Towards Self-Referential Analytic Assessment: A Profile-Based Approach to L2 Writing Evaluation with LLMs

Stefano Bannò, Kate Knill, Mark Gales
ALTA Institute, Department of Engineering, University of Cambridge (UK)
{sb2549,kmk1001,mjfg100}@cam.ac.uk

Abstract

Automated essay scoring (AES) research often relies on rank-based correlation metrics to validate analytic assessment. However, such metrics obscure both intrinsic intercorrelations among analytic dimensions that arise from the structure of writing proficiency itself and halo effects, whereby holistic impressions bleed into fine-grained component scores. As a result, high correlations may mask a system's true diagnostic behaviour. In this study, we propose a novel self-referential assessment evaluation framework that focuses on identifying intra-learner strengths and weaknesses rather than assessing inter-learner rankings. We conduct experiments on the publicly available ICNALE GRA, a uniquely dense second-language writing dataset annotated holistically and analytically by up to 80 trained raters. To obtain reliable reference scores, we apply two-facet Rasch modelling to calibrate rater severity and derive fair average scores across ten analytic aspects and holistic proficiency. We compare the analytic scoring performance of human operational raters and three large language models (LLMs) in a zero-shot setting. Our results show that LLMs tend to outperform single human raters in identifying relative weaknesses (negative feedback) across several proficiency aspects, while human raters remain stronger at identifying relative strengths (positive feedback). Overall, our findings highlight the limitations of rank-based evaluation for analytic assessment and demonstrate the value of intra-learner, profile-based methods for assessing and deploying LLMs in AES.

1 Introduction

Automated essay scoring (AES) has become a cornerstone of second language (L2) proficiency assessment, utilising computational systems to evaluate learner writing in educational contexts. (Klebanov and Madnani, 2022). AES systems have primarily targeted holistic assessment, focusing on assigning a single score that reflects a learner's overall writing proficiency. However, driven by the need for more actionable pedagogical insights, the field has recently expanded its focus to analytic scoring, i.e., the evaluation of distinct aspects of proficiency (Li and Ng, 2024). The advent of large language models (LLMs) has revolutionised this landscape. Recent work has successfully exploited LLMs for both holistic (Mizumoto and Eguchi, 2023) and analytic scoring (Bannò et al., 2024). This shift offers the potential for scalable, multi-dimensional assessment that was previously difficult to achieve with traditional supervised models.

Despite these advances, current evaluation practices predominantly rely on rank-based metrics, such as Pearson's correlation coefficient (PCC), Spearman's rank coefficient (SRC), or agreement metrics, such as Quadratic Weighted Kappa (QWK) (Yannakoudakis and Cummins, 2015). While these metrics effectively measure a system's ability to rank learners globally (i.e., inter-learner agreement), they may not be suitable to validate multi-dimensional analytic assessment. In addition to intrinsic intercorrelations among analytic dimensions, a major interfering factor is the halo effect (Thorndike, 1920; Engelhard, 1994), whereby a learner's general proficiency influences judgments across specific sub-scales. As we show in this study, an LLM can achieve high correlations on analytic aspects simply by acting as a proxy for holistic proficiency, without accurately detecting the nuances that distinguish different analytic proficiency dimensions. Consequently, high rank agreement may not guarantee diagnostic validity. To address this limitation, we propose shifting the analytic evaluation paradigm from *normative* (i.e., ranking learners against each other) to *self-referential* assessment, in which a learner's analytic skills are evaluated within the same learner profile. Self-referential analytic assessment focuses on intra-learner variability, identifying aspects where

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 174–189

a student performs significantly better or worse than their average proficiency (see Figure 1), thus offering a stricter and more pedagogically relevant test of an LLM's analytic assessment capabilities.

Ensuring the interpretability of such an evaluation requires a reference signal that is internally consistent, psychometrically calibrated, and representative of expert judgement under the known interdependence of analytic dimensions. Standard datasets often suffer from rater subjectivity and noise, which can obscure the genuine signal. We mitigate this by leveraging the publicly available ICNALE GRA (Ishikawa, 2024) (see Section 3.1), an L2 learner dataset annotated by 80 raters holistically and analytically. We employ Rasch modelling (Linacre, 1989) to rigorously filter raters based on infit statistics, constructing a fair average reference score derived solely from highly consistent human raters.

In this paper, we evaluate three LLMs in a zero-shot setting alongside operational human raters using this self-referential framework. Our contributions are as follows:

1. We demonstrate the limitations of traditional rank-based metrics in analytic assessment, showing that high correlations often mask a lack of diagnostic precision due to both intrinsic intercorrelations and the halo effect.

2. We introduce a novel self-referential analytic assessment evaluation method that disentangles absolute proficiency from relative profile deviations, creating a dual classification task for positive and negative feedback.

3. We report experimental results showing that, within this self-referential framework, LLMs outperform a single operational rater in identifying relative weaknesses (negative feedback), while human raters retain an advantage in identifying relative strengths (positive feedback).

2 Related work

Traditionally, AES systems have mainly focused on overall holistic assessment. This holistic orientation has characterised AES since its inception in the 1960s (Page, 1966) and has been predominant up to today in recent studies investigating L2 assessment with LLMs (Mizumoto and Eguchi, 2023; Yancey et al., 2023).

Research on automated analytic assessment and feedback emerged later, with only a few early exceptions (Page et al., 1997; Shermis et al., 2002). The advent of neural approaches led to a growing body of work targeting specific traits of written production, such as organisation, content, word choice, sentence fluency, and narrativity (Hussein et al., 2020; Mathias and Bhattacharyya, 2020; Ridley et al., 2021; Kumar et al., 2022; Do et al., 2023). As with holistic assessment, LLMs have also been successfully applied to analytic assessment. For instance, Naismith et al. (2023) investigated the use of GPT-4 to predict coherence-related scores. Bannò et al. (2024) explored zero-shot analytic assessment of L2 writing proficiency using GPT-4 across CEFR-aligned proficiency aspects (Council of Europe, 2020). Seßler et al. (2025) examined multiple open- and closed-source LLMs for analytic scoring of German student essays across ten traits, including content-related, syntactic, and formal dimensions. Wang et al. (2025) investigated LLMs' ability to conduct multi-dimensional analytic writing assessment by assigning scores and generating feedback comments across nine analytic aspects, finding that LLMs can produce generally reliable analytic assessments. In another recent work, Yoo et al. (2025) introduced a new dataset annotated across three analytic traits, i.e., content, language, and organisation, combining LLM-based scoring with synthetic data generation. Using the ICNALE GRA essays, Yamashita (2024) investigated the use of GPT-4 for analytic assessment along complexity, accuracy, and fluency dimensions. Their study primarily adopts a many-facet Rasch measurement framework to analyse rater severity, consistency, and potential biases when treating GPT-4 as an additional rater alongside humans.

