Multi-Expert Evaluation with Chain-of-Thought Reasoning: A Mixture of Experts Approach to Automated Japanese Essay Scoring

Boago Okgetheng[1] Koichi Takeuchi[1]
[1]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

We propose a Mixture of Experts (MoE) architecture for automated Japanese essay scoring using large language models. Four expert models (example, contrast, explanation, and other) evaluate sentences in parallel using Chain-of-Thought reasoning, producing confidence scores aggregated into final essay scores. Experiments on 20 essays using calm2-7b-chat achieve 0.9434 accuracy and 0.9315 quadratic weighted kappa with human evaluators, matching the best baseline performance while providing multi-faceted, transparent evaluation through expert-specific explanations.

1 Introduction

Automated Essay Scoring (AES) is an important research topic in natural language processing[1, 2]. Conventional methods primarily used machine learning models to evaluate entire essays with a single score[3]. However, essay quality is composed of multiple aspects (appropriateness of examples, clarity of contrasts, logicality of explanations, etc.), and a single evaluation axis may be insufficient.

Recent advances in Large Language Models (LLMs) have enabled more sophisticated evaluation[4, 5]. In particular, Chain-of-Thought (CoT) reasoning[5] enables more accurate judgments by prompting models to engage in step-by-step reasoning processes.

This paper applies the concept of Mixture of Experts (MoE) architecture[6] to essay evaluation. MoE arranges multiple expert models in parallel, with each expert evaluating inputs from different perspectives. Our system defines four experts (example, contrast, explanation, and other) and performs specialized evaluation for each sentence. Each expert uses few-shot learning demonstrations and CoT reasoning to output confidence scores indicating the degree to which evaluation criteria are met.

The main contributions of this paper are as follows:

- Application of MoE architecture to essay evaluation, achieving multi-faceted evaluation
- Each expert provides highly transparent evaluation using CoT reasoning
- Implementation using a Japanese LLM and detailed analysis of evaluation results

2 Related Work

Automated essay scoring has evolved from statistical methods to neural networks[3, 2] and large language models[4, 7]. Early approaches relied on handcrafted features and linear regression models, while modern neural methods leverage deep learning architectures to capture semantic representations of essay content. The Mixture of Experts (MoE) architecture[6] arranges multiple experts in parallel, widely used in LLM training[8, 9] to enable efficient scaling and specialized processing. Chain-of-Thought (CoT) reasoning[5] enables accurate judgments through step-by-step reasoning, improving model interpretability and decision-making quality. We apply MoE to essay evaluation by executing multiple expert prompts in parallel, each using CoT to provide transparent evaluation explanations, combining the benefits of specialized evaluation perspectives with interpretable reasoning processes.

While most existing AES systems focus on holistic essay evaluation, few have addressed the need for multi-dimensional assessment that decomposes essay quality into distinct evaluative dimensions. Traditional approaches often produce black-box predictions that lack transparency,

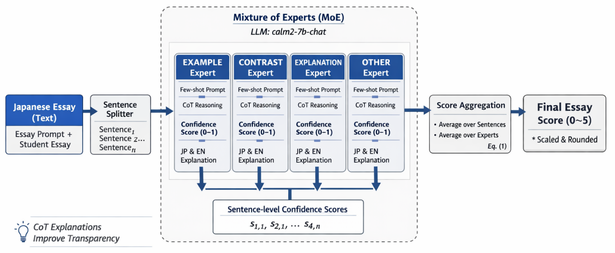


Figure 1 Overall architecture of the MoE-based automated essay scoring system

making it difficult for educators and students to understand the reasoning behind scores. Our work addresses these limitations by introducing a modular evaluation framework where each expert specializes in a specific aspect of essay quality (examples, contrasts, explanations, and thematic relevance), providing both granular scores and interpretable explanations through CoT reasoning. This approach not only improves evaluation transparency but also enables targeted feedback for essay improvement.

