Large Language Models For Second Language English Writing Assessments: An Exploratory Comparison

Zhuang Qiu, Peizhi Yan, Zhenguang Cai

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Large Language Models For Second Language English Writing Assessments: An Exploratory Comparison

Zhuang Qiu
The City University of Macau
zhuangqiu@cityu.edu.mo

Peizhi Yan
University of British Columbia
yanpz@ece.ubc.ca

Zhenguang Cai
The Chinese University of Hong Kong
zhenguangcai@cuhk.edu.hk

Abstract

The emergence of Large Language Models (LLMs) has ushered in a new era of innovation across various domains, including second language (L2) education. While attempts to incorporate LLMs into automated essay scoring (AES) systems in L2 settings are increasing, research on employing state-of-the-art LLMs, such as RoBERTa, Llama-3, and GPT-4o, in L2 proficiency assessment remains limited. This paper reports two exploratory studies comparing the performance of four LLMs in scoring L2 English essays. In the first study, RoBERTa was fine-tuned to grade IELTS essays. In the second study, GPT-4o and Llama-3-70B-Instruct were tasked with the same grading using prompt engineering. The models' performance was evaluated by comparing their predicted scores with official IELTS scores. Notably, the fine-tuned RoBERTa model and the GPT-4o modal both achieved a human-machine correlation exceeding 0.7. Overall, LLMs demonstrated promising potential in auto-grading IELTS writing tasks. Code is available at https://github.com/PON2020/IELTSWriting.

1 Introduction

The application of artificial intelligence (AI) in education has gained increasing attention since the establishment of the International AIED Society in 1997 (Zawacki-Richter et al., 2019). The advent of Large Language Models (LLMs) marks a significant leap in educational technology, where their potential to enhance content creation, improve student engagement, and personalize learning experiences is increasingly recognized by educators (Kasneci et al., 2023). Among the various applications of AI in education, assessment stands out as a key area with immense potential to drive substantial transformation (Cope et al., 2021). Automated Essay Scoring (AES) systems, which use computer programs to evaluate written prose (Shermis, 2003), have long been proposed as a practical solution to the labor-intensive task of manual essay grading in educational settings (Page, 1966). The field of AES has seen significant advancements with the introduction of machine learning approaches, with neural network models now representing the state-of-the-art (Lagakis and Demetriadis, 2021; Xie et al., 2022). However, a systematic review by Ramesh and Sanampudi (2022) highlights that while neural network models excel in recognizing text cohesion and coherence, they still exhibit limitations in understanding logical flow and sentence connections.

The introduction of LLM-powered chatbots, such as ChatGPT (OpenAI, 2023), has significantly improved performance in various natural language processing tasks, including resolving ambiguities (Ortega-Martín et al., 2023), addressing queries (Brown et al., 2020), and facilitating multilingual translation (Jiao et al., 2023). However, despite these advancements, their performance in tasks requiring logical reasoning (Liu et al., 2023a) and understanding implicit discourse relations (Chan et al., 2024) remains limited. These challenges raise important questions about the extent to which LLM-powered chatbots can be effectively utilized in AES.

Mizumoto and Eguchi (2023) employed ChatGPT to automatically score L2 English essays from a TOEFL test-taker database and compared the ChatGPT ratings with professional human ratings. Although the authors argued that ChatGPT can be effectively used as an AES tool, their results did not demonstrate ChatGPT's superiority over existing AES methods. This lack of an expected advantage might be attributed to aspects of their study design, including the omission of prompt-tuning ChatGPT for the specific grading task (Liu et al., 2023b) and using different scales for ChatGPT ratings compared to the benchmark human ratings. Similarly, Mansour et al. (2024) explored the ef-