Automated analytic scoring has largely inherited the evaluation metrics used for holistic assessment. The studies reviewed above typically rely on rank-based metrics, such as PCC or SRC, as well as agreement measures such as QWK (Yannakoudakis and Cummins, 2015), under a normative assessment framework, whereby learners are ranked against each other for each analytic proficiency aspect. In our work, we propose shifting the evaluation paradigm for analytic assessment from normative evaluation to self-referential interpretation of analytic profiles, in which analytic traits are evaluated within an individual learner's performance profile. In the psychometric literature, a

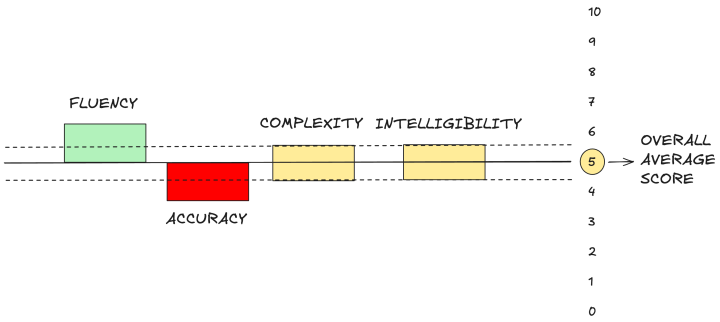


Figure 1: Illustration of proposed self-referential analytic assessment approach.

similar approach has been theorised as *ipsative* assessment (Cattell, 1944; Clemans, 1956). However, in some applied assessment traditions, the same term instead denotes comparison with an individual's previous performances (Hughes, 2011), rather than within-individual, self-referenced comparison across analytic dimensions. For this reason, we prefer the term *self-referential*.

Self-referential analytic assessment directly supports diagnostic feedback by identifying relative strengths and weaknesses within a learner, rather than across a population. To the best of our knowledge, this is the first study to introduce a self-referential evaluation framework for analytic assessment of L2 proficiency and to implement automated systems for this purpose. The closest conceptual precedent is Berninger and Abbott (2010), who employed within-learner comparisons across language modalities (listening, speaking, reading, and writing), operationalising relative strengths and weaknesses as deviations from a learner's own mean performance. While their approach focused on modality-level L1 language skills in a psychometric testing context and did not use any automatic systems, the present study applies a self-referential perspective within AES of L2 writing by comparing learners' performance across fine-grained analytic proficiency aspects.

3 Experimental setup

In this section, we first describe the dataset used in our experiments and outline the procedures for rater selection and score calibration. We then present details on LLM prompting and the method for extracting scores from the models. Next, we discuss the limitations of rank-agreement metrics when evaluating analytic assessment. Finally, we introduce our self-referential analytic assessment system.

3.1 Data

The International Corpus Network of Asian Learners of English (ICNALE) Global Rating Archive (GRA) (Ishikawa, 2020, 2024) is a publicly available L2 learner dataset derived from a subset of ICNALE (Ishikawa, 2011), containing learner essays and speeches, of which we use only the essay section for our experiments. A notable feature of this dataset is its English as a Lingua Franca (ELF) assessment perspective (Seidlhofer, 2005), which evaluates English as a communicative tool among non-native speakers instead of relying on native-speaker norms. Although the dataset is relatively small ($N = 140$), each essay has been annotated by 80 trained raters representing diverse first-language (L1) and occupational backgrounds.[1]

In addition to holistic scores, the essays have been annotated across three macro-aspects, which in turn comprise ten analytic rating aspects:

- **Language** (*Intelligibility*, *Complexity*, *Accuracy*, and *Fluency*);
- **Content** (*Comprehensibility*, *Logicality*, *Sophistication*, and *Purposefulness*);
- **Attitude** (*Willingness to communicate* and *Involvement*).

The analytic scores range from 0 to 10, whereas the holistic scores range from 0 to 100; both scales allow midpoint values.[2] Human raters were instructed to evaluate the essays holistically first, followed by analytic scoring.

To the best of our knowledge, this makes ICNALE GRA the only publicly available L2 learner

[1]Four essays were written by L1 speakers; we did not discard them from the dataset.

[2]We identified four anomalies in the original scores: for *Intelligibility*, a score of 1.2 (recoded as 1.0) and a score of 6.6 (recoded as 6.5); for *Complexity*, a score of 19 (recoded as 10); and for *Involvement*, a score of 8.6 (recoded as 8.5).

writing dataset that supports a fully crossed rater-essay design with more than two raters, comprising 123,195 ratings.[3]

3.1.1 Raters selection

In a previous study by the curator of the dataset, focusing specifically on the ICNALE ratings (Ishikawa, 2023), inter-rater reliability is evaluated using Cronbach's α (Cronbach, 1951). This statistic primarily reflects relative consistency, capturing the extent to which raters produce similar rank orderings of test-takers, while being insensitive to systematic differences in absolute scoring levels. As such, Cronbach's α is a measure of internal consistency rather than a direct measure of inter-rater reliability (Krippendorff, 2004; Lombard et al., 2006); it reflects the extent to which raters are measuring the same underlying construct.

For this reason, we instead adopt Krippendorff's α (Krippendorff, 2011), which is more commonly used for inter-rater reliability, as it explicitly quantifies absolute agreement among raters, i.e., the degree to which raters assign identical scores to the same test-taker. Moreover, this metric is well suited to ordinal data, multiple raters, and missing values (see Note 3). Table 1 reports inter-rater reliability in terms of Krippendorff's α: column *All* shows the estimate computed over all 80 raters, while column *Selected* reports the estimate on 12 raters, after applying the rater selection procedure described below. As can be seen, the low Krippendorff's α values for *All* indicate substantial disagreement among raters, which may be attributable to differences in severity, inconsistent category use, or divergent interpretations of the rating criteria.

Macro-aspect	Aspect	α	
		All	Selected
Language	Intelligibility	0.270	**0.531**
	Complexity	0.252	**0.542**
	Accuracy	0.278	**0.505**
	Fluency	0.256	**0.526**
Content	Comprehensibility	0.257	**0.538**
	Logicality	0.237	**0.499**
	Sophistication	0.225	**0.517**
	Purposefulness	0.221	**0.502**
Attitude	Willingness	0.202	**0.465**
	Involvement	0.172	**0.409**
	Holistic	0.297	**0.579**

Table 1: Krippendorff's α reliability estimates on ICNALE GRA.

To ensure the reliability of the rating data, we evaluate each of the 80 raters' consistency using the infit statistic derived from a two-facet Rasch model (Linacre, 1989), i.e., one model for each of the 11 analytic and holistic language aspects, with the two facets being rater severity and learner ability. To do this, we use the Many-Facet Rasch Model estimation function, `tam.mml.mfr`, within R's `TAM` (Test Analysis Modules) package.[4] To accommodate midpoints, analytic scores are fed into the model after multiplying them by 2, whereas holistic scores are binned into 11 categories from 0 to 10, as described in Ishikawa (2024, p. 33).

Infit measures how well a rater's scores match the model's expectations, giving more weight to items that are close to a person's ability level. Mathematically, for rater r and item i, the infit mean-square is computed as:

$$\text{Infit}_r = \frac{\sum_i w_{ri}(X_{ri} - E_{ri})^2}{\sum_i w_{ri}} \quad (1)$$

where X_{ri} is the observed rating, E_{ri} is the Rasch model-predicted rating, and w_{ri} is a weight reflecting the item's information (i.e., its expected variance). An infit value between 0.5 and 1.5 indicates that a rater's judgments are consistent with the model, whereas values substantially above 1.5 or below 0.5 indicate over- or under-discrimination, respectively (Linacre, 2002). First, we retained only raters with infit equal to or higher than 0.5 and equal to or less than 1.5, ensuring that all included raters provided reasonably consistent scores across items. Subsequently, we identified the intersection of acceptable raters across all language aspects, resulting in a subset of 12 raters (with 18,480 ratings in total) that satisfied the infit criterion for every aspect considered.

As can be observed in column *Selected* in Table 1, rater selection based on infit statistics leads to a substantial increase in observed reliability. This procedure identifies a subset of raters whose scoring behaviour falls within acceptable Rasch infit bounds and can therefore be used for score calibration and subsequent modelling. From this point onwards, all references to "raters" pertain exclusively to the subset of 12 selected raters.