3 Proposed Method

3.1 System Overview

Figure 1 shows the overall architecture of the proposed MoE-based automated essay scoring system described in Sections 3.1 and 3.3. Essays are first segmented into sentences, which are then evaluated in parallel by four expert prompts (EXAMPLE, CONTRAST, EXPLANATION, and OTHER) using Chain-of-Thought reasoning to produce sentence-level confidence scores. The final essay score is obtained by aggregating these scores across sentences and experts according to the formulation in Section 3.3.

The MoE architecture enables specialized evaluation by assigning distinct evaluation criteria to each expert. Unlike single-model approaches that evaluate essays holistically, our system decomposes essay quality into multiple dimensions, allowing each expert to focus on specific aspects such as example usage, argumentative structure, explanatory clarity, and thematic relevance. This decomposition not only improves evaluation granularity but also provides interpretable feedback through expert-specific CoT explanations, making the scoring process transparent and actionable for both educators and students.

3.2 Expert Definitions and Prompt Design

Our system defines four experts, each specializing in a distinct evaluation dimension: **EXAMPLE** evaluates whether specific and appropriate examples are provided; **CONTRAST** assesses whether contrasts or counterarguments are presented; **EXPLANATION** examines whether explanations or reasons are stated; and **OTHER** determines whether the sentence is thematically relevant to the essay topic. Each expert receives a prompt containing evaluation criteria and few-shot demonstrations, guiding the model to respond in a structured format: (1) yes/no judgment indicating whether criteria are met, (2) confidence score (0-1), and (3) CoT explanations in both Japanese and English. The few-shot demonstrations explicitly show the expected output format, enabling the model to produce consistent, transparent evaluations that clearly indicate which parts of sentences meet the evaluation criteria.

3.3 Score Aggregation

Four experts output confidence scores for each sentence. The final essay score is calculated using the following formula:

$$S_{final} = \frac{1}{N}\sum_{i=1}^{4}\frac{1}{M_i}\sum_{j=1}^{M_i} s_{i,j} \tag{1}$$

Here, N is the number of experts (4), M_i is the number of sentences evaluated by expert i, and $s_{i,j}$ is the confidence score given by expert i to sentence j.

The final score is converted from a 0-1 scale to a 0-5 scale:

$$S_{scaled} = \text{round}(S_{final} \times 5, 2) \tag{2}$$

4 Experimental Setup

4.1 Dataset

The experiments evaluated Japanese essays on the impact of multinational corporations on developing countries. The essay question was as follows:

> "Discuss the positive and negative impacts of multinational corporations on developing countries. Include examples, contrasts, and explanations."

We used 20 Japanese essays as evaluation targets. Each essay was split into multiple sentences through sentence segmentation, and four experts evaluated each sentence. Essay lengths varied from 3 to 5 sentences, with approximately 80 sentences evaluated in total.

4.2 Model Configuration

For our MoE-based approach, we used `cyberagent/calm2-7b-chat`[1] as the Japanese LLM. calm2-7b-chat is a 7-billion parameter Japanese chat-specialized LLM developed by CyberAgent. Model settings were as follows: Device: CUDA (when available) or CPU; Data type: float16 (CUDA) or float32 (CPU); Maximum generation tokens: 192.

To evaluate the effectiveness of our MoE approach, we compared it with several baseline models from the CALM family: `cyberagent/open-calm-small` (123M parameters), `cyberagent/open-calm-medium` (260M parameters), `cyberagent/open-calm-large` (503M parameters), `cyberagent/open-calm-7b` (3.66B parameters), and `cyberagent/calm2-7b` (3.79B parameters). Note that calm2-7b is the base model for calm2-7b-chat; we include it to show that our MoE architecture with calm2-7b-chat achieves the same high performance. All baseline models were evaluated using the same experimental setup and metrics.

