PsyScore: A Psychometrically-Aware Framework for Trait-Adaptive Essay Scoring and ZPD-Scaffolded Feedback

Wei Xia[1]*, Jin Wu[2,3]*, Haoran Shi[1], Xiangyu Wang[1], Chanjin Zheng[2]†
[1]Department of Educational Psychology, East China Normal University,
[2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University,
[3]School of Computer Science and Technology, East China Normal University
{51264118006, 52275901018}@stu.ecnu.edu.cn, chjzheng@dep.ecnu.edu.cn

Abstract

Effective Automated Essay Scoring (AES) are expected to support both reliable assessment and actionable instructional feedback. However, existing approaches often treat scoring and feedback as separate components: neural scoring models provide limited interpretability, while Large Language Model (LLM)-based feedback is typically insensitive to learners proficiency levels. To address this fragmentation, this work proposes **PsyScore**, a psychometrically-aware framework that integrates diagnostic assessment with instructional scaffolding through a shared latent ability representation. PsyScore comprises three key modules: a **Trait-Adaptive Neural IRT Scorer** that incorporates the Graded Partial Credit Model (GPCM) into a neural architecture, enabling the precise estimation of student ability while maintaining psychometric interpretability, a **ZPD-Scaffolded Feedback Generator**, which conditions multi-agent feedback strategies on the diagnosed ability parameter to adapt instructional focus across different proficiency levels, and a **Multi-Perspective Feedback Evaluation Strategy** that assesses feedback quality via pairwise preference judgments and student revision simulations. Experiments on the ASAP++ dataset demonstrate that PsyScore achieves competitive scoring performance while providing more pedagogically aligned feedback.

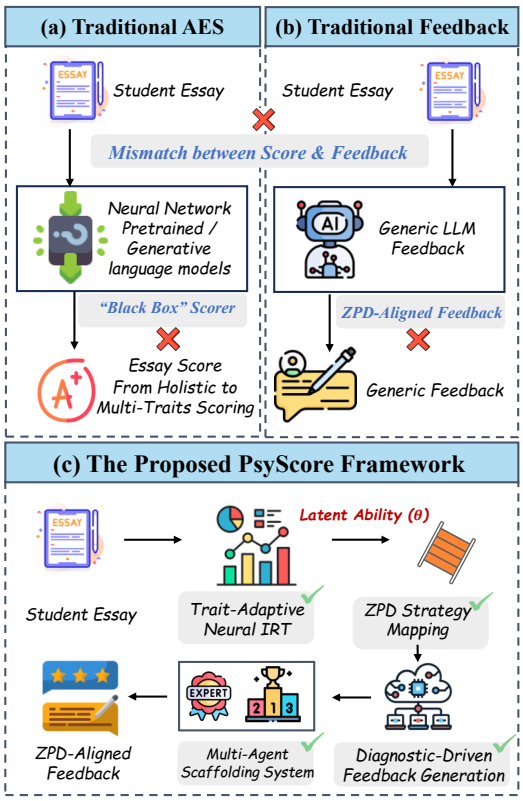


Figure 1: Comparison of traditional approaches and our proposed PsyScore framework. PsyScore aligns diagnostic scoring with scaffolded feedback.

1 Introduction

Automated Essay Scoring (AES) has become a central component of scalable digital education, evolving from early feature-engineering approaches (ElMassry et al., 2025) to deep representation learning and pre-trained language models (PLMs) (Taghipour and Ng, 2016; Dong et al., 2017; Faseeh et al., 2024; Kumar and Boulanger, 2020). Beyond predictive accuracy, an effective assessment system is expected to operate as a formative loop, in which reliable measurement directly informs targeted instructional scaffolding (Clark, 2012; Frank et al., 2018; Wang et al., 2020; Adair, 2024). With the rapid expansion of online learning platforms (Mashau and Nyawo, 2021; Li and Xing, 2025), the demand for timely, individualized, and actionable feedback has increased substantially.

Despite the strong performance of recent AES models (Ludwig et al., 2021a; ElMassry et al., 2025), current solutions often treat diagnostic scoring and pedagogical feedback as separate components (Huang et al., 2025). While recent pioneer-

* Co-first authors with equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 7768–7786

ing efforts such as IFlyEA and CEAES (Gong et al., 2021; Li and Pan, 2025) have begun to bridge this gap by modeling the synergy between scoring and review generation, these approaches typically lack a principled psychometric grounding. This fragmentation gives rise to three persistent limitations (Figure 1). First, most neural AES models function as opaque predictors (Misgna et al., 2024). Although they optimize standard objectives such as mean squared error (L. et al., 2023), they are weakly grounded in educational measurement theory (Voss, 2025), raising concerns about psychometric validity, interpretability, and fairness (Schaller et al., 2024; Shin et al., 2021; Jiang et al., 2023). Second, the separation between scoring and feedback restricts diagnostic utility. A single holistic score or even loosely coupled multi-trait predictions, fails to capture the structured nature of writing proficiency needed to support targeted instructional decisions (ONO et al., 2019; Imbler et al., 2022; Binbin et al., 2024; Stahl et al., 2024a). Third, while Large Language Models (LLMs) are capable of generating fluent feedback (Policar et al., 2025), such feedback is typically ability-agnostic. Without explicit modeling of learner proficiency or the Zone of Proximal Development (ZPD) (Chaiklin et al., 2003), LLM-generated comments frequently exhibit a cognitive mismatch, being overly procedural for advanced learners or overly abstract for novices (Jacobsen and Weber, 2025; Yang et al., 2025).

These limitations suggest a shared underlying cause: the absence of the principled latent representation that simultaneously supports measurement and instruction. Motivated by this observation, we explore the hypothesis that modeling student ability within a shared psychometric latent space can unify interpretable scoring and ZPD-aligned feedback. To this end, we propose **PsyScore**, a psychometrically-aware framework that integrates diagnostic assessment with instructional scaffolding through a latent representation (Figure 2).

PsyScore comprises two tightly coupled components. On the assessment side, it introduces a **Trait-Adaptive Neural IRT Scorer** that embeds Item Response Theory into a neural architecture to estimate interpretable student ability across multiple writing traits. This latent ability representation is then shared with a **ZPD-Scaffolded Feedback Generator**, which conditions pedagogical strategies on the diagnosed proficiency level within a multi-agent framework, enabling feedback that is aligned with learners cognitive readiness.

The contributions of this work are threefold:

- **Psychometric Calibration.** We propose a trait-adaptive initialization strategy for neural IRT models that aligns discrimination and difficulty parameters with psychometric priors.
- **Pedagogical Alignment.** We introduce a ZPD-scaffolded multi-agent feedback framework that conditions instructional strategies on diagnosed learner ability (θ).
- **Empirical Insight.** We conduct a dual-layer evaluation that analyzes the relationship between scoring performance and downstream learning outcomes across proficiency levels.

2 Related Work

2.1 From Holistic Scoring to Multi-Trait Analysis

AES research has evolved from early feature-engineering paradigms (Chen et al., 2014) to modern deep neural networks (Taghipour and Ng, 2016) and pre-trained language models (Ludwig et al., 2021b). While achieving high predictive accuracy, holistic scoring offers limited diagnostic value. To address this, recent frameworks like ArTS (Do et al., 2024b) and T-MES (Wang and Liu, 2025) adopt multi-task learning to evaluate essays along multiple dimensions. However, these models remain primarily predictive: trait scores are treated as parallel outputs rather than manifestations of a shared latent ability. This lack of psychometric grounding limits their interpretability and their capacity to guide principled instructional feedback.

