Estimating Japanese Essay Grading Scores with Large Language Models

Boago Okgetheng[1] Koichi Takeuchi[1]
[1]Graduate School of Environmental, Life, Natural Science and Technology

pcqm1k3t@s.okayama-u.ac.jp takeuc-k@okayama-u.ac.jp

Abstract

In Natural Language Processing (NLP), the role of Large Language Models (LLMs) has been transformative, particularly in automatic essay scoring. However, their application for Japanese essay scoring remains under-researched. This study investigates the effectiveness of the Open-Calm LLM family in grading Japanese essays, using a dataset of about 300 essays annotated by native Japanese educators, spanning four thematic categories with three types of prompts.

Our evaluation focused on two key metrics: Quadratic Weighted Kappa (QWK) and accuracy. The results highlighted the Open-Calm Large model as the standout performer, achieving an accuracy of 59% and a QWK score of 0.52. In contrast, the Open-Calm Small model showed lower efficacy, with 54% accuracy and a QWK of 0.32. Notably, essays from the 'Global' category received the highest accuracy rate of 63%. Performance also varied across different prompts, with Prompt 1 showing the highest accuracy at 62%, while Prompt 3 lagged at 50%.

These findings demonstrate the significant potential of LLMs in automated Japanese essay grading, emphasizing the importance of model choice based on essay type and category. This study contributes to the understanding of LLMs in educational assessment tools, showcasing their promising application in diverse linguistic contexts.

1 Introduction

The advancement of Large Language Models (LLMs) has significantly transformed the field of natural language processing (NLP), particularly in tasks like automatic essay scoring. While a multitude of models for English essay grading exist (e.g., [1, 2, 3]) leveraging datasets like ASAP [4][1)], the exploration of LLMs for Japanese essays has been limited. This is notable given the complexity of Japanese language, with its intricate grammar and unique idiomatic expressions, which poses substantial challenges not fully addressed by existing models such as BERT.

Our study addresses this gap by evaluating the Open-Calm series of LLMs[2)], including variants such as Open-Calm Small, Medium, Large, and 7b, for their effectiveness in Japanese essay grading. Utilizing a dataset [5] of about 300 Japanese essays covering a diverse range of topics and writing styles, this research aims to understand how well these models handle the nuances of Japanese essay scoring.

The key findings of our study indicate that the Open-Calm Large model shows superior performance with an accuracy of 59% and a Quadratic Weighted Kappa score of 0.52, outperforming the Open-Calm Small model which achieved 54% accuracy. These results reveal the strengths and limitations of LLMs in processing Japanese essays, contributing valuable insights to the field of automated essay grading. The study concludes with a discussion on the implications of these findings and future research directions, emphasizing the potential of LLMs in educational assessments for a variety of languages.

2 Essay Grading Models

The field of automated essay grading has evolved significantly over the past decades, marked by a transition from regression-based systems (like e-rator [6]) to more deep neural network models [1, 2, 7, 8, 3].

In early research of neural network models for essay grading task, the structure of neural networks is intensively studied (e.g., LSTM [1] and CNN [2]), recently, attention

1) https://www.kaggle.com/c/asap-aes
2) Open-Calm is published by CyberAgent https://huggingface.co/cyberagent/ (accessed January 10, 2024)

has been aimed at how to effectively apply pre-trained large language models to essay grading. The several studies have proposed essay grading models employing BERT for Japanese essay [8, 5] as well as English essay grading [7, 3], however, the performance are competitive among the other models [5, 7, 3]. On the other hand, with the recent success of GPT (generative pretrained transformer) [9] in NLP, language models with quite large parameter pre-trained on massive amount of text data are provided. Since the proposed GPTs are much larger than BERT in the parameter size[3], GPT can be expected to be effective for tasks with little data such as essay grading.

