LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have enabled zero-shot automated essay scoring (AES), providing a promising way to reduce the cost and effort of essay scoring in comparison with manual grading. However, most existing zero-shot approaches rely on LLMs to directly generate absolute scores, which often diverge from human evaluations owing to model biases and inconsistent scoring. To address these limitations, we propose LLM-based Comparative Essay Scoring (LCES), a method that formulates AES as a pairwise comparison task. Specifically, we instruct LLMs to judge which of two essays is better, collect many such comparisons, and convert them into continuous scores. Considering that the number of possible comparisons grows quadratically with the number of essays, we improve scalability by employing RankNet to efficiently transform LLM preferences into scalar scores. Experiments using AES benchmark datasets show that LCES outperforms conventional zero-shot methods in accuracy while maintaining computational efficiency. Moreover, LCES is robust across different LLM backbones, highlighting its applicability to real-world zero-shot AES.

1 Introduction

Automated essay scoring (AES) aims to assess the quality of written essays using natural language processing and machine learning techniques. AES has garnered significant attention as a means to reduce the cost relative to human grading and to ensure fairness (Uto, 2021; Do et al., 2023).

Most conventional AES methods focus on *prompt-specific* approaches[1], which train machine learning models or neural networks on scored essays tailored to each essay prompt (Alikaniotis et al., 2016; Dong et al., 2017; Yang et al., 2020; Xie et al., 2022; Shibata and Uto, 2022; Wang and Liu, 2025). However, this approach requires collecting large amounts of scored essays for every prompt, resulting in substantial costs. To address this issue, recent studies have proposed *cross-prompt* AES methods that leverage domain adaptation or domain generalization techniques (Ridley et al., 2021; Chen and Li, 2023; Do et al., 2023; Chen and Li, 2024; Li and Pan, 2025). In those techniques, models are trained on scored essays from source prompts and evaluated on a different, target prompt. Although these methods can maintain high score accuracy even when scored essays for the target prompt are scarce or unavailable, they still require a certain amount of scored essay data for training, leaving unsolved the fundamental challenge of satisfying data requirements.

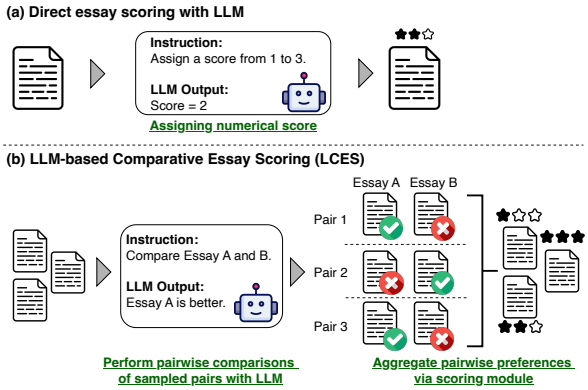

Figure 1: Comparison between (a) direct essay scoring using LLMs and (b) our proposed LCES framework.

In parallel to the above, large language models (LLMs) have recently demonstrated remarkable capabilities across various natural language processing tasks in zero-shot settings (Kojima et al., 2022), motivating efforts to apply them to AES without the use of scored essays. A typical zero-shot AES approach instructs an LLM with a rubric and essay to generate a numerical score (Mizumoto and Eguchi, 2023). A more advanced method first converts the original rubric defined in the dataset into

[1]Here, we use *prompt* to refer to the essay topic and *LLM prompt* to refer to instructions given to language models.

trait-level rubrics using LLMs, then employs LLMs to independently predict scores for each trait, and finally aggregates these scores to estimate the overall score (Lee et al., 2024). While these approaches are promising, they still have several limitations of direct score generation. They tend to be sensitive to the phrasing of LLM prompt and susceptible to model bias, and often exhibit grading behavior inconsistent with that of human raters (Zheng et al., 2023; Liu et al., 2024; Mansour et al., 2024; Li et al., 2025).

To address the issues above, we explore an alternative AES formulation. Instead of predicting absolute scores, we instruct the LLM to perform pairwise comparisons in which the better of two essays is determined. This approach is inspired by recent advances in LLM-based evaluation for natural language generation (Liu et al., 2024; Liusie et al., 2024a), dialogue systems (Park et al., 2024), and information retrieval (Qin et al., 2024), where pairwise comparisons have demonstrated stronger alignment with human preferences. Despite its promise, pairwise comparisons remain largely unexplored in the AES literature.

Against this backdrop, we propose **L**LM-based **C**omparative **E**ssay **S**coring (LCES), a novel framework for zero-shot AES that first collects pairwise comparisons using LLMs and then estimates continuous essay scores. As shown in Figure 1, LCES differs from conventional LLM-based scoring by shifting from direct score generation to relative preference modeling. To scale this approach to large essay datasets, we employ RankNet (Burges et al., 2005), which allows efficient training from pairwise comparisons without exhaustively enumerating all essay pairs. This mitigates the quadratic complexity in the number of items as is typically seen in pairwise comparisons (Liusie et al., 2024b).

Through comprehensive experiments using standard AES benchmark datasets, we demonstrate that LCES substantially outperforms existing zero-shot scoring methods. Moreover, LCES is robust to the choice of LLM and can be applied with virtually any model, making it well suited for practical deployment.

The contributions of this work are summarized as follows: (1) We introduce the first AES framework based on LLM-generated pairwise comparisons, addressing key limitations of direct score generation. (2) We leverage RankNet to convert LLM-generated preferences into continuous scores, enabling accurate and computationally efficient zero-shot AES. (3) Extensive experiments confirm that LCES outperforms conventional zero-shot AES baselines and is robust across different types of LLMs.