2.2 Item Response Theory in AES

Item Response Theory (IRT) provides a probabilistic framework linking latent ability to observed performance (Li et al., 2025; Samejima, 1969; Uto and Ueno, 2018; Wu and Zheng, 2025). Within AES, recent neuro-symbolic approaches have integrated IRT to enhance interpretability and fairness (Shin et al., 2021; Jiang et al., 2023). Notably, Shibata and Uto (2022) combined IRT with deep learning for analytic scoring, achieving robust multidimensional assessment. However, these prior neural IRT methods focus predominantly on score estimation as a regularization term

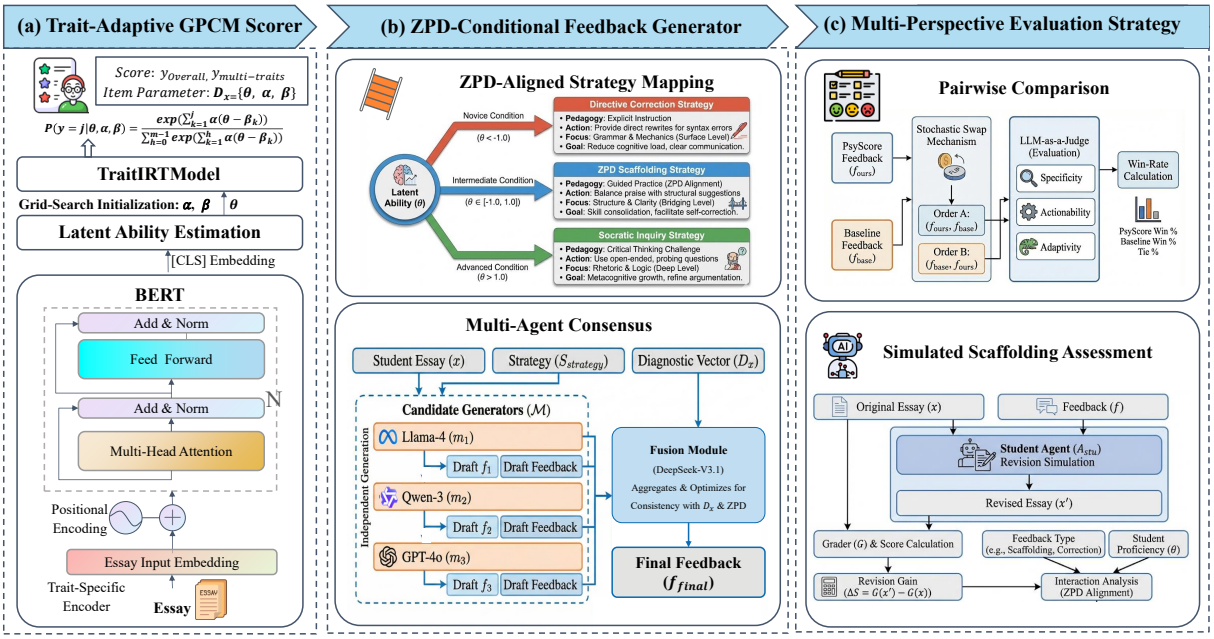


Figure 2: Overview of the PsyScore framework. (a) **Trait-Adaptive GPCM Scorer** estimates the student's latent ability (θ) and outputs a diagnostic vector (D_x). (b) **ZPD-Conditional Feedback Generator** synthesizes consensus feedback (f_{final}) by mapping θ to adaptive strategies via multi-agent fusion. (c) **Multi-Perspective Evaluation** validates quality via intrinsic LLM-based comparison and extrinsic simulated revision.

or multi-task objective. They typically suffer from calibration instability and, crucially, fail to utilize the estimated latent ability (θ) as a control signal for downstream tasks, leaving the critical loop between diagnosis and instructional intervention unconnected. While recent work such as SMART (Scarlatos et al., 2025) uses IRT-aligned simulated students primarily for item difficulty prediction, PsyScore framework extends the application of IRT latent traits to individualized diagnostic feedback generation, focusing on maximizing the pedagogical gain within the student's ZPD.

2.3 LLM-based Personalized Feedback

The transition from Seq2Seq models (Liu et al., 2024) to Large Language Models (LLMs) has enabled scalable, fluent feedback generation (Stahl et al., 2024b). However, current prompting strategies often lack explicit adaptation mechanisms, risking a cognitive mismatch where feedback is either too simplistic or overly complex for the learner. While benchmarks like PROF (Nair et al., 2024) and eRevise (Liu et al., 2025) assess feedback utility based on revision outcomes, they do not verify cognitive alignment. Consequently, existing systems often suffer from content homogenization (Padmakumar and He, 2024) and fail to operationalize Zone of Proximal Development principles, limiting their efficacy as adaptive instructional scaffolds.

3 The PsyScore Framework

3.1 Overall Architecture

PsyScore establishes an integrated framework that couples psychometrically grounded latent trait analysis with generative instructional scaffolding. As illustrated in Figure 2, the system pipeline consists of three mathematically integrated modules: (1) A **Trait-Adaptive Neural GPCM Scorer** for disentangled representation learning and latent parameter estimation. (2) A **ZPD-Conditional Feedback Generator** utilizing a "Generate-and-Fuse" mechanism with heterogeneous agents. And (3) a **Multi-Perspective Feedback Evaluation Strategy** quantifying both semantic alignment and pedagogical efficacy.

3.2 Trait-Adaptive Neural GPCM Scorer

This module functions as a probabilistic regressor that maps input text x to a psychometric latent space, which is a common strategy to improve robustness in high-stakes educational assessment.

3.2.1 Latent Ability Estimation

For a given writing trait t, a BERT-based encoder extracts a dense contextual representation $h_x \in$

$\mathbb{R}^d$. This vector is projected onto a scalar latent ability θ via a linear transformation. To align with the standard normal assumption of IRT ($\theta \sim \mathcal{N}(0,1)$) and ensure numerical stability within the GPCM exponent, we enforce a hard clipping constraint:

$$\theta = \text{Clamp}(W_\theta h_x + b_\theta, \text{min} = -3.0, \text{max} = 3.0) \quad (1)$$

This constraint restricts θ to the effective discrimination interval $[-3, 3]$, preventing gradient explosion during end-to-end training.

3.2.2 GPCM Probability Modeling

The constrained θ serves as the input to a Graded Partial Credit Model (GPCM) layer. The probability of assigning a discrete score category $k \in \{0, \ldots, K\}$ is modeled as:

$$P(y = k|\theta, a_t, \mathbf{b}_t) = \frac{\exp\left(\sum_{j=0}^{k} a_t(\theta - b_{t,j})\right)}{\sum_{c=0}^{K} \exp\left(\sum_{j=0}^{c} a_t(\theta - b_{t,j})\right)} \quad (2)$$

where $a_t \in \mathbb{R}^+$ is the learnable discrimination parameter, and $\mathbf{b}_t = \{b_{t,0}, \ldots, b_{t,K}\}$ represents the step difficulty thresholds. The model is optimized by minimizing the Negative Log-Likelihood (NLL) of the true labels y_{true}:

$$\mathcal{L} = -\sum_{i=1}^{N} \log P(y_i|\theta_i, a_t, \mathbf{b}_t) \quad (3)$$

3.2.3 Initialization and Ensemble Strategy

To resolve the non-convexity of the GPCM loss landscape, which has been noted as a practical challenge in neural IRT-based models, this work employ a grid-search initialization strategy. We pre-compute the optimal priors for a_{init} and b_{init} over the hyperparameter space $\mathcal{H} = \{0.5, 1.0, 1.5\} \times \mathbb{R}^K$ to prevent mode collapse. During inference, the system aggregates parameters from all k folds to construct a robust diagnostic vector $D_x = \{\bar{\theta}, \bar{a}, \bar{\mathbf{b}}\}$, which serves as the conditioning signal for the feedback module.

3.3 ZPD-Conditional Feedback Generator

This module operates as a conditional generation pipeline $F(x, D_x) \to Y$. It leverages diagnostic parameters to synthesize feedback via a multi-agent fusion architecture.

3.3.1 ZPD-Aligned Strategy Mapping

We define a deterministic mapping function $M : \mathbb{R} \to \mathcal{S}$ inspired by the notion of the ZPD, that translates the estimated ability $\bar{\theta}$ into a pedagogical control token:

$$S_{strategy} = \begin{cases} \text{Explicit Correction} & \text{if } \bar{\theta} < -1.0 \\ \text{Scaffolding} & \text{if } \bar{\theta} \in [-1.0, 1.0] \\ \text{Socratic Questioning} & \text{if } \bar{\theta} > 1.0 \end{cases} \quad (4)$$

Additionally, traits with high discrimination ($\bar{a} > 1.2$) are identified as "High-Information Dimensions" and are assigned higher weights in the prompt attention mechanism. This threshold is grounded in standard psychometric evaluation criteria, where discrimination values above 1.2 are categorized as "High" to "Very High", ensuring the selected traits possess sufficient statistical validity to distinguish between proficiency levels (Baker, 2001). The ZPD boundary values are designed to operationalize scaffolding theory within a computational framework (Wood et al., 1976).

3.3.2 Multi-Agent Consensus and Debiasing

To maximize pedagogical diversity, consistent with prior findings on diversity-oriented generation, we deploy a set of heterogeneous LLMs $\mathcal{M} = \{\texttt{Llama-4-Scout}, \texttt{Qwen3-235B-A22B-Instruct-2507}, \texttt{GPT-4o}\}$ as candidate generators. Each model $m \in \mathcal{M}$ independently generates a draft f_m conditioned on the essay x and strategy $S_{strategy}$. A superior model (`DeepSeek-V3.1`) functions as the Fusion Module. It aggregates the candidate set $\{f_m\}_{m=1}^{|\mathcal{M}|}$ into a final response f_{final} by optimizing for consistency with the diagnostic vector D_x:

$$f_{final} = \text{Fusion}(\{f_1, f_2, f_3\}, D_x) \quad (5)$$

This fusion step resolves semantic conflicts and ensures the linguistic complexity of the output strictly adheres to the learner's ZPD.

