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

Anonymous ACL submission

Abstract

This paper describes Japanese essay grading models with Generative Pre-trained Transformers (GPTs) in Japanese. Previous studies of essay grading show that neural network based models utilizing pre-trained language model such as BERT are effective for several essay data. With the recent rapid development of downloadable GPTs, which are trained on significantly larger datasets compared to BERT, it has become feasible to employ GPTs for the task of essay grading through fine-tuning with Low-Rank Adaptation (LoRA). Most models in previous studies have been applied to English essay data and evaluated for the accuracy, but it is not clear how much prediction accuracy can be achieved for Japanese essays, where linguistic resources are limited. Thus, we apply several Japanese GPTs into Japanese essay data with 12 prompt composed of 4 themes. The experimental results show that a model pre-trained from the beginning with Japanese data has higher accuracy than a model additionally pre-trained from multilingual Llama.

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 a sentence 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 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

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

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 handcrafted 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 handcrafted 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]. 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

A neural network-based essay scoring model is roughly divided into two parts: encoding a sentence to a vector and assigning scores. The pre-trained language models are employed for the encoding part. The employed Japanese pre-trained language models are Japanese BERT[4], open-calm models, calm2-7b models[5], Japanese StableLM Alpha models[6], and ELYZA[7]. In these models, the ELYZA models are built by applying continual pre-

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

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

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

[5]https://huggingface.co/cyberagent

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

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

training into Llama2 to train Japanese texts. The other models are trained with Japanese texts from the beginning.

Let an input essay document be s and its tokens be x_1 to x_n generated by the tokenizer of a pre-trained langauage model. When using BERT, the vector corresponding to [CLS] token is used as an embedding vector of s. On the other hand, when using the GPT model, the vector that outputs the next token[8] after the final token x_n is used as an embedding of s.

Our setup involved the following key components:

- **GPT Configuration:** We utilized a GPT model specifically configured for the Japanese language, ensuring that it is finely attuned to the linguistic characteristics unique to Japanese.
- **Early Stopping:** To prevent overfitting, we employed an early stopping mechanism. This approach allowed the model to cease training once the improvement in performance on the validation set plateaued, thereby 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 in a batch size of 8, we effectively simulated a larger batch size of 16. This method allowed for more stable and effective training of the model.
- **LoRA:** we apply LoRA implemented in PEFT by HuggingFace and the rank is set to 8.

3.2 Desing of the Loss Function

Since the proposed model is categorical classification model, the class is not independent, but order then we apply soft labeling (Diaz and Marathe, 2019) into the loss function. In training phase the loss for categorical model is cross entropy and give with one-hot labels. On the other hand the soft label gives k-th value as the following formula.

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

[8]The token that denotes the end of input document varies depending on the model. For open-calm-7b, the final token is '<|endoftext|>'.

The d_k stands for the teacher value for each k-th unit in the final layer of the classification model. The $\hat{k}$ denotes the correct category. By applying this, a large penalty is given when outputting results that are far from the correct answer class.

4 Experimental Setup

4.1 Dataset

The Japanese essay tests was conducted on Japanese university students, and then, Japanese essay 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 4. The length of the essays ranges from 100 characters to 800 characters. The essays are manually scored on 5-point scale for comprehension, logic, validity, and grammar. In this paper, we focus on comprehension scores to evaluate the essay scoring models. Table 1 shows the number of essays for each prompt. The P, ML, Num stand for Prompt number, Maximum Length of essay and number of essays.

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

This data was divided into training, development, and test data in a ratio of 8:1:1.

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. The table below illustrates how scores were allocated across five possible score levels (1 to 5) for each theme and prompt within the dataset. This 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.**

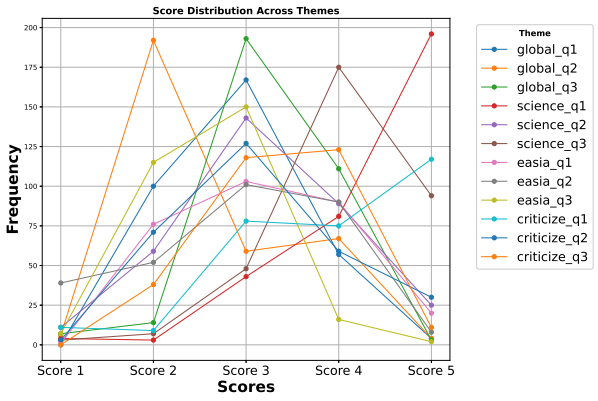

Figure 1: Scores Distribution per theme

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.

