Modeling Score Estimation for Japanese Essays with Generative Pre-trained Transformers

Boago Okgetheng and **Koichi Takeuchi**
Graduate School of Environmental, Life, Natural Science and Technology
Okayama University, Japan
pcqm1k3t@s.okayama-u.ac.jp takeuc-k@okayama-u.ac.jp

Abstract

This paper presents a study on Japanese essay grading using Generative Pre-trained Transformers (GPTs) in Japanese language. Previous research has demonstrated the effectiveness of neural network-based models, such as BERT, for essay grading across various datasets. With the advent of downloadable GPT models trained on significantly larger datasets compared to BERT, it has become feasible to employ these models for essay grading through fine-tuning with Low-Rank Adaptation (LoRA). Most existing models have focused on English essays and their accuracy, leaving a gap in understanding the performance on Japanese essays, which have limited linguistic resources. To address this, we apply several Japanese GPT models to a dataset comprising 12 prompts across 4 themes. The experimental results show that the model pre-trained exclusively on Japanese data, open-calm-medium, achieved an accuracy of 62.33% and a QWK of 0.5551. In comparison, the best-performing model additionally pre-trained on multilingual Llama, ELYZA-Llama-2-7b-fast, achieved an accuracy of 53.29% and a QWK of 0.3375. This study highlights the potential of GPT models for enhancing automated essay scoring in the Japanese context.

1 Introduction

Automated essay scoring (AES) is one of the most promising and rapidly evolving fields in educational technology owing to the growing opportunities of online lectures.

Previous studies first revealed neural network-based models such as LSTM and CNN are effective for essay tasks (Taghipour and Ng, 2016; Dong et al., 2017; Yi Tay and Minh C. Phan and Luu Anh Tuan and Siu Cheung Hui, 2018). A neural network-based essay scoring model is roughly divided into two parts: encoding an essay to a vector and assigning scores. After a pre-trained language model BERT (Devlin et al., 2019) has succeeded in improving the accuracy of benchmarks in NLP, some previous studies have applied simple BERT-based models into essay scoring task (Rodriguez et al., 2019; Mayfield and Black, 2020). The simple models were unable to improve the accuracy of existing neural network-based models. The newly proposed models, however, combining regression and ranking loss show improved performance comparing to the existing neural network-based models (Yang et al., 2020; Wang et al., 2022).

Thus, the previous studies have revealed pre-trained language models are effective for AES. In the recent advancements in Generative Pre-trained Transformers (GPTs) (Brown et al., 2020; OpenAI et al., 2023), which have much larger weight size and are trained on extensive datasets, several studies have explored the application of GPTs, both with and without fine-tuning (Mizumoto and Eguchi, 2023; Xiao et al., 2024). It has been observed that a prompt-based GPT model yields lower accuracy compared to the fine-tuned GPT-3.5 or BERT-based model (Xiao et al., 2024).

The findings of the models studied above have been often conducted on the commonly used English essay dataset ASAP (Hamner et al., 2012), but on the other hand, it is not clear how much prediction accuracy can be achieved for Japanese essays, where linguistic resources are limited. There are studies conducted on Japanese essay written by Japanese learners (Hirao et al., 2020; Obata et al., 2023); however, Japanese essay data (Takeuchi et al., 2021)[1] written by native Japanese speakers that can be used for research has recently been published, thus, in this paper, we conduct on the study of essay scoring model for Japanese.

Previous studies show that the fine-tuned language models based on BERT or GPT-3.5 are promising for AES task (Hirao et al., 2020; Xiao

[1]GSK2021-B https://www.gsk.or.jp/catalog/gsk2021-b/

et al., 2024). Thus, the middle size of downloadable GPT models such as Llama (Touvron et al., 2023) are worth to be applied into Japanese essay scoring task because of the following reasons: 1) API-based GPTs such as GPT-3.5 have limitations of learning while we can freely build an essay grading model that incorporate the downloaded GPT, 2) it is expected that linguistic knowledge within a GPT will contribute to solve the grading of Japanese essays, and 3) Low-Rank Adaptation (LoRA) (Hu et al., 2021) enables us to apply fine-tuning on a local GPU at a laboratory scale.