3.4 Multi-Perspective Feedback Evaluation Strategy

Evaluating open-ended instructional feedback requires moving beyond surface-level lexical overlap. We therefore establish a three-tier evaluation framework to holistically assess the generated feedback: **Preference-based Comparison**: Pairwise ranking by independent LLM judges to measure relative pedagogical quality. **Simulated Revision**: Measuring the normalized score gain achieved when a simulated student agent revises essays conditioned on the feedback. **Human Expert Evaluation**: Fine-grained assessment by education professionals along theoretically grounded dimensions.

3.4.1 Preference Comparison Evaluation with LLM Judges

To mitigate the inherent calibration biases of absolute scoring, we adopt a pairwise preference comparison paradigm. This approach compels the judge model to perform discriminative analysis between feedback generated by PsyScore-AEF (The PsyScore Automated Essay Feedback module) and baseline models, effectively capturing fine-grained qualitative differences.

This work designated `GPT-5` and `Gemini-3-pro` as independent judges to perform pairwise comparisons, selecting them for their superior discriminative performance. The evaluation process adheres to two rigorous protocols: (1) Anonymized Evaluation, a double-blind setup where models receive de-identified texts labeled solely as Response A and B; and (2) Positional Bias Mitigation, which employs stochastic position swapping to counteract the "lead bias" prevalent in LLMs by evaluating each pair twice in randomized orders to ensure consistent judgment.

Judges assign a "Win","Tie", or "Loss" across multiple pedagogical dimensions. We report the aggregate win rate to quantify the relative superiority of PsyScore-AEF in generating personalized, high-quality feedback.

3.4.2 Simulation-based Revision Assessment

To evaluate whether feedback provides cognitive scaffolding rather than direct answers, we implement a simulation-based revision protocol. We construct a "Student Agent" A_{stu} conditioned on ability-specific profiles aligned with the estimated latent trait θ. The agent is instructed to emulate a learner at the diagnosed proficiency level and revise the original essay x based strictly on the received feedback f, yielding a revised version x'.

To ensure measurement invariance across the draft-revision sequence, we utilize the fine-tuned PsyScore-AES as the uniform scoring function $S(\cdot)$. Given the heterogeneous score ranges across prompts (e.g., 0–3 vs. 0–60), absolute differences $\Delta S = S(x') - S(x)$ would introduce range-dependent bias. We therefore compute a normalized revision gain ΔS_{norm} to project improvements onto a unified $[-1, 1]$ scale:

$$\Delta S_{\text{norm}} = \frac{S(x') - S(x)}{S_{\text{max}} - S_{\text{min}}} \tag{6}$$

where S_{max} and S_{min} denote the prompt-specific score bounds. This metric quantifies the marginal pedagogical contribution of the feedback while eliminating confounding effects from disparate scoring scales.

3.4.3 Human Expert Evaluation

To validate the pedagogical efficacy of the generated feedback and to mitigate potential biases inherent in LLM-as-a-judge evaluations, we conducted a double-blind study involving three senior education experts. The experts assessed a stratified random sample of 80 essays along with their corresponding feedback across five theoretically grounded dimensions. The detailed rubric defining each dimension is provided in Appendix H.

4 Experiments and Results

4.1 Dataset

This work conduct experiments on the **ASAP++** dataset (Mathias and Bhattacharyya, 2018), utilizing a standard 6:2:2 data split (60% training, 20% validation, and 20% testing). This dataset extends the original ASAP corpus by providing trait-specific scores (e.g., Content, Organization,Conventions,Word Choice,Sentence Fluency) alongside holistic scores, enabling granular psychometric analysis.

4.2 Implementation Setup

All experiments were implemented using `PyTorch` and the `bert-base-uncased`(Devlin et al., 2019). To ensure reproducibility, we fixed the random seed to 42 for all data splitting and model initialization steps. The computational processes were executed on a single NVIDIA RTX A6000 GPU (48 GB of VRAM). The experiments leveraged CUDA 12.1 and cuDNN 8.9 to optimize GPU performance, with all operations carried out on a Linux-based environment (Ubuntu 22.04) to ensure reproducibility and stability of results.

4.2.1 Scoring Model Training

For the PsyScore-AES module, we fine-tune the `bert-base-uncased` backbone using a 5-fold cross-validation strategy; final results are reported as the ensemble average to minimize variance. The model is optimized with AdamW (learning rate 5e-6, batch size 16) for up to 30 epochs. We apply early stopping with a patience of 10 epochs, monitoring the Quadratic Weighted Kappa (QWK) on the validation set to prevent overfitting.

For the psychometric calibration, the grid search for GPCM parameters is performed within

Table 1: Main results on each prompt. The average QWK across all traits for each prompt is reported.

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG
STL-LSTM (Dong et al., 2017)	0.690	0.622	0.663	0.719	0.719	0.753	0.704	0.592	0.683
HISK (Cozma et al., 2018)	0.674	0.586	0.651	0.681	0.693	0.709	0.641	0.516	0.644
MTL-BiLSTM (Kumar et al., 2022)	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665
DualTrans (Cho et al., 2024)	0.712	0.671	0.690	0.760	0.714	0.740	0.748	0.620	0.707
ArTS (Do et al., 2024a)	0.708	0.706	0.704	0.767	0.723	**0.776**	0.749	0.603	0.717
SaMRL-large (Do et al., 2024b)	0.702	0.711	0.708	0.766	0.722	0.773	0.743	0.649	0.722
T-MES (Wang and Liu, 2025)	0.728	0.684	0.702	**0.771**	0.726	0.754	0.755	0.629	0.719
Shibata and Uto (2022)	0.667	0.642	0.655	0.669	0.658	0.651	0.644	0.638	0.653
PsyScore-AES (Our Model)	**0.744**	**0.723**	**0.732**	0.770	**0.750**	0.773	**0.760**	**0.730**	**0.747**
PsyScore-AES (w/o-IRT)	0.709	0.685	0.707	0.756	0.714	0.737	0.717	0.613	0.705

w/o IRT: *This denotes the ablation experiment in which the IRT module is removed from the proposed PsyScore-AES model.*

Table 2: Main results of each trait. The average QWK across all prompts for each trait is reported.

Model	Ovr.	Cont.	PA	Lang.	Nar.	Org.	Conv.	WC	SF	Style	Voice	AVG
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
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
MTL-BiLSTM	0.762	0.719	0.731	0.659	0.703	0.669	0.656	0.676	0.625	0.693	0.610	0.682
DualTrans	0.764	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.639
ArTS	0.778	0.726	0.732	0.660	0.704	0.682	0.668	0.674	0.663	0.689	0.619	0.690
SaMRL-large	0.774	0.730	0.750	0.702	0.730	0.685	0.686	0.679	0.675	0.693	0.590	0.699
T-MES	0.754	0.730	0.751	0.698	0.725	0.672	0.668	0.679	0.678	**0.721**	0.570	0.695
PsyScore-AES (Our)	**0.790**	**0.761**	**0.762**	**0.708**	**0.749**	**0.707**	**0.733**	**0.725**	**0.725**	0.683	**0.740**	**0.735**
PsyScore-AES (w/o-IRT)	0.735	0.739	0.726	0.676	0.717	0.650	0.671	0.680	0.653	0.619	0.636	0.682

the following spaces: discrimination prior $a_{init} \in \{0.5, 1.0, 1.5\}$ and difficulty range $b_{range} \in \{[-1, 1], [-2, 2], [-3, 3]\}$. We select the optimal configuration for each trait based on validation performance. The final trait-specific hyperparameters are detailed in Table 3.

4.2.2 Feedback Model Configuration

The PsyScore-AEF module implement a Generate-then-Fuse multi-agent system where the latent trait θ serves as a cognitive control signal. First, three heterogeneous agents (`Llama-4-Scout`, `Qwen3-235B-A22B-Instruct-2507`, `GPT-4o`) generate initial drafts to ensure broad pedagogical coverage within the learner's ZPD. Subsequently, an expert agent (`DeepSeek-V3.1`) synthesizes these drafts into a coherent response, filtering hallucinations and aligning the instructional scaffolding with the diagnosed ability θ.