2 Related Work

Automated Essay Scoring. Early AES systems were largely *prompt-specific*, beginning with handcrafted-feature-based models (Yannakoudakis et al., 2011) and later adopting neural networks (Dong et al., 2017; Xie et al., 2022). Because it is costly to collect scored essays for every new prompt, *cross-prompt* methods have been proposed to train models that generalize across prompts (Ridley et al., 2020; Chen and Li, 2023, 2024). Recently, *zero-shot* AES using LLMs has emerged (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Wang et al., 2024; Mansour et al., 2024; Lee et al., 2024), enabling score generation without the use of scored essays. Mizumoto and Eguchi (2023) used OpenAI's text-davinci-003 to score essays based on rubric and essay content. In a zero-shot framework called Multi-Trait Specification (MTS), Lee et al. (2024) instructed an LLM to generate trait-level rubrics and then used them to evaluate essays by scoring each trait individually and aggregating the results. Mansour et al. (2024) demonstrated that LLM-generated scores are highly sensitive to the instructions given to the model, raising concerns about reliability. While zero-shot AES offers a promising direction, its scoring accuracy still lags behind supervised prompt-specific and cross-prompt methods.

LLM-based Evaluation. With the growing zero-shot capabilities of LLMs, the *LLM-as-a-judge* paradigm (Zheng et al., 2023) has gained attention as a general framework for using LLMs in evaluation tasks. Although direct score generation is common, it often suffers from LLM prompt sensitivity (Li et al., 2025) and misalignment with human judgments (Liu et al., 2024). To improve reliability, recent studies in natural language generation (Liu et al., 2024; Liusie et al., 2024a), dialogue systems (Park et al., 2024), and information retrieval (Qin et al., 2024) have instructed LLMs to make pairwise comparisons in which the better of two candidates is selected. Compared with absolute scoring, this approach requires fewer reasoning steps by LLMs and yields more consistent and human-aligned judgments. However, it remains

underexplored in AES.

Comparisons to Scores. Converting pairwise comparisons into continuous scores, which can be interpreted as latent measures of item quality that explain observed comparisons, has been widely studied. The Elo rating system (Elo, 1978) updates scores iteratively based on match outcomes. The Bradley–Terry model (Bradley and Terry, 1952) estimates win probabilities using the difference in latent scores between items, which are inferred by maximizing the likelihood of the observed comparisons. RankNet (Burges et al., 2005) extended this idea by learning latent scores from input features via a neural pairwise loss function. We use RankNet to transform LLM-generated essay comparisons into latent scores, enabling accurate and computationally efficient zero-shot AES.

3 Proposed Method

We start with a set of unscored essays $\mathcal{D} = \{x_i\}_{i=1}^{N}$, where x_i denotes the ith essay and N is the total number of essays. The goal of LCES is to estimate a latent score $\hat{s}_i$ for each essay x_i, representing its relative quality within the set $\mathcal{D}$. Depending on the assessment objective, the estimated score $\hat{s}_i$ can be converted into a ranking $\hat{r}_i$ or a score $\hat{y}_i$ aligned with a predefined rubric.

LCES consists of three main steps: (1) **Pairwise comparison generation:** Sample essay pairs from $\mathcal{D}$ and use an LLM to judge which essay is better, or whether they are of equal quality, based on a given rubric; (2) **Latent score estimation:** Train a RankNet model on the comparison dataset to estimate a latent score $\hat{s}_i$ for each essay; and (3) **Output conversion:** Convert the latent score $\hat{s}_i$ into either a ranking $\hat{r}_i$ or a score $\hat{y}_i$, depending on the evaluation goal. Each step is described in detail in the following subsections.

3.1 Pairwise Comparison Generation

To generate pairwise comparisons, we use an LLM prompt template $\mathcal{T}$ that guides the LLM to evaluate two essays based on a given rubric. A simplified version of this template is shown in Figure 2, and the complete version can be found in Appendix A. Given essay prompt p, scoring rubric r, and two essays x_i and x_j, we construct the query $\mathcal{T}(p, r, x_i, x_j)$ by inserting each input into the corresponding placeholder in the template. Specifically, the placeholders `<prompt>`, `<rubric>`, `<essay1>`, and `<essay2>` are replaced

LLM prompt template $\mathcal{T}$

```
# 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: <prompt>
# Rubric Guidelines: <rubric>
# Essay 1: <essay1>
# Essay 2: <essay2>
Provide your reasoning and final decision.
Reasoning: (Your reasoning here)
Decision: (Either "Essay 1", "Essay 2", or "tie")
```

Figure 2: Simplified LLM prompt template $\mathcal{T}$ used for pairwise essay comparisons.

with p, r, x_i, and x_j respectively. To improve the reliability and interpretability of the comparisons, we use chain-of-thought prompting (Wei et al., 2022). This encourages the LLM to explain its reasoning before making a final decision. For the essay judged to be better, the LLM outputs a categorical label w_{ij}, which is one of "Essay 1", "Essay 2 ", and "tie ". We convert this to a numerical label c_{ij} by assigning scores of 1, 0, and 0.5 for "Essay 1", "Essay 2", and "tie", respectively.

LLMs can be sensitive to the order in which the two essays are presented (Zheng et al., 2023). To reduce this position bias, we query the LLM twice for each pair. One query presents the essays as (x_i, x_j), and the other as (x_j, x_i). Let c_{ij} be the numerical label from the first query and c_{ji} be the label from the second. We define the final debiased label $\tilde{c}_{ij}$ as follows:

$$\tilde{c}_{ij} = \begin{cases} c_{ij} & \text{if } c_{ij} = 1 - c_{ji} \\ 0.5 & \text{otherwise.} \end{cases} \quad (1)$$

If the two results are consistent, we retain the original label. If the results contradict each other or one of them indicates a tie, we treat the pair as a tie.

To apply this comparison procedure, we construct a set of essay pairs. Let $\mathcal{I} = \{(i, j) \mid i \neq j,\ i, j \in \{1, 2, \ldots, N\}\}$ be the set of all possible ordered essay pairs. Since comparing all $N(N-1)$ pairs is computationally expensive, we randomly sample a subset $\mathcal{I}_s \subset \mathcal{I}$ containing M pairs, where $M \ll N(N-1)$. For each sampled pair, we obtain a debiased label $\tilde{c}_{ij}$ as described above. This yields the pairwise comparison dataset $\mathcal{D}_{\text{pair}} = \{(x_i, x_j, \tilde{c}_{ij}) \mid (i, j) \in \mathcal{I}_s\}$, which is used to train the RankNet model.