[3] That is, (140 essays $\times$ 80 raters $\times$ 11 proficiency aspects, i.e., ten analytic and one holistic) $-$ 5 missing ratings. Five rating data points are missing from the original scores of the 80 raters; however, there are no missing values in the subset of 12 selected raters (see Section 3.1) used in our experiments.

[4] cran.r-project.org/web/packages/TAM

3.1.2 Score calibration

To obtain a fair average score $E(X_n)$ for a learner n, we fit a two-facet Rasch model for each language aspect, this time only with the 12 selected raters. The net cumulative difficulty barrier, $\Phi_{net,k}$, for achieving a score of k combines the average rater severity ($\bar{\lambda}$) and the global cumulative step difficulties (τ_m):

$$\Phi_{net,k} = \left(\sum_{m=1}^{k} \tau_m \right) + k \cdot \bar{\lambda} \tag{2}$$

where $\sum_{m=1}^{k} \tau_m$ is the cumulative step difficulty, defined as 0 for $k = 0$; τ_m is the estimated global step difficulty parameter (the difficulty of moving from score $m-1$ to m); k: the score category index ($k = 0, 1, 2, \ldots, K$); $\bar{\lambda}$ is the mean severity of all selected raters.

The probability of learner n achieving a score k is calculated as the exponent of the difference between the learner's ability and the net cumulative difficulty, normalised by the sum of probabilities across all possible categories:

$$P(X_n = k) = \frac{\exp\left(k\theta_n - \Phi_{net,k}\right)}{\sum_{c=0}^{K} \exp\left(c\theta_n - \Phi_{net,c}\right)} \tag{3}$$

where θ_n is the estimated ability of learner n and K is the maximum category index.

The final expected score $E(X_n)$ is the sum of the products of each category score s_k and its probability $P(X_n = k)$. Since, for the analytic aspects, the raw scores were multiplied by 2 prior to fitting the model to accommodate midpoints, the final result is divided by 2 for normalisation:

$$E(X_n) = \frac{1}{2} \cdot \sum_{k=0}^{K} s_k \cdot P(X_n = k) \tag{4}$$

For holistic scores, instead, the final result is

$$E(X_n) = \sum_{k=0}^{K} s_k \cdot P(X_n = k) \tag{5}$$

We refer to these final scores as *fair average scores*.

Table 2 reports the average Spearman's rank correlation coefficient (SRC) between the *average rater* (AR) and the reference fair average scores. In other words, we calculated the SRC between each of the 12 raters and the reference fair average scores, and reported the mean and standard deviation. However, in operational language assessment, essays are typically assigned to raters at random. To simulate this scenario, we randomly selected one rater for each essay. This procedure was repeated five times using different random seeds, and the results were averaged. We refer to this condition as the *operational rater* (OR).

	AR	**OR**
Intelligibility	$0.764_{\pm.05}$	$0.749_{\pm.02}$
Complexity	$0.770_{\pm.04}$	$0.737_{\pm.02}$
Accuracy	$0.765_{\pm.05}$	$0.735_{\pm.02}$
Fluency	$0.761_{\pm.03}$	$0.733_{\pm.02}$
Comprehensibility	$0.769_{\pm.05}$	$0.741_{\pm.02}$
Logicality	$0.748_{\pm.04}$	$0.723_{\pm.02}$
Sophistication	$0.761_{\pm.03}$	$0.741_{\pm.04}$
Purposefulness	$0.740_{\pm.05}$	$0.730_{\pm.02}$
Willingness	$0.718_{\pm.03}$	$0.696_{\pm.05}$
Involvement	$0.692_{\pm.07}$	$0.644_{\pm.05}$
Holistic	$0.788_{\pm.04}$	$0.775_{\pm.02}$

Table 2: SRC correlations of average rater (AR) and operational rater (OR) with fair average scores.

Notably, holistic scores exhibit the highest levels of agreement, as shown in Table 1, a pattern that is further supported by the SRC results in Table 2. This observation aligns with previous research indicating that holistic scoring is generally easier and more intuitive for human raters than analytic scoring, hence tending to yield higher inter-rater reliability than fine-grained analytic judgments (Weigle, 2002; Zhang et al., 2015). Furthermore, the *Attitude* aspects (i.e., *Willingness to communicate* and *Involvement*) are inherently more subjective and consequently more difficult for raters to evaluate, as reflected in their lower SRC and Krippendorff's α values compared with the other aspects.

3.2 LLM prompting and score extraction

We use GPT-4.1 (OpenAI, 2023), Qwen 2.5 72B (4-bit quantised) (Yang et al., 2024), and Llama 3.1 70B (4-bit quantised) (Llama Team, 2024) in a zero-shot setting to evaluate each language aspect in the essays, allowing us to compare both proprietary high-parameter and open-source, medium-size LLMs. The prompt template provided to the LLMs can be found in Appendix A. This contains an analytic rating prompt for each language aspect, originally used by human raters. Because the original analytic prompts (Ishikawa, 2024, pp. 34–36) were designed for both essays and speeches, we adapted them to include only content relevant to

essays. The revised prompts are provided in Appendix B.

To extract proficiency scores from the LLM, we use a weighted average approach (Ma et al., 2025; Bannò et al., 2025). For each analytic aspect, the LLM is prompted independently, producing a probability distribution over discrete proficiency levels via a softmax layer. Let $\mathbf{p} = [p_1, p_2, \ldots, p_K]$ denote this probability distribution over K ordinal levels, and let $\mathbf{v} = [v_1, v_2, \ldots, v_K]$ represent the numeric values assigned to each level (here, 0 through 9).[5] The weighted average score for the LLM prediction is then computed as:

$$\text{WAvg} = \sum_{k=1}^{K} p_k \cdot v_k \tag{6}$$

Intuitively, this procedure produces a continuous estimate of the score by weighting each possible level by the LLM's predicted probability.

3.3 Limitations of assessing analytic scoring using rank-based metrics

In real-world multi-aspect language assessment, a common characteristic is the high intercorrelation among scores assessing different aspects, as well as the strong correlation between analytic and holistic scores (Lee et al., 2008; Ono et al., 2019). This pattern is also observed in our study. Table 3 presents the SRC between predictions from various systems and the reference fair average scores for three aspects, each representing a macro-aspect: *Complexity* for Language, *Logicality* for Content, and *Willingness* for Attitude. We compare these with the correlations between holistic fair average scores and analytic fair average scores (first row). Additionally, we report the SRC correlations on analytic fair average scores when we prompt our systems for holistic scoring. As can be seen, holistic fair average scores correlate highly with analytic scores. As a result, when we use LLMs prompted to predict holistic scores, we observe strong correlations with analytic scores, in some cases even stronger than when we prompt them for the respective analytic aspect. For example, when prompted for holistic scores, GPT-4.1 shows a correlation with *Complexity* of 0.927 (vs 0.848) or Qwen 2.5 shows a correlation with *Willingness* of 0.754 (vs 0.457). Similarly, the operational rater (OR) achieves higher correlations with analytic scores when providing holistic ratings: $0.763_{\pm.02}$ for *Complexity*, $0.762_{\pm.01}$ for *Logicality*, and $0.750_{\pm.02}$ for *Willingness*, compared to $0.737_{\pm.02}$, $0.723_{\pm.02}$, and $0.696_{\pm.05}$, respectively, when performing analytic assessment. In addition to the presence of real intercorrelations among analytic scores, this pattern may be explained by a halo effect, in which the holistic judgments influence ratings across multiple analytic dimensions, effectively acting as a proxy for the learner's analytic ability.

