Data Mining and Knowledge Discovery (2026) 40:35
https://doi.org/10.1007/s10618-026-01193-z

Exploring zero-shot essay scoring: from feature-based to LLM-based approaches

Hongseok Choi[1] · **Myeong-Cheol Kang**[1] · **Jin Seong**[1] · **Jin-Xia Huang**[1]

Received: 7 October 2025 / Accepted: 5 February 2026

Abstract
Automated Essay Scoring (AES) plays a crucial role in educational assessment, yet building reliable systems often requires large amounts of prompt- and trait-specific labeled data. As a result, many studies rely heavily on a single public dataset, ASAP++, which limits their generalizability. In this paper, we explore diverse zero-shot AES approaches that are both prompt-agnostic and multi-trait, thereby reducing reliance on labeled data. Our study examines large language models (LLMs), BERT-like fine-tuned models, and simple feature-based methods. Furthermore, we propose two novel zero-shot frameworks: CAnUSe (Comparative Assessment and subsequent Uncertainty-aware Self-training) and EASY (Essay Assessment with Simple Yardsticks). CAnUSe combines LLM-based pairwise comparisons with uncertainty-aware self-training using a BERT-like single-essay scoring model, while EASY employs only three intuitive features-essay length, vocabulary diversity, and grammar quality. Evaluation on three benchmark datasets-ASAP++, Ellipse, and TOEFL11-shows that our frameworks achieve state-of-the-art performance, demonstrating strong generalizability across datasets. Remarkably, despite operating in a zero-shot setting, our models approach the performance of fully supervised models trained with approximately 10K labeled samples. Further analysis provides practical insights for building real-world AES systems.

Keywords Automated essay scoring · Zero-shot learning · Large language models · Feature-based methods · Uncertainty-aware self-training · Low-resource settings

Extended author information available on the last page of the article

Published online: 06 April 2026

1 Introduction

Automated Essay Scoring (AES) aims to evaluate and score essays automatically using technologies such as machine learning (ML) and natural language processing (NLP) (Almusharraf and Alotaibi 2023) techniques. In modern education, AES has become a crucial tool for efficiently and consistently assessing large volumes of student writing (Reilly et al. 2014; Cavalcanti et al. 2021). Unlike human evaluation, AES systems are not constrained by time and can provide immediate feedback, thereby supporting students' continuous learning (Ifenthaler 2022).

Over the years, researchers have explored various AES approaches (Ramesh and Sanampudi 2022). Feature-engineering methods extract diverse syntactic and lexical features to train ML models (Phandi et al. 2015; Li and Ng 2024). In contrast, deep learning approaches aim to automatically learn more complex patterns from essays and prompts (Dong et al. 2017; Yang et al. 2020; Wang et al. 2022; Li and Gao 2025; Cai et al. 2025). Recently, hybrid models that combine deep learning and feature engineering have become predominant (Tay et al. 2018; Li et al. 2020; Chen et al. 2025). While these methods have shown strong performance, they require large amounts of labeled data for training.

Typically, AES datasets consist of prompts, essays, and corresponding scores. In multi-trait settings, they include an overall score as well as trait-specific scores, such as content, organization, and conventions (Mathias and Bhattacharyya 2018). Since essays are often lengthy and evaluated across multiple traits, annotating such data is labor-intensive and time-consuming, thereby increasing the cost and complexity of dataset creation. Moreover, the diverse writing styles of novice learners demand strict adherence to well-defined rubrics during annotation (Ramesh and Sanampudi 2022).

These challenges have motivated researchers to explore data-efficient approaches to reduce annotation effort (Tao et al. 2022; He and Li 2024). Cross-prompt methods, in particular, have been widely studied in multi-trait settings (Ridley et al. 2021; Han et al. 2025). These methods aim to generalize to unseen prompts by leveraging existing annotated essay data, thereby reducing the need to construct new datasets whenever a new prompt is introduced. However, most current cross-prompt studies still assume that the training and test data come from the same distribution (e.g., many studies rely solely on the ASAP++ dataset) and require large amounts of labeled data for the source prompts (Mathias and Bhattacharyya 2018; Li et al. 2020; Do et al. 2023; Li and Ng 2024). Similarly, multi-trait settings often assume that both source and target prompts share the same set of traits. These constraints limit their applicability in real-world settings.

Recently, large language models (LLMs) have been explored as a promising means of zero-shot AES (Zhao et al. 2023; Askarbekuly and Aničić 2024). However, prior studies have mainly applied LLMs with a focus on prompting strategies, and the broader potential of LLMs for AES has not been fully explored (Mizumoto and Eguchi 2023; Lee et al. 2024; Stahl et al. 2024). Moreover, these LLM-based methods generally require large model sizes (over 8B parameters), raising concerns about their feasibility for real-time evaluation. Thus, these issues underscore the need for further improvement.

In this paper, we explore diverse zero-shot AES approaches, such as LLM-based methods, BERT-like fine-tuned models (Devlin et al. 2019), and simple feature-based methods. We further propose two novel frameworks: CAnUSe (Comparative Assessment and subsequent Uncertainty-aware Self-training) and EASY (Essay Assessment with Simple Yardsticks). In CAnUSe, LLMs first perform pairwise comparative assessment to generate pseudo-scores for unrated essays. The pseudo-labels are filtered using uncertainty metrics, and a BERT-like single-essay scoring model is subsequently fine-tuned through self-training on the high-confidence samples to improve inference efficiency and prediction performance with a smaller model. In contrast, EASY employs only three straightforward features: essay length, vocabulary diversity, and grammar quality.

We conduct extensive experiments on three public AES datasets: ASAP++, Ellipse, and TOEFL11 (Mathias and Bhattacharyya 2018; Blanchard et al. 2013; Crossley et al. 2023). The results show that CAnUSe achieves state-of-the-art performance across datasets, performing comparably to supervised baselines trained on approximately 10K labeled samples. Moreover, EASY demonstrates surprisingly strong performance despite its simplicity, even outperforming fully supervised feature-based methods. Additionally, our study analyzes key performance factors of the proposed frameworks to provide practical insights for developing real-world AES systems.

To sum up, our contributions are as follows:

- We explore diverse zero-shot AES approaches, covering LLM-based methods, BERT-like fine-tuned models, and simple feature-based methods, to examine their potential and limitations.
- We propose two novel zero-shot AES frameworks, CAnUSe and EASY. CAnUSe combines LLM-based comparative assessment with uncertainty-based self-training, while EASY employs only three straightforward features. Despite using no labels, CAnUSe achieves competitive performance compared to strong supervised baselines while offering better inference efficiency than LLM-based baselines. EASY, in turn, outperforms fully supervised feature-based methods while using fewer and simpler features.
- We conduct extensive experiments on three benchmark datasets-ASAP++, Ellipse, and TOEFL11-demonstrating the robustness and generalizability of our frameworks across diverse AES benchmarks. Further analysis provides practical insights for building real-world AES systems.

2 Related work

AES has been actively studied in the educational domain, but the number of publicly available datasets remains limited. Many studies have relied heavily on ASAP++ (Mathias and Bhattacharyya 2018), which restricts the evaluation of a models' generalization ability. Although other AES datasets like Ellipse and TOEFL11 have been introduced (Blanchard et al. 2013; Crossley et al. 2023), a standardized evaluation framework is still lacking. Early AES approaches mainly focused on

prompt-specific models, which suffer from poor performance on unseen prompts (Taghipour and Ng 2016; Uto 2021; Xie et al. 2022). To address this, cross-prompt methods have been proposed, often combining traditional linguistic features with deep learning (Li et al. 2020; Ridley et al. 2021). However, most existing cross-prompt studies have been confined to a single dataset, such as ASAP++, and have not been sufficiently explored on diverse AES datasets that vary in domains such as student populations and scoring rubrics (Ridley et al. 2021; Do et al. 2023; Li and Ng 2024). As a result, their true generalizability across datasets remains uncertain.

Recently, LLMs have been explored as a promising direction for AES. However, most existing studies rely on simple prompting strategies based on rubric descriptions or five-point Likert scales (Lee et al. 2024; Stahl et al. 2024). Concurrently, Shibata and Miyamura (2025) applied LLM-driven comparative assessment, followed by fine-tuning RankNet (Burges et al. 2005) to learn pairwise preferences, focusing on overall scoring. In this work, we demonstrate broader generalizability by covering the multi-trait setting and introduce two zero-shot frameworks-one LLM-based and the other feature-based-along with a detailed analysis for practical AES.

3 Proposed frameworks

3.1 CAnUSe: overall framework

Figure 1 illustrates our framework, CAnUSe, which takes a set of unrated essays (X_u) as input and proceeds in two stages: (1) **C**omparative **A**ssessment using an LLM, a**n**d (2) selection of high-confidence samples with an **U**ncertainty metric, followed by **Se**lf-training of a single-essay scoring (SES) model. These steps are detailed in the following sections.

3.1.1 Pairwise comparative assessment with LLMs

The first stage is the Zero-shot AES module. To assess essay x_i, we randomly select M essays from the unlabeled set X_u without replacement. Let the selected essay set for x_i be C_i. The LLM performs a comparative assessment between x_i and the essays in C_i. Given an essay prompt and two essays, the LLM is asked to determine which essay is better written in terms of a specific trait. For more precise comparison, a rubric for the trait is provided. Ultimately, the win rate of x_i is

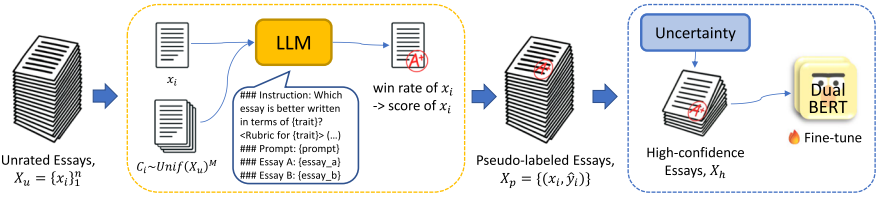

Fig. 1 Overall framework of CAnUSe

```
### Instruction: Which of these two essays is better written "overall"?
Just pick one. Consider the following rubrics in your evaluation:
<Rubric for Evaluating Overall Quality>
 - Good: The essay presents its ideas clearly and stays focused on its
purpose. The structure is logical, the tone is consistent, and the
writing reads smoothly overall.
 - Poor: The essay is unclear or unfocused, with weak structure or
inconsistent tone. The writing feels disjointed and fails to effectively
deliver its message.
### Answer format: A or B.
### Prompt: Write a letter to your local newspaper in which you state
your opinion on the effects computers have on people. Persuade the
readers to agree with you.\n
### Essay A: {essay_a}
### Essay B: {essay_b}
### Answer: Essay
```

Fig. 2 LLM prompt for comparative assessment of the overall trait in ASAP++. The essay prompt has been shortened for clarity. In *Answer*, *Essay* is used as a prefix token, which allows the LLM to generate either the A or B token as the next token

considered its relative score. This stage results in a set of pseudo-rated essays, X_p. An LLM prompt example is shown in Fig. 2, and more examples are provided in Appendix A.