4.2.3 Sampling Strategy and Data Validity

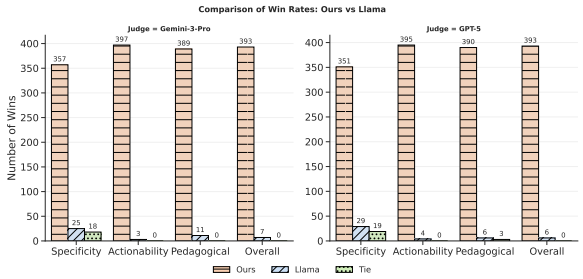
To balance evaluation depth with computational efficiency, we conducted the feedback analysis on a stratified random sample of $N = 400$ essays (50 per prompt) from the ASAP++ dataset. This subset preserves the original distribution of writing genresnarrative, persuasive, and expositoryensuring statistical representativeness. A negligible portion of cases ($< 1\%$) was excluded due to stochastic API invocation issues. Given this minimal attrition rate, the exclusions do not compromise the statistical validity of the results.

4.3 Main Results: Scoring Performance

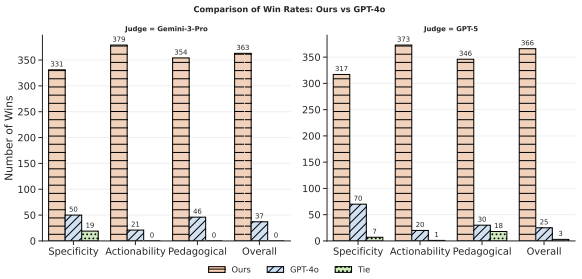
Tables 1 and 2 summarize the comparative scoring performance of PsyScore-AES against state-of-the-art baselines.

4.3.1 Prompt-level Comparison

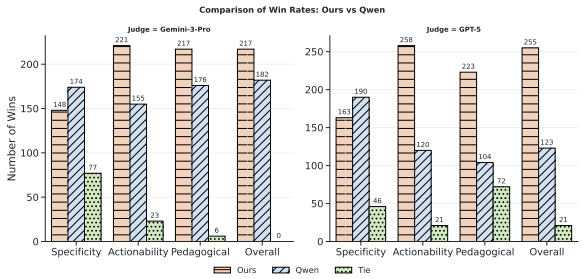
PsyScore-AES achieves a new state-of-the-art average Quadratic Weighted Kappa (QWK) of 0.747, surpassing the strongest baseline (SaMRL-large, 0.722) and ranking first in 6 out of 8 prompts. On the narrative-focused Prompt 8, our model attains a QWK of 0.730, compared with the previous best of 0.649. This improvement suggests that



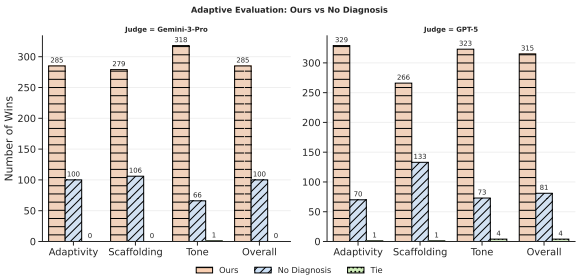
(a) PsyScore-AEF vs. Llama-4-Scout



(b) PsyScore-AEF vs. GPT-4o



(c) PsyScore-AEF vs. Qwen3-235B-A22B-Instruct-2507



(d) PsyScore-AEF vs. No IRT

Figure 3: **Pairwise preference evaluation results across four baselines.** The bars represent the number of wins awarded by judges. PsyScore demonstrates consistent superiority across both open-source (a-c) and closed-source (d) models, particularly in *Actionability* and *Adaptivity*.

Table 3: Optimal psychometric initialization parameters derived via grid search. Distinct configurations validate the necessity of trait-adaptive modeling.

Trait	a_{init}	b_{range}
Overall	0.5	$[-3, 3]$
Content	1.5	$[-2, 2]$
Organization	0.5	$[-3, 3]$
Word Choice	1.0	$[-3, 3]$
Sentence Fluency	1.0	$[-1, 1]$
Conventions	1.5	$[-1, 1]$
Prompt Adherence	1.5	$[-2, 2]$
Narrativity	1.5	$[-1, 1]$
Language	0.5	$[-3, 3]$
Style	1.5	$[-1, 1]$
Voice	0.5	$[-2, 2]$

the Trait-Adaptive Calibration strategy enhances model performance on prompts with higher narrative content.

4.3.2 Trait-level Assessment

At the trait level, PsyScore-AES ranks first in 10 out of 11 evaluated dimensions, achieving a mean QWK of 0.735. In the Voice dimension, which is typically more subjective, the model improves the QWK from 0.619 (ArTS) to 0.740. While slightly trailing T-MES in the *Style* dimension (0.683 vs. 0.721), PsyScore-AES maintains the highest overall performance across traits, indicating the effectiveness of disentangled representation learning in capturing multiple aspects of essay quality.

4.3.3 Ablation Analysis

We evaluate the contribution of the psychometric calibration by comparing the full model with a variant in which the IRT module is included without grid-search initialization of discrimination (a) and difficulty (b) parameters. Removing this initialization results in a performance drop from QWK 0.747 to 0.705. This observation highlights the importance of properly optimizing IRT priors to align neural representations with expected psychometric distributions.

4.4 Statistical Significance Analysis

We assess significance of QWK differences between PsyScore-MAES and BERT using the **Wilcoxon Signed-Rank Test** (one-tailed, $n = 5$ folds). This non-parametric choice avoids normality assumptions and leverages paired test splits. Under the one-tailed criterion, PsyScore-MAES outperforming BERT on all five folds yields $p = 1/2^5 = 0.03125$, the minimum attainable p-value. Full per-trait results with exact p-values are pro-

vided in Appendix E. Across prompts, PsyScore-MAES achieves statistically significant gains on the majority of traits, with the largest improvements on wide-range, fine-grained tasks (e.g., Prompts 7–8) and on core dimensions such as *Content* and *Organization*. Surface-level traits show smaller, often non-significant differences, indicating that the psychometric layer primarily enhances modeling of higher-order writing constructs.

4.5 Robustness Analysis of Hyperparameter Initialization

To assess the sensitivity of the PsyScore-MAES framework to IRT initialization parameters, we conducted experiments under nine distinct parameter configurations. The settings spanned discrimination values $a \in \{0.5, 1.0, 1.5\}$ and difficulty ranges $b_{range} \in \{[-1, 1], [-2, 2], [-3, 3]\}$. Table 4 and Table 5 report the mean Quadratic Weighted Kappa (QWK), standard deviation (σ), and Coefficient of Variation (CV) at the prompt level and trait level, respectively. In psychometric practice, a CV below 5% is generally regarded as indicative of high stability.

As shown in Table 4, the Coefficient of Variation (CV) across all prompts ranges from 0.80% to 3.27%, with an average of 1.67%, substantially below the 5% stability threshold. This confirms that PsyScore-MAES is highly robust to variations in IRT initialization, maintaining consistent scoring regardless of starting parameter values. Notably, performance and stability are not coupled: the framework adapts reliably across prompts with differing score distributions and task types.

Table 4: Performance statistics across different initialization settings (Prompt-level).

Prompt	Mean QWK	Std. (σ)	CV (%)
Prompt 1	0.730	0.016	2.290
Prompt 2	0.706	0.012	1.760
Prompt 3	0.714	0.014	1.970
Prompt 4	0.762	0.008	1.090
Prompt 5	0.711	0.017	2.390
Prompt 6	0.746	0.024	3.270
Prompt 7	0.746	0.013	1.800
Prompt 8	0.706	0.006	0.800
Average	**0.727**	**0.0121**	**1.67**

Table 5 reports trait-level statistics. The *Overall* trait exhibits the lowest CV (0.99%), while subjective dimensions such as *Style* and *Voice* show slightly higher variability (4.72% and 4.00%, respectively), remaining well within acceptable bounds. These results demonstrate that PsyScore-MAES yields reliable scoring outcomes across all traits, irrespective of initialization settings.

Table 5: Performance statistics across different initialization settings (Trait-level).

Trait	Mean QWK	Std. (σ)	CV (%)
Overall	0.777	0.008	0.990
Content	0.743	0.012	1.590
Organization	0.695	0.015	2.140
Word Choice	0.702	0.012	1.750
Sentence Fluency	0.700	0.011	1.530
Conventions	0.712	0.012	1.690
Prompt Adherence	0.739	0.017	2.280
Narrativity	0.717	0.015	2.020
Language	0.686	0.022	3.230
Style	0.666	0.032	4.720
Voice	0.712	0.029	4.000
Average	**0.714**	**0.012**	**1.710**

Table 6: Normalized revision gains and final scores stratified by student proficiency.