Consequently, evaluating analytic assessment quality using rank-based metrics such as SRC can be misleading:[6] these metrics primarily capture global agreement in learner ranking, and may obscure the model's ability to detect relative strengths and weaknesses within a single learner's profile. In other words, a system can achieve high correlations with analytic scores simply by reflecting holistic proficiency rather than accurately differentiating performance across individual aspects. It is nevertheless noteworthy that LLMs outperform OR on *Complexity*, *Logicality*, and *Holistic* (with the exception of Qwen 2.5), whereas on the more subjective dimension of *Willingness*, OR still achieves the best – albeit comparably lower – performance, as expected (see end of Section 3.1). The scatterplots for the three selected analytic aspects and holistic proficiency are shown in Figure 3 in Appendix C.

The next section addresses this issue by proposing an alternative approach to analytic assessment.

3.4 Implementing self-referential analytic assessment

We first standardise the reference analytic fair average scores. For each analytic aspect i, we compute

$$z_i = \frac{s_i - m_i}{\sigma_i} \tag{7}$$

where s_i denotes the fair average score, and m_i and σ_i are the mean and standard deviation of that score for i across essays.

For each essay, we then compute the average of the standardised analytic scores across all A aspects:

$$\mu = \frac{1}{A} \sum_{j=1}^{A} z_j \tag{8}$$

[5] We used a 0–9 scale because including 10 would require two tokens ("1" and "0"), making it difficult to extract a probability for this score.

[6] Due to the continuous nature of our scores, we cannot use QWK. However, because this also measures rank agreement, we expect it would yield similar patterns.

		Complexity	Logicality	Willingness	Holistic
	Fair Avg. Holistic	0.993	0.982	0.976	1.000
Prompted for analytic	**OR**	$0.737_{\pm.02}$	$0.723_{\pm.02}$	$0.696_{\pm.05}$	-
	GPT4.1	0.848	0.810	0.340	-
	Qwen2.5	0.753	0.769	0.457	-
	Llama3.1	0.814	0.824	0.554	-
Prompted for holistic	**OR**	$0.763_{\pm.02}$	$0.762_{\pm.01}$	$0.750_{\pm.02}$	$0.775_{\pm.02}$
	GPT4.1	0.927	0.921	0.910	0.936
	Qwen2.5	0.738	0.764	0.754	0.771
	Llama3.1	0.758	0.771	0.724	0.781

Table 3: SRC of selected aspects (*Complexity*, *Logicality*, *Willingness*) with fair average scores across different systems.

We define the difference between μ and the score for aspect i as

$$\Delta_i = \mu - z_i \tag{9}$$

This difference is used to construct a dual binary classification task corresponding to negative and positive feedback. Specifically, for each aspect i, we define the negative feedback target as $\Delta_i \geq \sigma_{\Delta_i}$ and the positive feedback target as $\Delta_i \leq -\sigma_{\Delta_i}$, where σ_{Δ_i} is the standard deviation of Δ_i across essays.[7]

For model predictions, let $\hat{s}_i$ denote the predicted score for aspect i. Predictions are standardised analogously:

$$\hat{z}_i = \frac{\hat{s}_i - \hat{m}_i}{\hat{\sigma}_i} \tag{10}$$

where $\hat{m}_i$ and $\hat{\sigma}_i$ are computed over the predicted scores for aspect i. The average of the standardised predictions across aspects is then

$$\hat{\mu} = \frac{1}{A} \sum_{j=1}^{A} \hat{z}_j \tag{11}$$

Finally, we compute the difference between $\hat{\mu}$ and the predicted score for aspect i

$$\hat{\Delta}_i = \hat{\mu} - \hat{z}_i \tag{12}$$

which serves as a continuous scoring signal for the feedback classification task. We report the best $F_{0.5}$ score for each aspect prioritising precision over recall to minimise false positives and avoid misleading learners about their proficiency strengths or weaknesses.

In order to be able to have a fair comparison of the results across the two binary tasks, we normalise the Precision and $F_{0.5}$ score using the method illustrated in Raina et al. (2023) setting the prevalence to 10% since we are working with different language aspects, each having a different original prevalence rate (see Tables 7, 8, 9, and 10 in Appendix C). To do this, we rescaled the proportion of negative cases so that the effective prevalence matched the target. Let N_+ and N_- denote the number of positive and negative samples, and let pr_{target} denote the desired prevalence. We first solved

$$\frac{N_+}{N_+ + \gamma\, N_-} = pr_{\text{target}} \tag{13}$$

for the scaling factor γ applied to the number of negatives. Assuming the classifier's false positive rate per negative instance remains constant, false positives scale proportionally by the same factor γ. The normalised precision P' is then obtained by recomputing precision using this adjusted false–positive count. Normalised recall is unaffected, and the normalised $F_{0.5}$ score is subsequently computed from P' and recall R.

Given this prevalence value, if we calculate the $F_{0.5}$ score for a random classifier, where β is 0.5, P' is 0.10, and R is 1:

$$F_\beta = (1 + \beta^2)\, \frac{P' \cdot R}{\beta^2 \cdot P' + R} \tag{14}$$

we obtain a score of ~ 12.19. Any results higher than this value contain information. Table 6 in Appendix C illustrates the percentage of essays receiving feedback across macro-aspects.

Before discussing the results, it is important to clarify the nature of the task. Unlike the inter-learner correlation analyses reported in Table 3, our proposed self-referential framework focuses on identifying aspects that are unusually weak or strong relative to an individual learner's overall

[7] A one-standard-deviation threshold has also been used in the aforementioned work by Berninger and Abbott (2010).

profile. As a result, systems may show high correlations while still differing substantially in how they distribute strengths and weaknesses across analytic aspects within the same essay.

4 Experimental results

Tables 4 and 5 show the results for GPT-4.1, Qwen 2.5 72B, Llama 3.1 70B, and OR on negative and positive feedback, respectively.

	GPT4.1	Qwen2.5	Llama3.1	OR
Int	**44.68**	31.96	40.92	$40.77_{\pm 15.37}$
Cpl	**49.20**	37.20	34.00	$30.76_{\pm 8.01}$
Acc	24.68	16.75	24.66	$\mathbf{26.10_{\pm 3.80}}$
Flu	**22.52**	17.70	20.51	$22.37_{\pm 5.45}$
Cpr	**37.71**	15.71	25.55	$34.10_{\pm 7.78}$
Lgc	**45.45**	36.52	28.18	$23.67_{\pm 7.67}$
Sph	20.52	**27.71**	20.20	$23.83_{\pm 6.11}$
Prp	38.20	46.13	**51.43**	$33.25_{\pm 9.22}$
Wil	16.37	12.72	20.11	$\mathbf{32.79_{\pm 7.61}}$
Inv	37.75	40.51	42.93	$\mathbf{45.87_{\pm 4.38}}$
Avg.	**33.71**	28.29	30.85	31.35

Table 4: **Negative** feedback results in terms of best $F_{0.5}$ across systems.