5 Experimental Results

5.1 Overall Performance

With Soft Labeling:

- **open-calm-medium:** It stands out with a high QWK of 0.5303 and the lowest RSME of 0.7243, indicating strong agreement with human raters and accuracy in score predictions as indicated in Table 2.
- **open-calm-large:** Table 2 shows that this model has the highest Accuracy of 0.6208, showing a good balance between precision and recall and correct predictions of essay scores while BERT has highest F1 score of 0.5056.

Without Soft Labeling:

- **calm2-7b:** This model shows top performance with the highest QWK of 0.5982, and the lowest RSME of 0.6957, suggesting that it is highly effective in scoring essays when soft labeling is not used.
- **open-calm-medium:** Notably, this model has the highest Accuracy of 0.6233 and a very competitive RSME of 0.7259, just slightly higher than **open-calm-large**, which has the lowest RSME of 0.7053 as shown in Table 2.

5.2 Category-wise Performance

With Soft Labeling Table 3

- **Criticize:** The **japanese-stablelm-instruct-alpha-7b-v2** model has the best QWK of 0.5239, but the best RSME is achieved by **open-calm-medium** at 0.7287.
- **Easia:** The **calm2-7b** model leads with the highest QWK of 0.5129, and RSME of 0.6259, indicating strong performance in this category.
- **Global:** open-calm-large scores highest in QWK (0.5593), which shows its strength in global content essays.
- **Science:** Both **open-calm-medium** and **open-calm-large** share the lead with an F1 score of 0.4515 and QWK of 0.7092. However, **open-calm-medium** has a slight edge with a lower RSME of 0.6604 compared to 0.6604 for open-calm-large.

Without Soft Labeling Table 4

- **Criticize:** **calm2-7b** excels with the highest QWK of 0.5831, as well as the lowest RSME of 0.7133, demonstrating its strong evaluative consistency in the "Criticize" category without soft labeling.
- **Easia:** The **calm2-7b** model again shows the best performance with the highest QWK of 0.5886, and the lowest RSME of 0.6280, indicating its robustness in understanding and scoring essays within this category.
- **Global:** The calm2-7b-chat model leads in QWK (0.5585), suggesting its effectiveness in the "Global" content essays. It also has the lowest RSME (0.6511), suggesting precise score predictions.

Table 2: Performance of GPT Models for all prompts

	With Soft labeling				Without Soft labeling			
Model Name	F1	QWK	Accuracy	RSME	F1	QWK	Accuracy	RSME
open-calm-small	0.2803	0.3417	0.5677	0.7855	0.2910	0.3848	0.5679	0.8112
open-calm-medium	0.3284	**0.530**3	0.5899	**0.7243**	0.3621	0.5551	**0.6233**	0.7259
open-calm-large	0.3502	0.5272	**0.6208**	0.7282	0.3772	0.5614	0.6219	0.7053
open-calm-7b	0.3072	0.4362	0.5963	0.7787	0.3370	0.5068	0.6089	0.7279
calm2-7b	0.3252	0.5288	0.6001	0.7417	0.3872	**0.5982**	0.6140	**0.6957**
calm2-7b-chat	0.3109	0.4512	0.5873	0.7761	0.3303	0.4994	0.6072	0.7332
japanese-stablelm-base-alpha-7b	0.2961	0.4201	0.5652	0.7933	0.3518	0.5367	0.6072	0.7332
japanese-stablelm-instruct-alpha-7b-v2	0.3372	0.4750	0.5886	0.7788	0.3362	0.4690	0.5918	0.7829
ELYZA-japanese-Llama-2-7b-instruct	0.2909	0.3760	0.5305	0.8980	0.3143	0.4501	0.5274	0.8365
ELYZA-japanese-Llama-2-7b-fast	0.2415	0.3105	0.5216	0.8884	0.2630	0.3375	0.5329	0.9217
ELYZA-japanese-Llama-2-7b	0.3372	0.4716	0.5930	0.7728	0.3526	0.4843	0.5768	0.8207
ELYZA-japanese-Llama-2-7b-fast-instruct	0.3115	0.4376	0.5481	0.7893	0.3260	0.4495	0.5520	0.8053
BERT	**0.5056**	0.4318	0.5602	0.7863	**0.4681**	0.3352	0.5450	0.8433

- **Science:** The **japanese-stablelm-base-alpha-7b** model achieves the highest Accuracy of 0.6061 and the best QWK of 0.7050, but open-calm-medium has the best RSME (0.6479), indicating its predictions are most closely aligned with actual scores.

  BERT achieved highest F1 score in Criticize, Easia and Global themes with scores of 0.4775, 0.5176 and 0.4699 respectively.

5.3 Prompt-wise Performance

With Soft Labeling Table 5

- **Prompt 1:** The **japanese-stablelm-instruct-alpha-7b-v2** model leads with an impressive QWK of 0.6881, and Accuracy of 0.6869, coupled with the lowest RSME of 0.6541, indicating its superior capability in accurately assessing essays based on the first prompt. BERT suppased with a relatively lower margin to have a better F1 score of 0.5605.

- **Prompt 2: open-calm-large** has the highest Accuracy (0.5876), while **calm2-7b-chat** has the highest QWK of 0.6963, suggesting nuanced understanding of the second prompt.

- **Prompt 3: japanese-stablelm-instruct-alpha-7b-v2** scores highest in QWK (0.4243) **BERT** in F1(0.5035), but **open-calm-large** leads in Accuracy (0.6300) and RSME (0.7100), reflecting its strong evaluative performance on the third prompt.

Without Soft Labeling Table 6

- **Prompt 1:** The **japanese-stablelm-instruct-alpha-7b-v2** model again dominates with the highest QWK of 0.7356, and Accuracy of 0.7355, while also maintaining a competitive RSME of 0.6070.

- **Prompt 2: ELYZA-japanese-Llama-2-7b-fast-instruct** has the highest QWK of 0.6920 and the lowest RSME of 0.6931, indicating it can effectively gauge the nuances of the second prompt.

- **Prompt 3: The open-calm-large** model performs best in Accuracy (0.6089) and has the lowest RSME (0.6834), while **calm2-7b** has the highest QWK (0.4373), showcasing its strong performance on the third prompt.

Conclusive Best Model Considering the consistent top-tier performance across multiple metrics and contexts, the **calm2-7b model** appears to be the most robust and versatile across themes and prompts, especially without soft labeling. Its ability to maintain high scores in F1 and QWK while also achieving the lowest RSME in several instances suggests that it could potentially offer the best overall performance for scoring Japanese essays.

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 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

Table 3: Performance of GPT Models across different themes with Soft Labeling

	Criticize				Easia				Global				Science			
Model Name	F1	QWK	Acc	RSME	F1	QWK	Acc	RSME	F1	QWK	Acc	RSME	F1	QWK	Acc	RSME
open-calm-small	0.2253	0.1956	0.5505	0.8631	0.2279	0.2648	0.6111	0.7109	0.2691	0.3683	0.5402	0.8029	0.3989	0.5381	0.5690	0.7653
open-calm-medium	0.3133	0.4983	0.6111	**0.7287**	0.3109	0.5071	0.6162	0.7157	0.2868	0.4257	0.4943	0.8156	**0.4515**	**0.7092**	**0.6667**	**0.6604**
open-calm-large	0.2707	0.3965	0.6010	0.7978	0.2930	0.4437	0.6465	0.6735	0.3857	**0.5593**	0.5690	0.7810	**0.4515**	**0.7092**	**0.6667**	**0.6604**
open-calm-7b	0.2581	0.3097	0.5707	0.7870	0.2756	0.4078	0.6364	0.7203	0.2831	0.4218	0.5460	0.8428	0.4119	0.6056	0.6322	0.7647
calm2-7b	0.2930	0.4464	0.5707	0.8354	0.3119	**0.5129**	**0.6919**	**0.6259**	0.3185	0.5045	0.5517	0.7646	0.3445	0.6515	0.5862	0.7409
calm2-7b-chat	0.2521	0.3361	0.5303	0.9231	0.2742	0.4136	0.6465	0.6922	0.3615	0.4732	**0.5920**	**0.7435**	0.3559	0.5819	0.5805	0.7456
jp-stablelm-base-alpha-7b	0.2182	0.1111	0.4949	0.9905	0.2657	0.4221	0.6566	0.6916	0.3207	0.4781	0.5172	0.7975	0.3798	0.6692	0.5920	0.6936
jp-stablelm-instruct-alpha-7b-v2	0.3395	**0.5239**	**0.6061**	0.7819	0.2799	0.3051	0.6162	0.8460	0.3247	0.4246	0.5575	0.7522	0.4048	0.6463	0.5747	0.7350
ELYZA-jp-Llama-2-7b-instruct	0.2346	0.2358	0.5202	0.9799	0.2933	0.4551	0.5960	0.7669	0.3356	0.2630	0.4655	1.0714	0.3710	0.5501	0.5402	0.7737
ELYZA-jp-Llama-2-7b-fast	0.2169	0.2616	0.5404	0.8682	0.2702	0.4420	0.6263	0.7245	0.1459	0.0265	0.4253	1.2031	0.3332	0.5119	0.4943	0.7576
ELYZA-jp-Llama-2-7b	0.3019	0.3841	0.5960	0.8578	0.2812	0.5123	0.6667	0.6597	0.3560	0.3992	0.5230	0.8460	0.4097	0.5907	0.5862	0.7276
ELYZA-jp-Llama-2-7b-fast-instr	0.2800	0.3636	0.5657	0.8660	0.2833	0.3906	0.5808	0.6990	0.3603	0.5048	0.5287	0.7593	0.3223	0.4914	0.5172	0.8329
BERT	**0.4871**	0.5217	0.5287	0.7982	**0.5146**	0.4122	0.5402	0.8157	**0.5315**	0.3755	0.6162	0.6884	0.4493	0.4117	0.5556	0.8429

Table 4: Performance of GPT Models across different themes without Soft Labeling

	Criticize				Easia				Global				Science			
Model Name	F1	QWK	Acc	RSME	F1	QWK	Acc	RSME	F1	QWK	Acc	RSME	F1	QWK	Acc	RSME
open-calm-small	0.2906	0.4172	0.6111	0.7863	0.2544	0.3008	0.5859	0.7885	0.2413	0.2945	0.5000	0.8919	0.3776	0.5268	0.5747	0.7780
open-calm-medium	0.3535	0.5565	0.6313	0.7415	0.3488	0.5293	0.6667	0.6878	0.3216	0.4878	0.5632	0.7766	0.4379	0.7025	**0.6494**	**0.6479**
open-calm-large	0.3153	0.5097	0.6212	0.7232	0.3576	0.4547	0.6364	0.6729	0.3980	0.5787	0.5805	0.7771	0.4379	0.7025	**0.6494**	**0.6479**
open-calm-7b	0.2691	0.3812	0.5909	0.7838	0.3448	0.5481	0.6667	0.6532	0.3144	0.4503	0.5747	0.7561	0.4196	0.6477	0.6034	0.7184
calm2-7b	0.3805	**0.5831**	0.5960	**0.7133**	0.3620	**0.5886**	0.6818	0.6280	0.3992	**0.5948**	0.5920	0.7239	0.4070	0.6264	0.5862	0.7177
calm2-7b-chat	0.3042	0.3961	0.6111	0.8073	0.2625	0.4212	0.5758	0.7394	0.4092	0.5585	**0.6149**	**0.6511**	0.3453	0.6216	0.5920	0.7780
jp-stablelm-base-alpha-7b	0.3136	0.4974	**0.6061**	0.7666	0.3350	0.5452	**0.7121**	**0.6200**	0.3308	0.3991	0.5172	0.8498	0.4277	**0.7050**	**0.6494**	0.6565
jp-stablelm-instruct-alpha-7b-v2	0.3040	0.4741	0.6010	0.8333	0.2701	0.2939	0.6111	0.8454	0.3976	0.5302	0.6034	0.6740	0.3731	0.5777	0.5517	0.7788
ELYZA-jp-Llama-2-7b-instruct	0.2624	0.3308	0.5354	0.8760	0.2631	0.3918	0.5455	0.8264	0.2663	0.4267	0.4483	0.9178	**0.4653**	0.6510	0.5805	0.7257
ELYZA-jp-Llama-2-7b-fast	0.1759	0.2653	0.4899	0.9575	0.3614	0.5472	0.6818	0.6578	0.1414	0.0922	0.4080	1.2287	0.3733	0.4455	0.5517	0.8428
ELYZA-jp-Llama-2-7b	0.3388	0.3823	0.5859	0.9083	0.3261	0.5496	0.6465	0.6962	0.3642	0.4552	0.5057	0.8407	0.3811	0.5501	0.5690	0.8378
ELYZA-jp-Llama-2-7b-fast-instr	0.2847	0.3056	0.5758	0.8807	0.2714	0.3680	0.5404	0.8019	0.3374	0.4675	0.4943	0.8378	0.4106	0.6569	0.5977	0.7007
BERT	**0.4775**	0.4473	0.5172	0.8282	**0.5176**	0.4297	0.5517	0.8339	**0.4699**	0.1996	0.5859	0.8333	0.4075	0.2642	0.5252	0.8775

Table 5: Performance of GPT models across different Prompts with soft Labeling

	Prompt 1				Prompt 2				Prompt 3			
Model Name	F1Score	QWK	Accuracy	RSME	F1	QWK	Accuracy	RSME	F1 Score	QWK	Accuracy	RSME
open-calm-small	0.3274	0.5068	0.6182	0.7332	0.3144	0.3920	0.5294	0.8052	0.2367	0.1263	0.5555	0.8182
open-calm-medium	0.4079	0.6524	0.6361	0.6999	0.3158	0.6270	0.5408	0.7362	0.2615	0.3116	0.5927	0.7369
open-calm-large	0.3684	0.5779	0.6447	0.7353	0.3741	0.6466	**0.5876**	0.7392	0.3082	0.3570	**0.6300**	**0.7100**
open-calm-7b	0.3857	0.5779	0.6361	0.7860	0.3358	0.6173	0.5530	0.7545	0.1999	0.1136	0.5998	0.7956
calm2-7b	0.3962	0.6237	0.6464	0.7607	0.3610	0.6925	0.5823	**0.7116**	0.2185	0.2703	0.5717	0.7527
calm2-7b-chat	0.3680	0.4767	0.6306	0.7812	0.3603	**0.6963**	0.5606	0.7388	0.2045	0.1806	0.5707	0.8083
japanese-stablelm-base-alpha-7b	0.4047	0.6579	0.6437	0.6664	0.2592	0.5212	0.5332	0.7659	0.2243	0.0813	0.5187	0.9476
japanese-stablelm-instruct-alpha-7b-v2	0.4352	**0.6881**	**0.6869**	**0.6541**	0.2564	0.3126	0.4932	0.9472	0.3201	**0.4243**	0.5857	0.7351
ELYZA-japanese-Llama-2-7b-instruct	0.2930	0.3468	0.5438	1.0113	0.3532	0.6463	0.5673	0.7700	0.2264	0.1349	0.4804	0.9126
ELYZA-japanese-Llama-2-7b-fast	0.2216	0.3078	0.5206	1.0370	0.3109	0.5356	0.5380	0.7680	0.1922	0.0881	0.5060	0.8600
ELYZA-japanese-Llama-2-7b	0.4342	0.6482	0.6739	0.6633	0.3322	0.6376	0.5543	0.7839	0.2451	0.1290	0.5507	0.8711
ELYZA-japanese-Llama-2-7b-fast-instruct	0.3030	0.3414	0.5858	0.8055	0.3563	0.6008	0.5586	0.7712	0.2751	0.3705	0.4999	0.7913
BERT	**0.5605**	0.4855	0.6393	0.7232	**0.4529**	0.4433	0.5024	0.9066	**0.5035**	0.3666	0.5389	0.7292

Table 6: Performance of GPT models across different Prompts without soft Labeling

	Prompt 1				Prompt 2				Prompt 3			
Model Name	F1 Score	QWK	Accuracy	RSME	F1 Score	QWK	Accuracy	RSME	F1 Score	QWK	Accuracy	RSME
open-calm-small	0.3056	0.4817	0.5534	0.8477	0.3050	0.4817	0.5500	0.7918	0.2623	0.1910	0.6003	0.7942
open-calm-medium	0.4031	0.6171	0.6609	0.7203	0.3947	0.6619	**0.6131**	0.7160	0.2883	0.3863	0.5960	0.7413
open-calm-large	0.4222	0.6235	0.6706	0.6570	**0.4297**	0.6721	0.5861	0.7754	0.2797	0.3885	0.6089	**0.6834**
open-calm-7b	0.4134	0.6511	0.6442	0.7317	0.2994	0.5709	0.5445	0.7299	0.2982	0.2985	**0.6381**	0.7221
calm2-7b	0.4752	0.6874	0.7128	0.6365	0.3424	0.6699	0.5375	0.7585	0.3440	**0.4373**	0.5917	0.6922
calm2-7b-chat	0.4661	0.6837	0.7106	**0.6013**	0.3023	0.5301	0.5449	0.8490	0.2225	0.2843	0.5398	0.7817
japanese-stablelm-base-alpha-7b	0.4592	0.7154	0.7079	0.6192	0.3158	0.5476	0.5419	0.8162	0.2803	0.3471	0.6138	0.7342
japanese-stablelm-instruct-alpha-7b-v2	0.4835	**0.7356**	**0.7355**	0.6070	0.2611	0.3939	0.4688	0.9581	0.2641	0.2774	0.5712	0.7835
ELYZA-japanese-Llama-2-7b-instruct	0.3259	0.5479	0.5546	0.8271	0.3868	0.5859	0.5568	0.8163	0.2300	0.2164	0.4707	0.8660
ELYZA-japanese-Llama-2-7b-fast	0.2554	0.3082	0.5573	1.0299	0.3243	0.5227	0.5759	0.8271	0.2093	0.1817	0.4654	0.9081
ELYZA-japanese-Llama-2-7b	0.4411	0.6489	0.6631	0.6951	0.3471	0.5042	0.5246	0.9254	0.2695	0.2999	0.5426	0.8417
ELYZA-japanese-Llama-2-7b-fast-instruct	0.2918	0.3452	0.5119	0.9161	0.3932	**0.6920**	0.5990	**0.6931**	0.2930	0.3113	0.5452	0.8066
BERT	**0.5558**	0.5128	0.6296	0.7501	0.3720	0.1839	0.4731	1.0328	**0.4765**	0.3089	0.5324	0.7469

Table 7: Prompt 3:Performance of long sentences with soft labeling

	Science				Criticize			
Model Name	F1 Score	QWK	Accuracy	RSME	F1 Score	QWK	Accuracy	RSME
open-calm-small	0.1676	0.0069	0.4394	0.9614	0.2318	0.2055	**0.5517**	0.7428
open-calm-medium	0.1655	0.2292	0.4545	0.8528	0.1333	0.0000	0.5000	0.8710
open-calm-large	0.2360	0.2028	0.5303	0.9374	**0.3616**	0.4262	**0.5517**	**0.6695**
open-calm-7b	0.1561	-0.0803-	0.5152	0.8961	0.1516	0.0204	0.5172	0.8610
calm2-7b	0.2258	0.3579	0.5000	**0.8528**	0.1506	0.0628	0.4310	0.7878
calm2-7b-chat_results	0.2535	0.3642	0.5000	0.8704	0.1333	0.0000	0.5000	0.8710
japanese-stablelm-base-alpha-7b	0.1265	-0.2170	0.2576	1.3540	0.1485	-0.0397	0.4138	1.0586
japanese-stablelm-instruct-alpha-7b-v2	0.2737	0.4094	0.4697	0.9211	0.2193	0.2685	0.5172	0.7311
ELYZA-japanese-Llama-2-7b-instruct	0.1481	-0.0233	0.3333	1.2851	0.1817	0.0752	0.4310	0.8200
ELYZA-japanese-Llama-2-7b-fast	0.1333	0.0000	0.5000	0.8439	0.1976	0.0796	0.4138	1.0422
ELYZA-japanese-Llama-2-7b	0.2094	0.1122	0.4697	1.0372	0.2034	-0.1177	0.4655	0.9738
ELYZA-japanese-Llama-2-7b-fast-instruct	0.1939	0.3438	0.4242	0.9455	0.1792	0.2991	0.4483	0.7428
BERT	**0.521**	**0.5865**	**0.5455**	0.7385	**0.5135**	**0.4612**	0.5172	0.7656