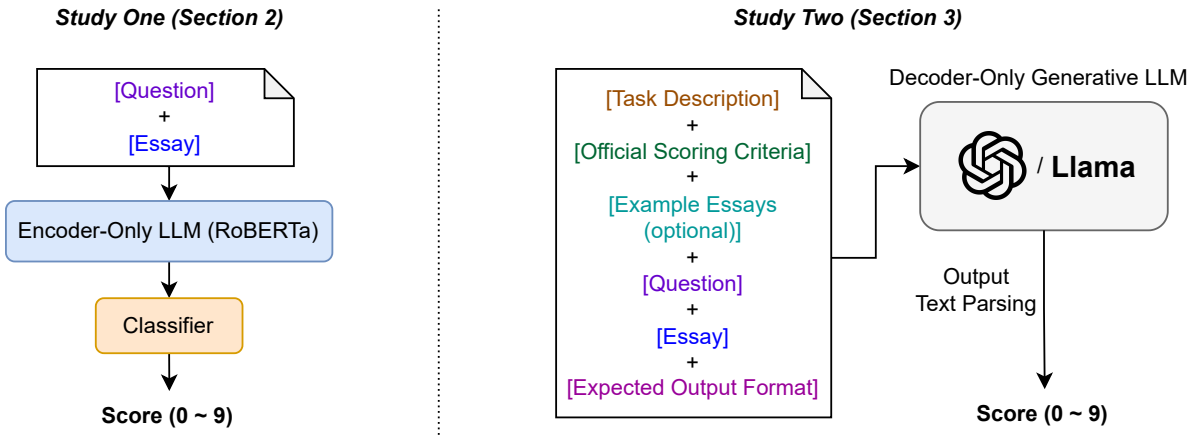

Figure 1: Workflows of Study One and Study Two. In Study One, we fine-tune the encoder-only LLM, RoBERTa, to classify IELTS essays by assigning scores. In Study Two, we use prompt engineering to guide the decoder-only generative LLMs, GPT-4o and Llama-3, in evaluating IELTS essays and generating scores.

fectiveness of prompt engineering in enhancing the performance of LLMs like ChatGPT and Llama-2 in AES. They found that while prompt engineering significantly impacts model performance, both LLMs still lag behind state-of-the-art AES models in terms of scoring accuracy, particularly when evaluated across different prompts. Sun and Wang (2024) took a more nuanced approach by employing fine-tuning and multiple regression techniques to develop a multi-dimensional scoring system for L2 English essays. Their study, which used the ELLIPSE Corpus and an official IELTS dataset, demonstrated that fine-tuned models like RoBERTa and DistilBERT could outperform existing AES methods in providing detailed, dimension-specific feedback on essays. Unlike the more holistic approach of Mizumoto and Eguchi (2023), Sun and Wang (2024)'s methodology emphasizes the need for multi-dimensional scoring to capture the varied aspects of language proficiency, such as vocabulary, grammar, and coherence. However, because official IELTS writing scores do not include a breakdown into sub-dimensional scores, they trained their models using AI-generated sub-scores, which raises questions about the validity of using model-generated scores to train other models.

We reported two exploratory studies that compared the performance of three LLMs in scoring L2 English essays. In Study One, we instructed RoBERTa (Liu, 2019) to grade IELTS essays using model fine-tuning. In Study Two, we instructed GPT-4o and Llama-3-70B-Instruct to perform the same task through prompt engineering. Figure 1 illustrates the workflows of both studies. Our project is similar to Sun and Wang (2024) in that both studies sought to explore the capabilities of LLMs in evaluating L2 English writings, using official IELTS datasets. However, our project diverges in several key aspects.

While Sun and Wang (2024) focused on fine-tuning models with AI-generated sub-scores, our approach incorporated only official IELTS scores to benchmark model performance, thereby ensuring alignment with real-world scoring criteria. Additionally, instead of relying solely on model fine-tuning, we employed prompt engineering with advanced LLMs like GPT-4o and Llama-3 to evaluate their ability to adapt to the scoring task without extensive retraining. This dual approach allows us to not only compare the efficacy of traditional fine-tuning against prompt engineering but also to assess the robustness of these models across different methodologies. By directly integrating real-world scoring standards and exploring alternative model training strategies, our research aims to provide a more comprehensive evaluation of LLMs in the context of L2 proficiency assessment.