Let $LLM(a, b)$ denote a binary comparison function, where $LLM(a, b) = 1$ if essay a is judged to be better written than essay b, and $LLM(a, b) = 0$ otherwise. To mitigate positional bias in LLM-based comparisons, each pair of essays is evaluated twice by swapping their input order. Given the comparative set C_i for essay x_i, we compute its win rate as follows:

$$\frac{1}{2|C_i|} \sum_{x_j \in C_i} \big(LLM(x_i, x_j) + (1 - LLM(x_j, x_i)) \big) \tag{1}$$

LLMs, pretrained on vast corpora, are known for their ability to perform well on various NLP tasks without task-specific fine-tuning. However, existing methods in the AES domain have typically relied on simpler prompting, such as five-scale Likert scoring, and thus failed to fully leverage LLMs' pretrained knowledge. In contrast, our comparative assessment prompts LLMs to compare essay pairs and aggregates the results to compute relative scores. Liusie et al. (2024) explored LLMs' comparative assessment in general domains.

3.1.2 Single-essay scoring module

Our comparative assessment requires multiple essay pairs for evaluating a single essay, which entails substantial computation. To address this limitation, we propose using pseudo-rated essays obtained from zero-shot comparative assessment as training data for a single-essay scoring (SES) model. This enables the final model to perform single-essay rather than pairwise scoring, offering fast inference with

a lightweight model. To select reliable training data from pseudo-rated essays, we utilize an uncertainty metric for an LLM, defined as:

$$-\frac{1}{|C_i|}\sum_{(x_i,x_j)\in C_i}\left|\big(l_a(x_i,x_j)+l_b(x_j,x_i)\big)-\big(l_a(x_j,x_i)+l_b(x_i,x_j)\big)\right| \tag{2}$$

where $l_a(x_i,x_j)$ and $l_b(x_i,x_j)$ denote the logits for tokens A and B, respectively, given the essay pair (x_i,x_j). The first term $\big(l_a(x_i,x_j)+l_b(x_j,x_i)\big)$ represents the score indicating how likely x_i is to beat x_j, and the second term $\big(l_a(x_j,x_i)+l_b(x_i,x_j)\big)$ represents the opposite. The absolute difference between these terms reflects the degree to which the LLM is confident in preferring one essay over the other, while also capturing the consistency of its judgment when the order of the same essay pair is swapped. We negate the confidence score to represent it as an uncertainty measure.

We train the SES model using only the $p\%$ least uncertain samples. To maintain score distribution, we sort the predictions by pseudo-scores into *B* bins and sample the $p\%$ samples from each bin, which we call Label Balancing. Additionally, to mitigate the impact of noisy pseudo-labels, we use a bagging method, training four models with randomly shuffled training and validation splits. We adopt DualBERT (Cho et al. 2024) as our SES model, as it has shown strong predictive performance in multi-trait settings.

We further explore few-shot settings to maximize the practical use of our frameworks, where we fine-tune DualBERT (Cho et al. 2024) on a mixture of a small amount of labeled data and pseudo-labeled data.

3.2 EASY: simple feature-based zero-shot AES

Apart from the aforementioned framework, we also propose **E**ssay **A**ssessment with **S**imple **Y**ardsticks (EASY), which employs simple features and can be applied in zero-shot settings. In essay scoring, a common assumption is that longer essays with more diverse vocabulary tend to be of higher quality. However, this relationship may not be strictly linear, as essay length and vocabulary diversity do not necessarily guarantee essay quality. Accordingly, we define the following formula:

$$Score_1 = \log(1+n_e)\cdot\log(1+n_v) \tag{3}$$

where n_e is the essay length (i.e., number of words) and n_v is the vocabulary size used in the essay.

Another assumption is that high-quality essays contain fewer grammatical errors. We estimate the degree of grammatical error using `grammarly/coedit-xl`, a pretrained T5-based LLM, which outputs a grammatically corrected version of the input text. The model is applied to each essay on a sentence-by-sentence basis. To split an essay into multiple sentences, we use the Python `spaCy` library. For each sentence, we compute the word error rate (WER) using the Python library `jiwer`. The overall grammar score of an essay is defined as follows:

$$Score_2 = \frac{1}{|S|} \sum_{s_i \in S} g\big(1 - WER(s_i, \hat{s}_i)\big)$$
$$g(x) = \begin{cases} x, & \text{if } x > 0.9 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $S = \{s_i\}_1^{n_s}$ is an essay with n_s sentences, s_i is an original sentence, and $\hat{s}_i$ is the version corrected by `coedit-xl`.

Finally, we combine $Score_1$ and $Score_2$ using a weighting hyperparameter α that reflects prior knowledge, as in $Score_1 \cdot \big(\alpha + (1-\alpha) \cdot Score_2\big)$. For instance, α can be set lower for the trait conventions than for content, since grammar plays a more critical role in conventions. To reduce the effect of outliers, we clip final test scores at the 5th and 95th percentiles.

4 Experimental

4.1 Datasets

To evaluate the generalizability of our frameworks, we conduct experiments on a range of AES datasets from diverse sources. Specifically, we use three AES benchmarks: ASAP++, Ellipse, and TOEFL11. Dataset statistics are presented in Table 1.

ASAP++: The ASAP++ dataset is a widely used benchmark that is an enriched version of the ASAP corpus released through a Kaggle competition (Mathias and Bhattacharyya 2018). It comprises 12,978 English essays written by U.S. students in grades 7–10 across eight prompts. ASAP++ includes trait-level annotations (e.g., content, organization, word choice, conventions) for eight prompts, enabling more fine-grained analysis in AES. In total, 11 traits, including the overall score, are annotated.

Ellipse: The English Language Learner Insight, Proficiency and Skills Evaluation (Ellipse) corpus is a publicly available dataset consisting of 6,482 essays written by English Language Learners (ELLs) in grades 8–12 (Crossley et al. 2023). The essays span 44 prompts and are annotated with both an overall score and six trait scores: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Ellipse includes a diverse student population with information on gender, race/ethnicity, grade level, and economic status.

Table 1 Statistics of the datasets

	ASAP++	TOEFL11	Ellipse
# Essays	13k	12.1k	6.5k
# Prompts	8	8	44
# Traits	11	1	7
Category	P, S, N	P	P
Language	Native	ELL	Native+ELL
Avg. Length	251	348	428
Avg. Essays/Prompt	1623	1513	147

P, S, and N denote essay categories: persuasive, source-dependent, and narrative, respectively

TOEFL11: The TOEFL11 corpus is a publicly available dataset of English essays written by non-native speakers from 11 different first-language (L1) backgrounds (Blanchard et al. 2013). It contains 12,100 essays written in response to TOEFL iBT independent writing prompts, balanced across eight topics and scored by human raters on a discrete three-level scale (low, medium, high), with only overall scores annotated.

For evaluation, we use the quadratic weighted kappa (QWK) (Cohen 1968), a metric commonly used in AES. Since TOEFL11, unlike ASAP++ and Ellipse, provides categorical labels-low, medium, and high-we scale the predicted scores to the range [1, 5] and map them into Low, Medium, and High categories using thresholds at 2.25 and 3.75, in line with prior work and consistent with the official TOEFL11 setting (Blanchard et al. 2013; Lee et al. 2024; Shibata and Miyamura 2025).

4.2 Implementation details

We conduct experiments using four instruction-tuned LLMs: Phi-3 (3.8B), Mistral (7B), Llama-3.1 (8B), and Qwen2.5 (7B).[1] To examine the effect of model size, we additionally evaluate larger variants, including Llama-3.1 (70B), Qwen2.5 models ranging from 3B to 72B, and Phi-3 (14B). NVIDIA H100 GPUs were used for all experiments. In the LLM comparative assessment, we used a symmetric set consisting of 16 randomly selected essays. For the SES module, we employed DualBERT and followed the same hyperparameter settings (Cho et al. 2024). We use half of the training data as validation data. Fine-tuning was performed using AdamW with a learning rate of 2e−5 and an epsilon of 1e−8 (Loshchilov and Hutter 2019); the dropout rate was set to 0.1 (Srivastava et al. 2014), the batch size to 16, and early stopping was applied with a patience of 10 within 100 epochs. We selected the top 10% of the least uncertain samples, according to the uncertainty metric, to self-train the SES model. In EASY, the `spaCy` library was used with the `en_core_web_sm` model for sentence splitting. In zero-shot settings, the weighting hyperparameter α was determined based on suggestions from Claude.[2] More details of the hyperparameter α are provided in Sect. 6.3. For ASAP++, we follow prior work and conduct experiments using a five-fold test split, while for TOEFL11 and Ellipse we evaluate on their official test sets. When fine-tuning the SES module, we split the data into training and validation sets at a 5:5 ratio-using only pseudo-labeled data in the zero-shot setting and a combination of pseudo-labeled and labeled data in the few-shot setting. In line with the existing literature, we conduct single-task learning on each prompt in ASAP++, and multi-task learning on TOEFL11 and Ellipse by combining prompts within each dataset (Jiang et al. 2023; Cho et al. 2024; Eltanbouly et al. 2025).

[1] The corresponding model IDs on https://huggingface.co/models are microsoft/Phi-3-mini-4k-instruct mistralai/Mistral-7B-Instruct-v0.3, meta-llama/Llama-3.1-8B-Instruct, and Qwen/Qwen2.5-7B-Instruct, respectively.

[2] https://claude.ai/chats.

4.3 Methods for comparison

LLM-based Scoring is categorized into two approaches: comparative assessment (**LLM-comp**) and absolute scoring (**LLM-abs**). LLM-comp follows the previously described pairwise comparison framework. In contrast, LLM-abs leverages rubric-guided prompting for absolute assessment. Specifically, we utilize the official scoring rubrics when available (e.g., TOEFL11 and Ellipse), with scores ranging from 1 to 5. When no official rubric is provided (e.g., ASAP++), we construct a simple rubric using ChatGPT.[3] For prediction, the model generates token logits corresponding to scores 1 through 5, from which we select the two tokens with the highest logits. The final score is then computed as a weighted sum based on their softmax-normalized probabilities. Examples of instruction prompts are provided in Appendix A.

