Evaluating the Impact of Continual Pre-Training on Japanese Essay Scoring Tasks

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

Abstract

This paper investigates whether continually pre-training Large Language Models on domain-specific reference texts can improve performance in Japanese Automated Essay Scoring tasks. We use a dataset covering multiple essay prompts related to four thematic areas— Globalization, Natural Science, Critical Thinking, and East Asian Economics. Each essay is scored on a five-point scale for Comprehensiveness. Models undergo two configurations: (1) direct fine-tuning on the scored essays, and (2) an additional continual pre-training phase using domain-specific texts prior to fine-tuning. Our findings indicate that most models benefit from this extra training, as evidenced by improvements in evaluation metrics such as the F1 Score, Quadratic Weighted Kappa, Accuracy, and Root Mean Squared Error. These results underscore the importance of domain adaptation for more accurate essay scoring.

1 Introduction

Automated Essay Scoring (AES) systems aim to assess the quality of written text using computational methods, thereby reducing the time and effort required for human grading [1]. Early AES systems often relied on feature engineering and statistical models, extracting linguistic features and employing regression or classification techniques to predict essay scores. However, recent advances in transformer-based language models have led to performance gains that surpass traditional approaches [2, 3, 4].

Despite the success of these models, an important limitation persists. Many of these architectures, such as BERT-based and GPT-based models, are pre-trained on large-scale corpora that may lack nuanced, domain-specific knowledge [5]. Consequently, their ability to handle specialized content is constrained, particularly in contexts like Japanese university admissions examinations, where essay topics can be both technical and diverse.

To address this challenge, this study explores continual pre-training—a process in which a language model is first pre-trained on massive, general-purpose corpora and then re-trained on narrower, domain-specific texts before the final fine-tuning phase. By continually pre-training on domain-relevant material, the model may acquire specialized vocabulary and contextual cues essential for effective assessment. We investigate whether such continual pre-training leads to higher scores in standard metrics such as the F1 Score, Quadratic Weighted Kappa, Accuracy, and Root Mean Squared Error (RMSE).

2 Related Work

Automated Essay Scoring systems were initially developed using feature-based and statistical approaches, which relied on handcrafted linguistic features such as word n-grams, part-of-speech tags, and discourse elements [1]. With the advent of deep learning, research began to shift toward end-to-end neural architectures, including Convolutional Neural Networks [6] and recurrent networks with Long Short-Term Memory [7, 8]. These approaches reduced the need for handcrafted features while capturing richer representations of text.

The introduction of transformer-based architectures revolutionized natural language processing. Models such as BERT and GPT have attained state-of-the-art results across tasks by leveraging large corpora for unsupervised pre-training and then applying supervised fine-tuning [2, 9]. However, pre-trained models may still suffer from domain mismatch. Several studies have highlighted the importance of domain adaptation or continual pre-training for specialized areas [5, 10].

For example, Hirao et al.[11] found that pre-training on nonnative Japanese data enhanced performance in scoring essays written by second-language learners. Similarly, domain-relevant text has been used to improve the performance of Automated Essay Scoring models [12]. Yet, applying continual pre-training specifically to Japanese university entrance topics remains under-explored, particularly with new Large Language Models of different parameter scales.

3 Dataset

3.1 Essay Prompts and Scores

We employ a dataset of Japanese essays written in response to prompts drawn from four thematic areas:

- Globalization
- Natural Science
- Critical Thinking
- East Asian Economics

Each theme contains several prompts (for example, subtopics focusing on international trade, environmental conservation, or economic interdependence). Essay lengths vary between 100 and 800 characters. All essays are labeled with a five-point score reflecting a single trait known as **Comprehensiveness**, which captures how thoroughly and coherently students have addressed the prompt.

3.2 Domain-Specific Texts

In addition to the scored essays, each theme is accompanied by reference documents that provide domain-specific background knowledge. These reference materials include academic articles, instructor-prepared sample responses, and explanatory texts. To facilitate continual pre-training, we compiled these domain-specific texts from all themes into a unified corpus. Specifically, the Globalization and Science themes each contribute approximately 2,600 characters, while the Criticize theme provides around 2,500 characters, and the Easia theme adds a more extensive 6,300 characters to the corpus.

4 Methodology

4.1 Models

We investigate the performance of several transformer-based models with varying parameter sizes and configurations:

1. Swallow-7b-hf
2. Swallow-7b-instruct-hf
3. Llama-3-Swallow-8B-v0.1
4. Llama-3-Swallow-8B-Instruct-v0.1
5. llm-jp/llm-jp-13b-v2.0 (for baseline comparison)

Each model includes a classification head on top of the language model to predict the essay score.

4.2 Experimental Design

4.2.1 Continual Pre-Training

To narrow the gap between general pre-training and the specialized context of Japanese university entrance examinations, we conduct an intermediate continual pre-training phase. The language models are exposed to a concatenated corpus of domain-specific texts using a next-token prediction objective. This is carried out for multiple epochs (two to five, depending on the model's size and memory constraints). We adopt the AdamW optimizer with a learning rate of 5×10^{-5} for stable convergence.

4.2.2 Fine-Tuning on Scored Essays

Once the model is continually pre-trained, it proceeds to a final fine-tuning phase on the labeled essay dataset. We convert the essay scoring task into a five-class classification problem, where each class corresponds to a specific score from one to five. We train the model for up to ten epochs, again using AdamW with a learning rate of 5×10^{-5}. A five-fold cross-validation setup is employed: 60% of essays are used for training, 20% for validation, and 20% for testing in each fold.

4.2.3 Evaluation Metrics

We report the following metrics on the test partition of each fold:

- **F1 Score**: The harmonic mean of precision and recall.
- **Quadratic Weighted Kappa (QWK)**: A measure of

Table 1 Comparison of Continual Pre-Training vs. Fine-Tuning Only

Model	F1	QWK	Accuracy	RMSE
Continual Pre-Training				
Swallow-7b-hf	0.7279	0.8244	0.8842	0.2308
Swallow-7b-instruct-hf	0.7170	0.8219	0.8803	0.2605
Llama-3-Swallow-8B-v0.1	0.8264	0.8160	0.8776	0.2440
Llama-3-Swallow-8B-Instruct-v0.1	0.8251	0.8004	0.8758	0.2603
Fine-Tuning Only				
Swallow-7b-hf	0.7223	0.8237	0.8843	0.2366
Swallow-7b-instruct-hf	0.7130	0.8178	0.8793	0.2434
Llama-3-Swallow-8B-v0.1	0.6899	0.7743	0.8625	0.3042
Llama-3-Swallow-8B-Instruct-v0.1	0.7009	0.7833	0.8674	0.3097
llm-jp/llm-jp-13b-v2.0	0.7108	0.7934	0.8648	0.2653

rating agreement that penalizes larger discrepancies more heavily.

- **Accuracy**: The percentage of exactly correct predictions.
- **Root Mean Squared Error (RMSE)**: The square root of the average squared differences between predicted and actual scores.

5 Results

Table 1 summarizes the performance of each model. "Fine-Tuning Only" denotes models that did not undergo continual pre-training on domain-specific texts, while "Continual Pre-Training" denotes those that received this additional training.

Models that received continual pre-training generally exhibit higher F1 scores and Quadratic Weighted Kappa values. For instance, Llama-3-Swallow-8B-v0.1 shows a notable jump in F1 from 0.6899 to 0.8264 when domain-specific pre-training is applied. The Swallow-7b-hf variants also see modest gains in both F1 and Quadratic Weighted Kappa, indicating that adding specialized content can benefit even mid-sized models.

6 Discussion

The most significant result from Table 1 is the performance boost observed in models that underwent continual pre-training. Exposure to reference texts filled with specialized terminology, context, and examples allows a model to better capture linguistic and conceptual cues relevant to the scored essays.

Although larger models have more capacity, the data and computational resources required for continual pre-training can be prohibitive. Smaller or mid-sized models can still produce competitive results when carefully aligned with the target domain, in line with prior research in parameter-efficient training methods [13].

Our study focuses on the Comprehensiveness dimension of essay scoring, but other traits—such as Logical Consistency and Grammar—may benefit similarly from domain-specific pre-training. Additionally, memory constraints limited the extent of our experiments on very large models. More efficient approaches to continuous adaptation (such as low-rank parameter updates) may help scale these methods to even larger models without sacrificing performance.

7 Conclusion

This paper demonstrated that continually pre-training Large Language Models on domain-specific Japanese texts can substantially enhance Automated Essay Scoring outcomes. By leveraging specialized reference materials before the final fine-tuning step, models achieved improved scores on metrics such as the F1 Score and Quadratic Weighted Kappa. These findings underline the value of bridging the gap between general-purpose pre-training and niche essay topics common in university-level entrance examinations. The results open avenues for future research on multi-trait scoring and efficient parameter adaptation techniques, contributing to more robust and context-aware essay evaluation systems.

Acknowledgement

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

References

[1] Yigal Attali and Jill Burstein. Automated Essay Scoring with e-rater V.2. **The Journal of Technology, Learning, and Assessment**, Vol. 4, No. 3, pp. 1–30, 2006.

[2] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1560–1569, Online, November 2020. Association for Computational Linguistics.

[3] Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. Many hands make light work: Using essay traits to automatically score essays. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Asso-**

ciation for Computational Linguistics: Human Language Technologies, pp. 1485–1495, Seattle, United States, July 2022. Association for Computational Linguistics.

[4] Shengjie Li and Vincent Ng. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7661–7681, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. Domain-adaptive neural automated essay scoring. In **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information**, pp. 1011–1020, 2020.

[6] Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In Jian Su, Kevin Duh, and Xavier Carreras, editors, **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1072–1077, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In Jian Su, Kevin Duh, and Xavier Carreras, editors, **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1882–1891, Austin, Texas, November 2016. Association for Computational Linguistics.

[8] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In Roger Levy and Lucia Specia, editors, **Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)**, pp. 153–162, Vancouver, Canada, August 2017. Association for Computational Linguistics.

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

[10] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6077–6088, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[11] Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. Automated essay scoring system for nonnative Japanese learners. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 1250–1257, Marseille, France, May 2020. European Language Resources Association.

[12] Robert Ridley, Liang He, Xin yu Dai, Shujian Huang, and Jiajun Chen. Automated cross-prompt scoring of essay traits. In **Proceedings of the AAAI Conference on Artificial Intelligence, 35(15)**, pp. 13745–13753, 2021.

[13] Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. Automated Chinese essay scoring from multiple traits. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3007–3016, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.