3.2 Latent Score Estimation

Using the pairwise comparison dataset $\mathcal{D}_{\text{pair}}$ generated in Section 3.1, we estimate a latent score $\hat{s}_i$ for each essay x_i. To this end, we employ

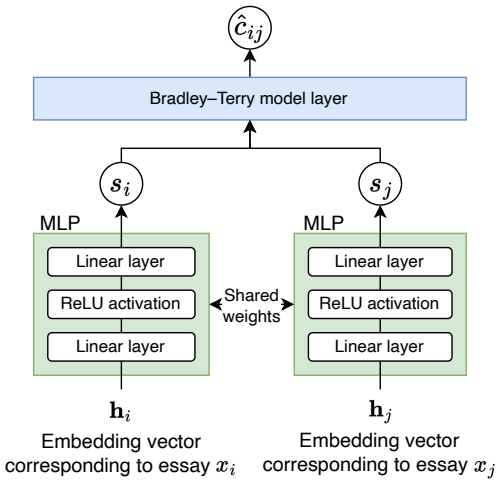

Figure 3: Architecture of the RankNet model used to estimate latent essay scores $\hat{s}_i$ from pairwise comparisons.

RankNet (Burges et al., 2005), a neural model designed to learn latent scores from pairwise preferences.

As shown in Figure 3, RankNet uses two parallel multi-layer perceptrons (MLPs) with shared weights. These form a scoring model, denoted by f, which maps an input essay representation to a scalar score. Specifically, we first convert each essay x_i into an embedding vector $\boldsymbol{h}_i$ using any suitable text embedding model, and then compute its score as $s_i = f(\boldsymbol{h}_i)$. Each MLP consists of two linear layers with a ReLU activation (Agarap, 2019) applied after the first layer.

Given two essays x_i and x_j, the model computes scores s_i and s_j using the shared network f, and estimates the probability that x_i is preferred over x_j as:

$$\hat{c}_{ij} = \sigma(s_i - s_j) = \frac{1}{1 + \exp(-(s_i - s_j))}. \quad (2)$$

Here, $\sigma(\cdot)$ denotes the sigmoid function. This formulation mirrors the Bradley–Terry model (Bradley and Terry, 1952) and enables probabilistic modeling of pairwise preferences.

The model is trained to minimize the discrepancy between the predicted preference $\hat{c}_{ij}$ and the debiased target label $\tilde{c}_{ij}$. We use the binary cross-entropy loss $\mathcal{L}$:

$$\mathcal{L} = -\frac{1}{M} \sum_{(i,j)\in\mathcal{I}_s} \left[\tilde{c}_{ij} \log \hat{c}_{ij} + (1 - \tilde{c}_{ij}) \log(1 - \hat{c}_{ij})\right].$$

Let $\mathcal{S} = \{s_i\}_{i=1}^N$ denote the set of latent essay scores. The optimized scores $\hat{\mathcal{S}} = \{\hat{s}_i\}_{i=1}^N$ are obtained by minimizing the loss: $\hat{\mathcal{S}} = \arg\min_{\mathcal{S}} \mathcal{L}$.

3.3 Output Conversion

The estimated latent scores $\hat{s}_i$ can be converted into standard AES outputs, such as numerical scores or rankings, depending on the evaluation goal.

To produce a score $\hat{y}_i$ within a rubric-defined range $[y_{\min}, y_{\max}]$, we apply a linear transformation to the latent scores:

$$\hat{y}_i = \frac{\hat{s}_i - s_{\min}}{s_{\max} - s_{\min}} \times (y_{\max} - y_{\min}) + y_{\min}, \quad (3)$$

where $s_{\min} = \min_i \hat{s}_i$ and $s_{\max} = \max_i \hat{s}_i$ are the minimum and maximum latent scores across all essays. If the rubric defines discrete score levels, the resulting $\hat{y}_i$ can optionally be rounded to the nearest valid level.

Alternatively, a ranking $\hat{r}_i$ can be obtained by sorting essays in descending order of their latent scores $\hat{s}_i$. This is useful in settings where only relative essay quality is required.

4 Experiments

We empirically evaluate the effectiveness of LCES through experiments using AES benchmark datasets, focusing on scoring performance and comparisons with existing methods.

4.1 Datasets

We utilized the following two benchmark datasets, which are commonly used in AES research (Taghipour and Ng, 2016; Chen and Li, 2023; Lee et al., 2024; Wang et al., 2024):

ASAP (Automated Student Assessment Prize) is a dataset released by the Kaggle competition[2]. It consists of 12,978 essays across eight different prompts, each with human-assigned scores.

TOEFL11 is a dataset of essays written by non-native English speakers taking the TOEFL iBT (Blanchard et al., 2013). It contains 12,100 essays across eight different prompts, each with human-assigned scores.

Table 1 summarizes the statistics of the ASAP and TOEFL11 datasets.

4.2 Baselines

We adopted the following two zero-shot AES methods as baselines for comparison with LCES:

[2]https://www.kaggle.com/c/asap-aes

Table 1: Statistics of the ASAP and TOEFL11 datasets. l/m/h denotes low/medium/high.

Dataset	Prompt	No. of Essays	Avg. Len.	Score Range
ASAP	1	1,783	427	2–12
	2	1,800	432	1–6
	3	1,726	124	0–3
	4	1,772	106	0–3
	5	1,805	142	0–4
	6	1,800	173	0–4
	7	1,569	206	0–30
	8	723	725	0–60
TOEFL11	1	1,656	342	l/m/h
	2	1,562	361	l/m/h
	3	1,396	346	l/m/h
	4	1,509	340	l/m/h
	5	1,648	361	l/m/h
	6	960	360	l/m/h
	7	1,686	339	l/m/h
	8	1,683	344	l/m/h

Vanilla. A direct scoring approach where the LLM generates a rubric-aligned score for each essay without pairwise comparison. It uses chain-of-thought prompting to elicit reasoning before scoring. We used the same LLM prompts and hyperparameters as Lee et al. (2024).

MTS. As described in Section 2, MTS (Lee et al., 2024) is a state-of-the-art zero-shot AES framework. The original implementation used GPT-3.5 to generate trait-level rubrics from the original rubric. In our experiments, we used GPT-4o instead because GPT-3.5 is no longer available. All other LLM prompts and hyperparameters followed the original implementation.