2 Study One

IELTS writing has two tasks, task 1 and task 2. These tasks assess different English writing skills. In task 1, candidates must describe visual information, such as graphs, charts, tables, or diagrams, in a minimum of 150 words. This task focuses on summarizing and reporting key patterns or comparing data. In task 2, candidates write an essay in response to a prompt, typically involving a discussion of an issue, argument, or problem. The essay

requires a clear position, supported by reasons and examples, with a minimum of 250 words. Both tasks are assessed based on criteria such as coherence, logical flow, grammar, and vocabulary. In our studies, we focus solely on task 2, as text-based language models face challenges in interpreting the non-textual data used in task 1. Additionally, the existing publicly available datasets only provide text data, further limiting the scope of analysis for task 1. In this study, we fine-tuned the pre-trained RoBERTa model (Liu, 2019), tailoring it for the automatic scoring of responses in official IELTS writing tests (task 2). Model performance was evaluated against the official score received from human examiners.

2.1 Dataset

In this study, we utilized two publicly available datasets of IELTS writing tests. The first dataset, referred to as the "Kaggle Dataset," is available on Kaggle[1]. It contains over 1,200 essays, including more than 500 essays for task 1 and approximately 700 essays for task 2, from the International English Language Testing System (IELTS). The dataset includes columns for the task index (task 1 or task 2), the prompt (task question), the essay, and the official score. It accurately reflects the real-world scoring criteria used in high-stakes language assessments. Since IELTS writing task 1 requires the interpretation of charts and tables which could be challenging for language models, our study focused on the data of task 2 only. We randomly split the task 2 data with a 7:3 training-testing ratio, resulting in 497 training samples and 214 testing samples.

The second dataset, referred to as the "HuggingFace Dataset," is available on HuggingFace[2] and contains only task 2 essays. This dataset includes a total of 10,324 essays, with columns for the prompt, the essay, comments, and the band score. We randomly split the data into a 9:1 training-testing ratio, resulting in 9,291 training samples and 1,033 testing samples.

The difference in training-testing ratios between the two datasets (7:3 for the Kaggle dataset and 9:1 for the HuggingFace dataset) was empirically determined based on the size of the respective datasets. The Kaggle dataset, comprising only 700 Task 2 essays, required a larger test set (30%) to ensure an adequate number of samples for a meaningful evaluation. In contrast, the HuggingFace dataset contains over 10,000 essays, allowing for a smaller test set (10%) while maintaining a sufficient number of samples for robust evaluation. To address the potential bias arising from the unequal distribution of band scores, we employed random splitting of the datasets to preserve the natural score distribution. Nevertheless, we recognize that imbalances in band score representation may still impact model performance, and we plan to explore techniques such as resampling or weighted loss functions in future work to mitigate these effects.

Figures 2 and 3 illustrate the distribution of scores in the Kaggle dataset and the HuggingFace dataset, respectively.

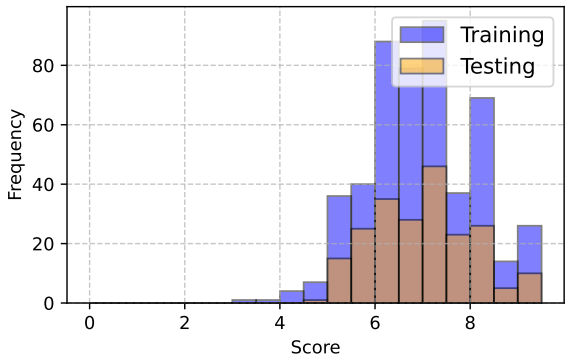

Figure 2: Score distribution of Kaggle dataset.

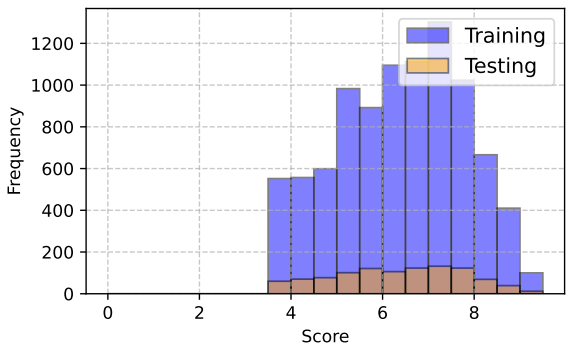

Figure 3: Score distribution of HuggingFace dataset.

