applied sciences

MDPI

Article

PROSE: Rubric-Aware Multi-Objective Fine-Tuning of Small Language Models for Automated Assessment of Second-Language Writing

Hua Xu [1], Jie Ding [2], Rui Xu [3] and Qiang Tang [4,*]

[1] Independent Researcher, Nashville, TN 37203, USA; xuhualexi@163.com
[2] Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA; jieding@iastate.edu
[3] College of Outstanding Engineers, Changsha University of Science and Technology, Changsha 410114, China; xr_cljs@stu.csust.edu.cn
[4] School of Artificial Intelligence, Anhui University of Science and Technology, Hefei 231131, China
* Correspondence: tangqiang@whu.edu.cn

Abstract

Automated assessment of second-language (L2) writing supports language teaching and testing by providing rapid, consistent scores across multiple rubric dimensions. We study whether small, openly available language models can be turned into reliable analytic raters, and we introduce PROSE, a rubric-aware, multi-objective fine-tuning method built on a compact DeBERTa-v3 encoder with low-rank adaptation. PROSE trains a regression read-out together with a soft-ordinal auxiliary head, and combines a squared-error term for calibration, a distance-aware soft-ordinal term that respects the ordered rubric bands, a concordance term that restores the spread of human ratings and improves agreement with human ratings, and a trait-relational coherence term. On the ELLIPSE corpus of L2 English essays, PROSE raises the quadratic weighted kappa (QWK) from 0.646 for a strong regression baseline to 0.661—a small but statistically reliable gain under a paired test across folds (mean difference +0.015—95% CI [0.007, 0.023], and paired t-test $p = 0.007$; $p = 0.030$ after correcting for the overlap between cross-validation training sets), with comparable error and improvements on all seven scoring dimensions, four of which remain significant after Holm correction, and on the Write & Improve (BEA-2019) proficiency data, it matches the strongest baseline on agreement and accuracy and clearly outperforms classification and pure ordinal training. Zero-shot prompting of three strong large language models (GPT-5.4, Claude-Sonnet-4-6, DeepSeek-Chat) reaches a kappa of only 0.13 to 0.21 on the same essays, and a distributional analysis shows that the tested large models produce compressed score distributions with means below the human mean, while the compact fine-tuned model reproduces the human score distribution. Further analyses show that the gain is stable under reasonable changes of the loss weights and holds for truncated and non-truncated essays and across gender, ethnicity, socioeconomic, and grade groups, although these analyses also reveal a shared grade-related bias and a reliance on length and lexical-diversity cues that call for caution in deployment. The results indicate that, for this task, a task-specific objective applied to a small supervised model is far more effective than general-purpose models used zero-shot; we are careful to attribute the gap to task-specific supervision and objective design rather than to model scale alone.

Keywords: automated essay scoring; second-language writing; language assessment; ordinal regression; parameter-efficient fine-tuning; large language models

Check for updates

Academic Editor: Marcos Gestal

Received: 8 September 2026
Revised: 19 October 2026
Accepted: 23 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Writing is one of the most demanding skills for second-language (L2) learners, and timely feedback on writing quality is central to language teaching, learning, and assessment. Human rating of learner essays is accurate but slow and expensive, and it is hard to keep consistent across large cohorts. Automated writing assessment, and automated essay scoring (AES) in particular, aims to reproduce expert judgements at scale so that learners receive frequent feedback and teachers can focus their attention where it matters most [1,2]. In the applied linguistics tradition, the value of such feedback is well documented: written corrective feedback helps learners notice the gap between their production and the target norm, and it supports measurable gains in accuracy over time [3–6]. A system that can score writing along interpretable rubric dimensions is therefore both a grading tool and a starting point for formative feedback.

Modern AES systems are built on pretrained language models, which map an essay to a score through a learned regression or classification head [7–10]. Two developments motivate a fresh look at the problem. First, encoder models such as BERT, RoBERTa, and DeBERTa [11–14] provide strong text representations that can be adapted cheaply with low-rank methods [15], which makes it feasible to train and deploy accurate raters on modest hardware. Second, the arrival of powerful instruction-tuned large language models (LLMs) has raised the question of whether scoring can be done by prompting alone, with no task-specific training [16–19]. The evidence so far is mixed. LLMs produce fluent, rubric-like justifications, yet studies of grammatical error correction and essay scoring report that they tend to be poorly calibrated against human minimal-edit or rubric references and often drift toward safe, central judgements [16,19].

This paper asks a focused question: given a small, openly available encoder, what is the training objective that turns it into a reliable analytic rater of L2 writing? We argue that three properties of the scoring task should be built into the loss rather than left to a generic regression head. (i) Rubric bands are *ordinal*: a band of 3.0 is closer to 3.5 than to 1.0, so the penalty for an error should grow with distance. (ii) Agreement is evaluated using an ordered-category metric such as quadratic weighted kappa, so the decoded value must stay *calibrated* as well as correctly ranked. (iii) Analytic rubrics are *relational*: the six ELLIPSE traits (cohesion, syntax, vocabulary, phraseology, grammar, conventions) are strongly correlated with each other and with the holistic score, and a rater should exploit this structure instead of scoring each trait in isolation.

We combine these three properties into PROSE (Proficiency Rating via Ordinal Soft-labels and trait Ensembling), a rubric-aware, multi-objective fine-tuning method. PROSE places a shared DeBERTa-v3 encoder with low-rank adapters under a set of per-trait heads, and it trains them with a distance-aware soft-ordinal objective, a squared-error regression term, a concordance term, and a trait-relational coherence term. Each component is simple and inexpensive, and each is motivated by one of the three properties above. We deliberately favour a compact and reproducible design over architectural novelty, because our aim is a method that works reliably and can be run in a day on a single GPU.

We evaluate PROSE on two openly available L2 writing datasets that require no restricted access: the ELLIPSE corpus of English-language-learner essays with six analytic traits and a holistic score [20], and the Write & Improve learner data from the BEA-2019 shared task, where each text carries a CEFR proficiency band [21]. On ELLIPSE, PROSE improves quadratic weighted kappa over regression, classification, and pure ordinal baselines, with comparable error and gains on all seven traits; on Write & Improve, it matches the strongest baseline and outperforms classification and pure ordinal training. A comparison with zero-shot prompting of three strong LLMs, evaluated on the same essays, shows a large gap in favour of the task-specific fine-tuned model and a distributional analysis over

the full sample explains the gap by showing that the tested LLMs produce compressed score distributions with means below the human mean.

We state at the outset the scope of our central comparison, because the point recurs below. Our fine-tuned models are trained on labelled ELLIPSE data, whereas the LLMs are prompted zero-shot; the comparison therefore measures the combined effect of task-specific supervision and of the objective we propose and does not isolate model scale as a causal factor. We draw the narrower conclusion that a small, task-specifically fine-tuned rater is the more accurate option that is inexpensive to train and run locally under the conditions tested, and we are explicit throughout about what the experiment does and does not show.

Our contributions are as follows:

1. We formulate automated analytic assessment of L2 writing as a rubric-aware, multi-objective learning problem, and we identify three task properties (ordinality, calibration, and trait relations) that a scoring objective should encode.
2. We propose PROSE, a compact method that realises these properties through a regression read-out with a soft-ordinal auxiliary head, together with a concordance term that improves agreement with human ratings and a trait-relational coherence term, all on a shared low-rank-adapted encoder.
3. On a multi-trait analytic corpus, we report statistically validated improvements (paired tests and effect sizes across folds) in quadratic weighted kappa over strong baselines, with comparable error metrics and a gain on each of the seven scoring dimensions; on a separate proficiency corpus, we detect no significant difference from the regression baseline; ablations isolate the effect of each component, and a loss-weight sensitivity analysis together with subgroup analyses by essay length and learner background, tests how robust the gain is.
4. We provide a controlled comparison against zero-shot prompting of GPT-5.4, Claude-Sonnet-4-6, and DeepSeek-Chat on an identical essay sample, with full prompt and decoding details, as well as a quantitative distributional analysis that clarifies why parameter-efficient fine-tuning of a small model remains the stronger option for this task.

The remainder of the paper is organised as follows. Section 2 reviews automated writing assessment, LLM-based scoring, and the ordinal and multi-task learning methods that inform our design. Section 3 presents PROSE. Section 4 describes datasets, metrics, and implementation. Section 5 reports results, ablations, the LLM comparison, and the distributional analysis. Section 6 discusses implications and limitations, and Section 7 concludes.

2. Related Work

2.1. Automated Writing Assessment

Early neural AES replaced hand-engineered features with recurrent and convolutional encoders trained end to end [7,8], and attention-based recurrent convolutional models improved the modelling of essay structure [9]. Transformer encoders later became the default backbone, and much of the recent effort has moved to robustness across prompts and to analytic, multi-trait scoring rather than a single holistic number [10]. Surveys of the field document a steady shift from feature engineering to representation learning and a growing interest in interpretability and fairness [1,2]. A recurring design question is how to treat the score itself: as a continuous target for regression, as a set of unordered classes, or as an ordered variable. Regression is the most common choice and is a strong baseline, but it optimises squared error rather than the agreement metrics for ordered scores on which systems are usually judged, and it tends to compress predictions toward the centre of the scale. Classification respects neither the order nor the metric. These tensions are precisely

what a rubric-aware objective can resolve, and they motivate the multi-objective design we adopt. Our work sits in this line but focuses on the training objective rather than the architecture: we keep the backbone deliberately small and ask what loss makes it a reliable analytic rater. The ELLIPSE corpus we use provides six analytic traits scored by trained raters and was released to support exactly this kind of study [20].

2.2. Large Language Models for Scoring and Feedback

Instruction-tuned LLMs can be prompted to score essays or to correct and explain learner errors. Studies in applied linguistics and natural language processing report that such models produce plausible rubric-style output, yet their agreement with human raters is uneven and sensitive to prompt design [16–18]. In grammatical error correction, a systematic comparison of open and commercial models found that they over-correct relative to gold minimal-edit references and therefore lag supervised systems on reference-based metrics, despite producing fluent rewrites [19]. These observations motivate our controlled comparison: rather than assuming that a larger model scores better, we measure zero-shot LLMs against a fine-tuned compact model on the same essays and the same rubric, while being explicit that this contrasts zero-shot general-purpose use with task-specific supervision. Related shared tasks such as BEA-2019 and the ERRANT error-annotation toolkit standardised evaluation for learner-writing systems and provide the CEFR-labelled data we use for our second benchmark [21,22].

2.3. Written Corrective Feedback and Assessment in Applied Linguistics

Research on written corrective feedback in second-language acquisition provides the pedagogical rationale for automated writing assessment. A long line of work shows that feedback which is specific and consistent helps learners notice the gap between their production and the target norm and supports gains in accuracy, particularly for well-defined features of the language [3,4]. The form and focus of feedback matter: studies comparing direct and indirect feedback, and focused against unfocused feedback, find that learners respond differently depending on how errors are flagged and explained [5,6]. Analytic assessment sits upstream of this process. Before a system can decide what feedback to give, it must form a reliable judgement of where a text stands on each dimension of quality, and it must do so consistently across many learners. Our work targets this judgement step, and it treats the analytic trait profile as the interpretable substrate on which formative feedback can later be built. This connects the machine learning problem of scoring to the applied-linguistics goal of feedback that is both accurate and actionable.

2.4. Ordinal Regression and Label Structure

Rubric scores are ordered categories, and treating them as unordered classes discards information. Ordinal regression addresses this by encoding order in the labels or in the model. Rank-consistent formulations enforce a monotone set of binary decisions and improve ordinal error [23], while soft label smoothing spreads probability mass to neighbouring categories in proportion to their distance from the target, which reduces large mistakes and calibrates the predictive distribution [24]. We adopt a distance-aware soft-ordinal objective in this spirit as an auxiliary head, and we pair it with a concordance term [25] so that the read-out score is directly encouraged to agree with human ratings in location, scale, and rank; we are careful to distinguish this differentiable surrogate from the quadratic weighted kappa [26] that we use to report agreement (Section 3). Recent work in Pattern Recognition continues to refine label-structure-aware learning, including label-semantic guidance for few-shot text classification [27] and prompting-based attention

for short-text classification [28], which supports the view that how labels and prompts are encoded is often decisive for text tasks.