4.3 Experimental Setup

LLMs. We conducted our evaluation using five distinct LLMs, namely, Mistral-7B (-instruct-v0.2) (Jiang et al., 2023), Llama-3.2-3B (-Instruct), Llama-3.1-8B (-Instruct) (Grattafiori et al., 2024), GPT-4o-mini (-2024-07-18), and GPT-4o (-2024-08-06) (OpenAI, 2024). All LLM inferences were performed with a temperature setting of 0.1.

Implementation Details. The number of sampled pairwise comparisons M was set to 5,000 to construct the $\mathcal{D}_{\text{pair}}$ dataset. Essay embedding vectors $\boldsymbol{h}_i$ were generated using OpenAI's `text-embedding-v3-large` model. Results obtained with alternative embedding models are presented in Appendix B. The RankNet model was trained for 100 epochs using the Adam (Kingma and Ba, 2015) optimizer with a learning rate of 0.001. The full set of hyperparameters is provided in Appendix C.

Rubrics. For pairwise comparisons within the ASAP dataset, we used the original scoring rubrics provided with the dataset. For the TOEFL11 dataset, consistent with previous studies (Mizumoto and Eguchi, 2023; Lee et al., 2024), we used the IELTS Task 2 Writing Band Descriptors as the evaluation rubric.

Evaluation Metrics. We evaluated model performance using two standard metrics in AES, namely quadratic weighted kappa (QWK) (Cohen, 1960) and the Spearman rank correlation coefficient. Following common practice in previous work (Taghipour and Ng, 2016; Alikaniotis et al., 2016; Dong et al., 2017; Do et al., 2023), we primarily report QWK. Results for Spearman correlations are provided in Appendix D. For the ASAP dataset, we followed Lee et al. (2024) and randomly sampled 10% of essays from each prompt for evaluation. For TOEFL11, we used the predefined test split consisting of 1,100 essays across eight prompts.

Scoring Strategy. For QWK-based evaluation, we rounded the predicted scores $\hat{y}_i$ to align with the score range of each prompt in the ASAP dataset. For the TOEFL11 dataset, we first converted the latent scores to a $[1, 5]$ scale by the linear transformation described in Section 3.3, and we then mapped them to low/medium/high categories using thresholds of 2.25 and 3.75, following the approach used in previous research (Blanchard et al., 2013; Lee et al., 2024).

4.4 Results and Discussion

The results in Table 2 show that LCES outperforms both MTS and Vanilla in most settings, achieving higher average QWK scores across models and prompts, with particularly large gains on the ASAP dataset. The only exception is TOEFL11 with Mistral-7B, where LCES performs worse than MTS. As shown later in Section 5.2, Mistral-7B exhibits a high inconsistency rate (51.4%) when essay order is reversed, suggesting that it has difficulty in reliably identifying the better essay. Notably, Mistral-7B achieves the highest performance under the MTS setting on TOEFL11, surpassing more recent or larger models such as GPT-4o and Llama-3.1-8B. This suggests that MTS and LCES may favor different model capabilities. While LCES underperforms MTS with Mistral-7B, it consistently outperforms MTS with all other LLMs, highlighting the general effectiveness of the LCES framework.

Table 2: QWK scores for each essay prompt in ASAP and TOEFL11. **Bold** indicates the best-performing method for each prompt. **P1-8** refers to Prompt 1 through Prompt 8.

Dataset	Model	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
ASAP	Mistral-7B	Vanilla	0.429	0.439	0.387	0.518	0.576	0.534	0.276	0.209	0.429
		MTS	0.546	0.479	0.481	**0.683**	0.706	0.519	0.501	0.175	0.511
		LCES	**0.600**	**0.603**	**0.690**	0.614	**0.729**	**0.792**	**0.591**	**0.315**	**0.617**
	Llama-3.2-3B	Vanilla	0.254	0.405	0.410	0.009	0.397	0.330	0.438	0.276	0.315
		MTS	0.197	0.452	0.353	0.507	0.460	0.462	0.146	0.190	0.346
		LCES	**0.555**	**0.608**	**0.647**	**0.603**	**0.717**	**0.756**	**0.580**	**0.612**	**0.635**
	Llama-3.1-8B	Vanilla	0.129	0.023	0.243	0.550	0.301	0.341	0.006	-0.042	0.194
		MTS	0.516	0.483	0.284	0.461	0.479	0.378	0.328	0.199	0.391
		LCES	**0.669**	**0.599**	**0.662**	**0.651**	**0.710**	**0.707**	**0.727**	**0.636**	**0.670**
	GPT-4o-mini	Vanilla	0.106	0.402	0.314	0.602	0.577	0.470	0.425	0.517	0.426
		MTS	0.472	0.386	0.448	0.552	0.708	0.419	0.479	0.412	0.485
		LCES	**0.537**	**0.602**	**0.679**	**0.638**	**0.709**	**0.737**	**0.614**	**0.521**	**0.630**
	GPT-4o	Vanilla	0.216	0.498	0.447	0.681	0.710	0.571	0.535	0.411	0.509
		MTS	0.380	0.547	0.513	0.621	0.500	0.515	0.421	0.432	0.491
		LCES	**0.531**	**0.592**	**0.702**	**0.626**	**0.747**	**0.766**	**0.669**	**0.593**	**0.653**
TOEFL11	Mistral-7B	Vanilla	0.235	0.128	0.174	0.106	0.050	0.046	0.106	0.222	0.133
		MTS	**0.634**	0.496	0.571	**0.607**	**0.603**	**0.573**	**0.578**	**0.689**	**0.594**
		LCES	0.415	**0.514**	**0.663**	0.519	0.508	0.496	0.532	0.644	0.536
	Llama-3.2-3B	Vanilla	0.184	0.117	0.291	0.195	0.149	0.206	0.067	0.149	0.170
		MTS	0.361	0.389	0.454	0.456	0.341	0.364	0.323	0.299	0.373
		LCES	**0.615**	**0.542**	**0.709**	**0.678**	**0.582**	**0.479**	**0.555**	**0.708**	**0.608**
	Llama-3.1-8B	Vanilla	-0.036	0.148	0.003	0.021	0.019	-0.023	-0.029	0.063	0.021
		MTS	0.368	0.408	0.407	0.311	0.351	0.285	0.335	0.379	0.356
		LCES	**0.597**	**0.570**	**0.727**	**0.697**	**0.652**	**0.550**	**0.558**	**0.717**	**0.633**
	GPT-4o-mini	Vanilla	0.094	0.202	0.182	0.107	0.041	0.101	0.126	0.124	0.122
		MTS	0.439	0.529	0.548	0.521	0.603	0.501	0.536	0.591	0.533
		LCES	**0.655**	**0.559**	**0.722**	**0.692**	**0.633**	**0.649**	**0.629**	**0.724**	**0.658**
	GPT-4o	Vanilla	0.206	0.208	0.365	0.189	0.211	0.245	0.226	0.252	0.238
		MTS	0.480	0.539	0.607	0.545	0.469	0.526	0.426	0.664	0.532
		LCES	**0.604**	**0.545**	**0.734**	**0.671**	**0.713**	**0.572**	**0.580**	**0.739**	**0.645**