Group	Count	PsyScore	No-IRT
Low ($\theta < -1$)	66	**0.1738**	0.0607
Mid ($\theta \in [-1, 1]$)	251	0.1139	0.0080
High ($\theta > 1$)	82	0.0111	0.0009

4.6 Main Results: Feedback Quality

4.6.1 Pairwise Comparison Result

Figure 3 presents the pairwise comparison results of PsyScore-AEF against strong baselines. **Actionability vs. General LLMs:** When compared to general-purpose models such as GPT-4o and Llama-3, PsyScore-AEF achieves win rates exceeding 90% in the *Actionability* dimension. This indicates that the Multi-Agent Fusion mechanism produces more concrete and executable feedback compared to standard generative models.

Adaptivity via Psychometric Diagnosis: Ablation analysis (Figure 3) demonstrates that removing the psychometric module results in a substantial reduction in both *Adaptivity* and *Tone*, with PsyScore retaining a win rate above 75%. This confirms that conditioning feedback generation on latent ability (θ) and discrimination (a) parameters is essential for aligning instructional scaffolding with the learner's ZPD.

Table 7: Expert Evaluation Results (Mean ± SD on a 1-5 Likert Scale).

Model	Accuracy	Actionability	Adaptivity	Specificity	Tone
GPT-4o	$3.10 \pm .45$	$2.88 \pm .77$	$2.79 \pm .53$	$2.55 \pm .61$	$3.15 \pm .63$
Llama-4	$2.88 \pm .52$	$2.42 \pm .66$	$2.66 \pm .68$	$2.34 \pm .50$	$2.62 \pm .69$
Qwen-3	$3.13 \pm .49$	$2.91 \pm .77$	$2.91 \pm .83$	$2.70 \pm .75$	$3.11 \pm .74$
PsyScore-No-IRT	$3.65 \pm .60$	$4.03 \pm .63$	$3.48 \pm .61$	$3.84 \pm .84$	$3.77 \pm .67$
PsyScore	$\mathbf{4.01 \pm .48}$	$\mathbf{4.13 \pm .68}$	$\mathbf{4.21 \pm .54}$	$\mathbf{4.17 \pm .70}$	$\mathbf{4.25 \pm .63}$

4.6.2 Simulation-based Revision Result

Table 6 reports normalized revision gains by student proficiency. PsyScore achieves **17.38%** gain for low-proficiency ($\theta < -1$) vs. 6.07% for No-IRT, confirming targeted scaffolding for foundational deficits. Mid-proficiency gains are 11.39% (vs. 0.80%), while high-proficiency gains are modest (1.11% vs. 0.09%) due to a ceiling effect. Ablation confirms that removing the psychometric layer (θ) severely degrades gains, particularly for low-ability learners, validating θ's role in calibrating feedback to cognitive readiness and avoiding over-assistance.

The resulting expert ratings, together with comprehensive inter-rater reliability statistics, are reported in Table7. PsyScore significantly outperforms all LLM baselines (GPT-4o, Llama-4, Qwen-3) on every dimension ($p < .001$, paired t-test with Bonferroni correction). The ablation variant (`PsyScore-No-IRT`) exhibits marked declines in *Adaptivity* and *Accuracy* ($p < .05$), confirming that IRT-based student modeling is indispensable for generating feedback that is both contextually appropriate and diagnostically precise. Inter-rater reliability analysis indicates substantial agreement among the three experts: Fleiss' κ reaches 0.733 for Specificity and 0.690 for Accuracy, while all pairwise weighted Cohen's κ values exceed 0.60. These metrics collectively validate the robustness and pedagogical validity of the expert assessments.

5 Discussion and Conclusion

5.1 Discussion

The IRT layer acts as a psychometric regularizer, constraining representations to improve robustness on high-variance traits like *Voice* (Prompt 8). This transforms AES from summative scoring into formative diagnosis. Simulated revision confirms ZPD-aligned scaffolding: PsyScore achieves a **17.38%** normalized gain for low-proficiency students ($\theta < -1$), far exceeding the baseline. Gains diminish for advanced learners due to a ceiling effect, underscoring that evaluation should prioritize long-term competency development over immediate text improvement.

5.2 Conclusion

PsyScore integrates psychometric calibration with ZPD-conditioned multi-agent feedback, advancing AES from static scoring to adaptive instructional scaffolding. Trait-adaptive IRT priors yield state-of-the-art scoring accuracy, while ability-aware feedback generation dynamically aligns support with learner proficiency, advocating a formative turn in educational AI.

Limitations

Computational Overhead. The multi-agent consensus mechanism trades off inference latency for pedagogical reliability. While acceptable for asynchronous feedback, this limits real-time classroom deployment. Future compression via knowledge distillation could preserve diagnostic fidelity while enabling interactive responsiveness.

Annotation Dependency. The trait-adaptive scorer requires fine-grained analytic labels (e.g., *Voice*, *Organization*), which are scarce in operational assessment settings. This restricts immediate generalization to holistic-only corpora. Semi-supervised trait disentanglement offers a promising path toward label-efficient psychometric modeling.

Ecological Validity of Simulation. The simulated revision protocol captures *competency gains* under idealized adherence, but cannot model motivational dynamics, epistemic trust, or cognitive fatigue that mediate real-world feedback uptake. Controlled classroom trials are necessary to validate whether scaffolding benefits persist under authentic instructional conditions.

References

Amy Adair. 2024. *AI-Driven Assessment and Scaffolding for Mathematical Modeling and Explanations During Science Investigations*. Ph.D. thesis, Rutgers The State University of New Jersey, School of Graduate Studies.

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Chen Binbin, Bao Lina, Zhang Rui, Zhang Jingyu, Liu Feng, Wang Shuai, and Li Mingjiang. 2024. A multi-strategy computer-assisted efl writing learning system with deep learning incorporated and its effects on learning: A writing feedback perspective. *Journal of Educational Computing Research*, 61(8):60–102.

Seth Chaiklin and 1 others. 2003. The zone of proximal development in vygotskys analysis of learning and instruction. *Vygotskys educational theory in cultural context*, 1(2):39–64.

H. Chen, J. Xu, and B. He. 2014. Automated Essay Scoring by Capturing Relative Writing Quality. *The Computer Journal*, 57(9):1318–1330.

Minsoo Cho, Jin-Xia Huang, and Oh-Woog Kwon. 2024. Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading. *ETRI Journal*, 46(1):82–95.

Ian Clark. 2012. Formative assessment: Assessment is for self-regulated learning. *Educational psychology review*, 24(2):205–249.

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, Melbourne, Australia. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, St. Julian's, Malta. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Ahmed M ElMassry, Nazar Zaki, Negmeldin AlSheikh, and Mohammed Mediani. 2025. A systematic review of pretrained models in automated essay scoring. *IEEE Access*.

Muhammad Faseeh, Abdul Jaleel, Naeem Iqbal, Anwar Ghani, Akmalbek Abdusalomov, Asif Mehmood, and Young-Im Cho. 2024. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics*, 12(21):3416.

Brian Frank, Natalie Simper, and James Kaupp. 2018. Formative feedback and scaffolding for developing complex problem solving and modelling outcomes. *European Journal of Engineering Education*, 43(4):552–568.

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. Iflyea: A chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Yue Huang, C. Palermo, Ruitao Liu, and Yong He. 2025. An early review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback generation. *Chinese/English Journal of Educational Measurement and Evaluation*.

Angenette C. Imbler, S. Clark, T. Young, and Erika Feinauer. 2022. Teaching second-grade students to write science expository text: Does a holistic or analytic rubric provide more meaningful results? *Assessing Writing*.

L. Jacobsen and Kira Elena Weber. 2025. The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of ai-driven feedback. *AI*.

Rui Jiang, Yijia Xue, and Dongmian Zou. 2023. Interpretability-aware industrial anomaly detection using autoencoders. *IEEE Access*, 11:60490–60500.

Stephen Krashen. 1982. Principles and practice in second language acquisition.

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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1485–1495. Association for Computational Linguistics.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA.

Hannah L., Jang E. E., Shah M., and Gupta V. 2023. Validity arguments for automated essay scoring of young students writing traits. *Language Assessment Quarterly*, 20(4-5):399–420.

Deming Li and Wei Xing. 2025. A comparative study on sustainable development of online education platforms at home and abroad since the twenty-first century based on big data analysis. *Education and Information Technologies*, pages 1–22.

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.

Xiaoyu Li, Shaoyang Guo, Jin Wu, and Chanjin Zheng. 2025. An interpretable polytomous cognitive diagnosis framework for predicting examinee performance. *Information Processing & Management*, 62(1):103913.

Yuanchao Liu, Jiawei Han, Alexander Sboev, and Ilya Makarov. 2024. GEEF: A neural network model for automatic essay feedback generation by integrating writing skills assessment. *Expert Systems with Applications*, 245:123043.