Several Japanese GPT models that are specifically pre-trained on Japanese texts are published; however, it is not clear which model is suitable for Japanese essay scoring task. The dataset includes Japanese essays to 12 prompts consists of 4 themes, which ranges in length from 100 to 800 characters. Therefore, in this paper, we clarify the performance of the several Japanese GPT models for the Japanese essay dataset and discuss the relations between GPTs and features of essays.

The contributions of this study are as follows: 1) it unveils Quadratic Weighted Kappa (QWK) and F1 scores achieved for Japanese essays using a Japanese GPT model, 2) it provides a comparative analysis of the performance across various Japanese GPT models employing Low-Rank Adaptation (LoRA) fine-tuning on Japanese essay datasets, and 3) it reveals that GPT models initially trained on Japanese texts outperform the model subjected to additional pre-training on multilingual Llama model using Japanese texts.

2 Previous Studies

In the initial phases of AES development, a variety of statistical models were employed. These included regression models that relied on hand-crafted features, exemplified by systems like e-rater (Attali and Burstein, 2006), as well as statistical approaches utilizing latent semantic indexing (LSI) (Deerwester et al., 1990; Ishioka and Kameda, 2006).

Neural network models that do not require hand-crafted features has been proposed and shown to be superior to previous models. Many studies used LSTM and CNN models (Taghipour and Ng, 2016; Dong et al., 2017; Yi Tay and Minh C. Phan and Luu Anh Tuan and Siu Cheung Hui, 2018), but there is also a study using word embedding and Support Vector Regression model (Cozma et al., 2018) that achieved an equivalent performance to the neural network-based models (Mayfield and Black, 2020).

Instead of learning sentence embedding directly from target data, pre-trained language models are employed (Rodriguez et al., 2019; Mayfield and Black, 2020; Yang et al., 2020; Wang et al., 2022; Mizumoto and Eguchi, 2023; Xiao et al., 2024; Hirao et al., 2020; Obata et al., 2023). Pre-trained models can be broadly divided into BERT (Rodriguez et al., 2019; Mayfield and Black, 2020; Yang et al., 2020; Hirao et al., 2020; Wang et al., 2022) and GPT (Mizumoto and Eguchi, 2023; Obata et al., 2023; Xiao et al., 2024). Although the initial model using BERT could not achieve high accuracy, it was shown that adding ranking to the loss function improved accuracy and outperformed neural network-based models (Yang et al., 2020; Wang et al., 2022). The prompt-based GPT model showed the limited performance compared to the linguistic feature-based model (Mizumoto and Eguchi, 2023; Obata et al., 2023) or fine-tuned GPT-3.5 model (Xiao et al., 2024). This indicates that significant large language model is not so effective for AES.

While most of the previous studies are conducted on English essay dataset, studies on Japanese essay are limited. Hirao et al. (2020) revealed that the BERT-based model is effective compared to the LSTM-based model on Japanese essay dataset[2]. The other Japanese essay dataset used in Obata et al. (2023) contains essays for one prompt[3]. Preliminary experiments have been conducted to predict scores for Japanese essay data by fine-tuning Japanese GPT models (Okgetheng and Takeuchi, 2024).

Thus, evaluating essay scoring models using a Japanese essay dataset—comprising essays of various lengths and themes, based on data available for research—is deemed valuable.

3 Methodology

3.1 Essay Scoring Model

The essay scoring model comprises two main modules: text encoding and score assignment. The encoding module leverages pre-trained language models to convert the input text into vector representations, while the score assignment module

[2]https://goodwriting.jp/wp/?lang=en

[3]That is included in I-JAS corpus https://www2.ninjal.ac.jp/jll/lsaj/.

utilizes these representations to predict scores. The models employed in this study include Japanese BERT, Open CALM, CALM2-7B, StableLM Alpha, and ELYZA, each designed specifically for handling Japanese texts.

Japanese BERT[4] is used for text encoding, where the vector corresponding to the [CLS] token serves as the embedding vector for the input essay. In contrast, decoder-only models such as Open CALM[5], CALM2-7B[6], Japanese StableLM Alpha[7], and ELYZA[8] are utilized for both encoding and score prediction. For these GPT-based models, the vector that predicts the next token after the final token of the input essay is used as the embedding vector.