	GPT4.1	Qwen2.5	Llama3.1	OR
Int	21.02	13.80	22.84	$\mathbf{35.59_{\pm 5.34}}$
Cpl	**74.07**	65.22	70.00	$41.58_{\pm 6.13}$
Acc	29.96	33.61	28.19	$\mathbf{34.75_{\pm 8.60}}$
Flu	19.44	18.69	22.38	$\mathbf{30.23_{\pm 4.04}}$
Cpr	23.65	14.11	23.74	$\mathbf{29.75_{\pm 3.13}}$
Lgc	21.18	37.63	**40.52**	$33.06_{\pm 3.64}$
Sph	37.74	**39.47**	35.71	$29.51_{\pm 6.83}$
Prp	25.68	27.23	37.64	$\mathbf{38.23_{\pm 5.27}}$
Wil	29.25	29.71	26.89	$\mathbf{32.81_{\pm 11.11}}$
Inv	37.68	38.98	**47.70**	$40.24_{\pm 7.25}$
Avg.	31.96	31.84	**35.56**	34.57

Table 5: **Positive** feedback results in terms of best $F_{0.5}$ across systems.

As can be observed, when examining individual analytic aspects under the negative feedback condition (Table 4), LLMs, particularly GPT-4.1, tend to achieve higher $F_{0.5}$ scores than the OR on several *Language* and *Content* dimensions. GPT-4.1 notably outperforms both the other models and OR on *Intelligibility*, *Complexity*, *Fluency*, *Comprehensibility*, and *Logicality*, indicating a stronger alignment with the reference signal in identifying aspects that deviate negatively compared to the overall average score. This suggests that LLMs are especially effective at detecting relative weaknesses in linguistically and structurally grounded dimensions. By contrast, OR attains higher scores on *Accuracy* and on the *Attitude* dimensions (i.e., *Willingness* and *Involvement*), for which all LLMs show comparatively lower performance. Given that *Attitude* aspects are inherently more subjective and have been shown to exhibit lower inter-rater agreement (see Tables 1 and 2), these differences likely reflect increased noise and variability in the reference labels rather than a systematic limitation of the self-referential assessment formulation or the models themselves. Within the group of LLMs, GPT-4.1 consistently leads across most *Language* and *Content* aspects, while Qwen 2.5 performs best on *Sophistication*, and Llama 3.1 achieves its highest scores on *Purposefulness*. However, overall, GPT-4.1 attains the highest average $F_{0.5}$ score for negative feedback.

For positive feedback (Table 5), a different pattern is observed. While LLMs continue to perform strongly on *Complexity*, *Logicality*, and *Sophistication*, the OR outperforms all models on most *Language* aspects, including *Intelligibility*, *Accuracy*, and *Fluency*, as well as on *Comprehensibility* and *Purposefulness*. Among the LLMs, Llama 3.1 achieves the highest overall average $F_{0.5}$ score under positive feedback, driven primarily by its performance on *Logicality*, *Purposefulness*, and *Involvement*. In contrast, GPT-4.1, while dominant in the negative feedback setting, shows comparatively weaker alignment with the positive feedback targets. Overall, we can observe that LLMs, especially GPT-4.1, exhibit stronger alignment with reference signals for identifying relative underperformance, whereas human raters show higher agreement with the reference in identifying relative strengths. For completeness, in addition to $F_{0.5}$, we also report the results in terms of Precision and Recall in Appendix C in Tables 7, 8, 9, and 10 for OR, GPT-4.1, Qwen 2.5, and Llama 3.1, respectively.

In addition to comparing LLMs with a single randomly selected operational rater per essay, we also report performance for ensembles of multiple raters, ranging in size from 2 to 12. Figure 2 reports the errorbars for feedback on *Logicality*. As can be seen, for this aspect, it takes an ensemble of three raters to outperform the best-performing models, i.e., GPT-4.1 on negative feedback and Llama 3.1 on positive feedback.

For completeness, we report the results for all the aspects and one, two, and three operational raters in Tables 11 and 12 in Appendix C. When averaging performance across aspects, an ensemble of two operational raters is sufficient to outperform the LLMs. However, considering individual aspects, as in Figure 2, larger ensembles are required.

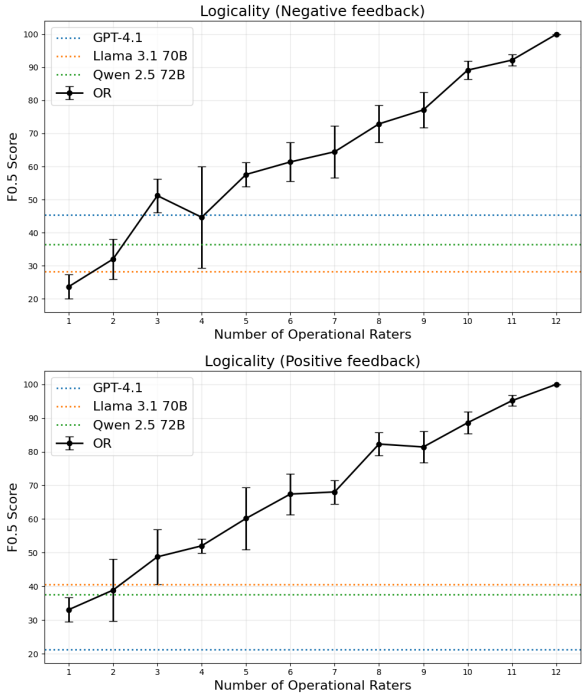


Figure 2: Errorbars on feedback of ensembles with increasing number of operational raters to GPT-4.1, Llama 3.1 70B, and Qwen 2.5 72B for *Logicality*.

5 Conclusions

This paper addresses the challenges of validating diagnostic analytic assessment in AES, particularly where rank-based metrics may overlook the influence of intercorrelations among analytic dimensions of proficiency as well as the halo effect. Through the introduction of a self-referential framework, we demonstrate how intra-learner evaluation serves as a crucial complement to standard inter-learner assessment.

Our results highlight divergent strengths: LLMs tend to outperform operational human raters in pinpointing relative weaknesses (negative feedback), while humans remain more adept at recognising relative strengths (positive feedback). While the primary focus of this paper is the proposed self-referential assessment framework, we acknowledge that our experiments are limited to zero-shot approaches. Future work will investigate alternative prompting strategies, including few-shot prompting and chain-of-thought reasoning.

Finally, while inter-learner rankings provide essential comparative data, we propose combining them with intra-learner diagnostics. This hybrid approach may offer a pathway to delivering more actionable and informative assessment and feedback to learners, teachers, and testers.

Limitations

First, we acknowledge that relying on a single dataset warrants caution; further experiments are therefore necessary to fully validate the effectiveness of our self-referential assessment framework. While the number of unique essays is relatively small ($N = 140$), the original dataset is exceptionally dense, containing over 120,000 ratings from 80 raters. This unique characteristic of the ICNALE GRA allowed us to implement Rasch modelling in a fully crossed rater–essay design, a rigorous approach not feasible with other L2 learner datasets.

Secondly, as the data focuses on Asian learners within an English as a Lingua Franca (ELF) framework, further research is required to determine if these findings generalise to learners from other L1 backgrounds or for other L2s.

Finally, regarding the self-referential framework, while the use of a one-standard-deviation threshold is a heuristic choice, it ensures a consistent data-driven approach and aligns with established methodology in psychometric literature (Berninger and Abbott, 2010). Moreover, it guarantees that a substantial proportion of learners receive both positive and negative feedback, as shown in Table 6 (Appendix C).

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. The authors would like to thank the ALTA Spoken Language Processing Technology Project Team for general discussions and contributions to the evaluation infrastructure.

References

Stefano Bannò, Rao Ma, Mengjie Qian, Siyuan Tang, Kate Knill, and Mark Gales. 2025. Natural Language-based Assessment of L2 Oral Proficiency using LLMs. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 189–193.