Zhexiong Liu, Diane Litman, Elaine L Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. 2025. eRevise+RF: A writing evaluation system for assessing student essay revisions and providing formative feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 173–190, Albuquerque, New Mexico. Association for Computational Linguistics.

Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021a. Automated essay scoring using transformer models. *Psych*, 3(4):897–915.

Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021b. Automated Essay Scoring Using Transformer Models. *Psych*, 3(4):897–915.

Pfano Mashau and Jabulani C Nyawo. 2021. The use of an online learning platform: A step towards e-learning. *South African Journal of Higher Education*, 35(2):123–143.

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)*, pages 730–734. European Language Resources Association (ELRA).

Haile Misgna, Byung-Won On, Ingyu Lee, and G. Choi. 2024. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58.

Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657, Miami, Florida, USA. Association for Computational Linguistics.

Masumi ONO, Hiroyuki YAMANISHI, and Yuko HIJIKATA. 2019. Holistic and analytic assessments of the toefl ibt® integrated writing task. *JLTA Journal*, 22:65–88.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *International Conference on Representation Learning*, volume 2024, pages 642–669.

P. G. Policar, Martin pendl, Toma Curk, and Blaz Zupan. 2025. Automated assignment grading with large language models: insights from a bioinformatics course. *Bioinformatics*, 41:i21 – i29.

Fumiko Samejima. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97.

Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. SMART: Simulated students aligned with item response theory for question difficulty prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25071–25094, Suzhou, China. Association for Computational Linguistics.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. pages 210–221.

Takumi Shibata and Masaki Uto. 2022. Analytic Automated Essay Scoring Based on Deep Neural Networks Integrating Multidimensional Item Response Theory.

Jinnie Shin, Qi Guo, and Mark J. Gierl. 2021. Automated essay scoring using deep learning algorithms. pages 37–47.

Lee S Shulman. 1986. Those who understand: Knowledge growth in teaching. *Educational researcher*, 15(2):4–14.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024a. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. *arXiv preprint*. ArXiv:2404.15845 [cs].

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024b. 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, Mexico City, Mexico. Association for Computational Linguistics.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Masaki Uto and Maomi Ueno. 2018. Item Response Theory Without Restriction of Equal Interval Scale for Raters Score. In Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict Du Boulay, editors, *Artificial Intelligence in Education*, volume 10948, pages 363–368. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Erik Voss. 2025. Comparison of traditional machine learning and neural network approaches for automated scoring of second language english essays. *Language Testing*, 42:369 – 396.

Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

Elaine Lin Wang, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana. 2020. erevis(ing): Students revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 44:100449.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

David J. Wood, Jérôme Seymour Bruner, and Gail P. Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry, and allied disciplines*, 17 2:89–100.

Jin Wu and Chanjin Zheng. 2025. Metacd: A meta learning framework for cognitive diagnosis based on continual learning. *Preprint*, arXiv:2512.22904.

Kaixun Yang, Mladen Raković, Dragan Gašević, and Guanliang Chen. 2025. Does theăprompt-based large language model recognize students' demographics andăintroduce bias inăessay scoring? In *Artificial Intelligence in Education*, pages 75–89, Cham. Springer Nature Switzerland.

A ASAP++ Dataset Details

The **ASAP++** dataset (Mathias and Bhattacharyya, 2018) extends the original Automated Student Assessment Prize (ASAP) corpus by providing fine-grained, trait-specific scores in addition to holistic essay ratings. It comprises eight distinct writing prompts spanning three genres: argumentative (persuasive essays), source-based responses, and narrative storytelling. Table 8 summarizes the prompt-level statistics, including the number of essays, average length, and overall score range for each task.

Each essay is evaluated along multiple analytic dimensions, the set of which varies by prompt. Commonly assessed traits include *Content*, *Organization*, *Word Choice*, *Sentence Fluency*, *Conventions*, *Prompt Adherence*, *Narrativity*, *Language*, *Style*, and *Voice*. Score ranges for individual traits differ from holistic scores; for instance, Prompt 8 assigns holistic scores on a 0–60 scale while trait scores follow a 2–12 range. A detailed mapping of trait dimensions and their corresponding score intervals is provided in Table 8.

Table 8: Basic statistics of the ASAP-AES dataset by prompt.

Prompt	Type	Count	Length	Range
Prompt 1	Argumentative	1,785	350	2–12
Prompt 2	Argumentative	1,800	350	1–6
Prompt 3	Response	1,726	150	0–3
Prompt 4	Response	1,772	150	0–3
Prompt 5	Response	1,805	150	0–4
Prompt 6	Response	1,800	150	0–4
Prompt 7	Narrative	1,569	300	0–30
Prompt 8	Narrative	723	650	0–60

B Initialization and Ensemble Strategy

Table 9 presents a comprehensive grid search over IRT parameter initialization strategies. We evaluate nine combinations of discrimination ($a \in \{0.5, 1.0, 1.5\}$) and difficulty ($b \in \{1.0, 2.0, 3.0\}$) parameters across all writing traits and prompts using two metrics: Quadratic Weighted Kappa (QWK) and Pearson Correlation Coefficient (PCC). The results demonstrate that moderate initialization values ($a = 1.0$, $b = 2.0$) generally yield optimal performance, with consistent stability across different writing dimensions. The table provides empirical evidence supporting our parameter selection strategy for the diagnostic feedback system.

C ZPD-Aligned Strategy Mapping

Table 10 presents our Zone of Proximal Development (ZPD) aligned pedagogical strategy mapping that dynamically selects teaching approaches based on the student's latent ability parameter (θ) estimated through our IRT model. The system implements three distinct instructional strategies: explicit directive instruction for struggling learners ($\theta < -1.0$), targeted scaffolding for developing learners ($-1.0 \leq \theta \leq 1.0$), and intellectual challenge for advanced learners ($\theta > 1.0$). This adaptive strategy selection forms the foundation of our diagnostic feedback generation pipeline, ensuring pedagogical appropriateness for each student's proficiency level.

D Multi-Agent Consensus and Debiasing

Table 11 details the three-stage prompting architecture used in our diagnostic feedback generation pipeline. The first stage generates personalized pedagogical instructions based on the student's latent ability parameter (theta) and item response theory metrics. The second stage employs multiple expert writing tutors to generate initial feedback drafts following these instructions. The final stage uses an expert editor model to synthesize the best elements from all drafts into a cohesive, pedagogically appropriate final feedback report. This multi-stage approach ensures both diagnostic precision and pedagogical effectiveness in the generated feedback.

E Full Statistical Significance Results

Table 12 presents the complete per-trait scoring performance of PsyScore-MAES compared against the BERT baseline across all eight prompts in the ASAP++ dataset. For each prompt and each evaluated writing trait, we report the mean Quadratic Weighted Kappa (QWK) and its standard deviation over five cross-validation folds, the absolute difference in mean QWK between the two models (Diff.), the exact one-tailed p-value derived from the Wilcoxon Signed-Rank Test, and an asterisk denoting statistical significance at the $p < 0.05$ level. Traits without a defined score range in a given prompt are omitted from the corresponding row group.

Table 9: The table reports metrics across all traits and prompts for nine different initialization combinations of discrimination (a) and difficulty (b).