Given an input essay document s with tokens x_1 to x_n generated by the tokenizer, the final token embedding is used for predicting the score. Specifically, for models like Open CALM, the vector corresponding to the token that denotes the end of the input document is used. Figure 1 illustrates the overall architecture of the essay scoring model.

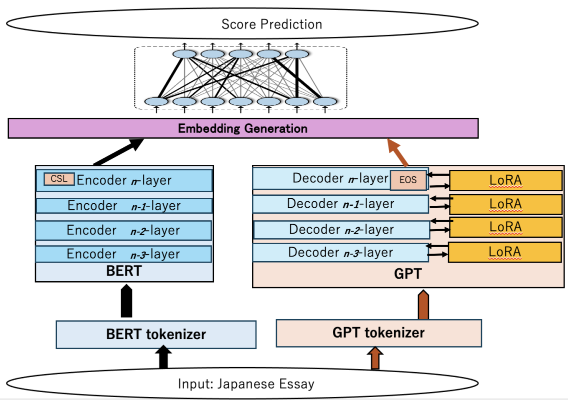

Figure 1: Methodology for the Neural Network-based Essay Scoring Model

3.2 Score Prediction from Embeddings

To predict the score from the embeddings, the final embedding vector (obtained either from the [CLS] token for BERT or the end-of-sequence token for GPT models) is passed through a fully connected neural network. This network consists of multiple layers that map the high-dimensional embeddings to a single score value representing the predicted essay score. The design of this neural network, including the number of layers and activation functions, is optimized to capture the nuanced relationships between the encoded text and the target scores.

3.3 Design of the Loss Function

Given that the proposed model is a categorical classification model where the classes are ordinal, we applied soft labeling(Diaz and Marathe, 2019) to the loss function. During the training phase, the loss for the categorical model is calculated using cross-entropy with one-hot labels. Soft labeling modifies the target labels such that the k-th value is calculated as follows:

$$d_k = \frac{exp(-|\hat{k} - k|)}{\sum_{i=1}^{K} exp(-|\hat{k} - i|)} \qquad (1)$$

Here, d_k represents the teacher value for each k-th unit in the final layer of the classification model, and $\hat{k}$ denotes the correct category. This approach assigns a larger penalty for predictions that are further from the correct answer, promoting better ordinal classification.

4 Experimental Setup

4.1 Dataset

The Japanese essay tests were conducted on Japanese university students, and the dataset consists of 12 prompts with 4 themes. In each theme, there are three prompts. The four themes are globalization (Global), natural science (Natural), East Asian economics (Easia), and critical thinking (Criticize). Each theme has three prompts from question 1 to 3. The length of the essays ranges from 100 characters to 800 characters.

The essays are manually scored on a 5-point scale for comprehension, logic, validity, and grammar. In this paper, we focus on comprehension scores to evaluate the essay scoring models. The essays were annotated by two Japanese-speaking raters, and the scores were averaged to obtain the final score for each essay.

The Japanese essay data is available to researchers and is provided by the Japanese Language Resource Association (GSK)[9]. Table 1 shows the number of essays for each prompt. In the table, 'P' stands for Prompt number, 'ML' represents the Maximum Length of an essay, and 'Num' indicates the number of essays.

[4] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[5] https://huggingface.co/cyberagent/open-calm

[6] https://huggingface.co/cyberagent/calm2-7b

[7] https://huggingface.co/stabilityai/japanese-stablelm-base-alpha-7b

[8] https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

[9] https://www.gsk.or.jp/en/

This dataset provides a diverse range of essay lengths and topics, enabling a comprehensive evaluation of the essay scoring models.

Table 1: Japanese essay data

Theme	P	ML	Num	Theme	P	ML	Num
	1	100	290		1	300	328
Criticize	2	400	290	Global	2	250	327
	3	800	290		3	300	327
	1	300	290		1	100	327
Easia	2	250	288	Science	2	400	325
	3	300	288		3	800	327

4.1.1 Example of Global Category Prompts: Japanese and English Versions

In the global category, the essay prompts challenge students to critically analyze various aspects of globalization. For example, Prompt 1 asks: **Japanese:** グローバリゼーションは、世界、または各国の所得格差をどのように変化させましたか。また、なぜ所得格差拡大、または縮小の現象が現れたと考えますか。300字以内で答えなさい。 **English:** How has globalization changed income inequality in the world or across countries? Also, why do you think the phenomenon of increasing or decreasing income inequality has appeared? Please answer within 300 characters.

Prompt 2 shifts focus to multinational corporations, asking: **Japanese:** 多国籍企業は、グローバリゼーションの進展の中でどのような役割を果たしましたか。多国籍業の具体例をあげて、250字以内で答えなさい。 **English:** What role have multinational corporations played in the development of globalization? Give a specific example of a multinational business and answer within 250 characters.

Lastly, Prompt 3 delves into cultural aspects, asking: **Japanese:** 文化のグローバリゼーションは、私たちの生活にどうのような影響を与えましたか。また、あなたはそれをどのように評価しますか。具体例をあげて、300字以内で答えなさい。 **English:** How has cultural globalization affected our lives? Also, how do you rate it? Give a specific example and answer within 300 characters.

4.2 Score Distribution Across Themes

The score distribution across different essay themes and prompts provides valuable insights into the grading trends and the level of challenge posed by each prompt. Figure 2 illustrates how scores were allocated across five possible score levels (1 to 5) for each theme and prompt within the dataset. This

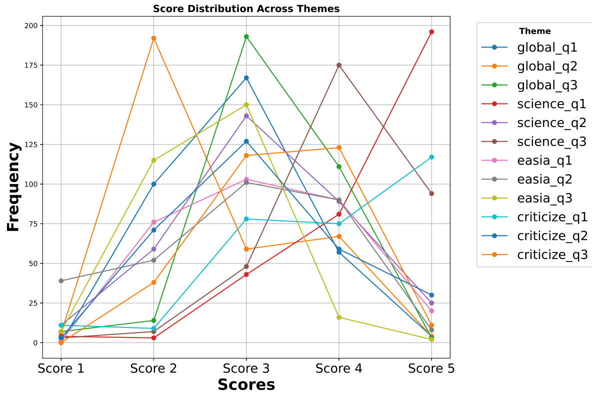

Figure 2: Scores Distribution per theme

distribution highlights the variability in grading across different prompts, with some prompts showing a higher concentration of scores in the middle ranges (Scores 2 and 3), while others have a significant number of essays scored at the higher end (Score 5), particularly in themes like **science_q1.**

4.3 Performance Measures

To evaluate the effectiveness of our model, we employed several performance metrics:

- Accuracy: This metric provided a straightforward measure of the model's ability to correctly predict the essay scores.
- Root Mean Square Error (RMSE): RMSE offered a quantitative measure of the model's prediction error, giving insights into the deviation of the predicted scores from the actual scores.
- Quadratic Weighted Kappa (QWK): QWK was used to assess the degree of agreement between the predicted and actual essay scores. This metric is particularly valuable in grading scenarios, as it accounts for the ordered nature of the rating scale.

4.4 Training Setup

Our setup involved the following key components:

- GPT Configuration: We utilized GPT models specifically configured for the Japanese language, ensuring that they are finely attuned to the linguistic characteristics unique to Japanese.

- Early Stopping: To prevent overfitting, we employed an early stopping mechanism. Training ceased once the improvement in performance on the validation set plateaued, ensuring the generalizability of the model.

- Gradient Accumulation: Recognizing the computational demands of training large language models, we implemented a gradient accumulation strategy. By setting the accumulation steps to 2 with a batch size of 8, we effectively simulated a larger batch size of 16, allowing for more stable and effective training.

- LoRA: We applied LoRA (Low-Rank Adaptation) implemented in PEFT (Parameter-Efficient Fine-Tuning) by HuggingFace with the rank set to 8.

- Training Configuration: Models were trained over a maximum of 10 epochs with early stopping criteria to prevent overfitting.

5 Experimental Results

In our experiments, we employed a 5-fold cross-validation technique to ensure the robustness and reliability of our results. Each model was trained with a batch size of 8, and we used a gradient accumulation step of 2, effectively making the batch size 16. The models were trained for a maximum of 10 epochs, with early stopping criteria to prevent overfitting.

The performance metrics used in our evaluation include F1 Score, QWK, Accuracy, and RMSE. These metrics provide a comprehensive evaluation of the models' capabilities in handling classification tasks, measuring the agreement between predicted and actual scores, assessing the proportion of correct predictions, and quantifying the average magnitude of prediction errors, respectively.

5.1 Overall Performance

Table 2 presents the overall performance of various models with and without soft labeling.

This table shows that models such as calm2-7b and open-calm-large perform consistently well across all metrics. Specifically, calm2-7b without soft labeling achieves the highest QWK (0.5982) and a relatively low RMSE (0.6957), indicating strong agreement with the true scores and precise predictions. In contrast, the F1 scores are generally higher for models without soft labeling, suggesting a better precision-recall balance when soft labels are not used.

5.2 Category-wise Performance

Table 3 illustrates the performance of different models across various essay categories with and without soft labeling. The results in this table are for the models that performed best in each category.

In the Criticize category, the calm2-7b model without soft labeling outperforms other models, achieving a QWK of 0.5831 and RMSE of 0.7133. The Easia category shows similar trends, with calm2-7b again performing best without soft labeling. For the Science category, the open-calm-medium model with soft labeling achieves the highest QWK of 0.7092, indicating strong performance in more technical essays.

5.3 Prompt-wise Performance

Table 4 provides the performance across different prompts with and without soft labeling. In this table, we are showing the results of the models that performed better than the others in each prompt.

For Prompt 1, the jp(Japanese)-stablelm-instruct-7b-v2 model without soft labeling achieves the highest QWK of 0.7356, indicating a strong agreement with human scoring. Prompt 2 shows the ELYZA-Llama-2-7b-fast-instruct model performing well, with balanced accuracy and F1 score. The calm2-7b model remains consistent across different prompts, showcasing its versatility.

5.4 Performance Comparison

Table 5 compares the performance of classification models with soft labeling, without soft labeling, and regression models.

Table indicates that regression models generally outperform classification models in terms of RMSE, indicating more precise error minimization. Soft labeling improves performance for medium and large models, but its benefits are less clear for small models. QWK and Accuracy metrics show balanced performance across all model types, with regression models slightly ahead in precision.

6 Discussions

The analysis of various models on the Japanese essay scoring task demonstrates that some models exhibit a high degree of proficiency within certain thematic areas. This is evidenced by their consistently strong performance across most evaluated

Table 2: Overall Performance of GPT Models

	Model	F1 Score	QWK	Accuracy	RMSE
With Soft Labeling	open-calm-small	0.2803	0.3417	0.5677	0.7855
	open-calm-medium	0.3284	**0.5303**	0.5899	**0.7243**
	open-calm-large	0.3502	0.5272	**0.6208**	0.7282
	open-calm-7b	0.3072	0.4362	0.5963	0.7787
	calm2-7b	0.3252	0.5288	0.6001	0.7417
	calm2-7b-chat	0.3109	0.4512	0.5873	0.7761
	jp-stablelm-alpha-7b	0.2961	0.4201	0.5652	0.7933
	jp-stablelm-instruct-7b-v2	0.3372	0.4750	0.5886	0.7788
	ELYZA-Llama-2-7b-instruct	0.2909	0.3760	0.5305	0.8980
	ELYZA-Llama-2-7b-fast	0.2415	0.3105	0.5216	0.8884
	ELYZA-Llama-2-7b	0.3372	0.4716	0.5930	0.7728
	ELYZA-Llama-2-7b-fast-instruct	0.3115	0.4376	0.5481	0.7893
	BERT	**0.5056**	0.4318	0.5602	0.7863
Without Soft Labeling	open-calm-small	0.2910	0.3848	0.5679	0.8112
	open-calm-medium	0.3621	0.5551	**0.6233**	0.7259
	open-calm-large	0.3772	0.5614	0.6219	0.7053
	open-calm-7b	0.3370	0.5068	0.6089	0.7279
	calm2-7b	0.3872	**0.5982**	0.6140	**0.6957**
	calm2-7b-chat	0.3303	0.4994	0.6072	0.7332
	jp-stablelm-alpha-7b	0.3518	0.5367	0.6072	0.7332
	jp-stablelm-instruct-7b-v2	0.3362	0.4690	0.5918	0.7829
	ELYZA-Llama-2-7b-instruct	0.3143	0.4501	0.5274	0.8365
	ELYZA-Llama-2-7b-fast	0.2630	0.3375	0.5329	0.9217
	ELYZA-Llama-2-7b	0.3526	0.4843	0.5768	0.8207
	ELYZA-Llama-2-7b-fast-instruct	0.3260	0.4495	0.5520	0.8053
	BERT	**0.4681**	0.3352	0.5450	0.8433

Table 3: Category-wise Performance of GPT Models

	Category	Model	QWK	RMSE	Accuracy	F1 Score
With Soft Labeling	Criticize	jp-stablelm-instruct-7b-v2	0.5239	0.7287	0.6061	0.3395
	Easia	calm2-7b	0.5129	0.6259	0.6919	0.3119
	Global	open-calm-large	0.5593	0.7810	0.5690	0.3857
	Science	open-calm-medium	0.7092	0.6604	0.6667	0.4515
Without soft labeling	Criticize	calm2-7b	0.5831	0.7133	0.5960	0.3805
	Easia	calm2-7b	0.5886	0.6280	0.6818	0.3620
	Global	calm2-7b-chat	0.5585	0.6511	0.6149	0.4092
	Science	jp-stablelm-alpha-7b	0.7050	0.6565	0.6061	0.4277

Table 4: Prompt-wise Performance of GPT Models

	Prompt	Model	QWK	RMSE	Accuracy	F1 Score
With Soft Labeling	1	jp-stablelm-instruct-7b-v2	0.6881	0.6541	0.6869	0.4352
	2	calm2-7b-chat	0.6963	0.7388	0.5606	0.3603
	3	open-calm-large	0.4243	0.7100	0.6300	0.3082
Without Soft Labeling	1	jp-stablelm-instruct-7b-v2	0.7356	0.6070	0.7355	0.4835
	2	ELYZA-Llama-2-7b-fast-instruct	0.6920	0.6931	0.5990	0.3932
	3	calm2-7b	0.4373	0.6922	0.5917	0.3440

Table 5: Performance Comparison using Classification Model with Soft Labeling (WS), Without Soft Labeling (WOS) and Regression Model (RM)

	Small			Medium			Large		
Metric	WS	WOS	RM	WS	WOS	RM	WS	WOS	RM
F1 Score	0.2803	0.2910	**0.5109**	0.3284	0.3621	**0.5552**	0.3502	0.3772	**0.5358**
QWK	0.3417	0.3848	**0.3872**	0.5303	**0.5551**	0.4521	0.5272	**0.5614**	0.3528
Accuracy	0.5677	**0.5679**	0.5441	0.5899	**0.6233**	0.5980	0.6208	**0.6219**	0.5882
RMSE	0.7855	0.8112	**0.6826**	0.7243	0.7259	**0.6511**	0.7282	0.7053	**0.6793**

metrics. Such results suggest that these models do better on predicting scores in that thematic area.

While BERT's performance was not the strongest, it did achieve commendable results in the F1 measure across all themes, indicating a balanced precision and recall in the classification task. However, in comparison to GPT models, BERT was surpassed in other key metrics, suggesting that while BERT is proficient in identifying relevant instances, GPT models may offer a more comprehensive understanding of the dataset, reflecting a deeper contextual grasp that extends beyond mere classification accuracy.

The analysis of prompt lengths in relation to essay difficulty reveals that longer prompts, such as Criticize prompt 3 and Science prompt 3, do not necessarily correlate with increased challenge levels. Contrastingly, Prompt 2 stands out, where despite its shorter length, human graders scored it as more difficult, indicating that the inherent complexity of a prompt and the resultant essay responses are not solely determined by length. This insight suggests that prompt difficulty could be influenced by the intricacy of the topic and the cognitive demands it places on the essay writers.

The research sought to gain deeper insights into the effectiveness of using a Regression Model (RM) for classification tasks and results were recorded in Table 5 for 3 GPT models (calm small, medium and large). In the Japanese essay scoring task, it was found that models employing the classification model with soft labeling (WS) generally had superior performance in terms of QWK compared to those using the classification model without soft labeling (WOS) and the regression model . This suggests that soft labeling models are better at accounting for the ordinal nature of the grading task. Although the regression models using Mean Square Error loss achieved the highest F1 Scores, this did not consistently extend to higher accuracy or QWK. Such findings indicate that while RM is proficient at minimizing the variance of the errors, it may not always translate into the most accurate categorization, especially when the task requires understanding the ordered grading system.

When evaluating the differences in the pre-training methods among the models in Table 2, the GPT models trained on Japanese texts from the beginning (i.e., open-calm, calm2-7b and jp-stable models) outperform the model subjected to continual pre-training on multilingual Llama model (i.e., ELYZA) for Japanese texts. Since there is only one model of continuous pre-trained model, however, this outcome presents intriguing prospects for future insights into pre-trained models.

7 Conclusions

In this paper, we have expanded the AES field by applying GPTs to Japanese essay grading—a linguistic domain previously underexplored due to limited resources. Our research demonstrates that Japanese-specific pre-trained GPT models, particularly when fine-tuned with LoRA, can effectively navigate the complex linguistic landscape of Japanese and provide accurate essay assessments. The research revealed that models pre-trained exclusively on Japanese corpora outperformed their counterparts fine-tuned from multilingual datasets, highlighting the importance of tailored linguistic training in automated essay scoring systems.

The calm2-7b model demonstrated exceptional capability, consistently achieving high scores across various evaluation metrics, including QWK and RMSE especially in Easia theme. Its robust performance across this topic underscores its suitability as a precise and reliable tool for the automated grading of Japanese essays in this thematic area.

This study not only contributes a significant finding to the field of educational technology but also opens avenues for the deployment of language-specific automated grading tools.

8 Limitations

The study faced limitations in data availability, model architecture, and computational resources, particularly GPU memory constraints, which may have impacted the training efficiency and model performance.

9 Ethical Considerations

Ethical considerations were rigorously adhered to, ensuring the protection of individual privacy. The dataset did not contain any personal information, guaranteeing the anonymity of all individuals involved. The data employed is publicly available, reinforcing the ethical integrity of our research.

Acknowledgements

Part of this study was supported by JSPS KAKENHI Grandt Number JP22K00530.

References

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment*, 4(3):1–30.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. arXiv:2005.14165.

Madalina Cozma, Andrei M. Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *R.E. Asher (Editor-in-Chief), The Encyclopedia of Language and Linguistics, Vol.6, Oxford: Pergamon Press*, pages 3168–3171.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(7):391–407.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4738–4746.

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 (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring.

Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. 2020. Automated essay scoring system for nonnative Japanese learners. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1250–1257, Marseille, France. European Language Resources Association.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv:2106.09685.

Tsunenori Ishioka and Masayuki Kameda. 2006. Automated Japanese Essay Scoring System based on Articles Written by Experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

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

Ayaka Obata, Takumi Tagawa, and Yuichi Ono. 2023. Assessment of ChatGPT's validity in scoring essays by foreign language learners of japanese and english. In *Proceeding of the 15th International Congress on Advanced Applied Informatics*.

Boago Okgetheng and Koichi Takeuchi. 2024. Estimating japanese essay grading scores with large language models. In *Proceedings of the 30th Annual Conference on Natural Language Processing (NLP)*, pages 643–647, Japan. NLP Society. This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. arXiv:2303.08774.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring. arXiv:1901.07744.

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, Austin, Texas. Association for Computational Linguistics.

Koichi Takeuchi, Masayuki Ohno, Kouta Motojin, Masahiro Taguchi, Yoshihiko Inada, Masaya Iizuka, Tatsuhiko Abo, and Hitoshi Ueda. 2021. Development of Essay Scoring Methods Based on Reference Texts with Construction of Research-Available Japanese Essay Data. In *IPSJ Journal Vol.62 No.9*, pages 1586–1604. (in Japanese).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. arXiv:2401.06431.

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: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Yi Tay and Minh C. Phan and Luu Anh Tuan and Siu Cheung Hui. 2018. SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text. In *Proceedings of the Thirty-Second AAAI Conference on Artifical Intelligence*, pages 5948–5955.