Stefano Bannò, Hari Krishna Vydana, Kate Knill, and Mark Gales. 2024. Can GPT-4 do L2 analytic assessment? In *Proc. of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Virginia W Berninger and Robert D Abbott. 2010. Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of educational psychology*, 102(3):635.

Raymond B Cattell. 1944. Psychological measurement: normative, ipsative, interactive. *Psychological review*, 51(5):292.

William Vance Clemans. 1956. *An analytical and empirical examination of some properties of ipsative measures (Psychometric Monograph No. 14)*. Psychometric Society, Richmond, VA.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion volume*. Council of Europe, Strasbourg.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

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, Toronto, Canada. Association for Computational Linguistics.

George Engelhard. 1994. Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2):93–112.

Gwyneth Hughes. 2011. Towards a personal best: A case for introducing ipsative assessment in higher education. *Studies in Higher Education*, 36(3):353–367.

Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5).

Shin'Ichiro Ishikawa. 2011. A new horizon in learner corpus studies: The aim of the ICNALE project. In *Corpora and language technologies in teaching, learning and research*, pages 3–11. University of Strathclyde Press.

Shin'Ichiro Ishikawa. 2020. Aim of the ICNALE GRA Project: Global Collaboration to Collect Ratings of Asian Learners' L2 English Essays and Speeches from an ELF Perspective. *Learner Corpus Studies in Asia and the World*, 5:121–144.

Shin'Ichiro Ishikawa. 2024. The ICNALE Global Rating Archives: A New Assessment Dataset for Learner Corpus Studies. *Learner Corpus Studies in Asia and the World*, 6:13–38.

Shin'Ichiro Ishikawa. 2023. Effects of Raters' L1, Assessment Experience, and Teaching Experience on their Assessment of L2 English Speech: A Study Based on the ICNALE Global Rating Archives. *LEARN Journal: Language Education and Acquisition Research Network*, 16(2):411–428.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated Essay Scoring*. Springer Nature.

Klaus Krippendorff. 2004. Reliability in Content Analysis. *Human Communication Research*, 30(3):411–433.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

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, Seattle, United States. Association for Computational Linguistics.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2008. Analytic scoring of TOEFL® CBT essays: Scores from humans and e-rater®. *ETS Research Report Series*, 2008(1):i–71.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: Recent successes and future directions. In *Proc. of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8114–8122. Survey Track.

John Michael Linacre. 1989. *Many-faceted Rasch measurement*. Ph.D. thesis, The University of Chicago.

John Michael Linacre. 2002. What do infit and outfit, mean-square and standardized mean. *Rasch measurement transactions*, 16(2):878.

Llama Team. 2024. The Llama 3 herd of models.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2006. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, 28(4):587–604.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J.F. Gales. 2025. Assessment of L2 Oral Proficiency using Speech Large Language Models. In *Interspeech 2025*, pages 5078–5082.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can Neural Networks Automatically Score Essay Traits? In *Proc. of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91. Association for Computational Linguistics.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proc. of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. 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.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Ellis B Page, John P Poggio, and Timothy Z Keith. 1997. Computer Analysis of Student Essays: Finding Trait Differences in Student Profile. In *Annual Meeting of the American Educational Research Association*, Chicago, IL.

Vatsal Raina, Nataliia Molchanova, Mara Graziani, Andrey Malinin, Henning Muller, Meritxell Bach Cuadra, and Mark Gales. 2023. Tackling Bias in the Dice Similarity Coefficient: Introducing nDSC for White Matter Lesion Segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proc. of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

Barbara Seidlhofer. 2005. English as a lingua franca. *ELT Journal*, 59(4):339–341.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472.

Mark D Shermis, Chantal Mees Koch, Ellis B Page, Timothy Z Keith, and Susanmarie Harrington. 2002. Trait ratings for automated essay grading. *Educational and Psychological Measurement*, 62(1):5–18.

Edward L Thorndike. 1920. A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29.

Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025. LLMs can Perform Multi-Dimensional Analytic Writing Assessments: A Case Study of L2 Graduate-Level Academic English Writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8637–8663, Vienna, Austria. Association for Computational Linguistics.

Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge Language Assessment. Cambridge University Press, Cambridge.

Taichi Yamashita. 2024. An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, 3(3):100133.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proc. of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of Automated Text Scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, Denver, Colorado. Association for Computational Linguistics.

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13439–13454, Vienna, Austria. Association for Computational Linguistics.

Bo Zhang, Yunnan Xiao, and Juan Luo. 2015. Rater reliability and score discrepancy under holistic and analytic scoring of second language writing. *Language Testing in Asia*, 5(1):5.

A Appendix A: Prompt

```
In the context of an examination
of English as a Lingua Franca
(ELF), a second language (L2)
learner of English is asked to
write an essay in response to the
following prompt:

Do you agree or disagree with the
following statement? Use reasons
and specific details to support
your opinion.

'It is important for college
students to have a part-time
job.'
```

`You have to score this essay by only considering the aspect of [ASPECT].`

`[ANALYTIC RATING PROMPT]`[8]

`Select a score from 0 (lowest) to 9 (highest). Only output the most suitable score without adding any comment or explanation.`

`Essay: [ESSAY]`

B Appendix B: Analytic rating prompts

Since ICNALE GRA contains essays and speeches, the original analytic rating prompts (Ishikawa, 2024) addressed both modalities. For our experiments with LLMs, we therefore tailored the prompts to retain only the content relevant to essays.

Language

Intelligibility

To which extent can you "decode", namely, verbally understand what is written? Factors such as spelling and sentence structure may influence it. Please note that intelligibility, which concerns the understandability of the language, should be discriminated from comprehensibility, which concerns the understandability of the content. You may sometimes find an essay that is intelligible but not comprehensible, such as a logically nonsense statement. Meanwhile, you may usually not find an essay that is comprehensible but not intelligible because if the text cannot be decoded, its content cannot be conveyed.

Complexity

To what extent do you think the writer uses morphologically and/or semantically complex words, phrases, expressions, constructions, and grammar? Complexity is seen at many levels of language. For example, "I speculate..." usually sounds more complex than "I think" (Vocabulary). "It is speculated that..." may sound more complex than "I speculate" (Voice, Construction). "If I were a bird" may sound more complex than "If I am a bird" (Subjunctive, Grammar).

Accuracy

To what extent do you think the sample is error-free in terms of vocabulary and grammar? In addition, you should examine the elements such as punctuation. Please note that you should ignore minor and only-once errors, which may be mistakes rather than errors. Please note that the standard for evaluation should be a proficient non-native ELF speaker, not an English native speaker.

Fluency

To what extent do you think the writer is fluent in the essays? Fluency needs to be evaluated in two ways: (a) fluency and (b) disfluency. If someone writes more, the fluency score should increase, while if s/he uses more disfluency markers, the score may decrease. Disfluency markers include unnecessary connectors (and, but, so because) and semantically empty phrases (such as "I think" most typically), etc. Please note that using these disfluency markers once or twice usually does not cause any problems in communication.

Content

Comprehensibility

To what extent can you understand the content of the essay? Please note that comprehensibility, which concerns the understandability of the content, should be discriminated from intelligibility, which concerns the understandability of the language. If a writer presents a logically reasonable idea, the score should increase.

Logicality

To what extent do you think the idea presented in the essay is logical and reasonable? You need to examine whether the reasons and the conclusions are logically connected.

Sophistication

To what extent do you think the ideas presented in the essay are well-sophisticated, critically thought, unique, original, and innovative?

