; this design is intended to reduce label noise in the test labels, rather than to simplify the task. The dataset statistics are shown in Table 2 for reference.

Dataset	Language	Expert Annotation	Dual Raters	Multi-trait	Fine-grained
EFCAMDAT (Geertzen et al., 2013)	EN	×	×	×	×
ICLE (Granger et al., 2009)	EN	✓	×	✓	×
TOEFL11 (Blanchard et al., 2013)	EN	✓	×	×	×
FCE (Yannakoudakis et al., 2011)	EN	✓	✓	×	×
ASAP++ (Mathias and Bhattacharyya, 2018)	EN	✓	✓	✓	✓
TOREE (Zhuang et al., 2024)	CN	✓	×	✓	×
CEDCC (Wu et al., 2023)	CN	✓	✓	✓	×
CFMS-34 (ours)	CN	✓	✓	✓	✓

Table 3: Comparison with existing essay scoring datasets.

3.2 Comparison with Existing Datasets

Table 3 compares CFMS-34 with representative English and Chinese essay scoring datasets. Like several prior datasets, CFMS-34 provides expert annotations, dual-rater scoring, and multi-trait labels. Its main distinction lies in the fine-grained rubric, which defines detailed sub-traits tailored to Chinese essay writing. This design enables more precise evaluation of writing quality across multiple dimensions and makes CFMS-34 a useful benchmark for multi-dimensional essay scoring.

3.3 Benchmark Performance of Frontier LLMs

To examine the difficulty of CFMS-34, we evaluate recent large language models on the overall scoring task, including GPT, Gemini, Doubao, Qwen, and DeepSeek series. We report quadratic weighted kappa (QWK) against human scores. The frontier-LLM baselines use the same zero-shot scoring prompt as Appendix D. We generate one score per sample using each provider’s default decoding settings and extract the numeric value from the [Overall Score] field.

Model	Mean Score	QWK
GPT-4.1 (OpenAI, 2025)	3.3475	0.4222
Gemini-3 (Google, 2025)	3.1937	0.4363
Doubao-Seed-1.6 (ByteDance, 2025)	2.0743	0.1595
Qwen3-max (Qwen Team, 2025)	3.4059	0.4201
DeepSeek-V3 (DeepSeek-AI, 2024)	3.5792	0.4130
DeepSeek-R1 (DeepSeek-AI, 2025)	3.0396	0.3271
Human (avg.)	3.64	0.57

Table 4: LLM performance on CFMS-34.

As shown in Table 4, current LLMs achieve only moderate agreement with human annotations, with the best QWK below 0.44. The Human QWK of 0.57 measures agreement between the two expert raters, whereas model QWKs are computed

against the resolved human scores; the two quantities are reported together for reference but are not strictly comparable. Their mean scores also reveal calibration issues, such as substantial under-prediction by Doubao-Seed-1.6 and slight over-prediction by several other models. These results suggest that CFMS-34 remains challenging for current LLMs and motivates structured reasoning methods beyond flat score prediction.

4 Method

As shown in Fig. 2, **HiFTS** formulates multi-trait AES as a unified autoregressive process that generates rubric-aware hierarchical feedback before trait-level and holistic scores. Training first uses teacher-generated CoT supervision for SFT (Section 4.1) and then applies GRPO with a composite reward (Section 4.2); inference further conditions decoding on a BERT-based global prior to improve score consistency (Section 4.3).

4.1 Warmup SFT for Hierarchical Feedback-to-Score Generation

Directly applying RL with only final score-level rewards may fail to incentivize the model to faithfully capture cross-trait dependencies within its reasoning traces. To address this, we introduce a warmup SFT stage prior to RL, designed to internalize the capacity for structured hierarchical reasoning. Using Gemini-3 (Google, 2025) as a teacher model, we construct CoT feedback following a top-down hierarchy: **global understanding of essay** → **analysis of a dimension** → **sub-trait evaluation**. To assure alignment of the generated feedback with the rubric guidelines and labeled scores, prompting the teacher model is explicitly conditioned on the essay, the rubric definitions, and the labeled scores, see Appendix C for full template.