Table 8: Prompt 3: Performance of long sentences without soft labeling

	Science				Criticize			
Model Name	F1 Score	QWK	Accuracy	RSME	F1 Score	QWK	Accuracy	RSME
open-calm-small	0.2272	0.1524	0.5152	0.9455	0.2240	0.1785	0.5345	0.7543
open-calm-medium	0.2606	0.3402	0.4848	0.9129	0.1716	0.1626	0.5000	0.8094
open-calm-large	0.1977	0.3255	0.5000	0.7977	0.2182	0.3119	0.5172	0.6948
open-calm-7b	0.1988	0.0408	0.5303	0.8876	0.2619	**0.3840**	**0.6207**	**0.6565**
calm2-7b	**0.4420**	**0.5291**	**0.5455**	**0.7385**	0.2095	0.2927	0.5000	0.7071
calm2-7b-chat_results	0.2238	0.1925	0.5152	0.8876	0.1763	0.1791	0.4138	0.8610
japanese-stablelm-base-alpha-7b	0.2409	0.2812	0.4848	0.9293	0.2267	0.2127	0.5345	0.7878
japanese-stablelm-instruct-alpha-7b-v2	0.2196	0.1815	0.5000	1.0000	0.1880	0.0596	0.4655	0.7987
ELYZA-japanese-Llama-2-7b-instruct	0.1909	0.0321	0.3939	1.0517	0.1218	0.0752	0.2931	0.9826
ELYZA-japanese-Llama-2-7b-fast	0.0857	0.0000	0.2727	1.3200	0.1981	0.3165	0.4483	0.7768
ELYZA-japanese-Llama-2-7b	0.2357	0.1338	0.4697	1.0372	0.1963	0.0977	0.3793	1.0586
ELYZA-japanese-Llama-2-7b-fast-instruct	0.2531	0.2115	0.4545	0.9535	0.2093	0.0932	0.4655	0.9377
BERT	0.4166	0.4384	0.4848	0.8348	**0.5169**	0.3767	0.5172	0.7656v

Table 9: Performance comparison using classification model with soft labeling (WS), without soft labeling (WOS) and Regression model (RM)

	Small			Medium			Large		
Metric	WS	W0S	RM	WS	W0S	RM	WS	W0S	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**

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 as results from Table 7 and Table 8. 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 9 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 with 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 RSME 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's scope was impacted by several key limitations encompassing data availability, model architecture, and computational resources. Architectural exploration was confined by resource constraints, inhibiting our ability to innovate beyond the existing pre-trained GPT models specifically optimized for Japanese text. Furthermore, computational limitations were encountered, particularly with GPU memory constraints that precipitated "CUDA out of memory" errors at higher batch sizes. This necessitated the use of gradient accumulation with a reduced batch size of 8 and a step size of 2 to mitigate memory issues, which may have constrained the models' learning capacity and the overall efficiency of the training process.

9 Ethical Considerations

In the development and evaluation of our models, ethical considerations were rigorously adhered to, ensuring the protection of individual privacy. The dataset utilized for this study did not contain any personal information, guaranteeing the anonymity of all individuals involved. Furthermore, the data employed is publicly available, reinforcing the ethical integrity of our research by using sources that are accessible and transparent.

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*.

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.