2.5. Multi-Task Learning and Parameter-Efficient Adaptation

Analytic scoring is naturally multi-task, since the traits share linguistic evidence. Multi-task learning improves data efficiency when tasks are related, but it requires care in how task objectives are balanced. Recent studies propose principled task weighting under uncertainty [29], progressively sparse multi-task architectures [30], and network Lasso formulations for personalised multi-task learning [31]. We take a lightweight route: a shared encoder with per-trait heads and a coherence term that links the traits to the holistic score, which captures the relational structure without a heavy architecture. For adaptation we use low-rank tuning [15], which updates a small fraction of parameters and has been extended to continual and subspace variants [32] and to curvature-aware, trust-region-guided updates [33]. Prompting is a similarly lightweight route, and self-supervised visual prompting has been shown to carry a detector across domains without touching the backbone [34]. Related structure-aware methods have proved effective on other prediction tasks, including multimodal named-entity recognition [35] and node classification in text-attributed graphs [36], as well as co-speech gesture generation through entropy-guided vector-quantised latents [37]. Low-rank tuning keeps our training fast and reproducible on a single GPU, in line with the goal of a method that can be run in a day.

3. Method

3.1. Overview

PROSE scores an essay along T ordered rubric dimensions. For ELLIPSE, $T = 7$ (the holistic Overall score plus six analytic traits), and each dimension takes one of $K = 9$ bands in $\{1.0, 1.5, \ldots, 5.0\}$. For Write & Improve, $T = 1$ and the label is one of $K = 6$ CEFR levels ordered from A1 to C2. Figure 1 shows the pipeline. A shared encoder produces one pooled representation of the essay. Two heads read this representation: a regression head that outputs the score for each dimension, and an auxiliary soft-ordinal head that predicts a distribution over the rubric bands and shapes the representation to be ordinally structured. Training combines four terms that target the properties identified in Section 2: a squared-error term for calibration, a distance-aware soft-ordinal term for ordinality, a concordance term that improves agreement with human ratings, and a trait-relational coherence term. At inference, the score of each dimension is read directly from the regression head.

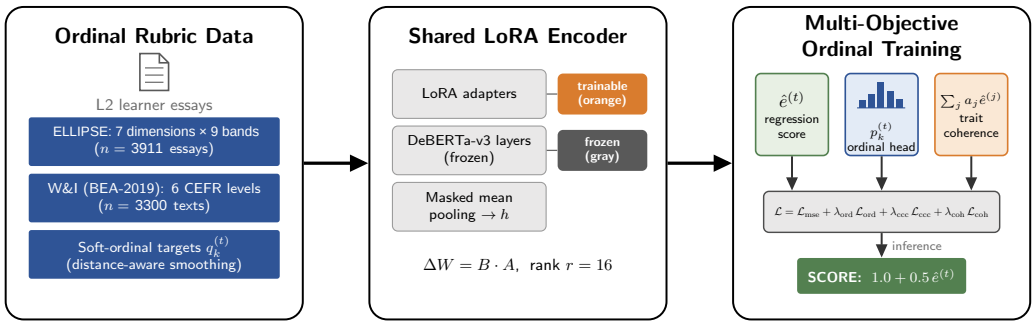


Figure 1. High-level conceptual overview of PROSE (see Figure 2 for the implementation-level architecture). A shared DeBERTa-v3 encoder with low-rank adapters maps the essay to a pooled representation, read by a regression head (the score used at inference) and an auxiliary soft-ordinal head. Training combines a squared-error term, a distance-aware soft-ordinal term, a concordance term that restores the spread of human ratings, and a trait-relational coherence term that ties the six analytic traits to the holistic Overall score.

3.2. Encoder and Heads

We encode the essay with a DeBERTa-v3 model [14] and adapt it with low-rank adapters (LoRA) [15] of rank $r = 16$ on the query, key, value, and dense projections, leaving the pretrained weights frozen. Each adapted projection with frozen weight $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is replaced by

$$Wx = W_0 x + \frac{\alpha}{r} BAx, \qquad A \in \mathbb{R}^{r \times d_{\text{in}}}, \quad B \in \mathbb{R}^{d_{\text{out}} \times r}, \tag{1}$$

where only A and B are trained, and α is a fixed scaling factor. Let $H \in \mathbb{R}^{L \times d}$ be the sequence of contextual token vectors for an essay of length L, and let $m \in \{0,1\}^L$ be the attention mask. We use masked mean pooling,

$$h = \frac{\sum_{i=1}^{L} m_i H_i}{\sum_{i=1}^{L} m_i}, \tag{2}$$

followed by dropout. Two linear heads read h. A regression head produces one value per dimension,

$$\hat{e}^{(t)} = (K-1)\,\sigma\big(w_r^{(t)\top} h + b_r^{(t)}\big) \in [0, K-1], \tag{3}$$

which is the score used at inference. An auxiliary ordinal head produces per-dimension logits over the K ordered bands,

$$z^{(t)} = W_o^{(t)} h + b_o^{(t)} \in \mathbb{R}^K, \tag{4}$$

which are normalised into a predictive distribution over the bands,

$$p^{(t)} = \text{softmax}\big(z^{(t)}\big); \tag{5}$$

this head is used only during training. We index bands by $k \in \{0, \ldots, K-1\}$ and write $y^{(t)}$ for the gold band index of dimension t. Figure 2 details the full architecture, from the tokenised essay through the adapted encoder and the two heads to the four loss terms defined in the following subsections.

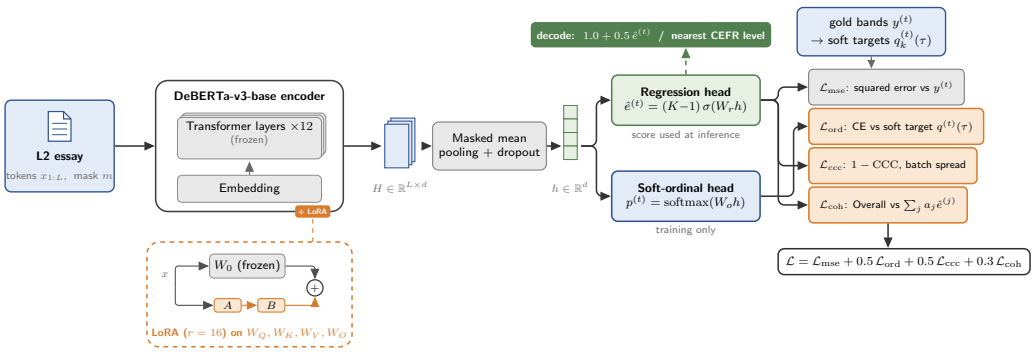


Figure 2. Implementation-level architecture of PROSE, complementing the conceptual overview of Figure 1. A frozen DeBERTa-v3-base encoder carries trainable low-rank adapters of rank 16 on the query, key, value, and output projections (inset). Masked mean pooling turns the token representations into a single vector h, which is read by the regression head, whose output is the score used at inference, and by the soft-ordinal head, which is active only during training. The regression output feeds the squared-error, concordance, and coherence terms; the ordinal head feeds the soft-ordinal term; and the four terms form the weighted training objective; the three rubric-aware terms (highlighted) are the contribution of this work.

3.3. Squared-Error Calibration

The regression head is trained to match the gold band with a squared-error term,

$$\mathcal{L}_{\text{mse}} = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{e}^{(t)} - y^{(t)}\right)^2. \tag{6}$$

This keeps the read-out score calibrated on the rubric scale, which is what an error metric such as RMSE and a quadratic weighted agreement metric both reward at the centre of the scale.

3.4. Distance-Aware Soft-Ordinal Auxiliary

Rubric bands are ordered, so a plain regressor ignores useful structure. We add an auxiliary objective that trains the ordinal head against a soft target that places mass on bands in proportion to their proximity to $y^{(t)}$,

$$q_k^{(t)} = \frac{\exp\left(-(k - y^{(t)})^2/\tau\right)}{\sum_{j=0}^{K-1}\exp\left(-(j - y^{(t)})^2/\tau\right)}, \tag{7}$$

where the temperature τ controls how quickly the target decays with distance [24]. The ordinal head is trained with the cross-entropy between this soft target and its predictive distribution,

$$\mathcal{L}_{\text{ord}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=0}^{K-1} q_k^{(t)} \log p_k^{(t)}. \tag{8}$$

Because the soft target $q^{(t)}$ is fixed with respect to the model parameters, this cross-entropy differs from the Kullback–Leibler divergence $\mathrm{KL}(q^{(t)} \,\|\, p^{(t)})$ only by the constant entropy of $q^{(t)}$, so the two are equivalent as training objectives; we use the cross-entropy form throughout and label the corresponding baseline SOrd (CE) in Section 5. Because the two heads share the encoder, this term is intended to shape the representation to respect the order of the bands and thereby to help the regression read-out, without changing how it is decoded. Empirically its contribution is small and not statistically established (Section 5), so we regard it as a low-cost regulariser rather than as a main source of the gain.

3.5. Concordance Objective

Minimising squared error tends to shrink the variance of the predictions toward the mean of the training scores, which compresses score differences at the extremes and can lower agreement after rounding. We counter this with a concordance term based on Lin's concordance correlation coefficient (CCC) [25]. For a batch, let $\hat{e}^{(t)}$ and $y^{(t)}$ denote the predicted and gold scores for dimension t across the batch, with means $\bar{p}, \bar{y}$, variances s_p^2, s_y^2, and covariance s_{py}. The coefficient is

$$\mathrm{CCC}^{(t)} = \frac{2\, s_{py}}{s_p^2 + s_y^2 + (\bar{p} - \bar{y})^2}, \tag{9}$$

and the corresponding loss averages its complement over the dimensions,

$$\mathcal{L}_{\text{ccc}} = \frac{1}{T}\sum_{t=1}^{T}\left(1 - \mathrm{CCC}^{(t)}\right). \tag{10}$$

The concordance term rewards predictions whose location, scale, and correlation match the gold scores, so it encourages the model to reproduce the spread of human ratings and to rank essays correctly. We stress that CCC is not identical to the quadratic weighted kappa (QWK): the two are distinct statistics with different constructions, and optimising CCC does not optimise QWK by definition. We use CCC only as a smooth, differentiable training surrogate whose location-, scale-, and correlation-matching behaviour is empirically aligned with the (non-differentiable, band-discretised) QWK we report; the ablation in Section 5 confirms this alignment, as removing $\mathcal{L}_{\text{ccc}}$ is the single change that most reduces test QWK.

3.6. Trait-Relational Coherence

For analytic scoring, the holistic Overall score is not independent of the traits. We encode this relation directly. Let $\hat{e}^{(1)}$ be the decoded Overall value and $\hat{e}^{(2)}, \ldots, \hat{e}^{(T)}$ the decoded trait values. We learn a convex combination of the traits,

$$a = \mathrm{softmax}(w), \qquad w \in \mathbb{R}^{T-1}, \tag{11}$$

where w is a free parameter learned jointly with the model, and require the Overall prediction to agree with the combination,

$$\mathcal{L}_{\mathrm{coh}} = \Big(\hat{e}^{(1)} - \sum_{j=2}^{T} a_{j-1}\, \hat{e}^{(j)}\Big)^2. \tag{12}$$

This term is active only when $T > 1$, that is, on ELLIPSE. It gives the model a soft consistency constraint that mirrors how a human rater expects the holistic impression to follow from the analytic profile. This is a modelling assumption rather than a property guaranteed by the ELLIPSE rubric; empirically, it is well motivated, as the six analytic traits correlate strongly with the Overall score in the corpus, but we do not claim the rubric defines Overall as a linear combination of traits. As the ablation shows, this term does not by itself improve aggregate agreement, and we retain it for the consistency it enforces at negligible cost rather than for a predictive gain (Section 5). The learned coefficients a are reported in Section 5.5: they stay almost exactly uniform in every fold, so in practice, the term ties Overall to the equally weighted mean of the six traits, and we do not read the coefficients as measures of trait importance.