Subsequently, the generated feedback is ap-

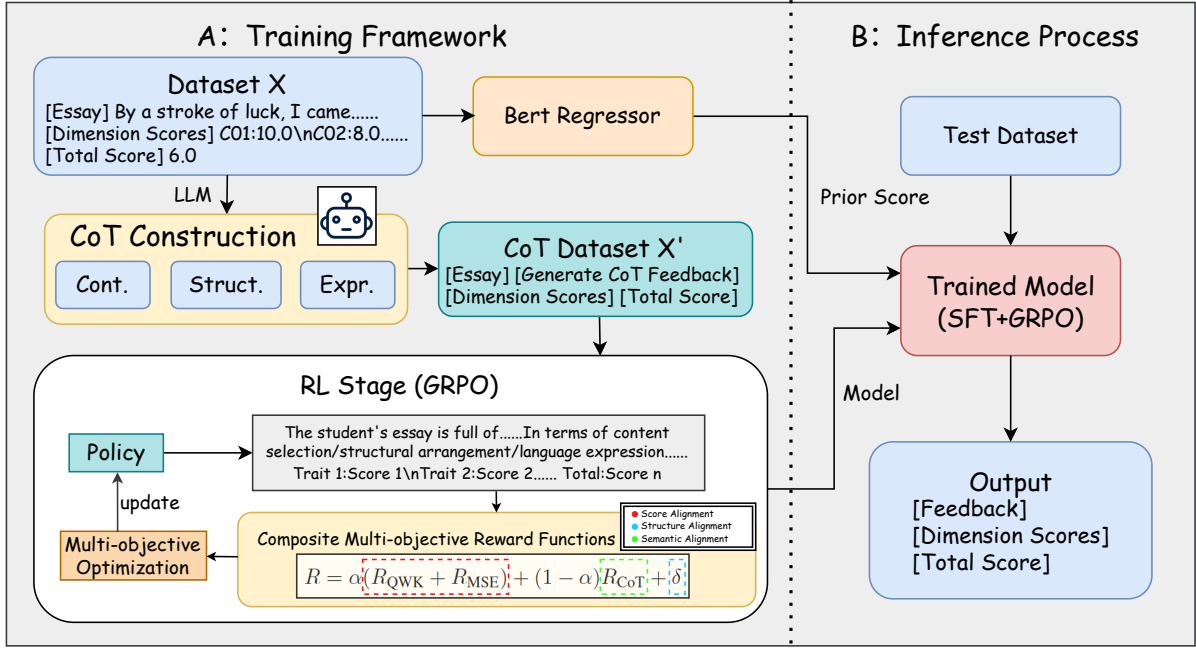


Figure 2: Overview of HiFITS. HiFITS first constructs hierarchical CoT feedback with an LLM teacher, then trains the student model with SFT and GRPO. During inference, a BERT-based global prior guides feedback generation and score prediction for more consistent multi-trait AES.

pendent with the labeled trait-level and overall scores to construct a unified feedback-to-score target sequence. We fine-tune the student model on these sequences using an instructional prompt that can frame the target feedback-to-score sequence as its natural continuation. The corresponding template is provided in Appendix D. To be concrete, the target sequence is the concatenation of the feedback and the labeled scores:

$$Y = [Y_{\text{CoT}}; Y_{\text{Score}}], \quad (1)$$

and we use the standard token-level cross-entropy loss during SFT:

$$\mathcal{L}_{\text{SFT}} = - \sum_{t=1}^T \log P_{\theta}(y_t | y_{<t}, X), \quad (2)$$

where X is the input essay and y_t is the target token at step t . To summarize, this warmup SFT stage ensures the student’s CoT trace tightly anchors its eventual scoring decisions, serving as an initialization point for subsequent RL stage.

4.2 Boosting Feedback Quality and Scoring Accuracy via Reinforcement Learning

To jointly optimize scoring precision, feedback quality, and formatting compliance, we define a comprehensive multi-objective reward function R ,

formulated as:

$$R = \alpha(R_{\text{QWK}} + R_{\text{MSE}}) + (1 - \alpha)R_{\text{CoT}} + \delta, \quad (3)$$

where $\alpha \in [0, 1]$ serves as a balancing coefficient that controls the trade-off between score-oriented alignment and feedback-oriented alignment. The structural term δ is a fixed rule-based bonus or penalty and does not introduce additional hyperparameters beyond α .

R_{QWK} captures both overall- and trait-level agreement with human annotations. We compute holistic QWK over a short sliding window of recent generations to stabilize the reward, and compute trait-level QWK over the traits:

$$R_{\text{QWK}} = \kappa(\hat{\mathbf{s}}_{t-h+1:t}, \mathbf{s}_{t-h+1:t}) + \kappa(\hat{\mathbf{z}}_t^{(m)}, \mathbf{z}_t^{(m)}) \quad (4)$$

where κ denotes QWK, $h = 64$ is the sliding-window size over the most recent generations, and $\hat{\mathbf{z}}_t^{(m)}$ and $\mathbf{z}_t^{(m)}$ denote the predicted and gold scores over the m traits.

R_{MSE} is defined as the negative mean squared error between the predicted and gold scores, thereby penalizing larger numerical deviations and encouraging accurate score prediction.

R_{CoT} measures semantic alignment between generated feedback and teacher CoT references. We compute this reward as the cosine similarity

between sentence embeddings of generated and teacher feedback:

$$R_{\text{CoT}} = \cos(f(\hat{c}_t), f(c_t)) = \frac{f(\hat{c}_t)^\top f(c_t)}{\|f(\hat{c}_t)\| \|f(c_t)\|}, \quad (5)$$

where $\hat{c}_t$ and c_t denote the generated and teacher feedback, respectively, and $f(\cdot)$ is instantiated as bge-small-zh-v1.5 for CFMS-34 and all-MiniLM-L6-v2 for ASAP++.

This formulation balances scoring accuracy and feedback quality without introducing multiple reward weights. The value of α is selected on the development set, as discussed in Section 6.5.