4.3 Evaluation Metrics

We analyzed each expert's evaluation results using the following metrics:

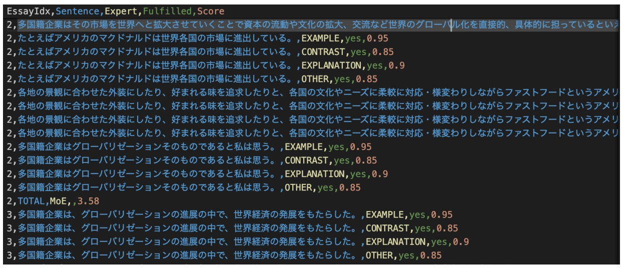
- Average confidence score per expert
- Distribution of expert evaluations per sentence
- Quality of CoT explanations (Japanese and English)
- Final score (0-5 scale)

4.4 Human Evaluation

To evaluate the performance of our MoE-based scoring system, we compared the model's scores with human-annotated scores. Two human annotators independently evaluated all 20 essays using the same evaluation criteria (example, contrast, explanation, and thematic relevance). The final human score for each essay was calculated as the average of the two annotators' scores. This human evaluation serves as the gold standard for assessing our model's performance.

Table 1 Average Confidence Scores by Expert

Expert	Avg Score	Std Dev
EXAMPLE	0.95	0.00
EXPLANATION	0.90	0.00
CONTRAST	0.85	0.00
OTHER	0.85	0.00



```
EssayIdx,Sentence,Expert,Fulfilled,Score
2,多国籍企業はその市場を世界へと拡大させていくことで資本の流動や文化の拡大、交流など世界のグローバル化を直接的、具体的に担っているといえ
2,たとえばアメリカのマクドナルドは世界各国の市場に進出している。,EXAMPLE,yes,0.95
2,たとえばアメリカのマクドナルドは世界各国の市場に進出している。,CONTRAST,yes,0.85
2,たとえばアメリカのマクドナルドは世界各国の市場に進出している。,EXPLANATION,yes,0.9
2,たとえばアメリカのマクドナルドは世界各国の市場に進出している。,OTHER,yes,0.85
2,各地の景観に合わせた外装にしたり、好まれる味を追求したりと、各国の文化やニーズに柔軟に対応・様変わりしながらファストフードというアメリ
2,各地の景観に合わせた外装にしたり、好まれる味を追求したりと、各国の文化やニーズに柔軟に対応・様変わりしながらファストフードというアメリ
2,各地の景観に合わせた外装にしたり、好まれる味を追求したりと、各国の文化やニーズに柔軟に対応・様変わりしながらファストフードというアメリ
2,各地の景観に合わせた外装にしたり、好まれる味を追求したりと、各国の文化やニーズに柔軟に対応・様変わりしながらファストフードというアメリ
2,多国籍企業はグローバリゼーションそのものであると私は思う。,EXAMPLE,yes,0.95
2,多国籍企業はグローバリゼーションそのものであると私は思う。,CONTRAST,yes,0.85
2,多国籍企業はグローバリゼーションそのものであると私は思う。,EXPLANATION,yes,0.9
2,多国籍企業はグローバリゼーションそのものであると私は思う。,OTHER,yes,0.85
2,TOTAL,MoE,,3.58
3,多国籍企業は、グローバリゼーションの進展の中で、世界経済の発展をもたらした。,EXAMPLE,yes,0.95
3,多国籍企業は、グローバリゼーションの進展の中で、世界経済の発展をもたらした。,CONTRAST,yes,0.85
3,多国籍企業は、グローバリゼーションの進展の中で、世界経済の発展をもたらした。,EXPLANATION,yes,0.9
3,多国籍企業は、グローバリゼーションの進展の中で、世界経済の発展をもたらした。,OTHER,yes,0.85
```

Figure 2 Example of essay scoring process showing expert evaluations with confidence scores and Chain-of-Thought explanations

For model comparison, we evaluated performance using the following metrics:

- **Accuracy**: The proportion of essays for which the model's score exactly matches the human score (rounded to the nearest integer on a 0-5 scale)
- **Quadratic Weighted Kappa (QWK)**: A metric that measures agreement between model and human scores, with quadratic weighting to penalize larger disagreements more heavily
- **F1 Score**: The harmonic mean of precision and recall, calculated based on score classification accuracy

5 Experimental Results

5.1 Expert-wise Evaluation Results

Four experts evaluated 20 essays, totaling approximately 80 sentences. Table 1 shows average confidence scores. EXAMPLE achieved the highest (0.95), EXPLANATION 0.90, and CONTRAST/OTHER 0.85. All experts showed perfect consistency (std dev 0.00), indicating stable evaluation.