Dimensions	**A=0.5 B=[-1, 1]**		**A=0.5 B=[-2, 2]**		**A=0.5 B=[-3, 3]**		**A=1.0 B=[-1, 1]**		**A=1.0 B=[-2, 2]**		**A=1.0 B=[-3, 3]**		**A=1.5 B=[-1, 1]**		**A=1.5 B=[-2, 2]**		**A=1.5 B=[-3, 3]**	
	QWK	PCC	QWK	PCC	QWK	PCC	QWK	PCC	QWK	PCC	QWK	PCC	QWK	PCC	QWK	PCC	QWK	PCC
Trait-wise Performance																		
Overall	0.775	0.782	0.774	0.784	0.783	0.785	0.775	0.782	0.785	0.786	0.781	0.784	0.779	0.782	0.778	0.781	0.780	0.782
Content	0.738	0.755	0.743	0.753	0.738	0.751	0.738	0.755	0.752	0.759	0.750	0.755	0.754	0.759	0.754	0.758	0.747	0.755
Organization	0.709	0.721	0.710	0.714	0.702	0.711	0.709	0.721	0.696	0.710	0.696	0.700	0.692	0.700	0.705	0.710	0.691	0.696
Word Choice	0.693	0.697	0.694	0.696	0.704	0.715	0.693	0.697	0.715	0.713	0.716	0.717	0.710	0.717	0.708	0.717	0.701	0.705
Sentence Fluency	0.717	0.720	0.703	0.721	0.691	0.700	0.717	0.720	0.700	0.704	0.703	0.708	0.708	0.711	0.697	0.713	0.705	0.714
Conventions	0.714	0.731	0.716	0.734	0.717	0.729	0.714	0.731	0.709	0.719	0.713	0.723	0.724	0.730	0.725	0.738	0.702	0.708
Prompt Adherence	0.740	0.760	0.739	0.751	0.745	0.754	0.740	0.760	0.754	0.758	0.746	0.751	0.743	0.765	0.751	0.761	0.743	0.752
Narrativity	0.720	0.732	0.722	0.740	0.720	0.734	0.720	0.732	0.706	0.722	0.726	0.737	0.731	0.750	0.727	0.734	0.722	0.740
Language	0.689	0.714	0.690	0.709	0.704	0.718	0.689	0.714	0.702	0.710	0.697	0.708	0.668	0.687	0.700	0.715	0.696	0.718
Style	0.641	0.686	0.693	0.726	0.668	0.682	0.641	0.686	0.684	0.703	0.679	0.702	0.685	0.712	0.681	0.686	0.680	0.705
Voice	0.701	0.692	0.738	0.733	0.742	0.756	0.701	0.692	0.711	0.710	0.723	0.738	0.705	0.732	0.641	0.697	0.727	0.737
Trait Avg	0.712	0.727	0.720	0.733	0.720	0.730	0.712	0.727	0.719	0.727	0.721	0.729	0.718	0.731	0.715	0.728	0.718	0.728
Prompt-wise Performance																		
Prompt 1	0.733	0.748	0.732	0.749	0.743	0.754	0.733	0.748	0.733	0.742	0.731	0.739	0.743	0.749	0.741	0.747	0.724	0.732
Prompt 2	0.712	0.720	0.707	0.713	0.722	0.727	0.712	0.720	0.703	0.714	0.717	0.721	0.711	0.718	0.704	0.717	0.697	0.706
Prompt 3	0.707	0.731	0.724	0.740	0.705	0.720	0.707	0.731	0.720	0.727	0.730	0.739	0.724	0.732	0.712	0.724	0.722	0.736
Prompt 4	0.759	0.775	0.763	0.777	0.764	0.770	0.759	0.775	0.764	0.775	0.762	0.769	0.762	0.773	0.772	0.784	0.768	0.771
Prompt 5	0.721	0.739	0.705	0.725	0.712	0.725	0.721	0.739	0.727	0.733	0.711	0.718	0.712	0.731	0.733	0.735	0.712	0.730
Prompt 6	0.741	0.764	0.752	0.766	0.765	0.772	0.741	0.764	0.760	0.767	0.757	0.767	0.736	0.761	0.761	0.771	0.758	0.771
Prompt 7	0.743	0.761	0.753	0.766	0.744	0.749	0.743	0.761	0.753	0.758	0.752	0.762	0.752	0.761	0.758	0.762	0.749	0.757
Prompt 8	0.706	0.700	0.707	0.707	0.697	0.712	0.706	0.700	0.709	0.709	0.710	0.711	0.712	0.713	0.697	0.709	0.713	0.715
Prompt Avg	0.728	0.742	0.730	0.743	0.731	0.741	0.728	0.742	0.734	0.741	0.734	0.741	0.732	0.742	0.735	0.744	0.730	0.740

Table 10: Zone of Proximal Development (ZPD) aligned pedagogical strategy mapping based on student latent ability (θ)

θ Range	Student Proficiency	Pedagogical Strategy
$\theta < -1.0$	Struggling Learner	Strategy: EXPLICIT INSTRUCTION (Directive) The student struggles with fundamentals. Action: Provide direct corrections for grammar and mechanics. Explanation: Keep explanations concise and rule-based. Goal: Fix errors that block communication.
$-1.0 \leq \theta \leq 1.0$	Developing Learner	Strategy: TARGETED SCAFFOLDING The student has emerging competence. Action: Balance praise with 1-2 key areas for improvement. Goal: Guide the student to the next proficiency level (ZPD).
$\theta > 1.0$	Advanced Learner	Strategy: INTELLECTUAL CHALLENGE (Facilitative) The student shows strong command. Action: Focus on rhetorical analysis, logic flow, and voice. Explanation: Use open-ended questions to trigger self-reflection. Goal: Refine style and depth of argument.

Table 11: Prompt templates used at different stages of the diagnostic feedback generation pipeline

Stage	Prompt Template
Diagnostic Instruction Generation	1. PEDAGOGICAL STRATEGY PROTOCOL > Student Latent Ability (Theta): <theta_value> > Required Approach: <strategy_description> CRITICAL INSTRUCTION: Must adopt defined 'Tone' and execute 'Action'. 2. FEEDBACK PRIORITY MAP - <trait> (a=<value>): [CRITICAL FOCUS/Standard/Secondary] 3. DIAGNOSTIC INTERVENTION POINTS - <trait>: <status> (Score: <score>)
Initial Feedback Generation	Role: Expert Writing Tutor implementing "Differentiated Instruction" [STUDENT ESSAY] <full essay text> [PEDAGOGICAL DIAGNOSIS & INSTRUCTIONS] <diagnostic instructions> Task: Write personalized feedback adhering to the defined Strategy. Output: Return feedback content directly.
Expert Synthesis	Role: Chief Editor of Educational Feedback [TRUTH - STUDENT PROFILE] - Latent Ability: <value> - Required Strategy: <strategy_desc> [DRAFT FEEDBACK FROM TUTORS] <multiple candidate drafts> Task: Combine best insights into ONE FINAL PERFECT FEEDBACK REPORT Fusion Rules: 1. Filter by Strategy: Ignore parts violating required strategy 2. ..... Output: Return ONLY the final synthesized feedback text.

Table 12: Performance comparison between PsyScore-MAES and the BERT baseline across prompts and traits.

Prompt	Trait	PsyScore-MAES	BERT	Diff.	p-value	Sig.
1	Overall	0.831 ± 0.007	0.794 ± 0.013	0.037	0.031	*
	Content	0.730 ± 0.012	0.716 ± 0.008	0.014	0.031	*
	Organization	0.696 ± 0.013	0.669 ± 0.016	0.028	0.031	*
	Conventions	0.706 ± 0.018	0.704 ± 0.010	-0.001	0.500	
	Word Choice	0.716 ± 0.014	0.705 ± 0.018	0.011	0.094	
	Sentence Fluency	0.689 ± 0.018	0.702 ± 0.008	-0.013	0.844	
2	Overall	0.680 ± 0.020	0.644 ± 0.021	0.036	0.063	
	Content	0.688 ± 0.019	0.659 ± 0.010	0.030	0.031	*
	Organization	0.689 ± 0.018	0.669 ± 0.014	0.020	0.031	*
	Conventions	0.721 ± 0.017	0.688 ± 0.020	0.033	0.031	*
	Word Choice	0.743 ± 0.011	0.708 ± 0.010	0.035	0.031	*
	Sentence Fluency	0.718 ± 0.008	0.691 ± 0.018	0.027	0.063	
3	Overall	0.707 ± 0.011	0.713 ± 0.014	-0.006	0.844	
	Content	0.715 ± 0.012	0.695 ± 0.013	0.020	0.031	*
	Prompt Adherence	0.740 ± 0.007	0.727 ± 0.008	0.013	0.031	*
	Language	0.689 ± 0.020	0.684 ± 0.018	0.005	0.406	
	Narrativity	0.743 ± 0.012	0.741 ± 0.018	0.002	0.313	
4	Overall	0.779 ± 0.003	0.774 ± 0.012	0.005	0.219	
	Content	0.772 ± 0.007	0.758 ± 0.023	0.015	0.406	
	Prompt Adherence	0.766 ± 0.012	0.755 ± 0.012	0.011	0.156	
	Language	0.708 ± 0.012	0.677 ± 0.017	0.031	0.031	*
	Narrativity	0.782 ± 0.008	0.753 ± 0.014	0.029	0.031	*
5	Overall	0.815 ± 0.006	0.797 ± 0.007	0.018	0.031	*
	Content	0.738 ± 0.014	0.720 ± 0.015	0.018	0.063	
	Prompt Adherence	0.700 ± 0.032	0.692 ± 0.008	0.004	0.313	
	Language	0.706 ± 0.004	0.684 ± 0.015	0.021	0.063	
	Narrativity	0.684 ± 0.001	0.654 ± 0.004	0.030	0.031	*
6	Overall	0.808 ± 0.010	0.801 ± 0.014	0.008	0.156	
	Content	0.824 ± 0.011	0.800 ± 0.015	0.024	0.031	*
	Prompt Adherence	0.783 ± 0.015	0.766 ± 0.016	0.016	0.063	
	Language	0.673 ± 0.018	0.682 ± 0.031	-0.009	0.781	
	Narrativity	0.708 ± 0.022	0.674 ± 0.012	0.034	0.063	
7	Overall	0.829 ± 0.006	0.779 ± 0.011	0.050	0.031	*
	Content	0.858 ± 0.007	0.708 ± 0.017	0.149	0.031	*
	Organization	0.679 ± 0.019	0.539 ± 0.031	0.140	0.031	*
	Conventions	0.710 ± 0.012	0.456 ± 0.029	0.253	0.031	*
	Style	0.664 ± 0.021	0.419 ± 0.022	0.245	0.031	*
8	Overall	0.791 ± 0.015	0.667 ± 0.031	0.125	0.031	*
	Content	0.666 ± 0.019	0.505 ± 0.052	0.161	0.031	*
	Organization	0.704 ± 0.017	0.532 ± 0.029	0.172	0.031	*
	Conventions	0.697 ± 0.043	0.535 ± 0.033	0.162	0.031	*
	Word Choice	0.674 ± 0.022	0.409 ± 0.058	0.265	0.031	*
	Sentence Fluency	0.678 ± 0.038	0.511 ± 0.029	0.167	0.031	*
	Voice	0.721 ± 0.017	0.484 ± 0.031	0.237	0.031	*