We adopt Group Relative Policy Optimization (GRPO) for reinforcement learning. For each prompt, a group of candidate responses is sampled and evaluated with the reward function above, and the policy is updated according to their group-relative rewards.

4.3 Robust Inference via Global Prior Guidance

Long-form hierarchical CoT generation can suffer from semantic drift, where the model gradually departs from its initial assessment as the reasoning chain unfolds. This is particularly problematic for multi-trait AES, since trait-level judgments should remain consistent with the overall writing quality.

To mitigate this issue, we introduce a lightweight global prior at inference time. We use a BERT-based regressor to estimate holistic essay quality. Given an essay X , the regressor produces a coarse prior score S_{prior} , which is inserted into the system prompt before decoding with the GRPO-aligned LLM as guidance.

Importantly, S_{prior} is not used as the final prediction. Instead, it serves as a soft semantic anchor that guides the model toward a plausible scoring region, while still allowing the LLM to generate rubric-grounded feedback and final scores autoregressively. This strategy introduces no additional decoding-time hyperparameters.

Formally, the generation process is conditioned on both the essay and the prior score:

$$P(Y_C, S \mid X, S_{\text{prior}}) = P(Y_C \mid X, S_{\text{prior}}) \cdot P(S \mid Y_C, X, S_{\text{prior}}), \quad (6)$$

where Y_C denotes the generated hierarchical feedback and S denotes the predicted trait-level and holistic scores.

By anchoring generation with an initial estimate of holistic quality, prior-guided decoding reduces inconsistencies between local trait analyses and final scoring decisions.

5 Experimental Setup

Datasets. We evaluate HiFITS on CFMS-34 and ASAP++ (Mathias and Bhattacharyya, 2018), covering both Chinese and English multi-trait essay scoring. For CFMS-34, we use the 20 core traits identified in our rubric analysis as the primary reasoning targets, with scores obtained by summing the two raters’ 0–5 annotations into a 0–10 scale for training and evaluation. The same 20-trait subset (Appendix A) is applied to all methods, including baselines, while the released data retain all 34 trait annotations. ASAP++ extends the ASAP benchmark with human-annotated trait scores for English essays across eight prompts, enabling evaluation of both holistic and trait-level scoring.

Evaluation Metrics. As shown in Table 5, we report quadratic weighted kappa (QWK) for scoring agreement: *Overall* denotes holistic QWK on CFMS-34, *Prompts* denotes the average holistic QWK over the eight ASAP++ prompts, and *Traits* denotes average trait QWK. We also report MSE on CFMS-34 for numerical calibration. For reasoning quality, we report *WinRate* from pairwise comparisons judged by DeepSeek-V3.2, used only between each SFT model and its GRPO-aligned counterpart on each dataset.

Training Details. We fine-tune Qwen2.5-7B and Qwen3-4B with a two-stage pipeline: supervised fine-tuning on hierarchical CoT data followed by GRPO-based alignment. For global prior guidance, we train a BERT-based regressor initialized from bert-base-chinese for CFMS-34 and bert-base for ASAP++ to estimate holistic essay quality. During GRPO, we sample $G = 4$ responses per input, set the KL coefficient to $\beta = 0.02$, and use a learning rate of 1×10^{-6} . For the composite reward, we set $\alpha = 0.8$ based on development-set performance and use a fixed structural bonus/penalty of $\delta = 0.1$. Results are reported from one fixed-seed run, with checkpoints selected by development-set holistic QWK.

Since QWK is a set-level metric, we compute the holistic QWK reward over a sliding window of the most recent $h = 64$ generations during GRPO. At inference time, the model generates structured

Method	Explain.	Overall	CFMS-34			ASAP++		
			Traits	MSE	WinRate	Prompts	Traits	WinRate
HISK	×	0.405	0.309	1.374	-	0.644	0.611	-
STL-LSTM	×	0.484	0.365	1.454	-	0.684	0.656	-
MTL-BiLSTM	×	0.519	0.396	1.306	-	0.665	0.638	-
ArTS	×	0.264	0.149	1.960	-	0.717	0.695	-
RMTS	✓	0.296	0.303	1.431	-	0.720	0.704	-
HiFITS-SFT _{Qwen2.5}	✓	0.528	0.387	1.447	Ref.	0.712	0.674	Ref.
HiFITS _{Qwen2.5}	✓	0.656	0.439	0.957	0.673	0.718	0.704	0.550
HiFITS-SFT _{Qwen3}	✓	0.557	0.395	1.379	Ref.	0.720	0.685	Ref.
HiFITS _{Qwen3}	✓	0.677	0.453	0.741	0.706	0.726	0.711	0.613

Table 5: Main scoring and feedback results on CFMS-34 and ASAP++.

outputs with [Analysis], [Trait Scores], and [Overall Score] fields, conditioned on the essay, rubric prompt, and, when enabled, the global score prior. We decode with temperature 0.6 and top- $p = 0.9$ for all runs.