3.7. Training Objective

The full objective is a weighted sum,

$$\mathcal{L} = \mathcal{L}_{\mathrm{mse}} + \lambda_{\mathrm{ord}}\, \mathcal{L}_{\mathrm{ord}} + \lambda_{\mathrm{ccc}}\, \mathcal{L}_{\mathrm{ccc}} + \lambda_{\mathrm{coh}}\, \mathcal{L}_{\mathrm{coh}}, \tag{13}$$

with $\lambda_{\mathrm{ord}} = 0.5$, $\lambda_{\mathrm{ccc}} = 0.5$, and $\lambda_{\mathrm{coh}} = 0.3$ in all experiments; the selection of these weights on validation data only is described in Section 4. Setting $\lambda_{\mathrm{ord}} = \lambda_{\mathrm{ccc}} = \lambda_{\mathrm{coh}} = 0$ recovers a plain regression model, and switching individual terms on and off isolates their effect; we study these ablations in Section 5. At inference, we map the read-out score back to the rubric scale. For ELLIPSE, whose bands run from 1.0 to 5.0 in steps of 0.5, the reported score is

$$s^{(t)} = 1.0 + 0.5\, \hat{e}^{(t)}, \tag{14}$$

and for Write & Improve, the predicted CEFR level is the nearest band index,

$$\hat{y} = \operatorname*{arg\,min}_{k \in \{0,\ldots,K-1\}} |\hat{e} - k|. \tag{15}$$

Algorithm 1 summarises one training step. Relative to a plain regression model, the added components introduce only an auxiliary ordinal head and a set of $T - 1$ coherence weights, so the parameter and compute overhead is small, and all of the extra parameters are trained together with the low-rank adapters.

Algorithm 1 PROSE training step

Input: batch of essays $\{x_b\}$ with gold band indices $\{y_b^{(t)}\}$; encoder f_θ with adapters; heads W_r, W_o; weights $\lambda_{\text{ord}}, \lambda_{\text{ccc}}, \lambda_{\text{coh}}$; temperature τ
$h_b \leftarrow \text{MeanPool}(f_\theta(x_b))$ // pooled representation
$\hat{e}_b^{(t)} \leftarrow (K-1)\,\sigma(W_r h_b)$; $z_b^{(t)} \leftarrow W_o h_b$ // regression read-out and ordinal logits
$\mathcal{L}_{\text{mse}} \leftarrow \text{mean}_{b,t}\,(\hat{e}_b^{(t)} - y_b^{(t)})^2$
build soft targets $q_b^{(t)}$ from $y_b^{(t)}$; $\mathcal{L}_{\text{ord}} \leftarrow \text{CE}(q, \text{softmax}(z))$
$\mathcal{L}_{\text{ccc}} \leftarrow \text{mean}_t\,(1 - \text{CCC}(\hat{e}^{(t)}, y^{(t)}))$ // over the batch
if $T > 1$ **then**
  $a \leftarrow \text{softmax}(w)$; $\mathcal{L}_{\text{coh}} \leftarrow \text{mean}_b\,(\hat{e}_b^{(1)} - \sum_{j=2}^{T} a_{j-1}\hat{e}_b^{(j)})^2$
else
  $\mathcal{L}_{\text{coh}} \leftarrow 0$
end if
$\mathcal{L} \leftarrow \mathcal{L}_{\text{mse}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{ccc}}\mathcal{L}_{\text{ccc}} + \lambda_{\text{coh}}\mathcal{L}_{\text{coh}}$
update adapters, heads, and w by AdamW on $\nabla\mathcal{L}$

4. Experimental Setup

4.1. Datasets

ELLIPSE. The English Language Learner Insight, Proficiency and Skills Evaluation corpus contains argumentative essays written by L2 English learners in grades 8 to 12, each scored by trained raters on a holistic Overall dimension and six analytic traits (cohesion, syntax, vocabulary, phraseology, grammar, and conventions) on a 1.0 to 5.0 scale in steps of 0.5 [20]. Our experiments use a study dataset of 3911 essays from ELLIPSE, evaluated with 5-fold cross-validation: each fold serves once as the test set, and within the remaining data, we hold out 10% as a validation set for model selection. We report the mean and standard deviation over the five folds. The cross-validation and internal-validation partitions are generated once with a fixed seed and are shared by every model, so all methods see identical splits.

Write & Improve (W&I). The Write & Improve learner data from the BEA-2019 shared task provides short learner texts, each annotated with a CEFR proficiency band [21]. We map the fine-grained labels to the six standard levels A1, A2, B1, B2, C1, and C2 and treat proficiency as a six-way ordinal target. We use the official training split of 3000 texts and evaluate on the 300-text development split, and we report the mean and standard deviation over three random seeds. This second dataset differs from ELLIPSE in genre, label granularity, and rating scheme, which tests whether the ordinal components of PROSE transfer beyond analytic trait scoring.

Sequence length and truncation. All models truncate the tokenised input to the first 512 subword tokens (Section 3). Because this can discard the end of a long essay, we measured the token-length distribution with the DeBERTa-v3 tokeniser (Table 1). On ELLIPSE, the median essay is 456 tokens, and 38.5% of essays exceed 512 tokens, so truncation is not negligible; on W&I, the texts are much shorter, and only 2.8% exceed the limit. We retain the 512-token window for comparability with standard encoder setups and because it is applied identically to every model, so it does not favour PROSE over the baselines; Section 5.5 compares performance on truncated and non-truncated essays, and we return to this as a limitation in Section 6.

Table 1. Token-length distribution (DeBERTa-v3 tokeniser, including special tokens) and the fraction of texts exceeding the 512-token window.

Corpus	Median	Mean	95th Percentile	Max	% >512
ELLIPSE	456	487	879	1428	38.5%
Write & Improve	204	218	422	2176	2.8%

4.2. Metrics

We report quadratic weighted kappa (QWK), the standard agreement metric for ordered scores [26], as the primary metric. For K bands it is defined as

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \qquad w_{ij} = \frac{(i-j)^2}{(K-1)^2}, \tag{16}$$

where O_{ij} counts essays with human band i and predicted band j, and E_{ij} is the agreement expected by chance from the marginal distributions. Continuous predictions are rounded to the nearest rubric band before computing QWK. We also report Pearson correlation, root mean squared error (RMSE), and mean absolute error (MAE) between predicted and gold scores. For ELLIPSE, all metrics are averaged over the seven dimensions, and we additionally report per-trait QWK. For W&I, we report QWK, accuracy, and MAE over the six ordinal levels.

Statistical testing. Because every model is trained and evaluated on identical folds, the design is paired. For the primary metric, we therefore compare PROSE against the regression baseline with a paired two-sided t-test over the five folds (and, as a distribution-free check, the Wilcoxon signed-rank test), and we report the mean fold-wise difference with a 95% confidence interval and the paired effect size (Cohen's d_z). We apply the same paired comparison to the ablation variants and to the per-trait results. With five folds, the smallest attainable two-sided Wilcoxon p-value is 0.0625, so we read the signed-rank test only as a consistency check on the sign of the difference and rely on the t-test and effect size for inference. Two further cautions apply. First, the five fold-wise estimates are not independent, because any two folds share three-quarters of their training data, and the plain paired t-test then understates the variance of the difference. We therefore also report the corrected resampled t-test of Nadeau and Bengio [38], which estimates the variance of the mean difference as $(1/k + n_{\text{test}}/n_{\text{train}})s_d^2$ rather than s_d^2/k, where s_d^2 is the sample variance of the fold-wise differences (here $k = 5$ and $n_{\text{test}}/n_{\text{train}} \approx 783/2816$). Second, the per-trait comparisons form a family of seven tests, so we adjust their p-values with the Holm step-down procedure [39]. We treat the mean-QWK comparison as the single pre-specified confirmatory test; the per-trait, ablation, sensitivity, and subgroup comparisons are exploratory and are reported to describe where the effect lies, not to establish further claims. With only five folds, all of these tests have limited power, and we interpret non-significant results as inconclusive rather than as evidence of no effect.

4.3. Baselines

We compare PROSE against three fine-tuned baselines that share the same encoder, adapters, pooling, and training schedule and differ only in the head and loss: a regression head trained with mean squared error (**Reg**), a classification head trained with cross-entropy over the bands (**Cls**), and a pure soft-ordinal head trained with $\mathcal{L}_{\text{ord}}$ alone (**SOrd (CE)**, the cross-entropy-against-soft-target objective of Section 3). We further compare against zero-shot prompting of three strong instruction-tuned LLMs, GPT-5.4, Claude-Sonnet-4-6, and DeepSeek-Chat, which receive the rubric and the essay and are asked to return the seven scores as JSON, with no task-specific training and temperature set to zero. These

baselines isolate the effect of the training objective under a fixed architecture; we discuss the relationship to externally published AES systems, which use different splits and targets, in Section 6.

4.4. Implementation

All fine-tuned models use DeBERTa-v3-base as the encoder unless stated otherwise. Table 2 summarises the model configuration and the split between frozen and trainable parameters, and Table 3 lists the training hyperparameters, which are shared by PROSE and all fine-tuned baselines. Within each fold, we select the checkpoint with the best QWK on that fold's held-out validation subset only; the test fold is never used for model selection. The selected epoch varies across folds (ranging from epoch 4 to 7 on ELLIPSE), confirming that selection is driven by validation performance rather than by a fixed schedule. No preprocessing, tokenisation, or label transformation is fitted on the data: we use the pretrained tokeniser as released and map labels to band indices by a fixed rule, so there is no statistic estimated across the pool before cross-validation. Experiments run on a single NVIDIA A800 GPU, and each cross-validation fold trains in about seven minutes. Quantitative compute figures are reported in Section 6.

Table 2. Model configuration. The output heads are sized for ELLIPSE ($T = 7$ dimensions, $K = 9$ bands); for Write & Improve, there is a single rubric dimension ($T = 1$, $K = 6$), and the coherence weights are dropped.

Component	Configuration
Backbone	DeBERTa-v3-base
Transformer layers	12
Hidden size d	768
Attention heads	12
Feed-forward size	3072
Backbone parameters (frozen)	≈184 M
Low-rank adapters (trainable)	≈2.7 M
Regression head (trainable)	$T \times (d + 1)$ parameters
Soft-ordinal head (trainable)	$T \times K \times (d + 1)$ parameters
Coherence weights (trainable)	$T - 1$ parameters
Trainable share of all parameters	≈1.5%

Table 3. Training hyperparameters, shared by PROSE and all fine-tuned baselines.

Hyperparameter	Value
Maximum sequence length	512 tokens
LoRA rank r	16
LoRA scaling α	32
Adapted projections	query, key, value, dense
Dropout	0.1
Optimiser	AdamW
Peak learning rate	2×10^{-5}
Learning-rate schedule	one-cycle
Epochs	8
Batch size	16
Weight decay	0.01
Precision	bfloat16
Soft-ordinal temperature τ	1.0
Loss weights $\lambda_{\text{ord}}, \lambda_{\text{ccc}}, \lambda_{\text{coh}}$	0.5, 0.5, 0.3
Model selection	best validation QWK

Hyperparameter selection. The loss weights $(\lambda_{\text{ord}}, \lambda_{\text{ccc}}, \lambda_{\text{coh}})$ were chosen by a small grid over $\{0.3, 0.5, 1.0\}$ on the fold-0 validation subset only, before any test fold was examined, and were then frozen at $(0.5, 0.5, 0.3)$ for all folds and for both datasets; no per-fold or per-dataset retuning was performed. We should be precise about what this guarantees. The fold-0 validation subset is a random 10% of the fold-0 training portion, so it is disjoint from the fold-0 test set but its essays do appear in the test sets of folds 1–4. Weight selection therefore never looked at a test fold, but for folds 1–4 it is not strictly nested. We consider the resulting optimism small for three reasons: the grid was coarse (three values per weight), the selection used validation QWK and never test scores, and PROSE also beats the regression baseline on fold 0, whose test essays had no contact with the selection (0.651 against 0.644). The sensitivity analysis in Section 5.5 adds a fourth: the result changes little across the grid, so the reported performance does not depend on a lucky choice of weights. The remaining hyperparameters (LoRA rank, scaling, learning rate, dropout, epochs, batch size, and the soft-ordinal temperature τ) were set to standard values and held fixed; we report a sensitivity analysis for τ and for encoder scale in Section 5, and a sensitivity analysis of the three loss weights in Section 5.5.