Lee et al. (2024) and Shibata and Miyamura (2025) investigated LLM-based approaches on the ASAP++ and TOEFL datasets, with their evaluations limited to overall scoring. **Vanilla-CoT** (Lee et al. 2024) is an LLM-abs variant that uses Chain-of-Thought prompting (Wei et al. 2022) to justify scores. **MTS** (Multi-Trait Specification) (Lee et al. 2024) creates trait-level rubrics, scores each trait separately, and aggregates them into the final score, aiming for greater reliability. **LCES** (LLM-based Comparative Essay Scoring) (Shibata and Miyamura 2025) also starts with comparative assessment, similar to our framework. Unlike ours, however, it leverages RankNet (Burges et al. 2005) in the second stage to learn pairwise preferences from the LLM outputs and compute final scores for test essays.

Fine-tuned Models: As baselines, we apply several fine-tuned models, including **BERT** (Devlin et al. 2019), **BigBird** (Zaheer et al. 2020), and **DualBERT** (Cho et al. 2024). BERT is a transformer-based pretrained model that has been widely adopted for various classification tasks. BigBird is an extended variant of RoBERTa (Liu 2019) with a maximum input length of 4096 tokens (compared to 512 tokens in BERT and RoBERTa), which makes it potentially suitable for AES due to the relatively long nature of essay data. DualBERT is a hierarchical architecture that processes essays at both the sentence and document levels, enabling it to better capture semantic and structural information (Cho et al. 2024). DualBERT has demonstrated strong performance on the ASAP++ dataset, particularly in multi-trait essay scoring settings.

SoTA Baselines: To provide an upper bound for AES performance, we report results from state-of-the-art (SoTA) *supervised* models fine-tuned on *full* training data ($\approx$*10K* labeled essays). We categorize them into two groups. (1) **Cross-prompt models** aim to generalize to unseen prompts using essays from seen prompts. **MOOSE** (Chen et al. 2025) demonstrated SoTA performance in cross-prompt settings of the ASAP++ dataset; it leverages multi-chunk BERT and linguistic features. In contrast, Li and Ng (2024) employ an extensive collection of 1,535 engineered features with simple correlation-based filtering and fine-tune a two-layer neural network on these features (hereafter referred to as **1.5K.feats. ft**). (2) **Prompt-specific SoTA Models**: **ArTS+RMTS** (Chu et al. 2024; Do et al.

[3] https://chatgpt.com/.

2024) and **SaMRL** (Do et al. 2024) are extensions of ArTS, a T5-based model (Raffel et al. 2020); ArTS+RMTS leverages rubric-guided rationales from LLMs, while SaMRL applies multi-reward reinforcement learning to boost performance. All prior studies on these SoTA baselines were evaluated only on the ASAP++ dataset; we use their reported results for comparison.

5 Results

5.1 Zero-shot setting

Tables 2 and 3 present QWK results on ASAP++, TOEFL11, and Ellipse, respectively.

LLM-abs-Z vs. LLM-comp-Z: Overall, LLMs perform better in comparative assessment (LLM-comp-Z) than in absolute assessment (LLM-abs-Z). Pairwise comparison is often easier for LLMs than assigning a score on a 5-point scale, as it reduces the decision space from five score tokens (1–5) to a binary choice (A vs. B).

EASY-Z: Despite relying on only three simple and intuitive features, EASY-Z achieves surprisingly competitive performance. On ASAP++, EASY-Z outperforms 1.5K.Feats.ft-which uses 1,535 engineered features and 10K labeled essays for training-by about 1%. On TOEFL11, EASY-Z also surpasses LLM-comp-Z, demonstrating its strong performance despite its simplicity.

CAnUSe-Z: CAnUSe-Z is trained on high-confidence pseudo-labeled samples extracted from LLM-comp-Z using an uncertainty metric. Across datasets, CAnUSe-Z consistently outperforms LLM-comp-Z, implying that (i) the uncertainty metric effectively distinguishes reliable predictions and (ii) training on a subset of high-confidence pseudo-labels is effective. A detailed analysis of uncertainty is presented in Sect. 6.2. Remarkably, without using any true labels, CAnUSe-Z achieves performance competitive with fully supervised SoTA models. On ASAP++, CAnUSe with Phi-3 reaches over 81.7% of the performance of ArTS+RMTS and SaMRL. On TOEFL11, CAnUSe with Llama-3.1 retains 94.3% of the performance of DualBERT, approaching the performance of BERT and BigBird trained on 10K labeled essays. On Ellipse, CAnUSe with Llama-3.1 retains 90.6% of the performance of DualBERT, surpassing BERT and BigBird trained on 4K labeled essays.

Table 4 compares our frameworks, CAnUSe and EASY, with the concurrent LLM comparison-based essay scoring approach (LCES) and several LLM-based baselines in the overall scoring setting, following the setup used in prior work (Lee et al. 2024; Shibata and Miyamura 2025). On ASAP++, LCES shows stronger performance than CAnUSe, whereas on TOEFL11, CAnUSe outperforms LCES, achieving the best performance across most prompts. The primary distinction between our framework and LCES lies in the way pseudo-labels are utilized. LCES applies chain-of-thought prompting and further trains a scorer model to learn pairwise preferences from pseudo-labeled essay pairs, whereas CAnUSe applies simple prompting and directly trains on high-confidence pseudo-labeled samples identified through an uncertainty metric. Section 6.1 analyzes the performance dif-

Table 2 Average QWK scores (×100) across prompts per trait on ASAP++; SD denotes the averaged standard deviation for five seeds

Method	Ovrl	Cnt	PA	Lng	Nar	Org	Cnv	WC	SF	Sty.	Voi.	Avg.↑ (SD↓)
Zero-shot Models												
`microsoft/Phi-3-mini-4k-instruct`												
LLM-abs-Z	41.6	36.6	29.9	45.8	29.7	31.0	23.4	42.2	43.5	34.8	30.6	35.4 (±0.7)
LLM-comp-Z	54.6	59.3	**64.7**	60.1	64.8	48.6	46.2	48.0	49.1	42.8	29.1	51.6 (±0.9)
CAnUSe-Z	**62.8**	**61.6**	63.8	61.0	65.0	**56.8**	**55.8**	55.6	59.6	**54.4**	**40.3**	**57.9** (±1.6)
`meta-llama/Llama-3.1-8B-Instruct`												
LLM-abs-Z	25.2	41.4	25.0	45.4	25.7	14.1	13.3	35.3	38.4	8.0	13.7	25.9 (±0.5)
LLM-comp-Z	49.1	54.6	60.1	59.5	62.6	45.2	41.9	45.9	50.2	39.5	24.1	48.4 (±1.1)
CAnUSe-Z	58.0	59.3	64.3	**62.8**	**66.3**	51.7	51.4	**56.0**	**60.2**	40.2	28.8	54.5 (±1.0)
`Zero-shot Feature-based Model`												
EASY-Z	55.8	58.5	60.6	58.1	61.2	50.6	49.9	52.2	54.4	43.9	31.8	52.5/55.7[a]
Supervised Models (fine-tuned on ≈10K labels)												
ArTS+RMTS	75.5	73.7	75.2	71.3	74.4	68.2	69.0	70.5	69.4	70.2	61.2	70.8 (±4.3)
SaMRL	75.4	73.5	75.1	70.3	72.8	68.2	68.5	68.8	69.1	71.0	62.7	70.5 (±1.3)
MOOSE[b]	65.0	65.1	64.9	62.4	66.5	65.2	60.4	63.4	64.3	–	–	64.1[a] (±1.8)
1.5K.Feats.ft[b]	69.8	59.2	61.7	55.6	63.7	47.8	43.9	45.9	45.2	–	–	54.8[a] (–)

Ovrl, Cnt, Org, WC, SF, Cnv, Nar, PA, and Lng denote the traits overall, content, organization, word choice, sentence fluency, conventions, narrativity, prompt adherence, and language, respectively. Bold and underlined text indicate the best and second-best performances among the zero-shot methods, respectively. [a]Denotes the average from Ovrl to SF. [b]Denotes cross-prompt methods, where '–' indicates unreported performance. The suffix 'Z' denotes the zero-shot setting. The results for Qwen2.5 and Mistral are provided in Appendix B

ferences between LCES and CAnUSe. We expect that combining the strengths of LCES with our uncertainty-based selection strategy could yield further improvements, which we leave for future work.

Additionally, EASY-Z shows strong performance, outperforming two LLM-based methods: Vanilla-CoT and MTS. This result suggests that even a few simple and intuitive features can provide a strong inductive bias for essay scoring.

5.2 Few-shot setting

Tables 5 and 6 present QWK results on ASAP++, TOEFL11, and Ellipse in few-shot settings, where the suffix of each method denotes the number of labeled examples used. Figure 3 shows the performance curves with respect to the number of labels for the overall trait. In this experiment, we use Llama-3.1 as the base LLM because it showed consistently high performance across datasets in Sect. 5.1.

LLM-based Methods: Interestingly, LLM-comp performs worse in the 5-shot setting than in the zero-shot setting across all datasets (see Fig. 3). We speculate that the long context in LLM-comp makes it harder for the model to identify the comparison target. In contrast, LLM-abs improves in both 5-shot and 20-shot settings on ASAP++, with greater gains at 20-shot, while on TOEFL11 and Ellipse it improves in both but achieves its best performance at 5-shot. We conjecture that a small number of examples provides a useful reference point for absolute assessment, whereas excessive examples can again degrade performance in certain datasets.

CAnUSe: CAnUSe benefits from additional labels, achieving further improvements with 20 labeled examples on ASAP++ and Ellipse. CAnUSe operates by using LLM-comp to generate pseudo-labels from unlabeled data and applying uncertainty-aware self-training to train DualBERT as its scoring component. While DualBERT tends to overfit in low-resource settings when trained alone, it demonstrates strong generalization when integrated into the CAnUSe framework (DualBERT-20 vs. CAnUSe-20), which provides sufficient pseudo-labels to support learning deeper features. As a result, CAnUSe consistently improves upon DualBERT across all datasets.

EASY: EASY-20 uses 20 labeled examples to tune the hyperparameter α in Sect. 3.2, while EASY-Z relies on prior knowledge for α without using labeled examples. As shown in Fig. 3, EASY improves with 20 examples on ASAP++, but decreases on TOEFL11 and Ellipse. This result reflects that EASY mainly captures surface-level features, which are sensitive to limited supervision. As a result, few-shot examples may also lead to suboptimal tuning.