Baselines. We compare HiFITS with representative essay scoring models, including HISK (Cozma et al., 2018), STL-LSTM (Dong et al., 2017), MTL-BiLSTM (Kumar et al., 2022), and ArTS (Do et al., 2024a). We also include RMTS (Chu et al., 2025), a rationale-enhanced multi-trait scoring model that uses trait-wise rationales generated by an external LLM. For ASAP/ASAP++, we follow Do et al. (2024b) and report published results for traditional baselines when the evaluation settings are comparable to ours.

6 Results and Analysis

6.1 Scoring Performance

Table 5 reports results on CFMS-34 and ASAP++. Although existing baselines remain competitive on ASAP++, their performance drops markedly on CFMS-34, suggesting that Chinese multi-dimensional essay scoring poses additional challenges. In contrast, HiFITS performs consistently across both datasets, with GRPO further improving CFMS-34 Overall QWK, Traits QWK, and MSE, as well as ASAP++ prompt-averaged and trait-level QWK. The Qwen3-based model with GRPO achieves the best results, reaching 0.677 Overall QWK and reducing MSE to 0.741. These results indicate that hierarchical CoT supervision provides a strong foundation for rubric-aware scoring, while reward-based alignment improves consistency and calibration.

6.2 Fine-grained Trait Analysis

Figure 3 shows trait-level QWK on the 20 evaluated CFMS-34 traits. GRPO-aligned models generally outperform their SFT counterparts, with the Qwen3-based variant achieving the highest average trait QWK and consistent gains on most dimensions. This indicates that reward-based alignment improves fine-grained rubric judgments in addition to holistic scoring.

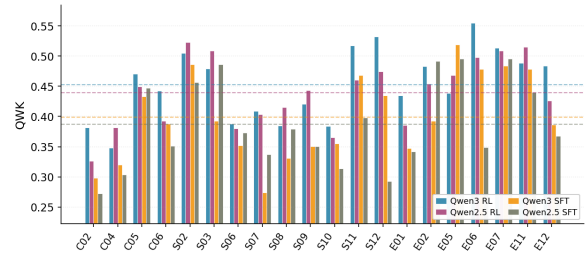


Figure 3: Trait-level QWK on CFMS-34, where dashed lines denote averages values.

Table 6 reports trait-level QWK on ASAP++ with Qwen2.5. PPO and GRPO improve most traits over the neural baselines, and prior guidance helps further; HiFITS attains the best mean QWK, indicating that the alignment strategy also generalizes to English multi-dimensional scoring.

6.3 Feedback Assessment

Beyond scoring accuracy, Table 5 shows that GRPO-aligned variants are preferred over their SFT counterparts in WinRate, especially on CFMS-34. This suggests that GRPO improves not only score prediction but also the coherence, rubric alignment, and logical consistency of feedback. The stronger WinRate of the Qwen3-based model further indicates that stronger base reasoning ability benefits from the proposed reward design.

Method	Over.	Cont.	PA	Lang.	Nar.	Org.	Conv.	WC	SF	Style	Voice	Mean
HISK	0.718	0.679	0.697	0.605	0.659	0.610	0.527	0.579	0.553	0.609	0.489	0.611
STL-LSTM	0.750	0.707	0.731	0.640	0.699	0.649	0.605	0.621	0.612	0.659	0.544	0.656
MTL-BiLSTM	0.764	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.638
ArTS	0.751	0.730	0.751	0.698	0.725	0.672	0.668	0.679	0.678	0.721	0.570	0.695
HiFITS-PPO w/o prior	0.751	0.715	0.730	0.662	0.708	0.645	0.640	0.652	0.625	0.690	0.565	0.671
HiFITS w/o prior	0.754	0.722	0.745	0.685	0.712	0.670	0.635	0.668	0.662	0.688	0.597	0.685
HiFITS-PPO	0.755	0.728	0.741	0.695	0.725	0.668	0.675	0.678	0.680	0.712	0.615	0.697
HiFITS	0.758	0.734	0.749	0.702	0.724	0.681	0.684	0.686	0.688	0.715	0.626	0.704

Table 6: Trait-level QWK on ASAP++ with Qwen2.5. The best result for each trait is shown in bold.

Flat feedback (DeepSeek-Reasoner)	Hierarchical CoT feedback (Ours)
<p>This essay revolves around the interaction between “me” and the hamster, with genuine emotions, but it is overall quite simple, lacking in content richness and loose in structure.</p> <p>The opening directly introduces the hamster’s appearance and habits, which is concise but lacks appeal.</p> <p>The main body describes playing, feeding, breeding, and changes in the relationship, but the material is organized haphazardly, with a lack of natural transitions between paragraphs—for example, from “I loved watching it eat sunflower seeds” to “it gave birth to nine babies,” resulting in a logical leap.</p> <p>Detail description is insufficient; for instance, the process of eating sunflower seeds is attempted but not vivid enough.</p> <p>While emotional expression is sincere, the depth of thought is lacking, failing to explore deeper insights.</p> <p>The ending suddenly confesses feelings, appearing somewhat rushed. The language is fluent but plain, with basically no rhetorical devices or varied expressions.</p> <p>Observation follows a certain order but lacks clear prioritization. Overall ...</p>	<p>L1 GLOBAL ESSAY UNDERSTANDING ⇒ The student essay “Little Hamster” is ...</p> <p>L2 DIMENSION ANALYSIS Content ⇒ The essay includes ... These materials are ... (C02), and they are ... theme (C04). The author ... in a ... order (E12). In particular, the description of ... is vivid and expressive (E11). The essay also elaborates ... engaging. However, the essay’s ... lacks deeper reflection (C05), and the integration between multi-perspective content and material selection ... (C06), while handling of transitional materials appears somewhat rigid (S10).</p> <p>Structure ⇒ In terms of structural organization, the essay begins with ... (S11), and the conclusion remains ... (S12). The content of ... (S06), and ... maintain a connection to topic sentences (S07). The preceding and subsequent content is also presented with ... (S02). However, transitions between ... (S03), resulting in ... (S08).</p> <p>Expression ⇒ ...</p>