Purposefulness

To what extent do you think the writer consistently and consciously pays attention to the purpose of the task? The participant was requested to persuade a supervisor to allow them to show their own opinion about part-time jobs for college students in an essay. You have to examine whether the participant fully understands the purpose of the task and consistently sticks to it. Purposefulness is closely related to task completion.

[8] See Appendix B.

Attitude

Willingness to communicate

To what extent do you think the writer is willing to communicate? It is possible that a participant with a limited L2 proficiency shows a high level of willingness to communicate (WTC), and it is also possible that a participant with a high L2 proficiency shows quite a low level of WTC. Factors such as the quantity of writing, the number of ideas s/he presents, and the use of various amplifiers (e.g., "very," "surely," "definitely," "I strongly believe," etc.) may represent the participant's WTC.

Involvement

To what extent do you think the participant tries to make the reader involved in his/her discourse rather than writing one-sidedly? The factors such as the use of the second-person pronouns (e.g., "You know," "as you see," "as you expect," etc.) and mentioning the reader are usually related to the degree of involvement.

Holistic

To what extent do you think the sample is close to an ideal ELF essay? Raters have to examine each sample and decide the score (0-100) based on the overall judgment of its quality as a professional ELF output. Please note that "9 points", for example, should be given to someone who you think is a 100% ideal professional ELF user, not to someone who you think is 100% close to English native speakers. Also, please note that the middle point is 5.

C Appendix C: Additional statistics and results

Table 6 reports the percentage of essays receiving positive and negative feedback across macro-aspects.

Figure 3 shows scatterplots comparing GPT-4.1 predictions with reference fair average scores for the three selected analytic aspects and holistic proficiency. For visualisation purposes, GPT-4.1's predictions were linearly calibrated using the same essay test data.

Tables 7, 8, 9, and 10 show the complete feedback results in terms of Precision, Recall, and $F_{0.5}$ for OR, GPT-4.1, Qwen 2.5 72B, and Llama 3.1 70B, respectively. Tables 11 and 12 show the feedback results in terms of $F_{0.5}$ for one random operational rater (OR), an ensemble of two random operational raters (OR2), and three random operational raters (OR3).

Macro-aspect	Negative	Positive	All
Language	45.71	47.14	77.86
Content	48.57	49.29	79.29
Attitude	27.14	31.43	54.29
All	81.43	87.14	92.86

Table 6: Percentage of essays receiving feedback across macro-aspects.

	Negative				Positive			
	P	**R**	$F_{0.5}$	**Prev.**	**P**	**R**	$F_{0.5}$	**Prev.**
Int	$47.54_{\pm 28.59}$	$37.89_{\pm 9.05}$	$40.77_{\pm 15.37}$	13.57	$43.39_{\pm 15.12}$	$29.17_{\pm 11.18}$	$35.59_{\pm 5.34}$	17.14
Cpl	$34.25_{\pm 13.44}$	$34.74_{\pm 13.56}$	$30.76_{\pm 8.01}$	13.57	$60.66_{\pm 23.40}$	$23.64_{\pm 7.27}$	$41.58_{\pm 6.13}$	15.71
Acc	$30.99_{\pm 14.06}$	$31.11_{\pm 19.26}$	$26.10_{\pm 3.80}$	19.29	$68.73_{\pm 25.64}$	$13.33_{\pm 4.86}$	$34.75_{\pm 8.60}$	17.14
Flu	$20.42_{\pm 6.94}$	$55.00_{\pm 18.71}$	$22.37_{\pm 5.45}$	20.00	$48.48_{\pm 26.92}$	$20.77_{\pm 10.20}$	$30.23_{\pm 4.04}$	18.57
Cpr	$51.23_{\pm 26.36}$	$25.45_{\pm 12.40}$	$34.10_{\pm 7.78}$	15.71	$33.10_{\pm 6.91}$	$24.17_{\pm 5.53}$	$29.75_{\pm 3.13}$	17.14
Lgc	$24.55_{\pm 9.37}$	$40.00_{\pm 24.58}$	$23.67_{\pm 7.67}$	15.0	$62.56_{\pm 21.50}$	$17.39_{\pm 13.19}$	$33.06_{\pm 3.64}$	16.43
Sph	$41.78_{\pm 30.56}$	$24.54_{\pm 26.44}$	$23.83_{\pm 6.11}$	15.71	$33.01_{\pm 9.76}$	$34.00_{\pm 23.54}$	$29.51_{\pm 6.83}$	14.29
Prp	$38.72_{\pm 8.42}$	$22.86_{\pm 9.71}$	$33.25_{\pm 9.22}$	15.00	$64.44_{\pm 29.71}$	$22.86_{\pm 10.17}$	$38.23_{\pm 5.27}$	15.00
Wil	$46.37_{\pm 28.31}$	$26.15_{\pm 11.77}$	$32.79_{\pm 7.61}$	18.57	$33.29_{\pm 13.41}$	$42.61_{\pm 20.83}$	$32.81_{\pm 11.11}$	16.43
Inv	$62.20_{\pm 9.49}$	$24.61_{\pm 7.54}$	$45.87_{\pm 4.38}$	9.29	$63.79_{\pm 30.62}$	$28.33_{\pm 14.04}$	$40.24_{\pm 7.25}$	17.14

Table 7: Feedback results in terms of best Precision, Recall, and $F_{0.5}$ (OR).

	Negative				Positive			
	P	**R**	$F_{0.5}$	**Prevalence**	**P**	**R**	$F_{0.5}$	**Prevalence**
Intelligibility	54.11	26.32	44.68	13.57	18.51	45.83	21.02	17.14
Complexity	73.89	21.05	49.20	13.57	100.00	36.36	74.07	15.71
Accuracy	21.42	62.96	24.68	19.29	31.52	25.00	29.96	17.14
Fluency	24.10	17.86	22.52	20.00	17.80	30.77	19.44	18.57
Comprehensibility	41.70	27.27	37.71	15.71	20.59	58.33	23.65	17.14
Logicality	100.00	14.29	45.45	15.00	18.44	52.17	21.18	16.43
Sophistication	18.50	36.36	20.52	15.71	37.21	40.00	37.74	14.29
Purposefulness	36.40	47.62	38.20	15.00	32.08	14.29	25.68	15.00
Willingness	15.78	19.23	16.37	18.57	32.02	21.74	29.25	16.43
Involvement	44.88	23.08	37.75	9.29	47.23	20.83	37.68	17.14

Table 8: Feedback results in terms of best Precision, Recall, and $F_{0.5}$ (GPT-4.1).

	Negative				Positive			
	P	**R**	$F_{0.5}$	**Prevalence**	**P**	**R**	$F_{0.5}$	**Prevalence**
Intelligibility	32.05	31.58	31.96	13.57	11.38	91.67	13.80	17.14
Complexity	36.15	46.11	37.20	13.57	100.00	27.27	65.22	15.71
Accuracy	13.88	96.30	16.75	19.29	34.94	29.17	33.61	17.14
Fluency	15.09	57.14	17.70	20.00	17.02	30.77	18.69	18.57
Comprehensibility	37.34	27.27	34.77	15.71	11.70	79.17	14.11	17.14
Logicality	33.50	57.14	36.52	15.00	53.06	17.39	37.63	16.43
Sophistication	37.34	13.64	27.71	15.71	66.67	15.00	39.47	14.29
Purposefulness	71.58	19.05	46.13	15.00	28.24	23.81	27.23	15.00
Willingness	10.47	92.31	12.72	18.57	36.11	17.39	29.71	16.43
Involvement	68.46	15.38	40.51	9.29	36.94	50.00	38.98	17.14

Table 9: Feedback results in terms of best Precision, Recall, and $F_{0.5}$ (Qwen 2.5 72B).