6 Analysis

6.1 Understanding the performance gap: CAnUSe vs. LCES

In this section, we analyze the factors contributing to the performance differences between CAnUSe and LCES. In particular, we focus on how differences

Table 3 Average QWK scores (×100) across prompts per trait on TOEFL11 and Ellipse

Dataset	TOEFL11	Ellipse							
Method	Ovrl (SD↓)	Ovrl	Coh	Syn	Voc	Phr	Grm	Cnv	Avg.↑ (SD↓)
Zero-shot Models									
`microsoft/Phi-3-mini-4k-instruct`									
LLM-abs-Z	41.4 (-)	22.8	23.1	15.6	12.3	9.3	15.8	16.2	16.4 (-)
LLM-comp-Z	45.7 (-)	44.7	41.0	39.5	38.3	36.5	36.9	38.0	39.3 (-)
CAnUSe-Z	63.1 (±1.1)	**61.5**	55.9	57.6	**61.3**	**59.5**	52.4	54.5	57.5 (±0.7)
`meta-llama/Llama-3.1-8B-Instruct`									
LLM-abs-Z	39.5 (-)	14.1	24.4	16.3	12.8	7.9	13.7	16.1	15.0 (-)
LLM-comp-Z	57.4 (-)	52.6	50.5	49.3	49.1	48.4	46.9	50.4	49.6 (-)
CAnUSe-Z	**68.1** (±0.3)	60.0	**57.3**	**58.2**	57.8	58.7	**54.2**	**59.4**	**57.9** (±0.3)
Zero-shot Feature-based Model									
EASY-Z	57.5 (-)	49.2	40.3	44.0	38.3	44.4	45.4	43.9	43.6 (-)
Supervised Models (fine-tuned on 10K labels in TOEFL11 and 4K labels in Ellipse)									
BERT-base	72.8 (±0.6)	60.5	47.6	55.8	48.6	52.8	57.0	52.2	53.5 (±1.4)
BigBird-base	72.9 (±1.2)	62.0	49.9	55.2	49.4	55.4	58.2	54.8	55.0 (±0.9)
DualBERT	72.2 (±2.0)	71.7	58.8	61.8	63.0	64.5	64.4	63.4	63.9 (±0.9)

Ovrl, Coh, Syn, Voc, Phr, Grm, and Cnv denote the traits overall, cohesion, syntax, vocabulary, phraseology, grammar, and conventions, respectively

Table 4 QWK scores (×100) for the overall trait on each essay prompt in ASAP++ and TOEFL11 using zero-shot methods

Dataset	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
`Zero-shot Feature-based Model`										
	EASY-Z (ours)	49.1	56.9	54.8	64.1	64.5	57.7	58.4	41.0	55.8
`meta-llama/Llama-3.1-8B-Instruct`										
ASAP++	Vanilla-CoT	12.9	2.3	24.3	55.0	30.1	34.1	0.6	−4.2	19.4
	MTS	51.6	48.3	28.4	46.1	47.9	37.8	32.8	19.9	39.1
	LCES	**66.9**	59.9	**66.2**	65.1	**71.0**	**70.7**	**72.7**	**63.6**	**67.0**
	CAnUSe-Z (ours)	54.8	**62.2**	56.5	**72.7**	68.5	52.1	55.0	42.0	58.0
`Zero-shot Feature-based Model`										
	EASY-Z (ours)	49.5	**62.0**	64.6	63.6	48.9	53.7	52.0	65.9	57.5
`meta-llama/Llama-3.1-8B-Instruct`										
TOEFL11	Vanilla-CoT	−3.6	14.8	0.3	2.1	1.9	−2.3	−2.9	6.3	2.1
	MTS	36.8	40.8	40.7	31.1	35.1	28.5	33.5	37.9	35.6
	LCES	59.7	57.0	72.7	69.7	65.2	55.0	55.8	**71.7**	63.3
	CAnUSe-Z (ours)	**66.8**	59.4	**77.6**	**75.8**	**67.2**	**66.0**	**60.1**	**71.7**	**68.1**

P1–P8 denotes Prompts 1 through 8. Bold and underlined text indicate the best and second-best performances, respectively

in LLM-based comparative assessments and dataset characteristics affect the final performance.

Table 5 Average QWK scores (×100) across prompts per trait on ASAP++ for few-shot methods

Method	Ovrl	Cnt	PA	Lng	Nar	Org	Cnv	WC	SF	Sty.	Voi.	Avg.↑(SD↓)
No Fine-tuning Methods: In-context Learning and Feature-based Methods.												
LLM-comp-5	47.5	51.3	61.2	56.8	62.3	28.6	31.8	35.6	36.0	43.4	2.8	41.6 (±2.9)
LLM-abs-5	56.0	58.4	45.9	59.9	54.6	52.0	44.7	59.3	62.5	45.9	43.8	53.0 (±1.3)
LLM-abs-20	**66.1**	**64.3**	56.1	**70.4**	52.1	56.1	51.2	**65.1**	**69.0**	56.0	**56.6**	**60.3** (±0.7)
EASY-20	62.0	58.1	58.2	57.5	60.6	53.6	51.7	53.9	54.6	52.6	34.1	54.3 (±1.5)
Fine-tuning Methods												
CAnUSe-20	64.3	63.4	**65.9**	63.1	**66.8**	**59.8**	**59.0**	60.1	63.0	**57.2**	38.0	60.1 (±1.3)
DualBERT-20	53.6	43.7	42.8	38.7	38.9	47.6	44.6	47.1	49.0	40.4	49.6	45.1 (±2.0)

The suffix denotes the number of labels used

6.1.1 Effect of LLM judgment quality in comparative assessment

Since both CAnUSe and LCES rely on LLM-based comparative assessments as their primary supervision signal, the quality of the initial LLM judgments is critical to final performance, as noise in the pseudo-labels can propagate to subsequent models.

One source of performance difference between LCES and CAnUSe lies in their prompting strategies (Shibata and Miyamura 2025). CAnUSe uses concise one- or two-line criteria to define good and poor essays, whereas LCES employs detailed official rubric guidelines, which are substantially longer-often exceeding 200–300 words. In addition, LCES incorporates chain-of-thought (CoT) prompting, further differentiating its comparative assessment setup from ours.

Table 7 reports the mean absolute error (MAE) of the LLM-based comparative assessments and the final QWK obtained by training second-stage models (RankNet for LCES and DualBERT for CAnUSe). Given an essay pair (x_i, x_j), the LLM outputs one of three comparison judgments for x_i: win (1), tie (0.5), or lose (0). Let $\hat{c}_{ij}$ and c_{ij} denote the LLM's judgment and the ground-truth label, respectively. We evaluate the accuracy of the LLM's judgments using MAE, computed as $\mathrm{MAE} = \frac{1}{N}\sum_{(i,j)} |\hat{c}_{ij} - c_{ij}|$, where N is the number of essay pairs. Following Shibata and Miyamura (2025), we set $N = 5000$.

The results in Table 7 highlight the performance gap between CAnUSe and LCES. On ASAP++, LCES achieves lower MAE in the LLM comparison stage and correspondingly higher QWK in the final performance. In contrast, on TOEFL11, CAnUSe benefits from more accurate LLM judgments. Figure 4 further supports this observation by showing a clear association between MAE of LLM-based comparisons and final performance on ASAP++.

In Appendix C, we further analyze the effect of different prompting strategies used in LLM-based comparative assessment.

6.1.2 Effect of score range and distribution

The second key difference between CAnUSe and LCES lies in their learning approaches after the LLM-based comparative assessment stage: RankNet for LCES and DualBERT for CAnUSe. In CAnUSe, each essay is assigned a pseudo score based on its win rate in the comparative assessment stage. Since the win rate is inherently a relative, zero-sum measure-where each comparison produces both a winner and a loser-the resulting pseudo scores reflect relative ranking, leading to a more balanced score distribution.

Moreover, our training setup samples pseudo-labels in a balanced manner (see Sect. 3.1.2), further encouraging the model to produce balanced score distributions at inference time. In contrast, LCES employs RankNet trained on randomly sampled essay pairs to learn relative preference relations. Given a random split where the training and test sets share the same label distribution, such random pair sampling implicitly preserves this assumption; consequently, the model tends to

Table 6 Average QWK scores (×100) across prompts per trait on TOEFL11 and Ellipse for few-shot methods

Dataset	TOEFL11	Ellipse							
Method	Ovrl (SD↓)	Ovrl	Coh	Syn	Voc	Phr	Grm	Cnv	Avg.↑(SD↓)
No Fine-tuning Methods: In-context Learning and Feature-based Methods.									
LLM-comp-5	38.4 (-)	43.9	38.8	39.2	39.5	38.0	35.4	32.0	38.1 (-)
LLM-abs-5	57.9 (-)	47.7	41.8	43.6	43.4	42.4	41.8	41.8	43.2 (-)
LLM-abs-20	49.9 (-)	40.3	32.1	37.3	33.3	41.4	38.8	34.5	36.8 (-)
EASY-20	59.2 (-)	42.7	33.1	42.1	39.2	43.9	40.1	38.7	40.0 (-)
Fine-tuning Methods									
CAnUSe-20	**67.8** (±0.8)	**63.0**	**57.8**	**58.3**	**60.4**	**60.4**	**54.2**	**60.3**	**59.2** (±0.7)
DualBERT-20	14.1 (±16.0)	27.3	25.9	35.0	35.6	33.1	36.6	27.9	31.6 (±9.8)

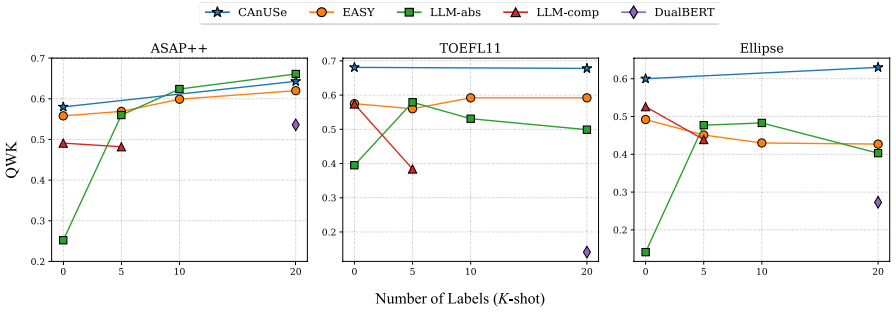

Fig. 3 Few-shot performance on ASAP++, TOEFL11, and Ellipse for the overall trait as a function of the number of labeled examples

Table 7 LLM comparative assessment quality (MAE, ×100) and final performance (QWK, ×100), averaged across prompts

	ASAP++		TOEFL11	
	LCES[a]	CAnUSe	LCES[a]	CAnUSe
QWK	64.6	58.0	60.8	68.1
MAE	24.5	26.5	24.5	23.2

[a]Denotes our reimplementation of LCES, as no code is provided in Shibata and Miyamura (2025), with minor differences due to inherent training randomness

reproduce the training data's distributional bias during inference, yielding score distributions similar to those of the test set.