To keep the comparison fair, every fine-tuned model shares the same data splits, encoder, adapters, optimiser, schedule, and number of epochs, and differs only in the head and loss. The cross-validation folds and the internal validation split are drawn with fixed seeds so that all models see the same partitions, and the reported figures are means over folds or seeds with standard deviations. For the concordance term, which is estimated over each mini-batch, we use the same batch size for all models so that the estimate is computed under identical conditions. The zero-shot large language models are evaluated on the first 400 essays of the fold-0 test set with a single rubric prompt and temperature zero. The exact protocol is as follows, so that the experiment can be reproduced. Each model is queried through an OpenAI-compatible chat interface with the system message carrying a brief rubric summary and the instruction to return a compact JSON object with the seven scores, and the user message carrying the essay text truncated to 6000 characters; decoding uses temperature 0 and a 200-token cap. The returned text is parsed by extracting the first brace-delimited JSON object and reading the seven fields; a malformed reply triggers up to three attempts in total, and any remaining failure falls back to the prior mean (3.0), though in practice, all 400 replies parsed successfully for every model. Parsed scores are clamped to $[1.0, 5.0]$ and rounded to the nearest half band, matching the rubric granularity. The model identifiers are `gpt-5.4`, `claude-sonnet-4-6`, and `deepseek-chat`, queried in a single evaluation run; the full prompt text and parser are released with the code (Supplementary Material S1 and Data Availability). Both datasets can be obtained from the official sources listed in the Data Availability Statement. Reproducing the saved row-index splits requires retaining the exact processed input ordering.

5. Results

5.1. Main Results

Table 4 reports the analytic scoring results on ELLIPSE, averaged over the seven rubric dimensions and five cross-validation folds. PROSE achieves a higher quadratic weighted kappa than the three fine-tuned baselines, ties the regression baseline on Pearson correlation, and yields numerically lower RMSE and MAE. The QWK improvement over the regression baseline, although modest in absolute terms (from 0.646 to 0.661), is consistent and statistically reliable: the mean fold-wise difference is +0.015 with a 95% confidence interval of $[0.007, 0.023]$, a paired t-test gives $t(4) = 5.1$, $p = 0.007$, and the paired effect size is large ($d_z = 2.3$); all five folds improve, so the Wilcoxon signed-rank test attains its $n = 5$ floor of $p = 0.0625$. RMSE and Pearson correlation show no statistically significant change

(RMSE −0.002, $p = 0.28$; Pearson unchanged, $p = 0.72$), so we claim an improvement in agreement with human ratings specifically, not in every metric. Table 5 collects these paired comparisons. Under the corrected resampled test, which accounts for the overlap between training folds, the QWK gain remains significant ($p = 0.030$). In practical terms, the gain is modest: pooled over all out-of-fold predictions, PROSE places 40.6% of scores in the exact human band against 40.1% for the regression baseline, and the share within 0.5 score points is the same (88.9%). The improvement is therefore best read as a consistent shift in agreement with human ratings rather than as a change that a single teacher would notice essay by essay. Section 5.5 checks that the gain holds in every fold and on subsets of the test data. The classification baseline is clearly weaker, which confirms that discarding the order of the bands wastes signal. The pure soft-ordinal model improves the error metrics over classification but does not by itself beat regression on kappa, which motivates the multi-objective design of PROSE. Figure 3a summarises the comparison and adds the root mean squared error on a second axis, which falls as the agreement rises.

Table 4. Analytic scoring on ELLIPSE (mean ± descriptive standard deviation over five folds, seven dimensions; descriptive standard deviations use divisor n). The upper block reports 5-fold cross-validation; the lower block reports zero-shot LLMs on the first 400 fold-0 test essays, with PROSE and Reg re-evaluated on that identical 400-essay subset for a like-for-like comparison (Section 5.2). The best fine-tuned values are shown in bold; bold in the model column identifies the proposed method. ↑: higher is better (QWK and Pearson); ↓: lower is better (RMSE and MAE). Italic rows separate the two evaluation settings.

Model	QWK ↑	Pearson ↑	RMSE ↓	MAE ↓
5-fold cross-validation (full test folds)				
Cls (CE)	0.536 ± 0.008	0.684 ± 0.005	0.481 ± 0.007	0.385 ± 0.004
SOrd (CE)	0.597 ± 0.010	0.693 ± 0.005	0.475 ± 0.005	0.382 ± 0.003
Reg (MSE)	0.646 ± 0.006	**0.709 ± 0.004**	0.471 ± 0.004	0.377 ± 0.002
PROSE (full)	**0.661 ± 0.007**	**0.709 ± 0.004**	**0.469 ± 0.004**	**0.375 ± 0.002**
Identical 400-essay fold-0 subset				
Reg (MSE), 400	0.616	0.688	0.472	0.379
PROSE (full), 400	0.627	0.687	0.475	0.382
GPT-5.4 (zero-shot)	0.206	0.504	0.996	0.851
DeepSeek-Chat (zero-shot)	0.184	0.512	1.036	0.894
Claude-Sonnet-4-6 (zero-shot)	0.128	0.548	1.358	1.250

Table 5. Paired comparison of PROSE and the regression baseline on ELLIPSE across the five folds. Δ is the mean fold-wise difference (PROSE − Reg) with its ordinary paired-t 95% confidence interval; p is from a paired two-sided t-test; p_{adj} is the Nadeau–Bengio corrected resampled t-test for the three mean-metric rows and the Holm-adjusted p-value for the seven per-trait rows; d_z is the paired effect size. Per-trait rows are for QWK and are exploratory.

Comparison	Reg	PROSE	Δ (95% CI)	p	p_{adj}	d_z
QWK (mean)	0.646	0.661	+0.015 [0.007, 0.023]	0.007	0.030	2.28
RMSE (mean)	0.471	0.469	−0.002 [−0.006, 0.002]	0.28	0.46	−0.56
Pearson (mean)	0.709	0.709	−0.000 [−0.003, 0.002]	0.72	0.81	−0.18
QWK Overall	0.713	0.719	+0.006 [−0.001, 0.013]	0.063	0.19	1.14
QWK Cohesion	0.577	0.599	+0.022 [0.010, 0.034]	0.007	0.037	2.24
QWK Syntax	0.639	0.655	+0.016 [0.014, 0.019]	<0.001	0.001	7.30
QWK Vocabulary	0.622	0.642	+0.020 [−0.007, 0.046]	0.11	0.19	0.93
QWK Phraseology	0.657	0.675	+0.018 [0.007, 0.029]	0.012	0.047	1.97
QWK Grammar	0.663	0.676	+0.013 [−0.002, 0.027]	0.074	0.19	1.08
QWK Conventions	0.650	0.661	+0.011 [0.006, 0.017]	0.004	0.022	2.73

Table 6 reports proficiency classification on the Write & Improve CEFR data, averaged over three seeds. This task is easier to rank than analytic trait scoring, because the CEFR

bands are well separated, and the quadratic weighted kappa is high (≈0.85–0.87) for all of the fine-tuned models, which leaves little room to separate the strongest methods. In this regime, PROSE is statistically indistinguishable from the regression baseline on agreement (0.865 against 0.866; paired t-test over the three seeds $p = 0.85$) and on accuracy (both 0.537), and it clearly outperforms classification and pure ordinal training, whose kappa values fall to 0.858 and 0.848. We therefore do not claim an improvement over the strongest baseline on this dataset: the observed difference is small and not statistically significant. We note that PROSE's QWK standard deviation across seeds is larger than the regression baseline's (0.007 against 0.001; see the per-seed values in Table 6), which we attribute to the additional loss terms interacting with the small 300-text development set; the conclusion we draw is the weaker one that the ordinal components can be applied to a different label scheme, with no statistically significant difference from regression detected in these runs. Figure 3b shows the same comparison.

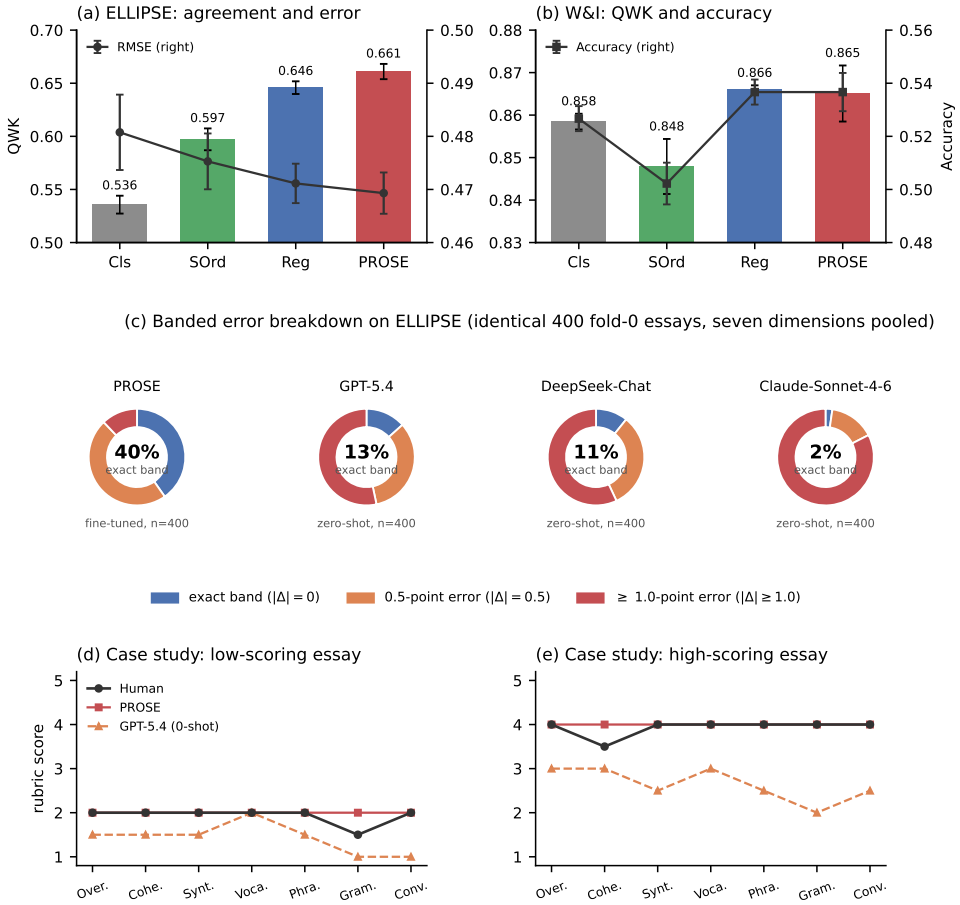


Figure 3. Results overview. (**a**) ELLIPSE: quadratic weighted kappa of the fine-tuned models (bars, standard-deviation error bars) with root mean squared error on the right axis. (**b**) Write & Improve: quadratic weighted kappa with accuracy on the right axis. (**c**) Banded error on ELLIPSE, pooled over the seven dimensions: share of scores with absolute errors of 0, 0.5, and at least 1.0 score points, respectively, for PROSE and the three zero-shot large language models; the fine-tuned and zero-shot results in this panel are computed on the same 400 fold-0 essays ($n = 400$ for every system). (**d**,**e**) Illustrative analytic profiles for a low-scoring and a high-scoring essay; PROSE follows the human profile, while the zero-shot model systematically assigns lower scores. Panels (**d**,**e**) are hand-picked examples (Section 5.6); the distributional evidence for all 400 evaluation essays is in Section 5.6.

Table 6. Proficiency classification on Write & Improve (CEFR, six ordinal levels). Per-seed QWK is shown for the three random seeds, followed by the mean ± descriptive standard deviation of each metric (divisor $n = 3$). Best mean result in bold. ↑: higher is better; ↓: lower is better.

	QWK per Seed					
Model	**s1**	**s2**	**s42**	**QWK ↑**	**Accuracy ↑**	**MAE ↓**
Cls (CE)	0.858	0.861	0.857	0.858 ± 0.002	0.527 ± 0.005	0.524 ± 0.004
SOrd (CE)	0.842	0.857	0.845	0.848 ± 0.006	0.502 ± 0.008	0.548 ± 0.010
Reg (MSE)	0.867	0.865	0.867	**0.866 ± 0.001**	0.537 ± 0.005	**0.517 ± 0.007**
PROSE (full)	0.865	0.873	0.857	0.865 ± 0.007	**0.537 ± 0.007**	**0.517 ± 0.008**

5.2. Comparison with Zero-Shot Large Language Models

The lower block of Table 4 reports three strong instruction-tuned models prompted zero-shot, together with PROSE and the regression baseline re-evaluated on the identical 400 fold-0 essays, so that all five systems are compared on exactly the same texts. On this common subset, PROSE reaches QWK 0.627 and the regression baseline 0.616; both are lower than their cross-validation means on this subset, but the ordering of the two fine-tuned models is the same as in cross-validation (Section 5.5), whereas the three zero-shot models reach only 0.13 to 0.21. All three zero-shot models trail the task-specific fine-tuned models by a wide margin on every metric, and their RMSE and MAE are more than twice as large. This is consistent with reports that large language models are fluent but poorly calibrated raters [16,19]. Section 6 and the distributional analysis in Section 5.6 explain the gap: the zero-shot models produce less variable scores and lower mean scores than the human raters. The result is a practical one, and we state it narrowly: under the conditions tested, a few-hundred-million-parameter encoder adapted with a small number of trainable weights on labelled data is a stronger analytic rater than the general-purpose models used zero-shot. Because the fine-tuned models are supervised on ELLIPSE while the LLMs are not, this comparison reflects task-specific supervision and objective design together and does not isolate model scale; the parameter counts of the commercial models are not publicly disclosed, so we avoid quantifying the scale difference.

5.3. Ablation Study

Table 7 isolates the contribution of each component of PROSE on ELLIPSE, and Figure 4 plots the same numbers. The concordance term is the main driver of the gain in quadratic weighted kappa: removing it lowers the kappa from 0.661 to 0.652, a paired drop of 0.009 that is statistically significant ($p < 0.001$, $d_z = 4.0$; $p = 0.004$ under the corrected resampled test) and accounts for most of the difference between PROSE and the plain regression baseline at 0.646. This matches the mechanism seen in the distributional analysis, where the concordance term keeps the spread of predictions close to that of the human ratings rather than letting squared error pull them toward the centre. The soft-ordinal auxiliary contributes a smaller gain, from 0.659 without it to 0.661, that is positive on four of five folds but does not reach significance under the paired test ($\Delta = +0.002$, $p = 0.07$; $p = 0.19$ under the corrected test); we therefore describe it as a small and consistent-in-sign gain rather than a statistically established one. In practical terms, a 0.002 change in QWK is negligible, and we do not claim that the soft-ordinal term is an important part of the method: it costs one extra linear head during training and nothing at inference, which is why we keep it, but the evidence for its benefit is weak. The term is better described as an inexpensive regulariser with no statistically significant effect detected in this experiment than as a source of the improvement, and the loss-weight analysis in Section 5.5 is consistent with this reading. The trait-relational coherence term has a neutral effect on aggregate agreement (removing it

changes QWK by +0.001, $p = 0.18$, i.e., not distinguishable from zero); we retain it because it enforces consistency between the analytic traits and the holistic score at no measurable cost, and we are explicit that it is included for this consistency rather than for a predictive gain; as Section 5.5 shows, its learned coefficients remain uniform, so the consistency it enforces is with the plain average of the traits. In short, the concordance term carries the statistically supported improvement, the soft-ordinal term adds a small consistent-in-sign effect, and the coherence term is retained for output consistency. The full combination stays within one standard deviation of the best tested variant on both agreement and error; these removal ablations do not establish whether any added term is sufficient on its own.

Table 7. Component ablation on ELLIPSE (mean ± descriptive standard deviation over five folds, using divisor n). Coh: trait-relational coherence; Ord: soft-ordinal auxiliary; CCC: concordance term. The last column reports the paired t-test p-value for the change in QWK relative to the full model. A minus sign denotes removal of a component. Bold identifies the proposed full model and its QWK. ↑: higher is better; ↓: lower is better.

Variant	QWK ↑	RMSE ↓	Pearson ↑	p vs. Full
PROSE (full)	**0.661 ± 0.007**	0.469 ± 0.004	0.709 ± 0.004	—
− CCC	0.652 ± 0.006	0.468 ± 0.004	0.708 ± 0.004	0.0009
− Ord	0.659 ± 0.006	0.468 ± 0.004	0.710 ± 0.003	0.07
− Coh	0.662 ± 0.007	0.468 ± 0.004	0.709 ± 0.004	0.18
Reg (all terms off)	0.646 ± 0.006	0.471 ± 0.004	0.709 ± 0.004	0.007

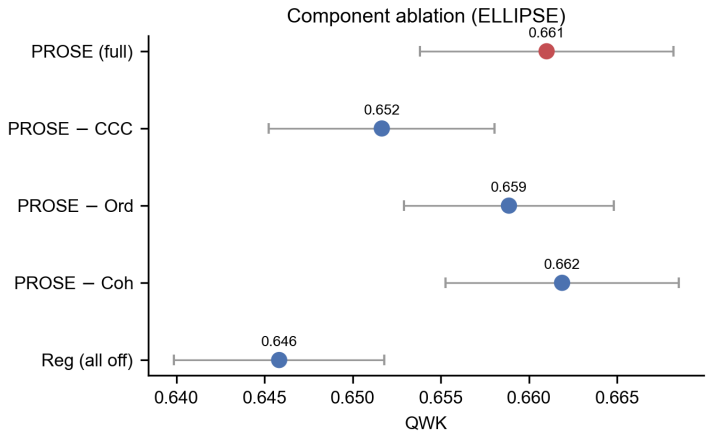


Figure 4. Component ablation on ELLIPSE. Removing the concordance term costs the most agreement (and is the only component whose removal significantly lowers QWK), removing the soft-ordinal term costs a little, and removing the coherence term changes the result within noise. The red point marks the full model; blue points mark the ablation variants and regression baseline; grey error bars show descriptive standard deviations across folds.

5.4. Per-Trait Agreement and Sensitivity

Figure 5 reports quadratic weighted kappa for each of the seven ELLIPSE dimensions. PROSE improves or matches the regression baseline on every trait, so the gains are consistent rather than concentrated on one dimension; the per-trait paired tests in Table 5 show that the improvement is individually significant on four traits (Cohesion, Syntax, Phraseology, Conventions), which remain significant after Holm correction (adjusted p from 0.001 to 0.047) and positive but not significant on the other three. We regard these per-trait results as exploratory. Table 8 examines two design choices. Using the same three folds (0–2) for every setting, varying the soft-ordinal temperature τ over $\{0.5, 1.0, 2.0\}$ changes QWK only slightly (0.658–0.659), while mean RMSE rounds to 0.467 for all three values. The default does not achieve the highest QWK on these folds, and these results do not establish a

performance advantage for a particular temperature. On the same folds, replacing the base encoder with DeBERTa-v3-large changes PROSE QWK from 0.658 to 0.660; the difference is small relative to the observed fold-to-fold variation. At the larger scale, PROSE and the regression baseline also have similar QWK, suggesting that comparable performance is possible with either tested encoder size.

Table 8. Sensitivity of PROSE on ELLIPSE (mean ± descriptive standard deviation, divisor $n = 3$, over the same folds 0–2): soft-ordinal temperature τ (top) and encoder scale (bottom). ↑: higher is better; ↓: lower is better.

Setting	QWK ↑	RMSE ↓
$\tau = 0.5$	0.659 ± 0.003	0.467 ± 0.002
$\tau = 1.0$ (default)	0.658 ± 0.005	0.467 ± 0.001
$\tau = 2.0$	0.659 ± 0.006	0.467 ± 0.001
Reg, DeBERTa-v3-base	0.647 ± 0.004	0.469 ± 0.003
PROSE, DeBERTa-v3-base	0.658 ± 0.005	0.467 ± 0.001
Reg, DeBERTa-v3-large	0.657 ± 0.005	0.464 ± 0.008
PROSE, DeBERTa-v3-large	0.660 ± 0.006	0.466 ± 0.004

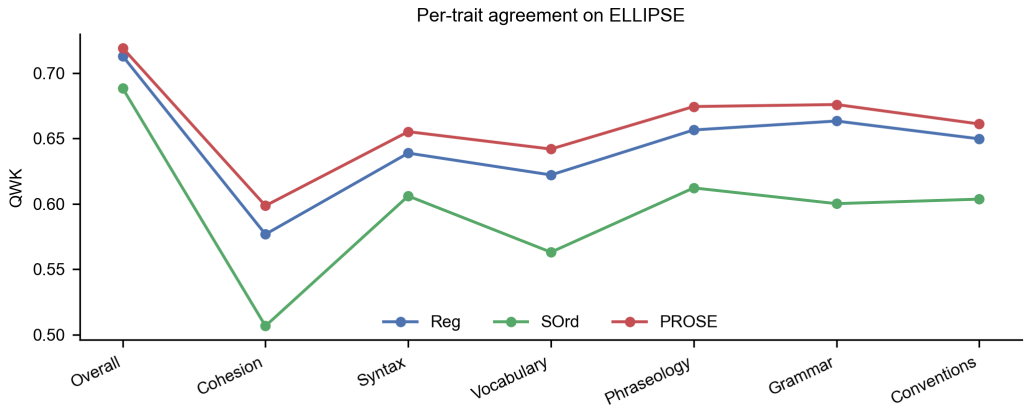


Figure 5. Per-trait quadratic weighted kappa on ELLIPSE. PROSE improves or matches the regression baseline on all seven dimensions; see Table 5 for the per-trait fold-wise differences, confidence intervals, and significance.

5.5. Robustness and Diagnostic Analyses

The analyses in this subsection use per-essay predictions from a re-run of PROSE and the regression baseline with the same code, seeds, and folds as the main experiments, extended only to save the test predictions, the token counts, and resource measurements. The re-run reproduces the original fold-wise QWK to within ±0.0006 (Table 9), so the conclusions below apply to the models reported in Table 4. Pooling the five test folds gives one out-of-fold prediction for each of the 3911 ELLIPSE essays.

Consistency across folds and evaluation subsets. Table 9 lists QWK by fold. PROSE is ahead of the regression baseline in all five folds, by between +0.007 (fold 0) and +0.022 (fold 3). On the 400-essay subset used for the LLM comparison, PROSE is also ahead (0.627 against 0.616; bootstrap 95% interval of the difference $[0.001, 0.022]$), as it is on the remaining 383 essays of fold 0 (0.673 against 0.669). To see how much a 400-essay sample can move the comparison, we drew 200 random 400-essay subsets from each test fold (1000 in total): PROSE has the higher QWK in 99.6% of them, with a mean difference of +0.015 and a 90% range of $[0.005, 0.025]$. A subset of this size can therefore shift the difference by about ±0.01, but the sign is stable.

Table 9. Fold-wise QWK on ELLIPSE for the regression baseline and PROSE (original runs), with the QWK of the re-runs used for the diagnostic analyses.

	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean
Reg (original)	0.644	0.653	0.644	0.637	0.651	0.646
PROSE (original)	0.651	0.664	0.658	0.659	0.673	0.661
Δ (PROSE − Reg)	+0.007	+0.010	+0.014	+0.022	+0.022	+0.015
Reg (re-run)	0.644	0.654	0.644	0.636	0.652	0.646
PROSE (re-run)	0.651	0.664	0.658	0.658	0.673	0.661

Sensitivity to the loss weights. Table 10 varies each loss weight while the other two stay at their default values and also crosses λ_{ord} and λ_{ccc} on a 3 × 3 grid over $\{0.25, 0.5, 1.0\}$ (Figure 6a). Every configuration is trained on folds 0–2, as for the temperature analysis in Table 8, and is compared with the default and the regression baseline on the same three folds. Every one of the ten settings beats the regression baseline on all three folds, and mean QWK stays between 0.653 and 0.664, against 0.658 for the default and 0.647 for the baseline, so the gain does not depend on a narrow choice of weights. The weights matter to different degrees. The coherence weight has no visible effect (0.658 at both 0.1 and 1.0). The soft-ordinal weight has little effect (0.655–0.658 from 0.25 to 1.0), in line with the ablation. The concordance weight is the one that matters: lowering it to 0.25 reduces QWK to 0.653–0.655, and raising it to 1.0 increases QWK to 0.660–0.664 at a slightly higher RMSE. The 3 × 3 grid shows no strong interaction between the ordinal and concordance terms: in each row, QWK rises with λ_{ccc}, and changing the ordinal weight moves QWK by at most 0.004 at a fixed concordance weight. The default is therefore not the best setting on these folds. A larger concordance weight would trade a little calibration for agreement, which matches the trade-off described for τ, but these exploratory comparisons use only three folds, and we do not re-tune the reported model.

Table 10. Sensitivity of PROSE to the loss weights on ELLIPSE (mean ± standard deviation over folds 0–2; the default and Reg rows are computed on the same three folds). Rows vary one weight from the default $(\lambda_{\text{ord}}, \lambda_{\text{ccc}}, \lambda_{\text{coh}}) = (0.5, 0.5, 0.3)$; "Folds > Reg" counts folds in which the configuration beats the regression baseline, and p is the paired t-test against the default. ↑: higher is better; ↓: lower is better.

Configuration	QWK ↑	RMSE ↓	Folds > Reg	p vs. Default
Default $(0.5, 0.5, 0.3)$	0.658 ± 0.006	0.467 ± 0.001	3/3	—
$\lambda_{\text{ord}} = 0.25$	0.658 ± 0.006	0.466 ± 0.002	3/3	0.12
$\lambda_{\text{ord}} = 1.0$	0.655 ± 0.006	0.468 ± 0.005	3/3	0.51
$\lambda_{\text{ccc}} = 0.25$	0.655 ± 0.005	0.466 ± 0.002	3/3	0.21
$\lambda_{\text{ccc}} = 1.0$	0.662 ± 0.007	0.469 ± 0.001	3/3	0.02
$\lambda_{\text{coh}} = 0.1$	0.658 ± 0.006	0.467 ± 0.001	3/3	0.59
$\lambda_{\text{coh}} = 1.0$	0.658 ± 0.005	0.467 ± 0.002	3/3	0.67
$\lambda_{\text{ord}} = \lambda_{\text{ccc}} = 0.25$	0.655 ± 0.007	0.465 ± 0.002	3/3	0.06
$\lambda_{\text{ord}} = 0.25, \lambda_{\text{ccc}} = 1.0$	0.664 ± 0.008	0.467 ± 0.002	3/3	0.03
$\lambda_{\text{ord}} = 1.0, \lambda_{\text{ccc}} = 0.25$	0.653 ± 0.006	0.470 ± 0.003	3/3	0.27
$\lambda_{\text{ord}} = \lambda_{\text{ccc}} = 1.0$	0.660 ± 0.004	0.470 ± 0.006	3/3	0.34
Reg (all weights 0)	0.647 ± 0.006	0.469 ± 0.003	—	0.03

Truncated and non-truncated essays. Of the 3911 essays, 1507 (38.5%) exceed 512 tokens and are truncated. Both models score them less accurately (Table 11): QWK falls from 0.656 to 0.634 for PROSE and from 0.640 to 0.620 for the regression baseline, and RMSE rises by about 0.03. The drop is similar for the two models, so the gain of PROSE over the

baseline is about the same for long essays (+0.015) as for short ones (+0.016; paired test across folds $p = 0.74$). Part of the drop is not caused by truncation itself: long essays have higher and less variable human scores (mean Overall 3.28 against 2.99), which lowers kappa for any rater. Among the truncated essays, the share of text that is cut off is not related to the absolute error (Spearman $\rho = 0.025$, $p = 0.32$). Performance is lower on essays subject to truncation for both models, but this observational comparison does not isolate the causal effect of removing the end of an essay; the relative advantage of PROSE is similar in the two length groups.

Learner subgroups. ELLIPSE records gender, race/ethnicity, socioeconomic status, and grade. Table 11 reports QWK and mean signed error (prediction minus human score, averaged over the seven dimensions) for every group with at least 100 essays; smaller groups are not reported because their estimates are too noisy. PROSE has a higher QWK than the regression baseline in every group, and in every group, both models over-score slightly, by 0.06–0.12 score points on average, with PROSE closer to zero. Raw differences in signed error between groups are not significant for gender, ethnicity, or socioeconomic status. Because groups differ in their score distributions, and any regressor over-scores weak writers and under-scores strong ones, we also tested the error after removing its linear dependence on the human Overall score. On this adjusted error, gender ($p = 0.21$) and socioeconomic status ($p = 0.95$) show no difference, ethnicity is borderline for PROSE ($p = 0.036$; regression baseline $p = 0.074$), with the largest adjusted error, +0.036 points, for the 286 White students, and grade shows a clear difference for both models ($p < 0.001$): after adjustment, grade 11 essays are over-scored by 0.038, and grade 12 essays under-scored by 0.044 points. The grade effect is observed for both models and is therefore not specific to the PROSE objective; its underlying cause is not isolated by this analysis, but it is the kind of systematic difference that a deployment would have to monitor.

Table 11. QWK (mean of seven dimensions) and mean signed error (prediction − human, in score points) by essay length and learner subgroup from the pooled out-of-fold predictions on all 3911 ELLIPSE essays. Groups with fewer than 100 essays are omitted. Italic rows identify grouping variables. ↑: higher QWK is better.

		QWK ↑		Signed Error	
Group	***n***	**Reg**	**PROSE**	**Reg**	**PROSE**
Essay length					
≤512 tokens	2404	0.640	0.656	+0.110	+0.087
>512 tokens (truncated)	1507	0.620	0.634	+0.092	+0.079
Gender					
Female	1721	0.662	0.676	+0.097	+0.080
Male	2190	0.631	0.646	+0.108	+0.087
Race/ethnicity					
Asian/Pacific Islander	482	0.654	0.662	+0.078	+0.067
Black/African American	321	0.642	0.659	+0.091	+0.078
Hispanic/Latino	2767	0.634	0.649	+0.107	+0.085
White	286	0.693	0.713	+0.119	+0.106
Socioeconomic status					
Economically disadvantaged	2709	0.626	0.642	+0.109	+0.088
Not economically disadvantaged	1202	0.684	0.695	+0.090	+0.073
Grade					
Grade 8	975	0.602	0.618	+0.108	+0.089
Grade 10	214	0.680	0.695	+0.091	+0.071
Grade 11	1377	0.656	0.672	+0.122	+0.107
Grade 12	1319	0.643	0.656	+0.082	+0.057

Feature-level behaviour. To see which properties of the text the predictions follow, we correlated the predicted and the human Overall scores with five indices distributed with ELLIPSE: number of words, sentences, and paragraphs, the measure of textual lexical diversity (MTLD), and the type–token ratio. PROSE follows the same cues as the human raters, but more strongly. Its predictions correlate with the number of sentences at $\rho = 0.47$ (human scores 0.38) and with MTLD at 0.45 (human 0.35), and the five indices together explain 35% of the variance of its predictions but only 23% of the variance of the human scores. The regression baseline behaves the same way. The prediction errors are only weakly related to these indices ($|\rho| \leq 0.054$), so the models are not simply rewarding length, but they lean on length and lexical diversity more than human raters do, and an essay that is long and lexically varied but weak in other respects may be over-scored. This analysis is global and correlational; it does not explain individual predictions. Checking a learned predictor against independent, interpretable evidence is common in other data-driven fields, for example, when candidate biomarkers mined from data are then validated experimentally [40], or when heterogeneous features are mapped into a common representation before learning [41]; an instance-level attribution analysis of PROSE along these lines is left for future work.

Learned coherence coefficients. Table 12 reports the coefficients a of the coherence term. In all five folds, they stay within 0.0006 of the uniform value $1/6 \approx 0.167$ at which they are initialised, and their fold-to-fold standard deviation is below 10^{-4}. They are stable, but this is because they hardly move: with the small learning rate used for all parameters, the term in practice ties Overall to the plain average of the six traits. This average is a reasonable target for ELLIPSE, where the six traits correlate almost equally with Overall ($r = 0.77$–0.81; normalised to sum to one, 0.163–0.171), but it differs from the least-squares weights of Overall on the traits (0.140–0.202), which give more weight to Cohesion, Syntax, and Vocabulary. The coefficients should therefore not be read as estimates of how much each trait contributes to the holistic score. What the term does achieve is consistency: the mean absolute gap between the predicted Overall score and the average of the predicted traits is 0.035 for PROSE against 0.062 for the regression baseline (human scores: 0.206). Learning the coefficients with a larger learning rate, or fixing them to empirically estimated weights, are simple alternatives that we did not test.

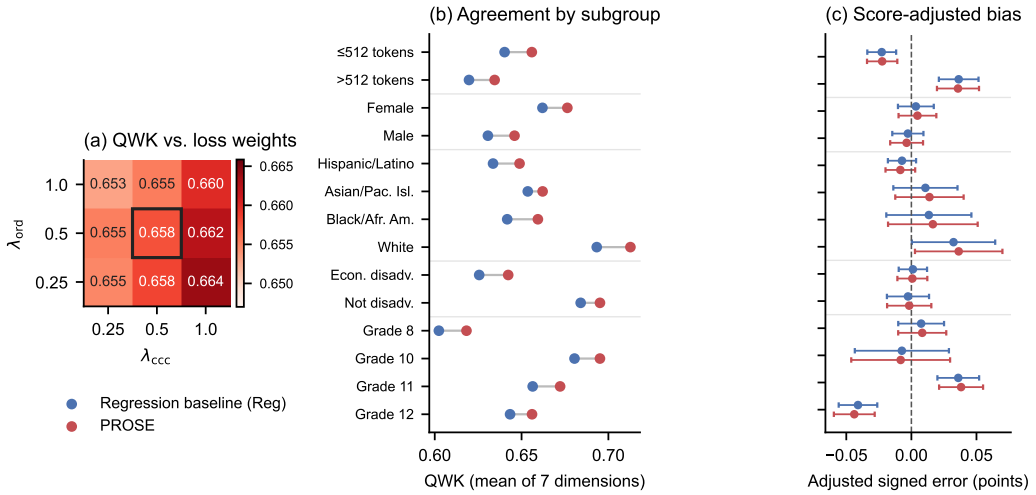


Figure 6. Robustness and diagnostic analyses on ELLIPSE. (**a**) Mean QWK over folds 0–2 on the $\lambda_{ord} \times \lambda_{ccc}$ grid ($\lambda_{coh} = 0.3$); the black rectangle marks the default setting, and the regression baseline reaches 0.647 on the same folds. (**b**) QWK of PROSE and the regression baseline by essay length and learner subgroup (pooled out-of-fold predictions). (**c**) Mean signed error after removing its linear dependence on the human Overall score, by subgroup; error bars are 95% confidence intervals.

Table 12. Learned coherence coefficients a (mean ± standard deviation over five folds) compared with the trait–Overall correlations in ELLIPSE, normalised to sum to one, and with the normalised least-squares weights of Overall on the six traits.

Trait	Learned a	Normalised Correlation	Normalised OLS Weight
Cohesion	0.1664 ± 0.00002	0.166	0.184
Syntax	0.1667 ± 0.00003	0.171	0.192
Vocabulary	0.1666 ± 0.00003	0.167	0.202
Phraseology	0.1672 ± 0.00004	0.169	0.140
Grammar	0.1664 ± 0.00002	0.164	0.141
Conventions	0.1668 ± 0.00005	0.163	0.141

5.6. Distributional Analysis and Illustrative Cases

We first assess score compression and downward bias quantitatively over the full evaluation sample, and only then show hand-picked essays as illustration. Table 13 reports, for the human raters, PROSE, and the three zero-shot models, the mean, standard deviation, and range of the predicted scores on the identical 400 fold-0 essays for every system, including the regression baseline, together with the ratio of the prediction standard deviation to the human standard deviation. The human Overall scores have a standard deviation of 0.61; PROSE reproduces 0.81 of this spread and keeps its mean close to the human mean (3.18 against 3.07); the regression baseline behaves similarly (0.80), whereas the three large models compress the spread to 0.54–0.66 of the human value *and* shift their mean well below the human centre, to between 1.84 and 2.42. The pattern is the same when pooled over all seven traits, where the spread ratio is 0.77 for PROSE, 0.73 for the regression baseline, and 0.58–0.61 for the zero-shot models. This is direct, selection-free evidence for the score compression and downward bias we describe: the zero-shot models both flatten the distribution and score too harshly, which is why their error is several times larger even though their outputs read as plausible. The banded-error breakdown in Figure 3c, computed on the identical 400 essays, tells the same story: PROSE places 40% of its scores in the exact human band and 88% within 0.5 score points, whereas the zero-shot models reach only 2–13% exact agreement and miss by 1.0 score point or more on roughly half to four-fifths of essays.

Table 13. Distribution of predicted scores on the identical 400 fold-0 ELLIPSE essays for all systems (Overall dimension). The standard-deviation ratio is the prediction standard deviation divided by the human standard deviation on the same data; a value near 1 indicates that the model reproduces the human spread. The zero-shot models compress the spread and shift the mean downward.

System	Mean	Std. Dev.	Min	Max	SD Ratio
Human raters	3.07	0.61	1.0	5.0	1.00
Reg (MSE)	3.18	0.49	1.7	4.3	0.80
PROSE	3.18	0.50	1.8	4.2	0.81
GPT-5.4 (zero-shot)	2.42	0.35	1.0	3.5	0.57
DeepSeek-Chat (zero-shot)	2.41	0.33	1.0	3.0	0.54
Claude-Sonnet-4-6 (zero-shot)	1.84	0.40	1.0	3.5	0.66

Table 14 shows three fold-0 test essays, one from the low, middle, and high end of the scale, with an excerpt of the learner's actual text and the band that each system assigns on every dimension. The essays are listed with their ELLIPSE text identifiers so that the cases can be located in the public corpus. These three essays are hand-picked, best-agreement illustrations and are not a basis for any general claim: within each band range, we chose an essay on which PROSE agrees closely with the human rater, so the table shows the

fine-tuned model near its best. We include them only to make the score-compression and downward-bias pattern concrete on readable text; the distributional evidence for all 400 evaluation essays for that behaviour is the distributional analysis of Table 13 and the banded-error breakdown of Figure 3c, neither of which involves any favourable selection. The pattern in these illustrative cases is that PROSE matches six of the seven human bands exactly in every case, while GPT-5.4 scores each essay between 0.50 and 1.29 score points too low on average, and the deficit is largest for the strongest writer.

Table 14. Illustrative, best-agreement case-study essays with text excerpts and per-dimension bands (Overall, Cohesion, Syntax, Vocabulary, Phraseology, Grammar, Conventions). These examples are selected for close PROSE–human agreement and are not representative of average behaviour; see Table 13 for the distributional evidence on all 400 evaluation essays. Bold labels identify score groups; italic text reproduces learner excerpts verbatim. Identifiers refer to the public ELLIPSE corpus. PROSE scores are rounded to the nearest half band; GPT-5.4 is prompted zero-shot with the rubric.

	Over.	Cohe.	Synt.	Voca.	Phra.	Gram.	Conv.
Low (ID 3F4A9ACFEDFB): *"for my uncle she death the blood cancer, she ti fight again for blood cancer for to manage for months agains to fail infectious venereal disease; she go to hospital …"*							
Human	2.0	2.0	2.0	2.0	2.0	1.5	2.0
PROSE	2.0	2.0	2.0	2.0	2.0	2.0	2.0
GPT-5.4	1.5	1.5	1.5	2.0	1.5	1.0	1.0
Middle (ID 009BCCC61C2A): *"positive attitude is the key to success. I agree because you can do anything as long as you put your mind and soul into it and motivation you can accomplish it …"*							
Human	3.0	3.0	3.0	3.5	3.5	3.0	3.0
PROSE	3.0	3.0	3.0	3.5	3.0	3.0	3.0
GPT-5.4	2.5	2.5	2.0	2.5	2.0	1.5	1.5
High (ID DC2E6750CDEC): *"Nowadays we live in a world where everything has to be perfect if you wanna fit into this society, therefore humans have made up an exteriotype of how people are supposed to dress, look, act and live …"*							
Human	4.0	3.5	4.0	4.0	4.0	4.0	4.0
PROSE	4.0	4.0	4.0	4.0	4.0	4.0	4.0
GPT-5.4	3.0	3.0	2.5	3.0	2.5	2.0	2.5

Figure 3c–e makes the behaviour of the systems concrete. Panel (c) reports banded error on the identical 400 fold-0 essays: PROSE places 40% of its scores in the exact human band and most of the remainder within 0.5 score points, whereas the three zero-shot models reach only 2% to 13% exact agreement and miss by 1.0 score point or more on the majority of essays. Panels (d) and (e) show the illustrative analytic profiles that the systems assign to the low- and high-scoring essays of Table 14. PROSE tracks the human profile across the seven dimensions, including the relative strengths and weaknesses within an essay, whereas the zero-shot model assigns lower scores than the human raters across both profiles. For the low-scoring essay, the human rater assigns band 2.0 on every dimension except Grammar, which receives 1.5; PROSE matches six of the seven bands exactly and misses Grammar by 0.5 score points, a mean absolute deviation of 0.07 score points, while the zero-shot model scores at or below the human band on every dimension and scores Conventions 1.0 score point too low, giving a mean absolute deviation of 0.50 score points. For the high-scoring essay, the gap widens. The human profile sits at band 4.0 with a single 3.5 on Cohesion, and PROSE again matches six of the seven bands exactly, with the same mean absolute deviation of 0.07 score points; the zero-shot model places the entire profile between bands 2.0 and 3.0 and has a mean absolute deviation of 1.29 score points, substantially underestimating the proficiency of this writer. These two essays are consistent with the full-sample statistics in

Table 13, which are the evidence we rely on: PROSE reproduces most of the human spread and keeps its mean near the human centre, while the large models compress the spread and shift the mean downward. In other words, the large models both flatten the distribution and score too harshly, which is why their error is several times larger even though their outputs read as plausible. This is exactly the failure mode that the concordance term is designed to avoid, and it explains why a small fine-tuned model gives more useful analytic feedback than a much larger general-purpose model used zero-shot.

6. Discussion

The central finding of this study is that, for automated analytic assessment of second-language writing, a small supervised model with a task-specific objective is a far more effective rater than general-purpose models used zero-shot, and that the gap is driven by task-specific supervision and objective design rather than by model scale. A compact encoder with a few million trainable adapter weights, trained with a rubric-aware multi-objective loss, is a substantially stronger analytic rater than three general-purpose large language models prompted zero-shot. The gap is not marginal. The fine-tuned model has higher QWK and Pearson correlation and lower RMSE and MAE; the ablations identify the concordance term as the principal supported source of the QWK improvement over plain regression. We are careful not to over-read this comparison: because our models are supervised on ELLIPSE while the LLMs are not, the experiment does not isolate scale as a causal factor, and a fairer test of scale alone would fine-tune the large models or provide them with in-context examples. What our experiment does establish is that, for a narrow rubric-bound task with abundant labelled data, task-specific adaptation of a small model is more accurate than zero-shot prompting of the tested large models and inexpensive to train and run locally.

The distributional analysis clarifies why the zero-shot models struggle. When asked to place an essay on a rubric they have not been trained to reproduce, the tested large models produce less variable scores and lower mean scores than the human raters. Score compression is also observed in a milder form for the squared-error regressor, and it is why plain regression, although a strong baseline, still benefits from the concordance term in PROSE, which explicitly rewards matching the spread of human ratings. The observation is consistent with the wider literature on large language models for learner writing, where fluent output masks weak calibration against human references [16,18,19]. It also suggests a complementary role rather than a competitive one: a small calibrated model can supply reliable scores, while a large model can supply the natural language explanation that turns a score into feedback.

Our baselines are deliberately controlled variants of a single encoder, which isolates the effect of the objective but does not place PROSE against the broader state of the art in AES. A direct numerical comparison with published ELLIPSE systems is difficult because reported results differ in data splits, target sets (holistic versus analytic), and preprocessing, and mixing such numbers uncritically would be misleading; for this reason, we restrict our quantitative comparisons to shared-encoder baselines evaluated on identical folds, and we treat the zero-shot LLMs as an external reference point. We regard a carefully matched comparison with recent multi-trait AES systems, re-run on our folds, as the most useful next step, and we release our folds and code to make such a comparison straightforward. We acknowledge plainly that, without such a comparison, our results do not show that PROSE is competitive with the strongest published AES systems. Our conclusions are limited to the controlled setting we tested: PROSE improves agreement with human ratings over shared-encoder baselines trained under identical conditions, and task-specific fine-tuning outperforms the zero-shot models we evaluated.

The finding also speaks to a broader assumption in current practice, that a larger general-purpose model is the natural first choice for any language task. For automated writing assessment, this assumption does not hold in the zero-shot setting we tested. The task is narrow, the label space is small and ordered, and what matters is agreement with a fixed rubric rather than open-ended generation. Under these conditions, a small model shaped to the task by its objective outperforms a much larger model used off the shelf, and it is far cheaper to run: only the adapters and the small output heads are trained, a cross-validation fold completes in about seven minutes on a single GPU, and inference costs one forward pass per essay, which is substantially cheaper than repeatedly querying a hosted model; we quantify the training budget, peak memory, and inference throughput in Table 15 and base the cost argument on the local forward-pass cost rather than on an undisclosed comparison of model sizes. Keeping the model local also avoids sending learner writing to an external service, so institutions with limited budgets, or settings where data cannot leave the premises, can still deploy an accurate analytic rater.

Table 15. Measured compute for the ELLIPSE experiments on a single NVIDIA A800 GPU (bfloat16). Times are wall-clock training time; inference adds a single forward pass per essay. Peak memory is the maximum memory allocated by PyTorch in GiB (2^{30} bytes), and throughput is end-to-end (tokenisation included) at batch size 16 and 512 tokens; both were measured in the re-runs described in Section 5.5 and are means over the five folds.

Quantity	Value
Trainable parameters (adapters + heads)	≈2.7 M (≈1.5% of the model)
Peak GPU memory, training (PROSE/Reg)	11.4 GiB/11.4 GiB
Peak GPU memory, inference	1.1 GiB
Inference throughput (PROSE/Reg)	≈165/≈166 essays s^{-1}
Training time per fold (PROSE)	≈6.7–7.1 min
Selected epoch (per fold, from validation)	4–7
Four main models × 5 folds (20 runs)	≈2.4 GPU-hours
Original ELLIPSE study (incl. ablations, sensitivity; 48 runs)	≈6.9 GPU-hours
Diagnostic re-runs and loss-weight sensitivity (40 runs)	≈9.0 GPU-hours
Combined ELLIPSE experiments (88 runs)	≈15.9 GPU-hours
Precision/hardware	bfloat16/NVIDIA A800

For teaching practice, analytic trait scores are the substrate of formative feedback: a learner gains more from knowing that grammar lags vocabulary than from a single holistic number, and written corrective feedback is most useful when it is specific and consistent [4,5]. A model that produces consistent per-trait scores at negligible cost can support frequent feedback in classrooms and self-study platforms, with human raters reserved for high-stakes decisions.

Several limitations qualify these conclusions. First, we evaluate on one corpus per label scheme; although the two corpora differ in genre and granularity, a broader study across more learner populations and first-language backgrounds would strengthen the claims. Second, PROSE produces scores rather than the explanatory comments that constitute full written corrective feedback; bridging from calibrated scores to useful textual feedback is an open problem that our results only set up. Third, our subgroup analysis (Section 5.5) is a first check rather than a fairness audit. It found no gender, ethnicity, or socioeconomic difference in residual bias that survives conditioning on the human score, apart from a borderline ethnicity effect, but it did find a grade-related bias shared by PROSE and the regression baseline, and several groups were too small to analyse. Our feature-level analysis is also correlational: it shows that predictions depend more strongly on length and lexical diversity than human scores do, but it does not explain individual predictions. Before any operational use, the model would need a proper fairness audit on the target population and an instance-level explanation method, and its scores should not be used for high-stakes

decisions without human review. Fourth, all inputs are truncated to 512 tokens, and 38.5% of ELLIPSE essays exceed this length (Table 1); although the truncation is applied identically to every model and so does not bias the comparison, it is associated with lower absolute performance on long essays (QWK about 0.02 lower for both models, Section 5.5), and a longer context window is a natural way to test for further gains. The archive records processed-row split indices but does not document a complete mapping from the original ELLIPSE release to the 3911-essay input table, which limits independent reconstruction of the exact study sample. Finally, our comparison with large language models is limited to zero-shot prompting; few-shot prompting, chain-of-thought rationales, or fine-tuning of the large models could narrow the gap, and such a study is needed before any claim about model scale as such could be made, at a cost in compute that our method avoids.

7. Conclusions

We studied automated analytic assessment of second-language writing and asked what turns a small, openly available language model into a reliable rater. We argued that three properties of the task, ordinality, calibration, and trait relations, should be built into the training objective, and we realised them in PROSE, a compact multi-objective method that places a regression read-out and a soft-ordinal auxiliary head on a shared low-rank-adapted DeBERTa-v3 encoder, and trains them with squared error, a distance-aware soft-ordinal term, a concordance term, and a trait-relational coherence term. On the ELLIPSE analytic corpus, PROSE improved quadratic weighted kappa over regression, classification, and pure ordinal baselines by a small but statistically reliable margin (paired t-test $p = 0.007$; corrected resampled test $p = 0.030$), with a gain on each of the seven scoring dimensions, significant on four after Holm correction; the gain is stable across reasonable loss weights, essay lengths, and learner subgroups, and the soft-ordinal and coherence terms contribute little to it; on the Write & Improve CEFR data, where agreement saturates, it matched the strongest baseline on agreement and accuracy and clearly beat classification and pure ordinal training. A controlled comparison on an identical essay sample showed that three strong large language models, prompted zero-shot, fall far short of the fine-tuned compact model, and a full-sample distributional analysis linked the gap to their compressed score distributions and lower mean scores. The practical message is that, for this task, task-specific supervision and a rubric-aware objective applied to a small model are more effective than zero-shot use of the tested large models and inexpensive to train and run locally; we draw no conclusion about model scale in isolation, which would require fine-tuning or in-context adaptation of the large models. These conclusions hold for the controlled comparisons we ran; we have not compared PROSE with externally published AES systems on our folds. Future work will extend the approach from scores to explanatory written corrective feedback, evaluate it across more learner populations, and study how a calibrated small rater can be paired with a large language model that supplies the explanation. A calibrated rater can also act as a reward signal for learning feedback-generation policies; search-based reinforcement learning with neural Monte Carlo tree search and self-play has handled structured problems such as first-order logic, combinatorial optimisation, and quantified Boolean formulas [42–44], and a similar approach could train a generator to produce feedback whose predicted effect on the trait profile is largest.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app162110683/s1, Supplementary Material S1 accompanies this manuscript as a reproducibility archive containing the code, environment specification, random seeds, fold indices, run configurations, LLM prompt and parser, saved per-run results, and available per-essay predictions.

Author Contributions: Conceptualization, H.X. and Q.T.; methodology, H.X.; software, H.X.; validation, H.X., J.D. and R.X.; formal analysis, J.D.; investigation, R.X.; data curation, J.D.; writing—original draft preparation, H.X.; writing—review and editing, J.D., R.X. and Q.T.; supervision, Q.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The ELLIPSE corpus is openly available at https://github.com/scrosseye/ELLIPSE-Corpus (accessed on 22 October 2026), and the Write & Improve data are distributed with the BEA-2019 shared task at https://www.cl.cam.ac.uk/research/nl/bea2019st/ (accessed on 22 October 2026). The accompanying Supplementary Material S1 provides the experimental code and saved results. It contains the training and evaluation scripts; the exact environment and package versions (Python 3.10.19, PyTorch 2.7.1 with CUDA 12.6, Transformers 5.17.0, PEFT 0.19.1, scikit-learn 1.7.2, SciPy 1.15.3, NumPy 1.26.4); the random seeds (split seed 42, training seeds 42, 1, and 2); the exact training, validation, and test indices of every ELLIPSE fold; the command lines and hyperparameter configuration of every run; the zero-shot LLM prompt, decoding settings, and output parser; the saved per-run result files and available per-essay predictions; and the scripts that compute the statistical tests and the analyses of Section 5.5. API keys are not included. The original corpora must be obtained separately from the sources above. Exact reconstruction of the 3911-essay processed ELLIPSE input requires the corresponding source-to-row selection mapping or the identically ordered processed input table.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ke, Z.; Ng, V. Automated Essay Scoring: A Survey of the State of the Art. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 6300–6308.
2. Ramesh, D.; Sanampudi, S.K. An automated essay scoring systems: A systematic literature review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [CrossRef] [PubMed]
3. Bitchener, J. Evidence in support of written corrective feedback. *J. Second. Lang. Writ.* **2008**, *17*, 102–118. [CrossRef]
4. Bitchener, J.; Knoch, U. Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *J. Second. Lang. Writ.* **2010**, *19*, 207–217. [CrossRef]
5. Ferris, D.R. Second Language Writing Research and Written Corrective Feedback in SLA: Intersections and Practical Applications. *Stud. Second. Lang. Acquis.* **2010**, *32*, 181–201.
6. Ferris, D.; Roberts, B. Error feedback in L2 writing classes: How explicit does it need to be? *J. Second. Lang. Writ.* **2001**, *10*, 161–184.
7. Taghipour, K.; Ng, H.T. A Neural Approach to Automated Essay Scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 1882–1891.
8. Alikaniotis, D.; Yannakoudakis, H.; Rei, M. Automatic Text Scoring Using Neural Networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany, 7–12 August 2016; pp. 715–725.
9. Dong, F.; Zhang, Y.; Yang, J. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL), Vancouver, BC, Canada, 3–4 August 2017; pp. 153–162.
10. Ridley, R.; He, L.; Dai, X.Y.; Huang, S.; Chen, J. Automated Cross-prompt Scoring of Essay Traits. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 13745–13753.
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019.
12. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
13. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, 2021. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
14. He, P.; Gao, J.; Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv* **2021**, arXiv:2111.09543.
15. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
16. Mizumoto, A.; Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. [CrossRef]

17. Yancey, K.P.; Laflair, G.; Verardi, A.; Burstein, J. Rating Short L2 Essays on the CEFR Scale with GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), Toronto, ON, Canada, 13 July 2023; pp. 576–584.
18. Naismith, B.; Mulcaire, P.; Burstein, J. Automated evaluation of written discourse coherence using GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), Toronto, ON, Canada, 13 July 2023; pp. 394–403.
19. Davis, C.; Caines, A.; Andersen, Ø.; Taslimipoor, S.; Yannakoudakis, H.; Yuan, Z.; Bryant, C.; Rei, M. Prompting open-source and commercial language models for grammatical error correction of English learner text. *arXiv* **2024**, arXiv:2401.07702.
20. Crossley, S.; Tian, Y.; Baffour, P.; Franklin, A.; Kim, Y.; Morris, W.; Benner, M.; Picou, A.; Boser, U. The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *Int. J. Learn. Corpus Res.* **2023**, *9*, 248–269. [CrossRef]
21. Bryant, C.; Felice, M.; Andersen, Ø.E.; Briscoe, T. The BEA-2019 Shared Task on Grammatical Error Correction. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA), Florence, Italy, 2 August 2019; pp. 52–75.
22. Bryant, C.; Felice, M.; Briscoe, T. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 793–805.
23. Cao, W.; Mirjalili, V.; Raschka, S. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognit. Lett.* **2020**, *140*, 325–331. [CrossRef]
24. Díaz, R.; Marathe, A. Soft Labels for Ordinal Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4733–4742.
25. Lin, L.I.K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268. [CrossRef]
26. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef] [PubMed]
27. Liu, X.; Gao, Y.; Zong, L.; Liang, W.; Xu, B. Guiding Prototype Networks with label semantics for few-shot text classification. *Pattern Recognit.* **2025**, *164*, 111497. [CrossRef]
28. Cai, Z.; Zhang, H.; Zhan, P.; Jia, X.; Yan, Y.; Song, X.; Xie, B. Multi-schema prompting powered token-feature woven attention network for short text classification. *Pattern Recognit.* **2024**, *156*, 110782. [CrossRef]
29. Aghajanzadeh, E.; Bahraini, T.; Mehrizi, A.H.; Yazdi, H.S. Task weighting based on particle filter in deep multi-task learning with a view to uncertainty and performance. *Pattern Recognit.* **2023**, *140*, 109587. [CrossRef]
30. Zhao, J.; Zhu, T.; Sun, L.; Du, B.; Wang, H.; Huang, L. Dual-mask: Progressively sparse multi-task architecture learning. *Pattern Recognit.* **2025**, *158*, 110950. [CrossRef]
31. Wang, J.; Fei, L.; Sun, L. Multi-level network Lasso for multi-task personalized learning. *Pattern Recognit.* **2025**, *161*, 111213. [CrossRef]
32. Li, X.; Guo, C.; Tang, X.; Jie, Y. RectLoRA: Subspace parameter-efficient fine-tuning for continual adaptation of LLMs and LVMs. *Pattern Recognit.* **2026**, *179*, 113887. [CrossRef]
33. Wang, Z.; Mo, M.; Xiao, X.; Liu, C.; Ma, C.; Zhang, Y.; Wang, X.; Krishnaswamy, S.; Wang, T. CTR-LoRA: Curvature-Aware and Trust-Region Guided Low-Rank Adaptation for Large Language Models. In *Proceedings of the ICASSP 2026–2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2026*; IEEE: Piscataway, NJ, USA, 2026; pp. 1396–1400.
34. Xiao, X.; Wang, Z.; Mo, M.; Liu, C.; Ma, C.; Li, Y.; Krishnaswamy, S.; Wang, X.; Wang, T. Self-Supervised Visual Prompting for Cross-Domain Road Damage Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 6–10 March 2026; pp. 3514–3524.
35. Jiang, C.; Wang, Y.; Xiong, B. Dual similarity enhanced hybrid orthogonal fusion for multimodal named entity recognition. *Pattern Recognit.* **2026**, *169*, 111940. [CrossRef]
36. Zhou, Y.; Ni, L.; Qu, Z.; Li, N.; Zhang, J.; Liang, Y. SASA: Bridging the semantic gap for node classification in text-attributed graphs using self-adversarial alignment. *Pattern Recognit.* **2026**, *180*, 113942. [CrossRef]
37. Xiao, Y.; Shu, K.; Zhang, H.; Yin, B.; Cheang, W.S.; Wang, H.; Gao, J. EGGesture: Entropy-Guided Vector Quantized Variational Autoencoder for Co-Speech Gesture Generation. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 6113–6122.
38. Nadeau, C.; Bengio, Y. Inference for the generalization error. *Mach. Learn.* **2003**, *52*, 239–281. [CrossRef]
39. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
40. Wang, Y.; Wang, J.; Zeng, T.; Qi, J. Data-mining-based biomarker evaluation and experimental validation of SHTN1 for bladder cancer. *Cancer Genet.* **2024**, *288–289*, 43–53. [CrossRef] [PubMed]
41. Kawahara, D.; Kishi, M.; Kadooka, Y.; Hirose, K.; Murakami, Y. Integrating radiomics and gene expression by mapping on the image with improved DeepInsight for clear cell renal cell carcinoma. *Cancer Genet.* **2025**, *292–293*, 100–105. [CrossRef] [PubMed]

42. Xu, R.; Kadam, P.; Lieberherr, K. First-order problem solving through neural MCTS based reinforcement learning. *arXiv* **2021**, arXiv:2101.04167.
43. Xu, R.; Lieberherr, K. Learning self-play agents for combinatorial optimization problems. *Knowl. Eng. Rev.* **2020**, *35*, e11. [CrossRef]
44. Xu, R.; Lieberherr, K. Towards tackling QSAT problems with deep learning and Monte Carlo tree search. In *Proceedings of the Intelligent Computing: Proceedings of the 2022 Computing Conference (Science and Information Conference)*; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2022; pp. 45–58.

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