F Preference Compare Evaluation

Table 13 presents the prompt template used for human-aligned preference evaluation between feedback variants. The prompt instructs an LLM to act as a pedagogical expert evaluating two feedback versions across three critical educational dimensions: specificity (citing concrete examples from the student essay), actionability (providing clear revision guidance), and pedagogical value (supporting student learning). The structured JSON output format ensures consistent, machine-readable evaluation results. This multi-dimensional assessment framework allows us to quantitatively compare the pedagogical quality of different feedback generation approaches while maintaining alignment with educational best practices.

G Simulated Scaffolding Assessment

Table 14 outlines the dual-role prompt templates used in our simulated pedagogical evaluation framework. The first template simulates a student who revises their essay strictly according to provided feedback, with explicit constraints ensuring revisions remain faithful to the feedback content. The second template simulates a professional essay grader who evaluates essay quality on a standardized 1.0-6.0 scale across multiple dimensions (Organization, Content, Grammar, and Style) with structured JSON output. This simulation framework allows us to quantitatively measure feedback effectiveness by tracking score improvements between original and revised essays, providing an automated yet pedagogically grounded evaluation of feedback quality.

H Human Expert Rubric

To address the limitations of automated evaluation in capturing pedagogical context and to validate the external validity of the experimental findings, we invited three experts with extensive experience in English language instruction to conduct a fine-grained human evaluation. The assessment employed a five-point Likert scale, with detailed criteria provided in Table 15.

Table 13: Prompt template used for multi-dimensional preference comparison between feedback variants

Evaluation Prompt Template

```
You are an expert Pedagogical Evaluator.

[STUDENT ESSAY]:
<essay content (smart truncated)>

[FEEDBACK A]:
<feedback variant A>

[FEEDBACK B]:
<feedback variant B>

[TASK]:
Compare Feedback A and Feedback B along THREE distinct dimensions.
For each dimension, choose a winner ("A", "B", or "Tie") and provide a brief reason.

1. Specificity: Does the feedback cite specific text from the essay?
2. Actionability: Does it give clear instructions on HOW to revise?
3. Pedagogical_Value: Is the feedback supportive and conducive to learning?

[OUTPUT FORMAT]:
Return ONLY a valid JSON object with these exact keys.
{
"specificity": {
"winner": "A" or "B" or "Tie",
"reason": "explanation"
},
"actionability": {
"winner": "A" or "B" or "Tie",
"reason": "explanation"
},
"pedagogical": {
"winner": "A" or "B" or "Tie",
"reason": "explanation"
},
"overall_preference": "A" or "B" or "Tie"
}
```

Table 14: Prompt templates used in the simulated pedagogical evaluation framework

Role	Prompt Template
Student Simulator	You are a student revising your essay based on a teacher's feedback. [YOUR ORIGINAL ESSAY]: <original essay text> [TEACHER'S FEEDBACK]: "<feedback content>" [TASK]: Revise your essay to improve its quality. **CRITICAL RULES**: 1. You must **ONLY** make changes that are suggested or implied by the feedback. 2. If the feedback is specific (e.g., "fix the intro"), focus on that. 3. If the feedback is generic (e.g., "write better"), try your best but do not rewrite the whole essay from scratch. 4. Do not add conversational text. Output ONLY the revised essay.

Essay Grader	You are a professional Essay Grader.
	[ESSAY]:
	<essay text>
	[TASK]:
	Rate this essay on a scale of 1.0 to 6.0 (increments of 0.5 allowed).
	Assess: Organization, Content, Grammar, and Style.
	- 1.0 = Very Poor
	- 6.0 = Excellent
	[OUTPUT FORMAT]:
	Return ONLY a valid JSON object.
	Do not simply copy the example values; determine the score based strictly on the essay quality.
	Example Structure:
	{
	"score":
	<FLOAT_BETWEEN_1_AND_6>,
	"reason":
	"<YOUR_SHORT_JUSTIFICATION>"
	}

Table 15: Human Expert Rubric for Writing Feedback Quality Assessment

Dim	Theoretical Grounding	Core Evaluation Focus	1 (Poor)	2 (Fair)	3 (Moderate)	4 (Good)	5 (Excellent)
1. Adaptivity	Zone of Proximal Development (Vygotsky, 1978)	Does the depth and complexity of the feedback align with the learner's current cognitive level and ZPD?	Severe mismatch. Feedback is either far too difficult or too trivial , completely ignoring individual proficiency.	Low alignment. Feedback feels rigid and lacks meaningful personalization for the specific learner.	Moderate. No obvious logical errors, but exhibits a "generic template" style; lacks targeted adaptation.	Good alignment. Dynamically adjusts semantic depth or linguistic complexity based on inferred student ability.	Perfect alignment. Precisely targets the ZPD. Provides scaffolding for low-proficiency learners and facilitative challenges for advanced learners.
2. Actionability	Feedforward Theory (Hattie and Timperley, 2007)	Does the feedback provide clear revision paths or scaffolding that the learner can immediately act upon?	Not actionable. Offers only vague criticism with no direction for change; advice is confusing.	Vague instructions. Suggestions are overly abstract, leaving the student uncertain where to begin.	Acceptable direction. Indicates a general area for revision but lacks concrete steps or methodological support.	Clear instructions. Explicitly identifies a revision path that the student can follow independently, though specific examples may be absent.	Fully scaffolded. Provides actionable directives alongside revision strategies, worked examples, or stepwise reasoning guidance.
3. Specificity	Cognitive Load Theory (Sweller, 1988)	Does the feedback cite specific textual evidence rather than offering vague evaluations that increase cognitive load?	Completely detached. Exhibits hallucination, referencing non-existent content, or is entirely off-topic.	Generic platitudes. Relies on boilerplate phrases with no connection to the actual text.	Slight relevance. Mentions the general topic or isolated keywords but fails to engage with textual details.	Concrete citation. Precisely references specific paragraphs, sentences, or phrases in the essay for targeted commentary.	Deep insight. Functions like a microscope, accurately citing original text and analyzing underlying logical or rhetorical issues with definitive evidence.
4. Accuracy	Pedagogical Content Knowledge (Shulman, 1986)	Are diagnostic information, linguistic corrections, and factual statements objectively correct and free of misleading content?	Severely erroneous. Provides incorrect knowledge, misidentifies correct usage as error, or introduces factual mistakes.	Obvious mistakes. Contains clear grammatical misjudgments or factual inaccuracies in the diagnosis.	Generally accurate. No fundamental errors, but may overlook deeper logical issues, or suggestions are debatable.	Professionally accurate. Diagnostic conclusions are precise; instructional suggestions fully align with linguistic conventions and norms.	Exemplary precision. Beyond surface correction, accurately identifies subtle issues such as pragmatic missteps or logical inconsistencies.
5. Supportive Tone	Affective Filter Hypothesis (Krashen, 1982)	Is the wording empathetic, aiming to reduce learner anxiety while stimulating motivation to revise?	Negative and cold. Robotic or condescending tone; feedback consists solely of negative criticism.	Lacks warmth. Reads like a machine-generated error log; purely task-oriented with no humanistic encouragement.	Polite and objective. Neutral in tone; acceptable to the learner but uninspiring and unlikely to boost motivation.	Positively encouraging. Employs a constructive approach, fostering a positive feedback loop.	Highly inspiring. Demonstrates empathy and adopts a collaborative stance that cultivates a growth mindset and strong revision willingness.