Open-Calm is one of the prevously published Japanese pre-trained models. The varying sizes of the Open-Calm models (Small, Medium, Large, and 7b) offer an array of computational approaches, ranging from less resource-intensive to more complex systems. This versatility could be key in addressing the nuanced requirements of Japanese essay grading. The existing literature indicates that while significant strides have been made in automated essay scoring, there is a clear necessity for further exploration and development of models that can more accurately interpret and evaluate non-English languages, particularly for educational purposes.

3 Experiment

3.1 Dataset

We use the Japaense Written Essay Data [4] as an experimental dataset. Our study utilized a dataset comprising 300 Japanese essays, which served as the foundation for evaluating the effectiveness of the Open-Calm models. These essays were categorized into four distinct themes: criticize, easia, global, and science. Each category was further divided into three types of prompts, labeled as q1, q2, and q3, providing a diverse range of topics and styles for comprehensive model evaluation. This variety was crucial in assessing the models' ability to adapt and accurately grade essays across different subject matters and writing complexities.

3) Japanese BERT small (https://huggingface.co/cl-tohoku/bert-base-japanese) has about 0.3B weights that is an equivalent to Open-Calm Small.

4) GSK2021-B is provieded from GSK https://www.gsk.or.jp/catalog/gsk2021-b/.

3.2 Evaluation Metric

F1 Score

In the domain of essay grading, a high F1 Score is indicative of a model's balanced grading capability, a crucial attribute for educational assessment tools.

Quadratic Weighted Kappa(QWK)

QWK measures the agreement between two ratings. For our study, it assesses the consistency between the model's scores and the scores assigned by human educators, considering the ordered nature of the grading scale. In the realm of essay grading, the QWK(κ) serves as a critical metric for evaluating the agreement between the scores assigned by automated grading models and those given by human educators.

Accuracy

In the context of automated essay grading, accuracy serves as a critical indicator of a model's grading performance. It quantifies the proportion of essays that are graded correctly, providing a straightforward measure of the model's effectiveness in aligning with human grading standards.

3.3 Experimental Setting

Early stopping

In our experimental setup, we implemented early stopping as a regularization technique to prevent overfitting. This method monitors the model's performance on the evaluation dataset. If the model's performance does not improve for a pre-determined number of epochs, the training is halted. This approach ensures that the model retains generalizability and does not learn the training data's idiosyncrasies too closely.

Mini Batch Size

Our model training employed a mini-batch size approach. We set the batch size to 8, using gradient accumulation to effectively simulate a batch size of 16. This method allows for more efficient memory usage while still reaping the benefits of training with larger batch sizes, which is particularly useful for stabilizing the learning process in large models.

Architecture

The architecture of the Open-Calm models is pivotal to their function. These models are based on a transformer architecture, renowned for its effectiveness in handling se-

quential data and its ability to capture long-range dependencies in text.

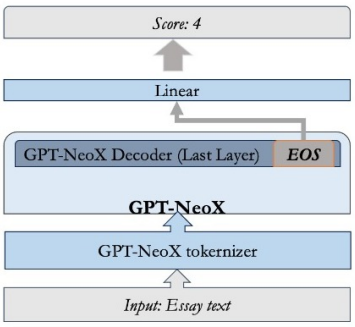

Figure 1 GPT-NeoX Model Architecture

The above diagram illustrates the basic structure of the Open-Calm model. At its core, the model comprises several layers of transformer blocks, each consisting of multi-head self-attention mechanisms and feed-forward neural networks. These components work in tandem to process input text, understand its context, and generate appropriate scores for the essays.

The unique aspect of the Open-Calm architecture is its adaptability and scalability, allowing it to handle various complexities within the Japanese language. Additionally, the use of the Lora Adapter and GPTNeox frameworks within this architecture enhances its language processing capabilities, making it well-suited for the task of essay grading.

Soft Labeling Method

In our study, essays are scored on a five- point scale, 1 to 5, treated as ordered classes. We employ ordered regression with a soft labeling approach [10]. Soft labels are assigned to the target outputs in the loss function, based on the following formula.

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

where d_k is the soft label for class k, K is the total number of classes, and k' is a given class. The comparison of scores with and without soft labels is shown in the table below:

Table 1 Target Output for each score

Score	Without Soft Labels (WO)	With Soft Labels (WL)
1	[1, 0, 0, 0, 0]	[0.6364, 0.2341, 0.0861, 0.0317, 0.0117]
2	[0, 1, 0, 0, 0]	[0.1915, 0.5206, 0.1915, 0.0705, 0.0259]
3	[0, 0, 1, 0, 0]	[0.0675, 0.1834, 0.4984, 0.1834, 0.0675]
4	[0, 0, 0, 1, 0]	[0.0259, 0.0705, 0.1915, 0.5206, 0.1915]
5	[0, 0, 0, 0, 1]	[0.0117, 0.0317, 0.0861, 0.2341, 0.6364]

3.4 Experimental Results

General Performance Across Models

This section interprets the performance metrics of the different models based on the F1 score, Quadratic Weighted Kappa (QWK), and accuracy, with and without soft labels. The comparison focuses on the Calm Small, Calm Medium, Calm Large, and Calm 7b Stable models.

- **Calm Small:** Exhibits the lowest overall performance among the four models. Notably, it has the highest accuracy without soft labels at 58.2%, but its F1 and QWK scores are relatively lower compared to other models.
- **Calm Medium:** Shows a noticeable improvement in performance over Calm Small, particularly in QWK scores. Its accuracy (59.7% with soft labels and 59.6% without) and F1 scores are better than Calm Small, indicating a more balanced performance.
- **Calm Large:** This model achieves the best QWK scores, both with (0.490) and without (0.532) soft labels, indicating superior grading consistency. The accuracy is slightly better than Calm Medium, with both soft label and non-soft label scores hovering around 59.7%.
- **Calm 7b Stable:** Although it doesn't reach the high QWK scores of Calm Large, Calm 7b Stable shows robust performance, particularly in QWK with soft labels (0.479). However, its accuracy is lower compared to Calm Large, at 58.2% with soft labels and 56.9% without.

Table 2 Average Performance Metrics Across Models

Model	WLF1	WOF1	WLQWK	WOQWK	WLA	WOA
Calm Small	0.226	0.304	0.243	0.427	0.533	0.582
Calm Medium	0.320	0.349	0.476	0.497	0.597	0.596
Calm Large	**0.325**	**0.355**	**0.490**	**0.532**	**0.598**	**0.597**
Calm 7b Stable	0.313	0.279	0.479	0.414	0.582	0.569

Analysis of Results: The data indicates a trend where increasing model complexity (from Calm Small to Calm 7b Stable) generally leads to improved accuracy and consistency in essay grading. The larger models, especially Calm Large, demonstrate a stronger capability in handling the nuances of Japanese essay grading, as shown by their higher QWK scores. However, the Calm 7b Stable model, despite its sophistication, does not outperform Calm Large in certain metrics, highlighting that increased complexity does not always translate to superior performance.

Performance by Category

- **Criticize Category:** All models performed comparably, with the Calm Large model slightly leading in accuracy. This suggests that while larger models may have a slight edge, the difference is not profound within this category.
- **Easia Category:** The Calm Medium model showed the highest accuracy, especially without soft labels. This indicates that the model's features are well-suited for essays within this theme.
- **Global Category:** The Calm Medium model again outperformed others in accuracy, particularly with soft labels, suggesting a robust capability in understanding and grading essays with global content.
- **Science Category:** The Calm Large model exhibited the highest accuracy, particularly with soft labels. It seems that the complexity of the Science category may benefit from the more extensive learning capacities of the larger models.