Table 7: A qualitative case study comparing flat feedback and hierarchical CoT feedback on the same essay.

Because pairwise WinRate relies on an LLM judge, we further report a rule-based grounding metric that checks whether the generated feedback is tied to rubric traits and essay evidence, calculated by: $0.5 \cdot \text{Coverage} + 0.3 \cdot \text{Trait-level Reasoning} + 0.2 \cdot \text{Reasoning Density}$. Coverage is the fraction of scoring traits mentioned in the feedback; trait-level reasoning checks whether each mentioned trait is supported by essay evidence and an evaluative judgment; reasoning density is the fraction of sentences that combine description, explanation, and evaluation. Table 8 shows that reward-aligned variants obtain higher grounding scores than SFT, and full HiFITS performs best.

Table 7 compares DeepSeek-Reasoner’s flat feedback with HiFITS’s hierarchical CoT on the

Method	Grounding
HiFITS-SFT	0.60
HiFITS-PPO w/o prior	0.72
HiFITS-PPO	0.74
HiFITS w/o prior	0.74
HiFITS	0.78

Table 8: Rule-based grounding scores of HiFITS-generated feedback on CFMS-34 with Qwen2.5.

same essay. While the flat rationale is relevant, the highlighted annotations indicate that it shifts abruptly across dimensions and lacks clear transitions. In contrast, HiFITS follows a global-to-local process, from overall understanding to rubric dimensions and sub-traits, with highlighted connec-

tive expressions linking different dimensions, yielding more cohesive, inspectable, and rubric-aligned feedback.

6.4 Ablation Study

Table 9 analyzes reward-based alignment and prior guidance. The BERT prior performs moderately, indicating that it is insufficient as a standalone scorer. However, during LLM inference, it provides coarse quality signals that stabilize hierarchical reasoning, serving as guidance rather than final prediction.

Method	Overall	Traits	WinRate
BERT prior	0.489	-	-
Qwen zero-shot	0.235	0.154	-
Qwen sft w/o CoT	0.432	0.193	-
HiFTS-SFT	0.528	0.387	Ref.
HiFTS-PPO w/o prior	0.572	0.417	0.603
HiFTS-PPO	0.647	0.407	0.651
HiFTS w/o prior	0.579	0.420	0.621
HiFTS w/ rand. prior	0.502	0.298	-
HiFTS w/ const. prior	0.411	0.252	-
HiFTS	0.656	0.439	0.673

Table 9: Ablation results on CFMS-34 with Qwen2.5. WinRate is computed against SFT feedback.

Comparing Qwen sft w/o CoT with HiFTS-SFT shows that hierarchical feedback-to-score generation improves scoring even without RL or the BERT prior. Both prior-free PPO and GRPO variants outperform HiFTS-SFT in Overall QWK, confirming the benefit of reward-based alignment. The GRPO variant also yields a higher WinRate than PPO, indicating better feedback quality. Adding the prior further improves Overall QWK and WinRate, with HiFTS achieving the best results. By contrast, random and constant priors degrade performance, showing that guidance must be essay-specific and meaningful.

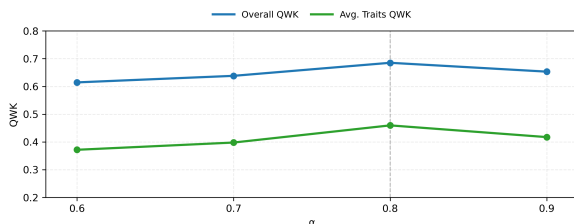


Figure 4: Effect of α on holistic and average trait QWK on the CFMS-34 development set.

6.5 Hyperparameter Analysis

Figure 4 analyzes the effect of the reward trade-off coefficient α . Smaller values place more emphasis on feedback alignment, while larger values focus more on score-oriented optimization. We select α on the CFMS-34 development set based on holistic and average trait QWK, which reflect the final agreement of the feedback-to-score process. Among the tested values, $\alpha = 0.8$ achieves the best trade-off and is used in the main experiments.

Table 10 analyzes the sliding-window size h in Eq. (4), which estimates holistic QWK over the most recent h generations. We vary only h on CFMS-34 with HiFTS and Qwen2.5, keeping all other settings fixed. Smaller windows yield less stable rewards, since QWK is a set-level metric. $h = 128$ performs slightly best, but $h = 64$ is close; we use $h = 64$ in the main experiments.