Moreover, LCES demonstrates strong robustness across backbone LLMs. On ASAP, the standard deviation of its average performance across five backbone models is just 0.021, compared to 0.072 for MTS and 0.122 for Vanilla. On TOEFL11, LCES similarly shows low variability, with a standard deviation of 0.048 across models, outperforming MTS (0.106) and Vanilla (0.079). These low inter-model variances indicate that LCES remains stable regardless of backbone choice, whereas MTS and Vanilla fluctuate more, making their performance less predictable.

In summary, LCES exhibited high zero-shot accuracy across diverse settings, with robustness to LLM choice and reliable scaling with model size. Even when the model struggles to distinguish essay quality, as with Mistral-7B on TOEFL11, LCES remains competitive, supporting its practicality in real-world AES scenarios.

5 Analysis

We present a set of analyses to further examine the effectiveness and properties of the proposed framework beyond overall performance metrics.

Table 3: Average QWK scores across all ASAP prompts for LCES and supervised learning baselines.

Method	Avg. QWK
Prompt-specific	
NPCR (Xie et al., 2022)	0.792
BERT-base-uncased (Devlin et al., 2019)	0.740
RoBERTa-base (Liu et al., 2019)	0.743
Cross-prompt	
PAES (Ridley et al., 2020)	0.678
PMAES (Chen and Li, 2023)	0.658
Zero-shot	
LCES (Llama-3.1-8B)	0.670

5.1 Comparison with Supervised Models

Although LCES is a zero-shot method, we also compare it with several supervised learning baselines on the ASAP dataset, as summarized in Table 3. We include both prompt-specific and cross-prompt models. The prompt-specific models are trained on 90% of the essays from a single prompt and evaluated on the remaining 10%, using the

Table 4: Average percentage of LLM judgments that change when the order of essay pairs is reversed, computed across all prompts in each dataset.

Model	ASAP (%)	TOEFL11 (%)
Mistral-7B	42.8	51.4
Llama-3.2-3B	28.8	39.0
Llama-3.1-8B	21.6	23.8
GPT-4o-mini	13.8	10.5
GPT-4o	10.4	17.0

same evaluation split described in Section 4.3. The cross-prompt models are trained on essays from all prompts except the one under evaluation, and are also evaluated on the same 10% split of the target prompt.

Specifically, we make comparisons against NPCR (Xie et al., 2022), which is reported to provide state-of-the-art results on ASAP, as well as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) fine-tuned on the same prompt-specific splits. We also include PAES (Ridley et al., 2020) and PMAES (Chen and Li, 2023), which are two strong cross-prompt baselines.

As shown in Table 3, LCES with Llama-3.1-8B, which achieved the highest overall performance among all tested LLMs in the zero-shot experiments (see Section 4.4), obtains QWK scores that are comparable to several supervised learning models. While NPCR, BERT, and RoBERTa still outperform LCES, the performance gap has significantly narrowed in comparison with previously reported zero-shot methods. In addition, LCES achieves performance on par with the strong cross-prompt baselines PAES and PMAES[3]. This level of performance is unprecedented among zero-shot AES methods. These results highlight the effectiveness of the proposed method in the absence of scored essays.

5.2 Position Bias

We measure the impact of position bias by calculating the percentage of pairwise comparisons that change when the order of essays is reversed. Table 4 shows the inconsistency rates for each LLM on the same comparison pairs used to construct $\mathcal{D}_{\text{pair}}$ for ASAP and TOEFL11. As expected, larger models such as GPT-4o exhibit lower inconsistency, suggesting greater robustness to position bias. In contrast, Mistral-7B shows a particularly high inconsistency rate of 51.4% on TOEFL11, indicating substantial sensitivity to essay order.

[3]PMAES was run with a smaller batch size due to GPU limitations (RTX 4090), which may have led to reduced performance.

Table 5: Average QWK on ASAP and TOEFL11 with and without position bias correction.

Dataset	Model	Avg. QWK	
		w/o Debias	w/ Debias
ASAP	Mistral-7B	0.611	0.617
	Llama-3.2-3B	0.630	0.635
	Llama-3.1-8B	0.661	0.670
	GPT-4o-mini	0.633	0.630
	GPT-4o	0.649	0.653
TOEFL11	Mistral-7B	0.510	0.536
	Llama-3.2-3B	0.588	0.608
	Llama-3.1-8B	0.628	0.633
	GPT-4o-mini	0.664	0.658
	GPT-4o	0.648	0.645

To assess the effect of position bias correction, we compare average QWK scores with and without the position bias correction. As shown in Table 5, models with higher inconsistency rates, such as Mistral-7B and Llama-3.2-3B, tend to benefit more from the correction. These results suggest that the proposed correction method is generally more effective for models with higher position inconsistency, whereas its effect is limited for models that already exhibit low inconsistency.

5.3 Comparison of Latent Score Conversion Methods

We evaluate the effectiveness of different latent score conversion techniques by comparing our RankNet-based approach with the Bradley–Terry model and the Elo rating system which are representative methods described in Section 2. The experiment examines how the number of pairwise comparisons M, ranging from 50 to 10,000, affects scoring accuracy, measured by QWK, on the ASAP and TOEFL11 datasets. This experiment adopts GPT-4o as the LLM, in view of its robust performance on both datasets.

Figure 4 illustrates the performance trends. As M increases, accuracy improves for all methods, highlighting the benefit of additional preference data. Among them, the RankNet-based approach consistently outperforms both the Bradley–Terry model and the Elo rating system across the entire range of M on both datasets. Notably, RankNet achieves high QWK scores even with relatively few comparisons (e.g., $M = 50$ or $M = 100$), demonstrating strong performance particularly in limited data scenarios. This advantage likely stems from RankNet's ability to incorporate textual features

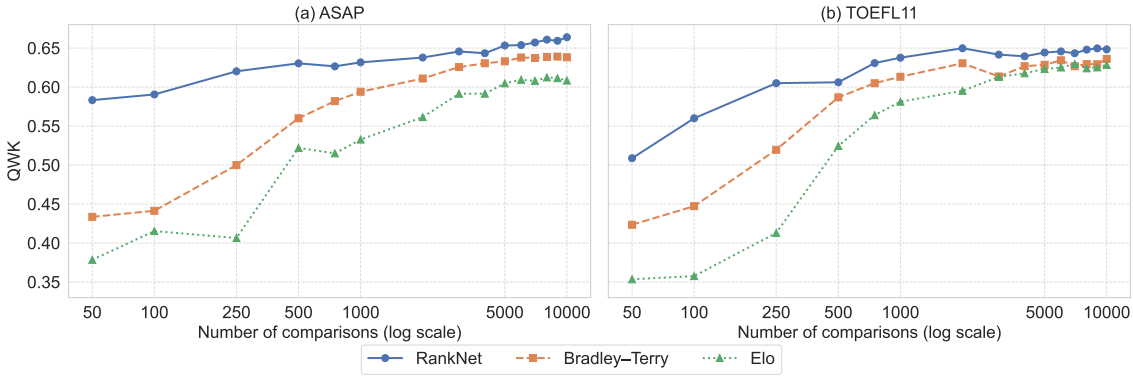

Figure 4: Relationship between the number of pairwise comparisons (log scale) and QWK scores. (a) ASAP dataset. (b) TOEFL11 dataset.

Table 6: QWK scores of LCES in *transductive* and *inductive* settings, evaluated using GPT-4o-mini.

Dataset	Setting	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
ASAP	Transductive	0.537	0.602	0.679	0.638	0.709	0.737	0.614	0.521	0.630
	Inductive	0.611	0.622	0.588	0.631	0.707	0.783	0.487	0.603	0.629
TOEFL11	Transductive	0.655	0.559	0.722	0.692	0.633	0.649	0.629	0.724	0.658
	Inductive	0.624	0.613	0.715	0.617	0.622	0.615	0.616	0.707	0.641

directly from essays, whereas the baseline methods rely solely on comparison outcomes.

These results suggest that RankNet is highly effective for pairwise-based essay scoring. Its superior accuracy and greater data efficiency make it well suited for practical settings where collecting extensive comparison data may be costly or infeasible.

5.4 Performance in the Inductive Setting

In LCES, pairwise preferences predicted by an LLM are used to train a scoring function, f. When all target essays are available at once, comparisons can be directly made among them. We refer to this as the *transductive* setting, which corresponds to the main experimental setup used throughout this paper. In contrast, the *inductive* setting assumes that new essays must be scored individually, without additional comparisons. Instead, the learned function f, trained on prior comparisons, is applied for score estimation. Because f maps essay embeddings to scalar scores, it can generalize to new essays without further pairwise information.

To simulate this scenario, we train the scoring function f on pairwise comparisons constructed from 90% of the essays in each dataset (ASAP or TOEFL11) and use it to predict scores for the remaining 10%. We use GPT-4o-mini for its computational efficiency and low API cost.

QWK scores in the inductive setting are close to those in the transductive setting, with 0.629 vs. 0.630 on ASAP and 0.641 vs. 0.658 on TOEFL11 (Table 6). These results demonstrate that f generalizes effectively to unseen essays. This ability to score new essays without constructing additional comparisons involving them makes LCES well suited for inductive scenarios. In contrast, models such as the Bradley–Terry model or Elo require the generation of new comparisons for each essay, leading to higher deployment overhead in inductive settings.

6 Conclusion

In this study, we presented LLM-based Comparative Essay Scoring (LCES), a zero-shot AES framework that leverages LLM-driven pairwise comparisons to address key limitations of direct score generation. LCES instructs an LLM to judge which of two essays is better, and then trains a RankNet model to estimate continuous essay scores.

Experimental results on two benchmark datasets, namely, ASAP and TOEFL11, demonstrate that LCES consistently outperforms existing zero-shot methods in scoring accuracy. It maintains strong performance even with a limited number of comparisons and is robust to the choice of LLM. Moreover, LCES can be applied in inductive settings without requiring additional comparisons for new essays. These properties make LCES well suited for real-world AES applications.

Limitations

Despite its advantages, LCES has several limitations. First, it relies on pairwise preference labels generated by an LLM, which may contain noise or inconsistencies. These imperfect labels directly affect the quality of learned scoring function f.

Second, while LCES tends to perform reliably when provided a sufficient number of comparisons M, it remains unclear how to determine an appropriate value of M. This limits the ability to systematically control scoring quality.

Third, LCES maps latent relative scores to an absolute scale via linear transformation, assuming sampled comparisons span the full score range. If low- or high-scoring essays are missing, the transformation may yield inaccurate absolute scores. While ranking performance would remain unaffected, this can reduce alignment with human judgment in tasks requiring precise or rubric-specific scoring.

Finally, the zero-shot nature of LCES means that, without labeled data, its performance cannot be quantitatively assessed. For practical deployment, this necessitates manually scoring a subset of essays to establish a benchmark for evaluation.

Acknowledgments

This research was conducted as part of an internal project at Deloitte Touche Tohmatsu LLC. The authors would like to thank Tomoaki Geka and Tomotake Kozu for helpful discussions and support. We also used ChatGPT to assist in clarifying the structure and wording of certain parts of the manuscript.

Ethics Statement

The primary ethical consideration for LCES is its reliance on LLMs for pairwise essay comparisons. LLMs may inherit and perpetuate biases present in their extensive training data. While our method includes a debiasing step for position bias, other latent biases could potentially influence the fairness of evaluations across different student demographics or writing styles. We recommend further auditing for such biases before any high-stakes deployment of LCES.