For a fair comparison, we re-implemented LCES using the same LLM outputs as CAnUSe, thereby controlling for differences in first-stage supervision. Table 8 reports the comparison between CAnUSe and the re-implemented LCES under this setting. When the LLM signals are identical, LCES performance on ASAP++ substantially degrades compared to its original results (see Table 4), suggesting that LCES is sensitive to the quality of LLM-generated supervision. We further report

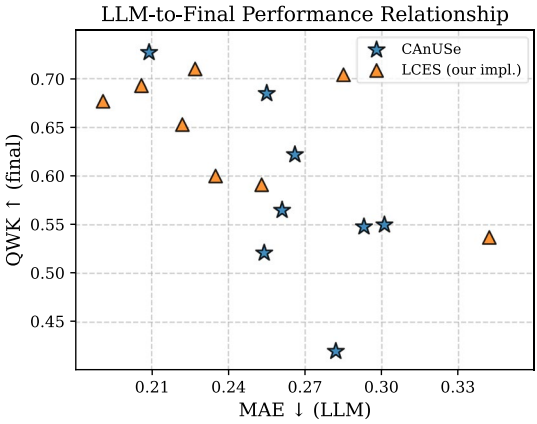

Fig. 4 Relationship between LLM comparison quality (MAE) and final performance (QWK) obtained from the subsequent models: each point corresponds to a prompt on ASAP++

Table 8 QWK scores and entropies (×100) of predicted score distributions for each prompt on ASAP++ and TOEFL11

Dataset	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
ASAP++	LCES w. our LLM	**60.8**	44.1	55.6	64.6	59.3	**57.6**	**59.3**	**57.0**	57.3
	↪*Entropy*	79.3	77.3	70.1	78.8	68.8	62.3	88.4	90.6	79.6
	CAnUSe	54.8	**62.2**	**56.5**	**72.7**	**68.5**	52.1	55.0	42.0	**58.0**
	↪*Entropy*	92.2	87.8	84.6	91.1	91.4	89.0	95.9	94.6	90.8
	Score Range (bins)	2–12 (11)	1–6 (6)	0–3 (4)	0–3 (4)	0–3 (4)	0–4 (5)	0–30 (31)	0–60 (61)	-
	True Entropy	72.7	64.1	84.1	95.0	84.9	82.6	82.9	70.1	77.0
TOEFL11	LCES w. our LLM	62.4	**63.4**	70.3	59.9	61.3	54.2	53.0	65.9	61.3
	↪*Entropy*	81.2	84.5	90.2	73.2	78.7	82.1	80.3	87.4	82.2
	CAnUSe	**66.8**	59.4	**77.6**	**75.8**	**67.2**	**66.0**	**60.1**	**71.7**	**68.1**
	↪*Entropy*	95.4	88.2	97.3	91.9	89.3	83.9	91.9	95.2	91.6
	Score Range (bins)	1–5 (5) ⟶ {*low, medium, high*} (3)								
	True Entropy	81.2	82.4	91.8	83.9	85.2	84.0	83.1	89.0	85.1

LCES w. our LLM denotes the re-implemented LCES trained using the same LLM outputs as CAnUSe

the entropy of the predicted score distributions for analysis. Overall, CAnUSe consistently exhibits higher entropy than LCES, indicating that it produces more balanced score predictions for each prompt. In contrast, LCES tends to concentrate predictions within a narrower range of scores. Notably, on ASAP++, the prompts where LCES outperforms CAnUSe generally have wider official rubric score ranges, i.e., a larger number of score bins (e.g., prompts P1, P7, and P8).

Figure 5 (left) shows two representative cases on ASAP++, P2 and P8, where CAnUSe respectively outperforms and underperforms LCES. Although the ground-truth score distributions for both prompts are heavily concentrated around the middle score range, P2 has only six score bins, whereas P8 has 61. In both cases, CAnUSe produces more balanced predictions across score bins, while LCES

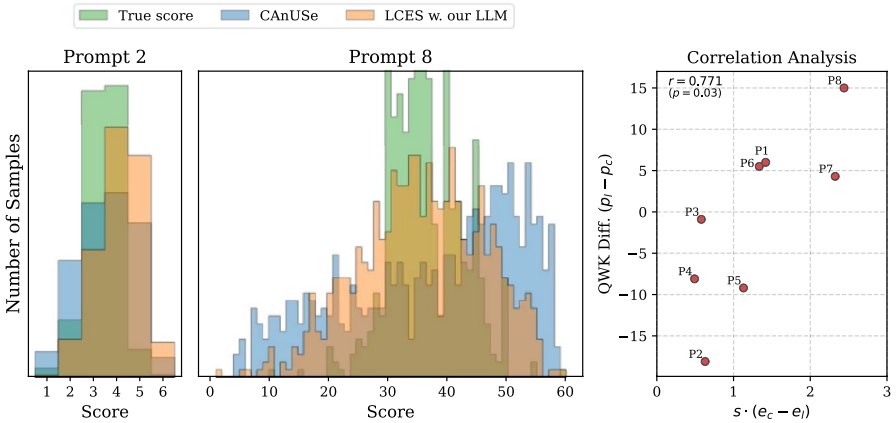

Fig. 5 (Left) Two representative examples of predicted score distributions on ASAP++ prompts. (Right) Relationship between score range–weighted distributional differences and QWK performance differences between CAnUSe and LCES

Table 9 Average QWK scores (×100) across traits on ASAP++, TOEFL11, and Ellipse for different uncertainty sampling strategies

Dataset	ASAP++			TOEFL11			Ellipse		
p%least uncertain	*10*	*30*	*50*	*10*	*30*	*50*	*10*	*30*	*50*
LLM-comp *w/* bal	64.6	59.8	56.0	72.2	61.4	60.7	63.5	**59.2**	**55.6**
+ CAnUSe-Z	**54.5**	**56.1**	**55.8**	**68.1**	**67.4**	**67.3**	**57.9**	**57.9**	**57.6**
LLM-comp *w/o* bal	**72.5**	**64.8**	**59.4**	**85.0**	**75.6**	**67.6**	**64.6**	58.5	55.5
+ CAnUSe-Z	45.7	50.4	52.1	57.9	59.2	61.5	39.1	46.7	51.5

"bal." denotes label-balanced sampling during self-training. Bold text indicates the larger score between the *w/* bal. and *w/o* bal. settings

yields predictions similar to the ground-truth distribution by concentrating in the mid-score region. This behavior stems from LCES's implicit assumption that the training and test distributions are the same. While this assumption can be advantageous when the two distributions are similar, it becomes a weakness under distribution shift.

Accordingly, LCES appears to perform better in settings with many score bins and a skewed true score distribution. Figure 5 (right) supports this tendency by showing a strong correlation between $s \cdot (e_c - e_l)$ and $p_l - p_c$, where s denotes the number of score bins, e_c and e_l denote the entropies of the predicted score distributions of CAnUSe and LCES, respectively, and p_c and p_l denote their QWK performances.

This analysis explains why CAnUSe outperforms LCES on TOEFL11 in most cases. In TOEFL11, all prompts have a score range of 1–5, and the scores are further mapped into three categories (low, medium, high) through post-processing. In this setting, performance is less influenced by learned distributional bias and instead depends more on instance-level inference.

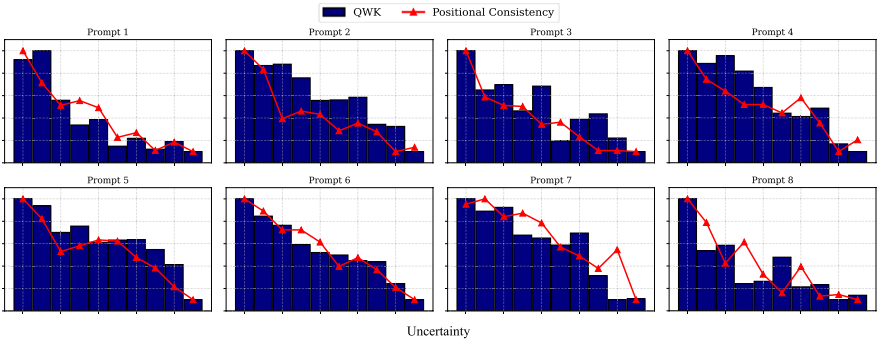

Fig. 6 QWK performance by uncertainty levels on ASAP++ for the overall trait across Prompts 1–8. QWK scores and positional consistency are normalized for clarity

Table 10 Average positional consistency (×100) for LLM-comp-Z and LLM-comp-5

	ASAP++	TOEFL11	Ellipse
LLM-comp-Z	**73.8**	**72.9**	**70.6**
LLM-comp-5	49.7	34.1	34.4

6.2 Effect of uncertainty estimation

Uncertainty-based sampling for high-confidence labels is the main factor driving the performance gains of CAnUSe over LLM-comp. This section examines the role of uncertainty estimation using Llama-3.1 as the base LLM. Table 9 reports the average QWK results across traits on ASAP++, TOEFL11, and Ellipse when selecting the $p\%$ least uncertain samples from the outputs of LLM-comp, and Fig. 6 illustrates performance trends with respect to uncertainty on ASAP++. Overall, the uncertainty metric aligns well with prediction reliability: low-uncertainty samples yield higher performance, while high-uncertainty samples yield lower performance.

Label balancing and **sampling ratio**: Table 9 also presents the results of sampling strategies with and without balancing label distributions. Interestingly, for low-uncertainty sample sets, QWK scores are generally higher when sampling without balancing than with balancing across datasets. However, in CAnUSe, which is fine-tuned on these samples, QWK scores are consistently higher under the balancing strategy across all datasets. This result suggests that label distribution in pseudo-labeled samples plays an important role in effective self-training. Additionally, in LLM-comp *w/* bal., $p = 10$, which uses fewer pseudo-labels, yields performance comparable to $p = 50$ across all datasets, indicating the effectiveness of the uncertainty metric.

Positional consistency: We further measure positional consistency for each uncertainty bin in Fig. 6. Positional consistency refers to the proportion of cases in LLM-comp where swapping the input order of two essays leads to a corresponding reversal in the output (i.e., from A to B or B to A). As shown in Fig. 6, positional consistency aligns closely with uncertainty and QWK. This result implies that positional consistency can serve as an early indicator of LLM-comp performance in zero-shot settings, where no labels are available for validation. This property

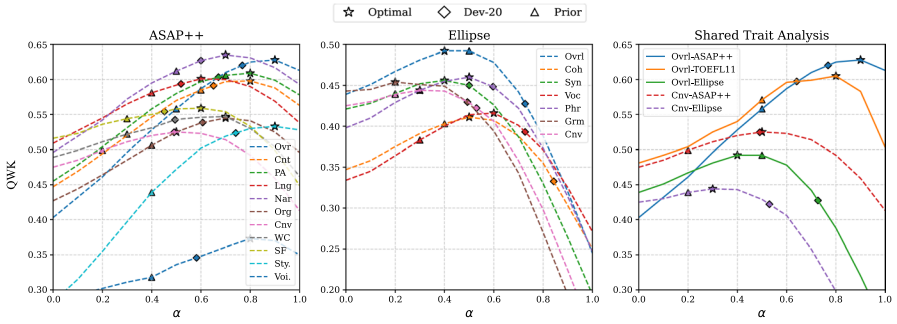

Fig. 7 QWK performance of EASY on ASAP++ and Ellipse for all traits, and on shared traits across datasets (including TOEFL11), as a function of the hyperparameter α

of LLM-comp also explains why LLM-comp-5 in Sect. 5.2 performs worse than LLM-comp-Z. Table 10 reports the average positional consistency for LLM-comp-Z and LLM-comp-5.