	Negative				Positive			
	P	**R**	$F_{0.5}$	**Prevalence**	**P**	**R**	$F_{0.5}$	**Prevalence**
Intelligibility	67.98	15.79	40.92	13.57	19.96	54.17	22.84	17.14
Complexity	34.67	31.58	34.00	13.57	100.00	31.82	70.00	15.71
Accuracy	23.66	29.63	24.66	19.29	30.92	20.93	28.19	17.14
Fluency	19.16	28.57	20.51	20.00	23.34	19.23	22.38	18.57
Comprehensibility	28.43	18.18	25.55	15.71	20.28	75.00	23.74	17.14
Logicality	26.46	38.10	28.18	15.00	44.17	30.43	40.52	16.43
Sophistication	22.96	13.64	20.20	15.71	100.00	10.00	35.71	14.29
Purposefulness	59.50	33.33	51.43	15.00	44.04	23.81	37.64	15.00
Willingness	21.78	15.38	20.11	18.57	31.14	17.39	26.89	16.43
Involvement	40.85	52.85	42.93	9.29	61.70	25.00	47.70	17.14

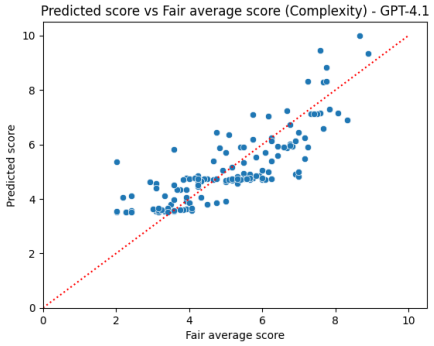
Table 10: Feedback results in terms of best Precision, Recall, and $F_{0.5}$ (Llama 3.1 70B).

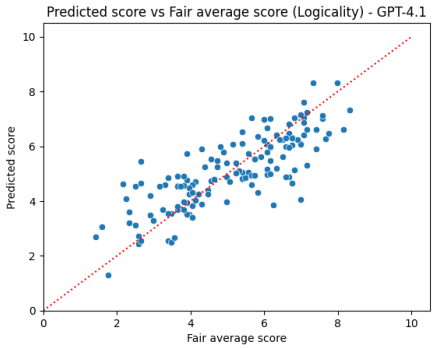
	OR	OR2	OR3	GPT4.1	Qwen2.5	Llama3.1
Int	$40.77_{\pm 15.37}$	$34.06_{\pm 3.71}$	$45.30_{\pm 12.82}$	44.68	31.96	40.92
Cpl	$30.76_{\pm 8.01}$	$41.77_{\pm 9.13}$	$47.32_{\pm 6.78}$	49.20	37.20	34.00
Acc	$26.10_{\pm 3.80}$	$33.71_{\pm 3.47}$	$44.75_{\pm 8.62}$	24.68	16.75	24.66
Flu	$22.37_{\pm 5.45}$	$41.12_{\pm 18.23}$	$44.13_{\pm 6.46}$	22.52	17.70	20.51
Cpr	$34.10_{\pm 7.78}$	$38.42_{\pm 14.26}$	$41.35_{\pm 9.32}$	37.71	15.71	25.55
Lgc	$23.67_{\pm 7.67}$	$32.01_{\pm 6.07}$	$51.20_{\pm 5.14}$	45.45	36.52	28.18
Sph	$23.83_{\pm 6.11}$	$23.64_{\pm 5.44}$	$40.02_{\pm 9.57}$	20.52	27.71	20.20
Prp	$33.25_{\pm 9.22}$	$44.81_{\pm 8.60}$	$59.79_{\pm 9.10}$	38.20	46.13	51.43
Wil	$32.79_{\pm 7.61}$	$33.16_{\pm 4.57}$	$38.74_{\pm 10.10}$	16.37	12.72	20.11
Inv	$45.87_{\pm 4.38}$	$49.22_{\pm 15.24}$	$52.20_{\pm 10.90}$	37.75	40.51	42.93
Avg.	31.35	37.19	46.48	33.71	28.29	30.85

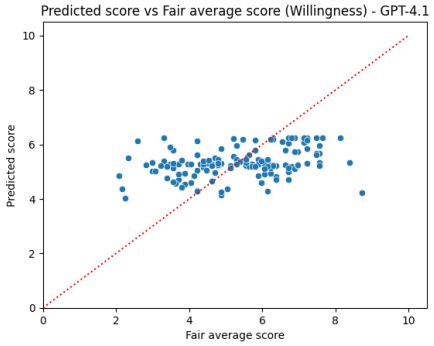
Table 11: Negative feedback results in terms of best $F_{0.5}$. OR, OR2, OR3 vs GPT4.1, Qwen2.5, Llama3.1.

	OR	OR2	OR3	GPT4.1	Qwen2.5	Llama3.1
Int	$35.59_{\pm 5.34}$	$35.59_{\pm 5.60}$	$39.92_{\pm 5.71}$	21.02	13.80	22.84
Cpl	$41.58_{\pm 6.13}$	$36.92_{\pm 5.66}$	$56.44_{\pm 11.69}$	74.07	65.22	70.00
Acc	$34.75_{\pm 8.60}$	$45.15_{\pm 7.61}$	$50.28_{\pm 8.58}$	29.96	33.61	28.19
Flu	$30.23_{\pm 4.04}$	$33.46_{\pm 7.41}$	$42.79_{\pm 5.08}$	19.44	18.69	22.38
Cpr	$29.75_{\pm 3.13}$	$30.29_{\pm 6.46}$	$45.63_{\pm 10.32}$	23.65	14.11	23.74
Lgc	$33.06_{\pm 3.64}$	$38.87_{\pm 9.27}$	$48.76_{\pm 8.13}$	21.18	37.63	40.52
Sph	$29.51_{\pm 6.83}$	$28.29_{\pm 4.51}$	$36.85_{\pm 9.96}$	37.74	39.47	35.71
Prp	$38.23_{\pm 5.27}$	$35.08_{\pm 5.54}$	$45.71_{\pm 13.48}$	25.68	27.23	37.64
Wil	$32.81_{\pm 11.11}$	$50.18_{\pm 10.59}$	$50.21_{\pm 12.03}$	29.25	29.71	26.89
Inv	$40.24_{\pm 7.25}$	$51.79_{\pm 6.22}$	$61.79_{\pm 6.49}$	37.68	38.98	47.70
Avg.	34.57	38.56	47.84	31.96	31.84	35.56

Table 12: Positive feedback results in terms of best $F_{0.5}$. OR, OR2, OR3 vs GPT4.1, Qwen2.5, Llama3.1.







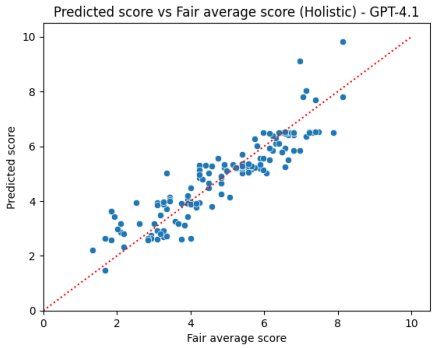


Figure 3: Scatterplots between fair average scores and GPT-4.1 scores for selected aspects.