The experiments in this study were conducted using publicly available and established benchmark datasets (ASAP and TOEFL11). No new personally identifiable information was collected or used in this research.

We envision LCES as an assistive tool to support human graders, reducing their workload and potentially improving consistency, rather than as a complete replacement for human judgment. Given its zero-shot nature, thorough validation of LCES on specific target prompts and scoring rubrics is crucial before its application in real-world educational assessments to ensure reliability and prevent potential negative impacts on students.

References

Abien Fred Agarap. 2019. Deep learning using rectified linear units (ReLU). *Preprint*, arXiv:1803.08375.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 715–725.

Daniel Blanchard, Joel R. Tetreault, Derrick Higgins, A. Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013:15.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1489–1503.

Yuan Chen and Xia Li. 2024. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12775–12786.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay

trait scoring. In *Proceedings of Findings of the Association for Computational Linguistics, the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1538–1551.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 153–162.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics, the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 181–198.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *Preprint*, arXiv:2411.16594.

Xia Li and Wenjing Pan. 2025. KAES: Multi-aspect shared knowledge finding and aligning for cross-prompt automated scoring of essay traits. In *Proceedings of the 39th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, volume 39, pages 24476–24484.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *Proceedings of the First Conference on Language Modeling*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024a. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–151, St. Julian's, Malta.

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. 2024b. Efficient LLM comparative assessment: A product of experts framework for pairwise comparisons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6835–6855.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2777–2786.

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

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. PairEval: Open-domain dialogue evaluation with pairwise comparison. *Preprint*, arXiv:2404.01015.

Zhen Qin, Rolf Jagerman, Kai Hui, and 1 others. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics, the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1504–1518.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *Preprint*, arXiv:2008.01441.

Robert Ridley, Liang He, Xin yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the Thirty-Fifth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, volume 35, pages 13745–13753.

Takumi Shibata and Masaki Uto. 2022. Analytic automated essay scoring based on deep neural networks

integrating multidimensional item response theory. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2917–2926.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):1–26.

Jiong Wang and Jie Liu. 2025. T-MES: Trait-aware mix-of-experts representation learning for multi-trait essay scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236.

Yupei Wang, Renfen Hu, and Zhe Zhao. 2024. Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on linguistically-informed counterfactuals. In *Findings of the Association for Computational Linguistics, the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8906–8925.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733.

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

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics, the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1560–1569.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 180–189.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

A LLM Prompts

This section describes the LLM prompt templates used to elicit pairwise preferences from LLMs during the comparison step in Section 3.1. We design separate LLM prompts for the ASAP and TOEFL11 datasets to reflect their target populations and scoring rubrics. Each LLM prompt includes a system message defining the evaluator's role and a user message with the task context, rubric, and two essays. The model is instructed to return a brief justification and a final decision in structured JSON format for automated parsing. Our LLM prompt format is based on the template introduced by Lee et al. (2024).

A.1 ASAP

System Prompt

As an English teacher, your primary responsibility is to evaluate the writing quality of essays written by middle school students on an English exam. During the assessment process, you will be provided with a prompt and an essay. First, you should provide comprehensive and concrete feedback that is closely linked to the content of the essay. It is essential to avoid offering generic remarks that could be applied to any piece of writing.
To create a compelling evaluation for both the student and fellow experts, you should reference specific content of the essay to substantiate your assessment.
Next, your task is to determine which essay, Essay 1 or Essay 2, scores higher, or if they score the same, please respond with "tie". The evaluation criteria can be part of an overall rubric or separate evaluation criteria. Regardless of the type of rubric, please determine which essay achieves a higher overall score.

User Prompt

Prompt
{prompt}
Rubric Guidelines
{rubric}
Note
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER).

The relevant entities are identified in the text and then replaced with a string such as "PERSON", "ORGANIZATION", "LOCATION", "DATE", "TIME", "MONEY", "PERCENT", "CAPS" (any capitalized word) and "NUM" (any digits). Please do not penalize the essay because of the anonymizations.
Essay1
{essay1}
Essay2
{essay2}
Provide your reasoning and final decision in json format:
{ "reasoning": "Your reasoning in one sentence here.", "preference": "essay1" or "essay2" or "tie" }

A.2 TOEFL11

System Prompt

As an English teacher, your primary responsibility is to evaluate the writing quality of essays written by second language learners on an English exam. During the assessment process, you will be provided with a prompt and an essay.
First, you should provide comprehensive and concrete feedback that is closely linked to the content of the essay. It is essential to avoid offering generic remarks that could be applied to any piece of writing. To create a compelling evaluation for both the student and fellow experts, you should reference specific content of the essay to substantiate your assessment.
Next, your task is to determine which essay, Essay 1 or Essay 2, scores higher, or if they score the same, please respond with "tie". The evaluation criteria are based on four assessment categories. Use these categories to comprehensively evaluate and compare the essays, and decide which one achieves a higher overall score.

User Prompt

Prompt
{prompt}
Rubric Guidelines
{rubric}
Essay1
{essay1}
Essay2
{essay2}
Provide your reasoning and final decision in json format:
{ "reasoning": "Your reasoning in one sentence here.", "preference": "essay1" or "essay2" or "tie" }

Table 7: QWK scores of LCES with different embeddings (using GPT-4o).

Embedding Model	ASAP	TOEFL11
text-embedding-v3-large	0.653	0.645
text-embedding-v3-small	0.668	0.630
BERT-base-uncased	0.658	0.663
RoBERTa-base	0.655	0.601

Table 8: Hyperparameters for RankNet.

Hyperparameter	Value
Batch size	4096
Dropout rate	0.3
Hidden units	256
Weight decay	0.01

B Embedding Models

We compare four pretrained embedding models used to convert essays into fixed-length vectors for RankNet. Two of them are OpenAI models: `text-embedding-v3-large` (3072 dimensions) and `text-embedding-v3-small` (1536 dimensions), both of which were designed for semantic similarity tasks. The other two are BERT-base and RoBERTa-base. For these models, we use the [CLS] token from the final hidden layer as the essay representation.