6.3 Effect of α in EASY

EASY has a single hyperparameter, α, which controls the relative weight of $Score_1$ and $Score_2$. Figure 7 shows performance curves with respect to α for all traits on ASAP++, Ellipse, and TOEFL11. The optimal α is denoted by a star, the α selected using 20 validation examples by a diamond, and the α based on prior knowledge by a triangle. As shown in Fig. 7, on ASAP++, a few validation examples helped identify a better α than prior knowledge, whereas this was mostly not the case on Ellipse. Indeed, a small number of examples may lead to overfitted hyperparameters.

The optimal α varies substantially across datasets: on ASAP++, most values range between 0.5 and 0.9, while on Ellipse they fall between 0.2 and 0.6. Although the datasets share the same traits, the optimal hyperparameters can differ due to distinct scoring rubrics. This result suggests that while EASY can provide a useful inductive bias, it has limitations in adapting to dataset-specific characteristics.

Additionally, the interpretation of α in EASY is consistent with its intended meaning: lower α indicates that grammar-related aspects carry more weight, aligning with traits such as grammar or conventions, whereas higher α indicates that content aspects are more important, corresponding to traits such as content or overall. For instance, on ASAP++, the optimal α values for overall and content traits are 0.9 and 0.8, respectively, while conventions is 0.5. Similarly, on Ellipse, the optimal α values for grammar and conventions are 0.2 and 0.3, which are lower than that of the overall trait (0.4).

6.4 Effect of LLM size

Figure 8 shows QWK scores for the overall trait on ASAP++ as a function of LLM size. In general, larger LLMs are expected to exhibit stronger inference ability. However, in LLM-comp, increased model size does not consistently lead to better

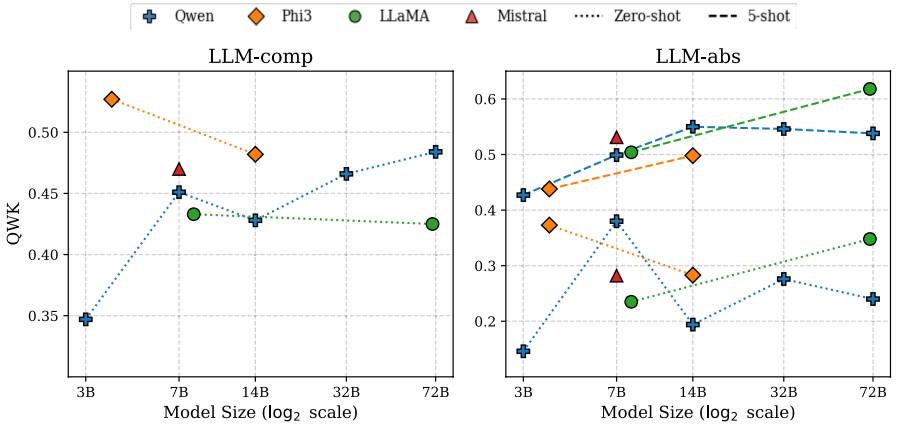

Fig. 8 QWK scores for the overall trait on ASAP++ as a function of LLM size. Due to the substantial computational overhead of large models, we used 100 samples per prompt in this experiment

performance, whereas in LLM-abs-5, larger models typically yield improvements. Moreover, large LLM-abs-5 models often outperform LLM-comp-Z.

We hypothesize that comparative assessment reduces the difficulty of absolute assessment, which requires aligning an essay with appropriate rubrics, by reformulating it into a binary classification task (choosing A or B). Consequently, smaller LLMs may already be sufficiently pretrained to handle comparative assessment effectively.

Computational cost: The most computationally expensive component of CAnUSe is the LLM-based comparative assessment stage. To evaluate N essays, this stage requires $O(N \times M)$ LLM calls, where M denotes the number of comparisons per essay. This cost is higher than that of LLM-based *absolute* assessment, which scales as $O(N)$. In our experiments, evaluating 100 essays took an average of 3.52 min using approximately 30 GB of VRAM on a single H100 GPU with a batch size of 16, 16×2 comparisons per essay (via pairwise swapping), and an average input length of approximately 800 tokens. The computation time increases linearly with the number of input tokens; for example, for 100 essays, each additional 100 tokens increases the runtime by about 0.44 min.

Importantly, this comparative assessment is performed only once during training. Afterward, CAnUSe trains a single-essay scoring model, enabling real-time essay scoring with $O(N)$ complexity; with a batch size of 16, it can evaluate 100 essays in under one second using approximately 4 GB of VRAM.

7 Key insights for zero-shot and few-shot AES

In this section, we summarize the key empirical insights derived from our experiments on zero-shot and few-shot AES.

Insight 1: In LLM-based zero-shot AES, comparative assessment (LLM-comp) is generally more accurate than absolute assessment (LLM-abs) (Sect. 5.1). In a

zero-shot setting, absolute assessment is difficult without concrete examples and requires aligning essays with score-specific rubrics, whereas comparative assessment simplifies the task to a binary choice of which essay is better. However, in practical settings where absolute scores are required, comparative assessment should be incorporated into improved frameworks.

Insight 2: When few labeled examples are provided, LLM-comp can paradoxically suffer from performance degradation (Sect. 5.2). We speculate that the added examples and the resulting longer input context can obscure the comparison target in LLM-comp. In contrast, LLM-abs generally benefits from few-shot supervision, as the examples serve as concrete references for evaluation, leading to improved performance.

Insight 3: In zero-shot settings, the size of LLMs has a limited impact on performance for both comparative and absolute assessment. However, in few-shot absolute assessment, larger models exhibit clear performance gains (Sect. 6.4). This suggests that LLM-abs can effectively leverage increased model capacity when few-shot examples are provided, as these examples serve as concrete references for evaluation.

Insight 4: Uncertainty estimation for LLM predictions is effective for extracting reliable pseudo-labels (Sect. 6.2). In CAnUSe, the win-rate of each essay derived from LLM-comp is treated as the target score, and low-uncertainty samples are used to self-train a single-essay scoring model, DualBERT, resulting in faster inference and improved accuracy compared to LLM-comp, as well as better generalization than DualBERT trained alone.

Insight 5: In comparative assessment, positional consistency shows a strong correlation with uncertainty (Sect. 6.2). This metric, ranging from 0 to 1, reflects the stability of LLM predictions with respect to input order. Positional consistency can serve as a simple yet informative diagnostic for assessing LLM reliability in unlabeled zero-shot settings.

Insight 6: Even simple surface-level features-such as essay length, vocabulary diversity, and grammar quality-can be fairly effective in zero-shot AES (Sect. 5.1). When properly combined, these features can achieve performance comparable to complex feature-based models, highlighting the practical value of lightweight and interpretable approaches.

8 Conclusion

Building large-scale essay datasets with reliable annotations remains challenging in real-world educational settings, underscoring the need for zero-shot AES methods. However, prior LLM-based approaches have been limited, often relying solely on prompting strategies, focusing only on overall scoring, and depending heavily on the ASAP++ dataset.

In this paper, we explored various zero-shot AES methods in a multi-trait setting and proposed two new frameworks: CAnUSe and EASY. CAnUSe performs comparative assessment with an LLM to generate pseudo-labels and then leverages an uncertainty metric to select high-confidence samples for training a scorer model.

EASY, in contrast, relies only on three simple features-essay length, vocabulary size, and grammar score-yet provides a strong inductive bias.

Experiments on ASAP++, TOEFL11, and Ellipse show that CAnUSe achieves state-of-the-art zero-shot results, approaching the performance of fully supervised models. EASY, despite using no labeled data and only three features, demonstrates competitive results and in some cases even surpasses LLM-based methods.

Nonetheless, CAnUSe has limitations: it requires considerable computation during the comparative assessment stage to obtain pseudo-labels and relies on a large number of unlabeled samples. For future work, we plan to reduce inference costs by adopting Elo-style matching for comparisons and decreasing reliance on unlabeled data by integrating comparative and absolute assessment strategies.

Appendix A: prompt examples

Figures 9 and 10 present prompt examples for the overall trait in LLM-comp and LLM-abs across datasets, respectively. Full prompt examples for all traits are available at https://github.com/hongking9/exploring_zero_shot_aes. Figure 11 presents examples of few-shot prompting in the 3-shot setting (Fig. 10).

<ASAP++>

```
### Instruction: Which of these two essays is better written "overall"? Just pick one. Consider the following rubrics in your evaluation:
<Rubric for Evaluating Overall Quality>
  - Good: The essay presents its ideas clearly and stays focused on its purpose. The structure is logical, the tone is consistent, and the writing reads smoothly overall.
  - Poor: The essay is unclear or unfocused, with weak structure or inconsistent tone. The writing feels disjointed and fails to effectively deliver its message.
### Answer format: A or B.
### Prompt: {essay_prompt}
{few_shot_examples}

### Essay A: {essay_a}
### Essay B: {essay_b}
### Answer: Essay
```

<TOEFL11>

```
### Instruction: Which of these two essays is better written "holistic"? Just pick one. Consider the following rubrics in your evaluation:
<Rubric for Evaluating Holistic>
  - Good: The essay effectively addresses the topic with clear organization, detailed support, logical flow, and consistent language use, showing a strong grasp of vocabulary, syntax, and coherence despite minor errors.
  - Poor: The essay is poorly organized, underdeveloped, or off-topic, with minimal support or clarity, and marked by frequent language errors that hinder communication and understanding.
### Answer format: A or B.
### Prompt: {essay_prompt}
{few_shot_examples}

### Essay A: {essay_a}
### Essay B: {essay_b}
### Answer: Essay
```

<Ellipse>

```
### Instruction: Which of these two essays is better written "overall"? Just pick one. Consider the following rubrics in your evaluation:
<Rubric for Evaluating Overall Quality>
  - Good: The essay demonstrates strong command of language with varied sentence structures, precise vocabulary, and well-organized ideas; grammar and usage are mostly accurate, with only minor errors that do not hinder understanding.
  - Poor: The essay shows limited or inconsistent control of language, frequent grammatical errors, weak organization, and vocabulary that often fails to convey meaning, resulting in communication breakdowns.
### Answer format: A or B.
### Prompt: {essay_prompt}
{few_shot_examples}

### Essay A: {essay_a}
### Essay B: {essay_b}
### Answer: Essay
```