2.2 Methods

We model the IELTS writing scoring as a multi-class sequence classification problem, where each sequence consists of both the prompt and the corresponding essay. The scores are discretized into 19 distinct classes, ranging from 0 to 9, including half-point increments (e.g., 0, 0.5, 1.0, 1.5, ..., 8.5,

[1] https://www.kaggle.com/datasets/mazlumi/ielts-writing-scored-essays-dataset

[2] https://huggingface.co/datasets/chillies/IELTS-writing-task-2-evaluation

9.0). To implement this model, we began with a pre-trained RoBERTa model and added a new classification head to the output vector corresponding to the special "[CLS]" token. The classification head comprises two linear layers: the first layer (hidden layer) with 768 neurons applies a tanh (hyperbolic tangent) activation function, while the second layer, which serves as the output layer, uses a softmax activation function to produce the final class probabilities.

To fine-tune the model, we used the Adam optimizer with an initial learning rate of 2×10^{-5}, which was decreased by 20% after each training epoch, with a minimum learning rate of 10^{-6}. The training loss function we use is the cross-entropy loss for multi-class classification. All model parameters were included during training. The model was trained for a total of 20 epochs with a batch size of 16. The training was performed on one Nvidia A6000 GPU. The GPU memory usage is around 12.7GB. Fine-tuning on Kaggle and HuggingFace datasets took around 4 minutes and 74 minutes respectively. The scripts for this study are publicly available via GitHub[3]

2.3 Results

We first compared two different training schemes: one where only the classifier parameters were trained ("classifier only") and another where all model parameters were trained. Table 1 presents the testing results for both the Kaggle and HuggingFace datasets. As shown in the table, training all parameters leads to an improved correlation. Specifically, we observed a 12% improvement on the Kaggle dataset and a 4% improvement on the HuggingFace dataset. These results suggest that fine-tuning all RoBERTa model parameters yields better outcomes for the IELTS writing scoring task. Therefore, we included all model parameters in the training process in our subsequent fine-tuning experiments. Figures 4 and 5 visualize the model (all parameters were fine-tuned) predictions in comparison to the ground-truth (human-evaluated) scores.

Dataset	Training Scheme	Correlation	RMSE
Kaggle	Classifier Only	0.651	0.830
Kaggle	All Parameters	**0.731**	0.784
HuggingFace	Classifier Only	0.707	0.757
HuggingFace	All Parameters	**0.735**	0.770

Table 1: Testing results with different training schemes.

[3] https://github.com/PON2020/IELTSWriting

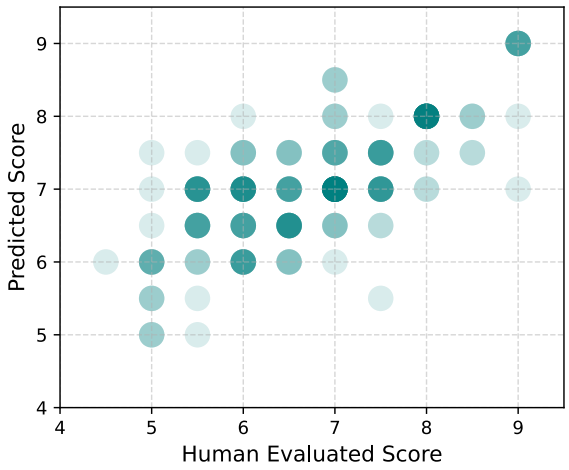

Figure 4: Scatter plot showing the predictions of the model trained on the Kaggle dataset versus the human-evaluated scores on the Kaggle test dataset.

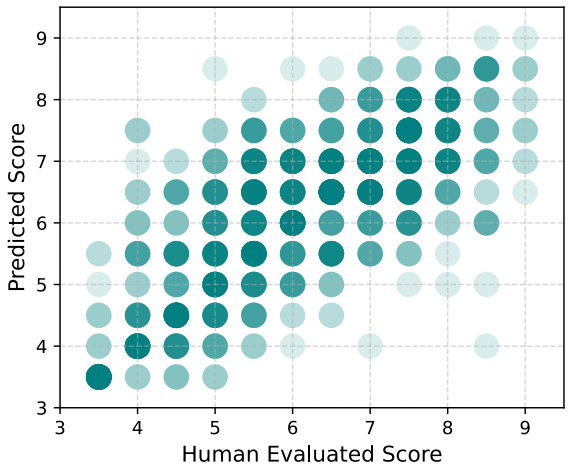

Figure 5: Scatter plot showing the predictions of the model trained on the HuggingFace dataset versus the human-evaluated scores on the HuggingFace test dataset.