Window size	Overall	Traits
16	0.612	0.422
32	0.641	0.432
64	0.656	0.439
128	0.659	0.443

Table 10: Ablation of the sliding-window size h for the holistic QWK reward on CFMS-34 with Qwen2.5.

7 Conclusion

In this work, we introduced CFMS-34, an expert-annotated Chinese multi-trait AES dataset with dual ratings, holistic scores, and 34 fine-grained rubric-based traits, providing a challenging benchmark for interpretable writing assessment beyond English settings. We also proposed HiFTS, a unified autoregressive framework that first generates global-to-local rubric-grounded feedback and then derives holistic and trait-level scores. HiFTS is trained to learn teacher-generated hierarchical feedback and further aligned with GRPO under a composite reward, while a lightweight BERT-based global prior anchors generation at inference time to improve scoring stability.

Experiments on CFMS-34 and ASAP++ show that HiFTS improves holistic and average trait QWK, reduces MSE on CFMS-34, and produces more coherent rubric-aligned feedback. Ablation results further verify the benefits of GRPO alignment and prior-guided decoding, suggesting that feedback-to-score reasoning is a promising direction for interpretable multi-trait essay scoring.

Limitations

This work has two main limitations. First, HiFITS adopts a predefined global-to-local autoregressive reasoning order for generating hierarchical feedback and scores. Although this structure is effective in our experiments and aligns with rubric-based assessment, exploring alternative generation orders or more flexible feedback organizations may bring further improvements. Second, our evaluation is based on CFMS-34 and ASAP++, covering Chinese and English multi-trait AES settings. To better assess the scalability and educational impact of HiFITS, future studies are needed on more writing genres, learner populations, and real classroom interactions, where students and educators can directly engage with the generated rubric-grounded feedback.

Ethics Statement

Use of Scientific Artifacts We use ASAP++ as a publicly available AES benchmark, consistent with its intended research use, and use pretrained models and API-based LLMs under their respective licenses or service terms. The creators of the datasets, models, and baselines used in this work are cited in the corresponding sections.

CFMS-34 Data, Annotators, Privacy, and Intended Use CFMS-34 is constructed for research on multi-trait automated essay scoring and rubric-grounded feedback generation. The essays were collected from routine educational assessment settings with authorization from the responsible educational institution, and appropriate consent and/or guardian consent procedures were followed according to institutional requirements. The expert raters were Chinese language education specialists from our collaborating education team rather than crowdworkers; they annotated essays according to the predefined rubric described in Section 3 and Table 1, and their participation followed the corresponding institutional or collaboration arrangements. All essays were anonymized before annotation, modeling, and release preparation, including removal of direct personal identifiers and screening for potentially identifying content. We release only anonymized essay texts and score annotations under research-use terms, and do not release raw examination records or handwriting images that may contain additional personal information. The dataset and system are intended for research and

educational support, not as the sole basis for high-stakes educational decisions without human oversight.

AI Assistants in Writing AI assistants were used solely to improve the clarity and coherence of the writing. The ideas and experiments presented in this paper are original to the authors.

Acknowledgements This work is supported by the National Natural Science Foundation of China (62677001), and the Fundamental Research Funds for the Central Universities, Peking University.

References

- Yigal Attali and Jill Burstein. 2006. [Automated essay scoring with e-rater® v.2](#). *The Journal of Technology, Learning and Assessment*, 4(3).
- Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. [TOEFL11: A corpus of non-native english](#). ETS Research Report RR-13-24, Educational Testing Service.
- ByteDance. 2025. Seed1.6 technical introduction. https://research.doubao.com/zh/seed1_6.
- SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. [Rationale behind essay scores: Enhancing S-LLM’s multi-trait essay scoring with rationale generated by LLMs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5811–5829.
- Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. [Automated essay scoring with string kernels and word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509.
- Sourya Dipta Das, Yash A. Vadi, and Kuldeep Yadav. 2024. [Transformer-based joint modelling for automatic essay scoring and off-topic detection](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16751–16761.
- DeepSeek-AI. 2024. [Deepseek-v3 technical report](#). *arXiv preprint arXiv:2412.19437*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Nature*, 645:633–638.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- 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, Yunsu Kim, and Gary Geunbae Lee. 2023. [Prompt- and trait relation-aware cross-prompt essay trait scoring](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551.
- 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.
- Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. 2026. [Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring](#). *Expert Systems with Applications*, 319:132119.
- 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.
- Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. [TRATES: Trait-specific rubric-assisted cross-prompt essay scoring](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20528–20543.
- James Fiaccio, David Adamson, and Carolyn Rose. 2023. [Towards extracting and understanding the implicit rubrics of transformer based automatic essay scoring models](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 232–241.
- Peter Foltz, Darrell Laham, and T. Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*.
- Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, and 1 others. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadilla Proceedings Project, pages 240–254.
- Google. 2025. Gemini 3 developer guide. <https://ai.google.dev/gemini-api/docs/gemini-3>.
- S Granger, E Dagneaux, F Meunier, and M Paquot. 2009. The international corpus of learner english version 2. handbook and cd-rom. louvain-la-neuve: Presses universitaires de louvain.
- Keno Harada, Lui Yoshida, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2026. [Automated refinement of essay scoring rubrics for language models via reflect-and-revise](#). In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pages 771–789.
- Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. [Many hands make light work: Using essay traits to automatically score essays](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495.
- 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. 2025. [Graph-based multi-trait essay scoring](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33325–33351.
- Xia Li and Wenjing Pan. 2025. [CEAES: Bidirectional reinforcement learning optimization for consistent and explainable essay assessment](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26267–26279.
- Sandeep Mathias and Pushpak Bhattacharyya. 2018. [ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Sandeep Mathias and Pushpak Bhattacharyya. 2020. [Can neural networks automatically score essay traits?](#) In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91.
- OpenAI. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- OpenAI. 2025. GPT-4.1 Model Documentation. <https://developers.openai.com/api/docs/models/gpt-4.1>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*.
- Qwen Team. 2025. Qwen3-max: Big is good. <https://qwen.ai/blog?id=qwen3-max>.

- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.
- Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. [Automated cross-prompt scoring of essay traits](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:13745–13753.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. [Exploring LLM prompting strategies for joint essay scoring and feedback generation](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298.
- 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.
- Jiong Wang and Jie Liu. 2025. [T-MES: Trait-aware mix-of-experts representation learning for multi-trait essay scoring](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236.
- 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.
- Yun Wang, Zhaojun Ding, Xuansheng Wu, Siyue Sun, Ninghao Liu, and Xiaoming Zhai. 2026a. [Autoscore: Enhancing automated scoring with multi-agent large language models via structured component recognition](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):40898–40906.
- Zhengyang Wang, Sanwoo Lee, Jiaxin Wang, Chenxi Miao, Weikang Li, and Yunfang Wu. 2026b. Trait-aware policy optimization for autoregressive multi-trait essay scoring. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*.
- Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. [A multi-task dataset for assessing discourse coherence in Chinese essays: Structure, theme, and logic analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688.
- Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. [Rating short L2 essays on the CEFR scale with GPT-4](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.
- 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.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.
- Wang Yupei and Hu Renfen. 2021. [A prompt-independent and interpretable automated essay scoring method for chinese second language writing](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1202–1217.
- Shiman Zhao, Siyuan Liu, and Zhiqi Shen. 2026. [Multi-knowledge Enhanced Graph Neural Network for Multi-trait Essay Scoring](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28733–28741.
- Xinlin Zhuang, Hongyi Wu, Xinshu Shen, Peimin Yu, Gaowei Yi, Xinhao Chen, Tu Hu, Yang Chen, Yupei Ren, Yadong Zhang, Youqi Song, Binxuan Liu, and Man Lan. 2024. [TOREE: Evaluating topic relevance of student essays for Chinese primary and middle school education](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5749–5765.

A Selection of 20 Core Traits

Although the original CFMS-34 annotations contain 34 fine-grained traits (Table 1), we select the 20 traits that exhibit the highest correlation with the overall score as the core reasoning subset for modeling and evaluation.

This design is motivated by two considerations: (1) in real-world grading scenarios, holistic scores are typically dominated by several key traits, while certain traits contribute marginally and may introduce noise; (2) incorporating too many traits substantially increases the length of the reasoning chain and the complexity of structural constraints, thereby reducing stability in rubric adherence.

Specifically, we compute the Pearson correlation coefficient between each trait score and the overall score on the training set, and select the top 20 traits with the highest correlations.

The selected traits are listed as follows:

C02: Able to select concrete and sufficient supporting materials
C04: Able to select and organize materials around the central theme
C05: Able to produce writing with originality and unique insights (depth of thought)
C06: Able to approach the topic from multiple perspectives and select appropriate materials
S02: Able to clearly present preceding and subsequent content
S03: Able to use transitional words and sentences to connect adjacent paragraphs
S06: Each paragraph is closely connected to the main theme
S07: Each sentence is closely connected to the topic sentence of the paragraph
S08: Paragraph-to-paragraph relationships are coherent and tightly connected
S09: Able to elaborate on key parts of the essay in detail
S10: Able to summarize or briefly address transitional content
S11: Able to craft an appropriate introduction
S12: Able to provide a concise and effective conclusion
E01: Sentences are fluent and semantically complete
E02: Word choice is appropriate, vivid, and expressive
E05: Comprehensive use of multiple rhetorical devices
E06: Comprehensive use of multiple modes of expression
E07: Able to employ diverse expressive techniques
E11: Attention to details of characters and events
E12: Observation is structured, prioritized, and sequential

B Per-trait Score Statistics and Inter-rater Agreement

Table 11 reports per-trait score statistics and dual-rater agreement on all 951 CFMS-34 essays. Mean and standard deviation are computed over the pooled 0–5 ratings of both experts. Exact is the proportion of identical ratings, Within-1 is agreement within one point, and QWK is quadratic weighted kappa. The 20 traits used in the main experiments (Appendix A) are marked with †.