Table 3 Performance of Open-Calm models across different categories with and without soft labels (WL and WO respectively)

Model Name	Criticize		Easia		Global		Science	
	WLA	WOA	WLA	WOA	WLA	WOA	WLA	WOA
Calm Small	0.494	0.515	0.525	0.617	0.579	0.620	0.535	0.577
Calm Medium	0.540	0.538	**0.621**	**0.630**	**0.654**	**0.633**	0.577	**0.584**
Calm Large	**0.552**	**0.573**	0.613	0.603	0.633	0.628	**0.627**	**0.584**
Calm 7b	0.512	0.490	0.607	0.603	0.641	0.640	0.569	0.544

Overall Performance: The Calm Medium and Calm Large models consistently showed high accuracy across categories, suggesting these models strike a good balance between computational complexity and grading performance.

Performance by Prompt

- **Prompt 1:** Calm 7b excels with the highest WL accuracy (66.11%), but Calm Large leads in WO accuracy (64.40%). Calm Small, although lower in WL accuracy, is competitive in WO accuracy.
- **Prompt 2:** Calm Medium dominates with the highest accuracies in both WL (62.87%) and WO (63.82%). Calm 7b's performance drops compared to its lead in Prompt 1.
- **Prompt 3:** Calm Large and Calm 7b show close WO accuracies, with Calm Large slightly ahead. Calm Small lags in WL accuracy but improves in WO accuracy.

Overall Observations: Calm Large consistently performs well across all prompts. Calm Medium and Calm 7b show strong results in particular contexts but lack overall consistency. Calm Small, less effective in WL accuracy, fares better in WO accuracy, suggesting its suitability for definitive categorizations. The choice of using soft labels significantly affects model performance, with certain models like Calm Small performing better without soft labels. Calm Large stands out for its stable performance, emphasizing the importance of model selection based on the nature of prompts and scoring methodology.

Table 4 Performance of Open-Calm models by prompt with and without soft labels (WL and WO respectively)

Model Name	Prompt 1		Prompt 2		Prompt 3	
	WLA	WOA	WLA	WOA	WLA	WOA
Calm Small	0.568	0.599	0.512	0.5571	0.520	0.591
Calm Medium	0.629	0.638	**0.569**	**0.573**	0.555	0.579
Calm Large	0.630	**0.644**	0.578	0.559	**0.588**	**0.589**
Calm 7b	**0.661**	0.598	0.535	0.530	0.551	0.580

4 Conclusion

Our research has affirmed the Open-Calm series' effectiveness in automated Japanese essay scoring, with the Open-Calm Large model standing out for its superior performance. Discrepancies in accuracy across different prompts and essay categories highlighted the importance of context in model selection. Despite these promising results, the study acknowledges limitations, such as the dataset's scope. For future work, we aim to compare the Open-Calm models with other formidable LLMs like Calm2, Swallow, etc., which possess even greater size and parameter complexity. This next step will provide a clearer understanding of how model scale correlates with grading proficiency. Advancing this research will contribute to the refinement and practical application of LLMs in the realm of educational assessments.

Acknowledgement

Part of this study was supported by JSPS KAKENHI Grandt Number 22K00530.

References

[1] Kaveh Taghipour and Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pages 1882–1891, 2016.

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

153–162, 2017.

[3] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 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, 2022.

[4] Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. The hewlett foundation: Automated essay scoring. Technical report, Kaggle, 2012.

[5] Koichi Takeuchi, Masayuki Ohno, Kouta Motojin, Masahiro Taguchi, Yoshihiko Inada, Masaya Iizuka, Tatsuhiko Abo, and Hitoshi Ueda. 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, 2021. (in Japanese).

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

[7] Elijah Mayfield and Alan W Black. 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, 2020.

[8] Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. Automated essay scoring system for nonnative japanese learners. In **Proceedings of the 12th Conference on Language Resources and Evaluation**, pages 1250–1257, 2020.

[9] OpenAI. GPT-4 Techinical Report. Technical report, 2024.

[10] Raul Diaz and Amit Marathe. Soft Labels for Ordinal Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2019.