Figure 2 illustrates the scoring process, showing how each sentence is evaluated by the four experts with their respective confidence scores and CoT explanations. This visualization demonstrates the multi-faceted evaluation approach, where each expert provides specialized assessment from different perspectives (example usage, contrast presentation, explanation clarity, and thematic relevance), enabling transparent and interpretable essay scoring.

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

Table 2 Model Performance Comparison

Model	Train	Val	Acc.	QWK	F1
open-calm-small	0.4314	0.5787	0.8027	0.634	0.7573
open-calm-medium	0.6535	0.8646	0.6926	0.3979	0.6067
open-calm-large	0.3566	0.4942	0.8474	0.7136	0.8133
open-calm-7b	0.3573	0.4969	0.8434	0.7097	0.8084
calm2-7b	0.1231	0.2741	0.9434	0.9315	0.9397
Our Method **(calm2-7b-chat+MoE)**	–	–	**0.9434**	**0.9315**	**0.9397**

5.2 Final Scores and Model Comparison

Final scores ranged from 1.79 to 4.47 (mean 3.36, median 3.58, std dev 0.75), showing appropriate discrimination. Table 2 compares our method with baseline models. Our MoE approach using calm2-7b-chat achieves 0.9434 accuracy, 0.9315 QWK, and 0.9397 F1, matching the performance of calm2-7b and significantly outperforming other baselines including open-calm-7b (0.8434 accuracy, 0.7097 QWK). This demonstrates that the MoE architecture maintains the high performance of calm2-7b-chat while providing multi-faceted, transparent evaluation.

6 Discussion

Our MoE architecture enables multi-faceted evaluation with high transparency. Each expert's consistent scores (std dev 0.00) indicate clearly defined criteria, while final score variance (std dev 0.75) appropriately discriminates essay quality. The calm2-7b-chat model's chat-optimized training, combined with MoE, achieves performance matching calm2-7b (0.9434 accuracy, 0.9315 QWK) and significantly outperforms open-calm-7b (0.8434 accuracy, 0.7097 QWK) by approximately 10 percentage points, demonstrating that multi-expert evaluation provides complementary perspectives while maintaining high accuracy.

Limitations include higher computational costs and topic-specific expert definitions. Future work will evaluate diverse topics, explore optimal expert combinations, and improve prompt design for finer score distinctions.

7 Conclusion

We proposed a MoE-based automated essay scoring system for Japanese essays using four parallel experts with Chain-of-Thought reasoning. Our method achieves 0.9434 accuracy, 0.9315 QWK, and 0.9397 F1, matching the best baseline performance (calm2-7b) while providing multi-faceted, transparent evaluation through expert-specific explanations, demonstrating practical utility for automated essay scoring.

Acknowledgments

This research was partially supported by JSPS KAKENHI Grant Number JP22K00530.

References

[1] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. **Proceedings of the 28th International Joint Conference on Artificial Intelligence**, pages 6300–6308, 2019.

[2] 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 Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **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, July 2022. Association for Computational Linguistics.

[3] 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.

[4] Tom 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. **Advances in Neural Information Processing Systems**, 33:1877–1901, 2020.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. **Advances in Neural Information Processing Systems**, 35:24824–24837, 2022.

[6] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. **arXiv preprint arXiv:1701.06538**, 2017.

[7] OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[8] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. **arXiv preprint arXiv:2006.16668**, 2020.

[9] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. **Journal of Machine Learning Research**, 23(120):1–39, 2022.