C Example of CoT Generation Prompt Design

The following prompt is used to guide the teacher model to generate rubric-grounded hierarchical feedback. The scoring rubric follows the core trait definitions in Appendix A.

You are an experienced Chinese language teacher with many years of essay grading experience. Please read the student essay and the given trait scores (0 to 10), and generate a hierarchical evaluative comment organized into three levels:

[Hierarchical Structure]

1. Content: Analyze indicators related to material selection, theme alignment, originality, depth of thought, and multi-perspective thinking.
2. Structure: Analyze paragraph organization, transitions, logical sequencing, coherence, and the handling of the introduction and conclusion.
3. Expression: Analyze sentence fluency, word choice, rhetorical devices, modes of expression, expressive techniques, descriptive details, and observation order.

[Core Requirements]

- Scope of Analysis: You must analyze only the indicators explicitly provided with scores in the input, such as C02, S02, or E01. The use of generalized category codes, such as Cxx, Sxx, or Exx, is strictly prohibited.
- Indicator Anchoring: Whenever a specific evaluation point is discussed, append the corresponding indicator code in parentheses at the end of the sentence, without spaces, e.g., (C02).
- Structured Output: The analysis must strictly follow the order Content -> Structure -> Expression.
- Style Constraint: Encouraging or congratulatory language, such as "Keep up the good work", is strictly prohibited.
- Formatting Constraint: Use natural language for analysis without bold formatting, bullet points, or special symbols in the final response.

[Scoring Rubric Reference]: Appendix A.

Use only the provided core trait definitions and their corresponding scores.

D Example of Scoring System Prompt Design

You are a primary school Chinese language teacher with many years of grading experience. Your task is to read the student essay and strictly follow the provided scoring rubric and evaluation procedure to generate a hierarchical evaluative comment.

Trait	Mean	Std	Exact	Within-1	QWK
Overall	3.64	0.74	0.59	0.97	0.57
C01	4.08	0.86	0.40	0.87	0.25
C02 [†]	3.96	0.84	0.43	0.87	0.32
C03	3.63	0.81	0.45	0.90	0.34
C04 [†]	3.81	0.79	0.50	0.89	0.33
C05 [†]	3.23	0.86	0.49	0.93	0.52
C06 [†]	3.59	0.79	0.52	0.93	0.41
S01	3.56	0.89	0.50	0.90	0.50
S02 [†]	3.42	0.80	0.52	0.96	0.52
S03 [†]	3.41	0.82	0.52	0.95	0.54
S04	3.70	0.83	0.49	0.92	0.45
S05	3.71	0.91	0.47	0.88	0.45
S06 [†]	3.66	0.90	0.49	0.89	0.48
S07 [†]	3.38	0.85	0.46	0.92	0.46
S08 [†]	3.49	0.86	0.53	0.94	0.56
S09 [†]	3.55	0.89	0.46	0.89	0.43
S10 [†]	3.29	0.94	0.40	0.84	0.32
S11 [†]	3.40	0.94	0.45	0.88	0.45
S12 [†]	3.43	0.90	0.47	0.90	0.46

Trait	Mean	Std	Exact	Within-1	QWK
E01 [†]	3.88	0.84	0.48	0.90	0.42
E02 [†]	3.36	0.87	0.52	0.92	0.50
E03	3.44	0.83	0.53	0.92	0.50
E04	3.68	0.81	0.50	0.91	0.42
E05 [†]	2.94	0.82	0.51	0.93	0.41
E06 [†]	3.02	0.82	0.47	0.91	0.35
E07 [†]	2.92	0.79	0.49	0.92	0.36
E08	3.67	0.81	0.55	0.91	0.46
E09	2.56	0.78	0.69	0.96	0.64
E10	3.58	0.79	0.52	0.90	0.38
E11 [†]	3.53	0.84	0.54	0.90	0.45
E12 [†]	3.69	0.83	0.47	0.87	0.32
Cv01	4.21	0.85	0.31	0.79	-0.01
Cv02	4.22	0.82	0.36	0.87	0.17
Cv03	3.94	0.88	0.52	0.90	0.49
Cv04	3.55	0.93	0.36	0.77	0.12

Table 11: Per-trait score statistics and inter-rater agreement on CFMS-34 (951 essays). [†] denotes the 20-trait subset used in the main experiments.

[Scoring Rubric Reference]: Appendix A.

[Evaluation Procedure]

1. Analysis: First, provide a detailed analysis of the essay across all fine-grained dimensions according to the scoring rubric. The analysis must follow the hierarchical logic of each dimension.
2. Sub-scores: Second, provide the specific scores (0-10) for all 20 fine-grained dimensions.
3. Overall Score: Finally, provide the total score of the essay.

[Output Format]

The output must contain exactly the following three top-level tags:

1. [Analysis]: Provide detailed chain-of-thought analysis and hierarchical evaluative comments.
2. [Trait Scores]: Provide all fine-grained scores line by line (e.g., C02:7.0).
3. [Overall Score]: Provide the final overall score of the essay.

Strict adherence to the [Output Format] requirements is mandatory.