Fig. 9 Prompt examples for the overall trait in LLM-comp across datasets

<ASAP++>

```
### Instruction: You are an English teacher. Evaluate the following essay based on the rubric provided below and assign an overall score from 1 to 5.
<Rubric for Evaluating Overall Quality>
- 5 (Excellent): The essay is exceptionally clear, well-structured, and highly engaging throughout.
- 4 (Good): The essay is mostly clear and focused, with a logical structure and consistent tone.
- 3 (Fair): The essay has a generally clear purpose but shows some issues in organization or clarity.
- 2 (Poor): The essay is often unclear or disorganized, with noticeable problems in tone and flow.
- 1 (Very Poor): The essay lacks clarity and structure, making the message difficult to understand.
### Answer format: score (score range = [1, 2, 3, 4, 5]). Do not include explanations or comments.
### Prompt: {essay_prompt}
{few_shot_examples}

### Essay: {essay}
### Answer:
```

<Ellipse>

```
### Instruction: You are an English teacher. Evaluate the following essay based on the rubric provided below and assign an overall score from 1 to 5.
<Rubric for Evaluating Overall Quality>
- 5: Native-like facility in the use of language with syntactic variety, appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication.
- 4: Facility in the use of language with syntactic variety and range of words and phrases; controlled organization; accuracy in grammar and conventions; occasional language inaccuracies that rarely impede communication.
- 3: Facility limited to the use of common structures and generic vocabulary; organization generally controlled although connection sometimes absent or unsuccessful; errors in grammar and syntax and usage. Communication is impeded by language inaccuracies in some cases.
- 2: Inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or unsuccessful. Communication impeded in many instances by language inaccuracies.
- 1: A limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication impeded in most cases by language inaccuracies.
### Answer format: score (score range = [1, 2, 3, 4, 5]). Do not include explanations or comments.
### Prompt: {essay_prompt}
{few_shot_examples}

### Essay: {essay}
### Answer:
```

<TOEFL11>

```
### Instruction: You are an English teacher. Evaluate the following essay based on the rubric provided below and assign an overall score from 1 to 5.
<Rubric for Evaluating Overall Quality>
- 5: An essay at this level largely accomplishes all of the following:
· effectively addresses the topic and task
. is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details
· displays unity, progression, and coherence
· displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors

- 4: An essay at this level largely accomplishes all of the following:
· addresses the topic and task well, though some points may not be fully elaborated
· is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details
· displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections
· displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning

- 3: An essay at this level is marked by one or more of the following:
· addresses the topic and task using somewhat developed explanations, exemplifications, and/or details
· displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
. may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning
· may display accurate but limited range of syntactic structures and vocabulary

- 2: An essay at this level may reveal one or more of the following weaknesses:
· limited development in response to the topic and task
· inadequate organization or connection of ideas
· inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task
· a noticeably inappropriate choice of words or word forms
· an accumulation of errors in sentence structure and/or usage

- 1: An essay at this level is seriously flawed by one or more of the following weaknesses:
· serious disorganization or underdevelopment
· little or no detail, or irrelevant specifics, or questionable responsiveness to the task
· serious and frequent errors in sentence structure or usage
### Answer format: score (score range = [1, 2, 3, 4, 5]). Do not include explanations or comments.
### Prompt: {essay_prompt}
{few_shot_examples}

### Essay: {essay}
### Answer:
```

Fig. 10 Prompt examples for the overall trait in LLM-abs across datasets

Fig. 11 Examples of few-shot prompting in the 3-shot setting

<LLM-comp>

```
### Essay A: {essay_a1}
### Essay B: {essay_b1}
### Answer: Tie.

### Essay A: {essay_a2}
### Essay B: {essay_b2}
### Answer: Essay A.

### Essay A: {essay_a3}
### Essay B: {essay_b3}
### Answer: Essay B.
```

<LLM-abs>

```
### Essay: {essay_1}
### Answer: 3.

### Essay: {essay_2}
### Answer: 2.

### Essay: {essay_3}
### Answer: 5.
```

Appendix B: zero-shot performance of Qwen2.5 and Mistral

Tables 11 and 12 present QWK results on ASAP++, TOEFL11, and Ellipse for Qwen2.5 and Mistral.

Appendix C: effect of prompting variations in comparative assessment

For prompting, LCES employs official rubric guidelines originally designed for absolute scoring, such as Likert-type scales Shibata and Miyamura (2025). Figure 12 shows an example of the prompt used in LCES. These official rubric guidelines often exceed 200–300 words, with some prompts reaching over 1,000 words; we therefore also evaluate a shortened version of approximately 100 words per prompt, as long contexts may hinder LLM performance.

Using `meta-llama/Llama-3.1-8B-Instruct`, we evaluate three variants of the prompting scheme used in LCES: Prompt Style 1 (PS-1; full rubric guidelines (RGs) with chain-of-thought (CoT)), which corresponds to the original setting Shibata and Miyamura (2025), Prompt Style 2 (PS-2; full RGs without CoT), and Prompt Style 3 (PS-3; shortened RGs with CoT). We refer to the prompting scheme used in CAnUSe as Good/Poor-based prompting (PS-4), which does not include CoT; examples are shown in Fig. 9.

As shown in Table 13, we do not observe significant performance differences among the LCES prompt variants (PS-1, PS-2, and PS-3), suggesting that the presence of CoT and the length of rubric guidelines have limited impact on performance in this setting. In contrast, PS-4 exhibits slightly different behavior: CAnUSe's Good/Poor-based prompting performs slightly better on TOEFL11 but slightly worse than the LCES-style prompting variants (PS-1–3) on ASAP++. These results suggest that the style of the prompting template plays a more dominant role than the presence of CoT or the length of rubric guidelines.

Table 11 Average QWK scores (×100) across prompts per trait on ASAP++ for Qwen2.5 and Mistral

Method	Ovrl	Cnt	PA	Lng	Nar	Org	Cnv	WC	SF	Sty.	Voi.	Avg.↑(SD↓)
Zero-shot Models												
`Qwen/Qwen2.5-7B-Instruct`												
LLM-abs	35.9	42.8	28.6	53.1	26.9	17.5	22.2	46.1	52.7	10.1	39.8	34.1 (±1.4)
LLM-comp	46.5	53.0	61.2	56.8	61.3	39.9	37.8	41.6	41.7	34.2	22.9	45.2 (±1.2)
CAnUSe-Z	56.9	59.2	64.9	61.9	66.7	49.3	51.1	53.0	56.3	42.2	36.2	54.3 (±1.6)
`mistralai/Mistral-7B-Instruct-v0.3`												
LLM-abs	36.3	36.8	39.0	37.3	41.0	40.9	34.6	32.1	29.6	36.3	25.6	35.4 (±1.1)
LLM-comp	50.9	56.8	62.1	57.4	61.9	45.7	43.8	48.6	48.2	42.5	26.3	49.5 (±1.2)
CAnUSe-Z	62.2	61.3	62.5	60.2	64.0	55.8	55.6	55.6	59.1	55.6	40.4	57.5 (±1.2)

Table 12 Average QWK scores (×100) across prompts per trait on TOEFL11 and Ellipse for Qwen2.5 and Mistral

Dataset	TOEFL11	Ellipse							
Method	Ovrl (SD↓)	Ovrl	Coh	Syn	Voc	Phr	Grm	Cnv	Avg.↑(SD↓)
Zero-shot Models									
`Qwen/Qwen2.5-7B-Instruct`									
LLM-abs-Z	44.4 (-)	38.8	38.4	38.2	35.6	33.4	35.9	38.0	36.9 (-)
LLM-comp-Z	48.2 (-)	45.2	46.5	42.8	41.8	42.2	43.1	42.1	43.4 (-)
CAnUSe-Z	65.9 (±1.2)	59 0	55.7	55.9	55.4	58.1	54.2	57.9	56.6 (±0.3)
`mistralai/Mistral-7B-Instruct-v0.3`									
LLM-abs-Z	30.4 (-)	21.9	27.4	29.5	28.1	23.9	22.5	19.7	24.7 (-)
LLM-comp-Z	40.9 (-)	43.2	40.0	38.4	40.2	36.4	36.4	38.4	39.0 (-)
CAnUSe-Z	57.7 (±1.6)	61.3	55.8	56.9	56.5	57.1	53.0	57.2	56.8 (±0.5)

```
# Instruction:
Read the following two essays and evaluate them based on rubric
guidelines. Then, indicate which essay is better overall. If both essays
are judged to be of the same score, evaluate them as “tie”
# Prompt: Write a letter to your local newspaper in which you state
your opinion on the effects computers have on people. Persuade the
readers to agree with you.
# Rubric Guidelines:
Score Point 1: An undeveloped response …
Score Point 2: An under-developed response …
…
# Essay 1: {essay_1}
# Essay 2: {essay_2}
Provide your reasoning and final decision.
Reasoning: (Your reasoning here)
Decision: (Either "Essay 1", "Essay 2", or "tie")
```

Fig. 12 An example of the prompt used in Shibata and Miyamura (2025)

Table 13 MAE↓ ($\times 100$) results for different prompting strategies in LLM-based comparative assessment

Dataset	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
ASAP++	PS-1: Full RGs with CoT	22.7	25.3	23.5	19.1	20.6	22.2	28.5	34.2	24.5
	PS-2: Full RGs without CoT	25.1	25.9	24.8	20.4	22.2	23.1	27.2	25.5	24.3
	PS-3: Shortened RGs with CoT	23.7	25.0	23.7	18.8	21.2	22.1	29.9	34.4	24.8
	PS-4: Good/Poor-based (our setting)	29.3	26.6	26.1	20.9	25.5	25.4	30.1	28.2	26.5
TOEFL11	PS-1: Full RGs with CoT	24.3	25.6	24.1	23.7	25.0	23.4	24.3	25.9	24.5
	PS-2: Full RGs without CoT	23.2	25.4	23.4	22.9	24.7	23.1	24.7	25.6	24.1
	PS-3: Shortened RGs with CoT	23.2	25.7	23.8	22.0	25.7	23.0	23.4	24.5	23.9
	PS-4: Good/Poor-based (our setting)	22.6	24.2	23.1	22.5	23.7	22.9	22.5	24.0	23.2

MAE is computed as described in Sect. 6.1.1

Acknowledgements This work was supported by Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

Author contributions H.C. curated the data, contributed to the conceptualization and methodology of the study, implemented the software, and wrote the original draft of the manuscript. M.-C.K. and J.S. carried out the investigation and contributed to the review and editing of the manuscript. J.-X.H. supervised the project, managed its administration, acquired funding, and also participated in the review and editing of the manuscript.