Table 7 shows the average QWK scores on ASAP and TOEFL11 using GPT-4o for pairwise comparisons. For ASAP, the choice of embedding model has little impact on performance overall. For TOEFL11, we observe slightly more variation, but all models yield consistently high accuracy. These results suggest that LCES is robust to the choice of embedding encoder.

C Hyperparameters

Table 8 shows the hyperparameters used for training the RankNet model described in Section 3.2. The model consists of two linear layers with a ReLU activation and a dropout layer applied between them. Weight decay is applied as part of the

Table 9: The Spearman rank correlation coefficient for each prompt in ASAP and TOEFL11. **Bold** indicates the best-performing method for each prompt.

Dataset	Model	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
ASAP	Mistral-7B	Vanilla	0.511	0.511	0.439	0.658	0.527	0.418	0.379	0.459	0.488
		MTS	0.593	0.468	0.612	0.729	0.739	0.555	0.566	0.306	0.571
		LCES	**0.616**	**0.678**	**0.745**	**0.784**	**0.811**	**0.806**	**0.684**	**0.632**	**0.719**
	Llama-3.2-3B	Vanilla	0.068	0.109	0.452	-0.033	0.276	0.142	0.209	0.076	0.162
		MTS	0.205	0.528	0.500	0.712	0.606	0.527	0.210	0.276	0.445
		LCES	**0.665**	**0.693**	**0.725**	**0.767**	**0.741**	**0.738**	**0.589**	**0.684**	**0.700**
	Llama-3.1-8B	Vanilla	0.005	0.050	0.451	0.618	0.424	0.429	0.061	-0.090	0.245
		MTS	0.538	0.580	0.546	0.723	0.731	0.543	0.570	0.366	0.574
		LCES	**0.702**	**0.685**	**0.723**	**0.809**	**0.754**	**0.710**	**0.724**	**0.719**	**0.728**
	GPT-4o-mini	Vanilla	0.394	0.472	0.464	0.730	0.668	0.545	0.435	0.580	0.536
		MTS	0.560	0.523	0.509	0.672	0.763	0.565	0.498	0.555	0.580
		LCES	**0.588**	**0.678**	**0.736**	**0.817**	**0.761**	**0.727**	**0.636**	**0.693**	**0.705**
	GPT-4o	Vanilla	0.468	0.518	0.525	0.787	0.729	0.557	0.546	0.549	0.585
		MTS	0.417	0.642	0.639	0.771	0.557	0.576	0.502	0.608	0.589
		LCES	**0.578**	**0.682**	**0.750**	**0.833**	**0.812**	**0.776**	**0.713**	**0.713**	**0.732**
TOEFL11	Mistral-7B	Vanilla	0.272	0.126	0.185	0.145	0.030	0.042	0.141	0.241	0.148
		MTS	**0.717**	**0.587**	**0.674**	0.649	**0.703**	**0.634**	**0.640**	**0.740**	**0.669**
		LCES	0.470	0.565	0.638	**0.665**	0.560	0.495	0.562	0.681	0.579
	Llama-3.2-3B	Vanilla	0.204	0.144	0.339	0.205	0.182	0.229	0.080	0.161	0.193
		MTS	0.649	0.572	0.720	0.644	0.532	**0.549**	0.608	0.563	0.604
		LCES	**0.663**	**0.628**	**0.748**	**0.722**	**0.636**	0.505	**0.627**	**0.721**	**0.656**
	Llama-3.1-8B	Vanilla	-0.077	0.166	-0.002	-0.005	-0.004	-0.047	-0.034	0.095	0.012
		MTS	0.665	0.609	**0.791**	0.686	0.647	0.542	0.622	0.663	0.653
		LCES	**0.751**	**0.668**	0.759	**0.755**	**0.723**	**0.582**	**0.690**	**0.767**	**0.712**
	GPT-4o-mini	Vanilla	0.131	0.252	0.261	0.172	0.044	0.123	0.151	0.177	0.164
		MTS	0.684	0.655	**0.781**	0.716	0.727	0.645	0.650	0.715	0.696
		LCES	**0.753**	**0.674**	0.757	**0.745**	**0.753**	**0.684**	**0.695**	**0.769**	**0.729**
	GPT-4o	Vanilla	0.257	0.244	0.440	0.239	0.253	0.270	0.258	0.323	0.285
		MTS	0.675	0.655	**0.802**	0.713	0.728	**0.628**	0.635	0.727	0.695
		LCES	**0.748**	**0.712**	0.768	**0.733**	**0.779**	0.614	**0.699**	**0.784**	**0.730**

Adam optimizer configuration.

D Evaluation by Spearman Rank Correlation Coefficient

In addition to the primary metric QWK, we report Spearman rank correlation coefficients to evaluate the ordinal consistency between predicted and gold-standard scores. This metric is especially relevant in applications where preserving the relative ranking of essays is more important than matching exact scores. Compared with the baseline methods, LCES generally achieves higher Spearman correlations across most prompts and LLMs (Table 9), supporting its strength in maintaining rank order.

E Agreement Rate

To further validate the reliability of LLM-generated pairwise comparisons, we measure the agreement rate between LLM decisions and human annotations on a subset of evaluation pairs. We report results for two metrics (Table 10): **All**, which reflects agreement across all pairs including ties, and **Excl. Ties**, which excludes cases where the gold-standard label indicates a tie. The latter focuses on pairs where a clear score difference exists and thus better captures the LLM's ability to detect meaningful distinctions.

Table 10: Agreement rates (%) between LLMs and human evaluators in pairwise comparisons.

Model	ASAP		TOEFL11	
	All	Excl. Ties	All	Excl. Ties
Mistral-7B	55.9	58.0	52.1	41.2
Llama-3.2-3B	56.3	65.0	54.5	60.1
Llama-3.1-8B	60.3	71.6	57.6	76.6
GPT-4o-mini	59.9	75.1	55.9	86.6
GPT-4o	64.3	80.0	57.8	83.0

Better-performing LLMs such as GPT-4o and Llama-3.1-8B show higher agreement rates, particularly when ties are excluded. These results are consistent with the final scoring performance in terms of both QWK and Spearman correlation, sup-

porting the use of agreement rate as an indicator of pairwise comparison quality.