Next, we explored combining both the Kaggle and HuggingFace training sets to fine-tune the model (all parameters were fine-tuned) and tested the trained model on the Kaggle testing set, HuggingFace testing set, and the combined Kaggle and HuggingFace testing sets. The evaluation results are shown in Table 2. A notable finding from this experiment is a significant improvement in the testing results on the HuggingFace dataset, with a 6% increase in correlation (from 0.735 to 0.779), compared to the model trained only on the HuggingFace training set. However, we did not observe a significant change in the results on the Kaggle dataset. Figures 6 and 7 visualize the model predictions in comparison to the ground-truth (human

evaluated) scores.

Testing Dataset	Correlation	RMSE
Kaggle	0.647	0.872
HuggingFace	0.779	0.897
Kaggle + HuggingFace	0.771	0.893

Table 2: Model performance with training on combined data and testing on different datasets.

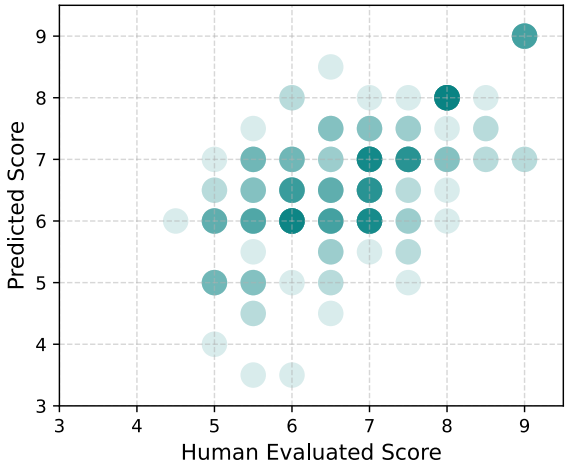

Figure 6: Scatter plot showing the predictions of the model trained on the combined dataset versus the human-evaluated scores on the Kaggle test dataset.

We believe this difference is largely due to the fact that the HuggingFace dataset contains approximately 9,300 training samples, which is about 19 times larger than the Kaggle dataset. Moreover, since the Kaggle and HuggingFace datasets contained non-overlapping questions (prompts), the combined dataset likely introduced more variety, enabling the model to generalize better on the HuggingFace testing set, which dominates in size, while maintaining performance on the smaller Kaggle dataset. This experiment highlights the importance of a large and inclusive dataset in achieving robust model performance.

Finally, we conducted an experiment to explore cross-dataset training and testing. In this experiment, we tested the model trained on the Kaggle dataset using the HuggingFace dataset, and vice versa. The results, shown in Table 3, indicate weak performance in both cases, with correlations around 0.4. As previously mentioned, the Kaggle and HuggingFace datasets cover non-overlapping questions (prompts), which likely causes the model to overfit on a single dataset. This outcome is somewhat expected, given that both datasets are still relatively small compared to those typically used for

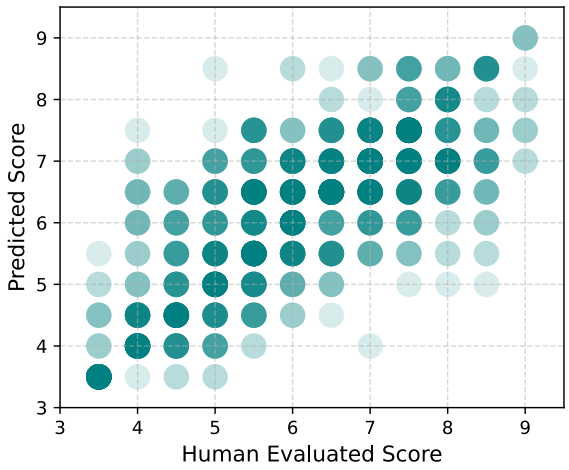

Figure 7: Scatter plot showing the predictions of the model trained on the combined dataset versus the human-evaluated scores on the HuggingFace test dataset.

training LLMs.

Training Dataset	Testing Dataset	Correlation	RMSE
Kaggle	HuggingFace	0.426	1.735
HuggingFace	Kaggle	0.386	1.488

Table 3: Model performance with training and testing on different datasets.

3 Study Two

Instead of fine-tuning LLM classifiers for AES tasks, this study evaluated the feasibility of utilizing current LLM-powered chatbots to evaluate L2 English writing through careful prompting. We compared the performance of different models across different prompts, specifically examining the impact of including or excluding example essays in the prompts. The data and script for this study are publicly available via GitHub[4].

3.1 Models

We selected GPT-4o (OpenAI, 2024) and Llama-3-70B-Instruct (Dubey et al., 2024), the most up-to-date versions from the GPT and Llama families at the time of this project. To ensure that both models evaluated the essays consistently with IELTS standards, we explicitly set their roles as "well-trained IELTS examiners". For GPT-4o, this was achieved using the OpenAI API's role-based messaging system, where the model was instructed in the system role to adopt the perspective of an

[4] https://github.com/PON2020/IELTSWriting

experienced IELTS examiner. For Llama-3-70B-Instruct, we included the instruction of "act as an IELTS examiner" in the system message.

3.2 Dataset

We took the Kaggle dataset from Study One, and selected a random subset of 400 task 2 essays for the auto-grading task. Each record in the dataset included the essay prompt, the essay itself, and the official IELTS score.

3.3 Prompt Design

We crafted detailed prompts to instruct the models to evaluate essays according to the official IELTS band descriptors. The prompts emphasized key scoring criteria such as task achievement, coherence and cohesion, lexical resource, and grammatical range and accuracy. Importantly, the study design included two variations of the prompts: (1) With Example Essays: These prompts included scoring criteria and example essays corresponding to various band scores to guide the model's understanding and evaluation process. (2) Without Example Essays: These prompts provided the same scoring criteria but excluded the example essays, allowing us to compare the impact of including such examples on the models' scoring performance.

3.4 Score Generation and Validation

For each essay in the dataset, the models were tasked with generating a score based on the provided prompt. To ensure the reliability of the scores, the models generated two independent scores for each essay. These scores were averaged if they differed by no more than two points. If the scores diverged by more than two points, the essay was re-evaluated up to three times to achieve consistency. Only scores within the valid range of 0 to 9 were considered, and any invalid or missing scores were flagged and handled accordingly.

3.5 Performance Evaluation

The models' performances were evaluated by calculating the Root Mean Square Error (RMSE) between the model-generated scores and the official IELTS scores, providing a measure of the models' accuracy. Additionally, we computed the Pearson correlation coefficient to assess the linear relationship between the model-generated scores and the human ratings. The study also compared the performance of each model across the two prompt variations (with and without example essays) to determine the impact of this variable on scoring accuracy.

Model	Prompt Type	Correlation	RMSE
GPT-4o	with example	0.71	1.05
GPT-4o	without example	0.72	1.13
Llama-3	with example	0.56	1.25
Llama-3	without example	0.63	0.99

Table 4: Summary of Model Performance in Study Two

3.6 Results

As shown in Table 4, GPT-4o in general outperformed Llama-3 in this task. When example essays of band level 9-3 were included into the prompt, the correlation between GPT-4o's scores and official examiners' scores was 0.71, and the RMSE in predicting official IELTS writing score was 1.05. These figures did not change much when we excluded example essays from the prompt. When GPT-4o was prompted with IELTS writing band descriptors without example essays, the correlation between human and model score was 0.72 and the RMSE of predicting official IELTS writing score was 1.13. Different from GPT-4o, Llama-3's performance in the grading task was noticeably influenced by the two types of prompts. When example essays were included in the prompt, the correlation between Llama-3 scores and official examiners' scores was 0.56, and the RMSE in predicting official IELTS writing score was 1.25. Interestingly, the performance of Llama-3 improved noticeably when example essays were removed from the prompt. When it was prompted with IELTS writing band descriptors without example essays, the correlation between human and Llama-3 score was 0.63 and the RMSE of predicting official IELTS writing score was 0.99.

4 Discussion

This study explored the application of the state-of-the-art LLMs in the automated scoring of L2 English writing, specifically using the Cambridge IELTS dataset due to its well-established reliability and validity as a measure of English proficiency (Schoepp, 2018). The use of this dataset not only provided a robust foundation for our experiments but also ensured that our findings were grounded in a widely recognized assessment standard. In evaluating model performance, we selected RMSE as our primary metric. RMSE was chosen for its interpretability within the context of the IELTS grading scale. Specifically, an RMSE value of less than 1

indicates that, on average, the model's predicted scores deviate from the true IELTS scores by less than one point on a 9-point scale. This metric is particularly useful for educators and assessment professionals who are accustomed to the IELTS scoring system. However, the use of RMSE also presents a challenge when comparing our results with those of other studies, such as Sun and Wang (2024), which used QWK as their performance measure. The difference in metrics complicates direct comparisons, particularly with studies that employed different datasets and scoring scales (such as Mansour et al. (2024); Mizumoto and Eguchi (2023). Future work should consider reporting multiple performance metrics to facilitate broader comparisons across different AES studies.

The fine-tuning experiments in Study One underscore the crucial role of the training dataset in determining the model's performance. In contrast, the generative LLM models (GPT-4o and Llama-3-70B-Instruct) used in Study Two are significantly larger than the encoder-only RoBERTa model and were trained on vast datasets to develop general capabilities. Despite lacking prior knowledge of the new datasets and tasks, these generative models exhibit promising performance. Future research could explore bridging the gap between encoder-only and generative LLMs by leveraging the fine-tuning efficiency of encoder-only models and the generalization strength of generative models, potentially achieving even better performance in AES.

In Study Two, both GPT-4o and Llama-3-70B-Instruct were tasked with grading the same subset of 400 task 2 essays under different prompt conditions. Each model graded the dataset only once per prompt type, serving as a proof of concept for the potential of generative AIs in mimicking human educators in the scoring of L2 English writing. The findings suggest that, with appropriate prompt engineering, these models can achieve a level of grading consistency and accuracy that aligns with human scoring. However, the experiment also underscores the need for further exploration into how different prompts and model configurations affect scoring outcomes. Understanding the distribution of RMSE and correlation coefficients across various models and prompt types will be a key focus of our future research. This will help us refine the prompt engineering process and optimize model performance.

In summary, our findings provide insights into the ongoing exploration of LLMs in AES, specifically within the context of IELTS – a domain that has not been widely explored with the most up-to-date models. While the usefulness of LLMs in AES has been previously demonstrated, our study uniquely tests the capabilities of the latest models, such as GPT-4o, Llama-3-70B-Instruct, and RoBERTa, in grading IELTS essays. Both fine-tuning and prompt engineering emerged as effective approaches. The results from our studies also underscore the importance of dataset selection. Future work will focus on a deeper exploration of how factors such as model type, prompt design, and dataset characteristics influence the performance and reliability of LLMs in AES tasks.

References

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Bill Cope, Mary Kalantzis, and Duane Searsmith. 2021. Artificial intelligence for education: Knowledge and its assessment in ai-enabled learning ecologies. *Educational philosophy and theory*, 53(12):1229–1245.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

W. Jiao, W. Wang, J. T. Huang, X. Wang, and Z. Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Paraskevas Lagakis and Stavros Demetriadis. 2021. Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6. IEEE.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023a. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023b. Gpt understands, too. *AI Open*.

Y Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? *arXiv preprint arXiv:2403.06149*.

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. 2023. Chatgpt. https://www.openai.com/chatgpt. Accessed: 2024-08-27.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-08-27.

M. Ortega-Martín, Ó. García-Sierra, A. Ardoiz, J. Álvarez, J. C. Armenteros, and A. Alonso. 2023. Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

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

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Kevin Schoepp. 2018. Predictive validity of the ielts in an english as a medium of instruction environment. *Higher Education Quarterly*, 72(4):271–285.

MD Shermis. 2003. Automated essay scoring: A cross-disciplinary perspective.

Kun Sun and Rong Wang. 2024. Automatic essay multi-dimensional scoring with fine-tuning and multiple regression. *arXiv preprint arXiv:2406.01198*.

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.

Olaf Zawacki-Richter, Victoria I Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):1–27.