Data availability The datasets analysed during the current study are publicly available. The ASAP++ dataset is available via Mathias and Bhattacharyya (2018), the TOEFL11 dataset is available via Blanchard et al. (2013), and the Ellipse dataset is available via Crossley et al. (2023).

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Almusharraf N, Alotaibi H (2023) An error-analysis study from an efl writing context: human and automated essay scoring approaches. Technol Knowl Learn 28(3):1015–1031

Askarbekuly N, Aničić N (2024) Llm examiner: automating assessment in informal self-directed e-learning using chatgpt. Knowl Inf Syst 66(10):6133–6150

Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on machine learning, pp 89–96

Blanchard D, Tetreault JR, Higgins D, Cahill A, Chodorow M (2013) Toefl11: a corpus of non-native English. ETS Res Rep Ser 2013:15

Cavalcanti AP, Barbosa A, Carvalho R, Freitas F, Tsai Y-S, Gašević D, Mello RF (2021) Automatic feedback in online learning environments: a systematic literature review. Comput Educ Artif Intell 2:100027

Cho M, Huang J-X, Kwon O-W (2024) Dual-scale bert using multi-trait representations for holistic and trait-specific essay grading. ETRI J 46(1):82–95

Chu S, Kim J, Wong B, Yi M (2024) Rationale behind essay scores: enhancing s-llm's multi-trait essay scoring with rationale generated by llms. arXiv preprint arXiv:2410.14202

Cai K, Kong L, Zhou J, Liang D, Qu W (2025) Exploring structure-aware representation learning for automated essay scoring. Knowl Inf Syst 1–25

Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213

Crossley S, Tian Y, Baffour P, Franklin A, Kim Y, Morris W, Benner M, Picou A, Boser U (2023) The English language learner insight, proficiency and skills evaluation (ellipse) corpus. Int J Learn Corpus Res 9(2):248–269

Chen P-K, Tsai B-W, Wei SK, Wang C-Y, Wang J-C, Huang Y-T (2025) Mixture of ordered scoring experts for cross-prompt essay trait scoring. In: Che W, Nabende J, Shutova E, Pilehvar MT (eds) Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers), pp 18071–18084. Association for Computational Linguistics, Vienna, Austria. https://doi.org/10.18653/v1/2025.acl-long.884 . https://aclanthology.org/2025.acl-long.884/

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423 . https://aclanthology.org/N19-1423

Do H, Kim Y, Lee GG (2023) Prompt- and trait relation-aware cross-prompt essay trait scoring. In: findings of the association for computational linguistics: ACL 2023, pp 1538–1551. Association for Computational Linguistics, Toronto, Canada. https://doi.org/10.18653/v1/2023.findings-acl.98 . https://aclanthology.org/2023.findings-acl.98

Do H, Kim Y, Lee G (2024) Autoregressive score generation for multi-trait essay scoring. In: Graham Y, Purver M (eds) Findings of the association for computational linguistics: EACL 2024, pp 1659–1666. Association for Computational Linguistics, St. Julian's, Malta. https://aclanthology.org/2024.findings-eacl.115/

Do H, Ryu S, Lee G (2024) Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In: Al-Onaizan Y, Bansal M, Chen Y-N (eds) Proceedings of the 2024 conference on empirical methods in natural language processing, pp 16427–16438. Association for Computational Linguistics, Miami, Florida, USA. https://doi.org/10.18653/v1/2024.emnlp-main.917 . https://aclanthology.org/2024.emnlp-main.917

Dong F, Zhang Y, Yang J (2017) Attention-based recurrent convolutional neural network for automatic essay scoring. In: Levy R, Specia L (eds) Proceedings of the 21st conference on computational natural language learning (CoNLL 2017), pp 153–162. Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/K17-1017 . https://aclanthology.org/K17-1017

Eltanbouly S, Albatarni S, Elsayed T (2025) TRATES: trait-specific rubric-assisted cross-prompt essay scoring. In: Che W, Nabende J, Shutova E, Pilehvar MT (eds) Findings of the association for computational linguistics: ACL 2025, pp 20528–20543. Association for Computational Linguistics, Vienna, Austria. https://doi.org/10.18653/v1/2025.findings-acl.1054. https://aclanthology.org/2025.findings-acl.1054/

He J, Li X (2024) Zero-shot cross-lingual automated essay scoring. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), pp 17819–17832. ELRA and ICCL, Torino, Italia. https://aclanthology.org/2024.lrec-main.1550

Han D, Roh D, Han E, Song H, Yi MY (2025) Fine-grained multi-prompt essay scoring with multi-level disentanglement. Data Min Knowl Discov 39(5):67

Ifenthaler D (2022) Automated essay scoring systems. Handbook of open, distance and digital education. Springer, Singapore, pp 1–15

Jiang Z, Gao T, Yin Y, Liu M, Yu H, Cheng Z, Gu Q (2023) Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), pp 12456–12470. Association for Computational Linguistics, Toronto, Canada. https://doi.org/10.18653/v1/2023.acl-long.696 . https://aclanthology.org/2023.acl-long.696

Lee S, Cai Y, Meng D, Wang Z, Wu Y (2024) Unleashing large language models' proficiency in zero-shot essay scoring. In: Al-Onaizan Y, Bansal M, Chen Y-N (eds) Findings of the association for computational linguistics: EMNLP 2024, pp 181–198. Association for Computational Linguistics, Miami, Florida, USA. https://doi.org/10.18653/v1/2024.findings-emnlp.10 . https://aclanthology.org/2024.findings-emnlp.10

Li X, Chen M, Nie J-Y (2020) Sednn: shared and enhanced deep neural network model for cross-prompt automated essay scoring. Knowl-Based Syst 210:106491. https://doi.org/10.1016/j.knosys.2020.106491

Li S, Gao Y (2025) Using join learning of multi-level features for automated essay scoring. Neural Comput Appl 37(16):10197–10214

Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International conference on learning representations. https://openreview.net/forum?id=Bkg6RiCqY7

Liu, Y (2019) Roberta: a robustly optimized bert pretraining approach, vol 364. arXiv preprint arXiv:1907.11692

Liusie A, Manakul P, Gales M (2024) LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In: Graham Y, Purver M (eds) Proceedings of the 18th conference of the European chapter of the association for computational linguistics (volume 1: long papers), pp 139–151. Association for Computational Linguistics, St. Julian's, Malta. https://doi.org/10.18653/v1/2024.eacl-long.8. https://aclanthology.org/2024.eacl-long.8/

Li S, Ng V (2024) Conundrums in cross-prompt automated essay scoring: making sense of the state of the art. In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers), pp 7661–7681. Association for Computational Linguistics, Bangkok, Thailand. https://doi.org/10.18653/v1/2024.acl-long.414 . https://aclanthology.org/2024.acl-long.414

Mathias S, Bhattacharyya P (2018) ASAP++: enriching the ASAP automated essay grading dataset with essay attribute scores. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1187

Mizumoto A, Eguchi M (2023) Exploring the potential of using an ai language model for automated essay scoring. Res Methods Appl Linguist 2(2):100050. https://doi.org/10.1016/j.rmal.2023.100050

Phandi P, Chai KMA, Ng HT (2015) Flexible domain adaptation for automated essay scoring using correlated linear regression. In: Màrquez L, Callison-Burch C, Su J (eds) Proceedings of the 2015 conference on empirical methods in natural language processing, pp 431–439. Association for Computational Linguistics, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1049 . https://aclanthology.org/D15-1049

Ridley R, He L, Dai X-Y, Huang S, Chen J (2021) Automated cross-prompt scoring of essay traits. Proc AAAI Conf Artif Intell 35(15):13745–13753. https://doi.org/10.1609/aaai.v35i15.17620

Ramesh D, Sanampudi SK (2022) An automated essay scoring systems: a systematic literature review. Artif Intell Rev 55(3):2495–2527

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67

Reilly ED, Stafford RE, Williams KM, Corliss SB (2014) Evaluating the validity and applicability of automated essay scoring in two massive open online courses. Int Rev Res Open Distrib Learn 15(5):83–98. https://doi.org/10.19173/irrodl.v15i5.1857

Stahl M, Biermann L, Nehring A, Wachsmuth H (2024) Exploring LLM prompting strategies for joint essay scoring and feedback generation. In: Proceedings of the 19th workshop on innovative use of nlp for building educational applications (BEA 2024), pp 283–298. Association for Computational Linguistics, Mexico City, Mexico. https://aclanthology.org/2024.bea-1.23

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(56):1929–1958

Shibata T, Miyamura Y (2025) Lces: zero-shot automated essay scoring via pairwise comparisons using large language models. arXiv preprint arXiv:2505.08498

Taghipour K, Ng HT (2016) A neural approach to automated essay scoring. In: Su J, Duh K, Carreras X (eds) Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1882–1891. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1193 . https://aclanthology.org/D16-1193

Tay Y, Phan M, Tuan LA, Hui SC C. Skipflow: incorporating neural coherence features for end-to-end automatic text scoring. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32. AAAI Press, New Orleans, Louisiana, USA

Tao Q, Zhong J, Li R (2022) Aesprompt: self-supervised constraints for automated essay scoring with prompt tuning. In: SEKE, pp 335–340

Uto M (2021) A review of deep-neural automated essay scoring models. Behaviormetrika 48(2):459–484

Wang Y, Wang C, Li R, Lin H (2022) On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In: Carpuat M, Marneffe M-C, Meza Ruiz IV (eds) Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 3416–3425. Association for Computational Linguistics, Seattle, United States. https://doi.org/10.18653/v1/2022.naacl-main.249 . https://aclanthology.org/2022.naacl-main.249

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D et al (2022) Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 35:24824–24837

Xie J, Cai K, Kong L, Zhou J, Qu W (2022) Automated essay scoring via pairwise contrastive regression. In: Proceedings of the 29th international conference on computational linguistics, pp 2724–2733. International Committee on Computational Linguistics, Gyeongju, Republic of Korea. https://aclanthology.org/2022.coling-1.240

Yang R, Cao J, Wen Z, Wu Y, He X (2020) Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In: Cohn T, He Y, Liu Y (eds) Findings of the association for computational linguistics: EMNLP 2020, pp 1560–1569. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.141 . https://aclanthology.org/2020.findings-emnlp.141

Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L et al (2020) Big bird: transformers for longer sequences. Adv Neural Inf Process Syst 33:17283–17297

Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z et al (2023) A survey of large language models. arXiv preprint arXiv:2303.18223

Authors and Affiliations

Hongseok Choi[1] · Myeong-Cheol Kang[1] · Jin Seong[1] · Jin-Xia Huang[1]

✉ Hongseok Choi
hongking9@etri.re.kr

Myeong-Cheol Kang
kangmc1@etri.re.kr

Jin Seong
real_castle@etri.re.kr

Jin-Xia Huang
hgh@etri.re.kr

[1] Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea