Multi-Expert Japanese Essay Scoring: MoE Architecture with Rubric-Augmented Prompts and LoRA-Efficient Fine-Tuning

BOAGO OKGETHENG, Okayama University, Japan
KOICHI TAKEUCHI, Okayama University, Japan

Automated essay scoring (AES) in Japanese has traditionally relied on single-classifier approaches with limited performance. This article presents a multi-expert Japanese essay scoring system that combines Mixture of Experts (MoE) with parameter-efficient LoRA fine-tuning and rubric-augmented prompts, applying MoE to Japanese AES in a low-resource setting. We apply a dense MoE head that replaces the standard single linear classifier with multiple lightweight expert MLPs and a small routing network that performs input-dependent expert selection. Under 5-fold cross-validation on each GSK prompt separately, with that prompt's rubric included in the input, our CALM2-7B MoE model achieves 92.7% accuracy and 0.915 Quadratic Weighted Kappa (QWK), improving over a controlled CALM2-7B single-head baseline that uses the same rubric input. The MoE gain appears primarily when rubric templates are present and is model-dependent. The results show that this routing head improves scoring over a single-head baseline while remaining parameter-efficient through LoRA adaptation.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning approaches*.

Additional Key Words and Phrases: Automated essay scoring, japanese language models, mixture of experts, LoRA fine-tuning, educational technology

ACM Reference Format:
Boago Okgetheng and Koichi Takeuchi. 2026. Multi-Expert Japanese Essay Scoring: MoE Architecture with Rubric-Augmented Prompts and LoRA-Efficient Fine-Tuning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 25, 11, Article 91 (November 2026), 22 pages. https://doi.org/10.1145/3849087

1 Introduction

Automated essay scoring (**AES**) in Japanese has progressed from feature-engineered regressors to neural encoders [16] and, most recently, to LoRA-tuned LLMs [7] that map a single pooled representation to a score via one classifier head. While the most recent AES research continues to explore innovative methodologies [17, 21], the fundamental architectural constraint of single-head classification persists, limiting the field's ability to capture the nuanced, multi-faceted nature of essay quality assessment. Recent Japanese AES work, such as fine-tuning GPT-style models with LoRA on the GSK dataset of 12 prompts across four themes, follows this single-encoder, single-head pattern and reports moderate agreement (e.g., Open-CALM-Medium ≈ 62% accuracy and QWK ≈ 0.56) [12] when treating the last-token embedding as the essay representation and supervising with cross-entropy/soft-labels. While [14] demonstrated that combining generative

This research was supported in part by JSPS KAKENHI Grant Number JP22K00530.
Authors' Contact Information: Boago Okgetheng (corresponding author), Graduate school of Environmental, Life, Natural Science and Technology, Okayama University, Okayama, Okayama Prefecture, Japan; e-mail: pcqm1k3t@s.okayama-u.ac.jp; Koichi Takeuchi, Graduate school of Environmental, Life, Natural Science and Technology, Okayama University, Okayama, Okayama Prefecture, Japan; e-mail: takeuc-k@okayama-u.ac.jp.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2375-4699/2026/11-ART91
https://doi.org/10.1145/3849087

pre-trained language models with soft labels can improve scoring accuracy, and [13] showed that continual pre-training on domain-specific data enhances performance, these approaches still employ single classifier heads that limit their ability to capture the multi-faceted nature of essay quality assessment. However, this approach fundamentally constrains the model's ability to capture the nuanced, multi-faceted nature of essay quality assessment, where different aspects of writing quality such as content sufficiency, coherence/logic, and writing style require specialized evaluation criteria. This limitation is particularly acute in Japanese AES, where the mixed-script nature of Japanese writing (kanji, hiragana, and katakana) creates evaluation criteria that benefit from specialized expert processing rather than monolithic assessment approaches. Despite the success of Japanese-specific pre-training [12], current systems still employ single classifier heads that miss the opportunity to leverage linguistic insights for expert specialization.

The Japanese AES literature reveals several critical insights that inform our approach. First, the success of Japanese-specific pre-training [12] demonstrates that Japanese writing has unique characteristics that require specialized processing. However, current approaches treat these characteristics as a monolithic "Japanese-ness" rather than recognizing that different aspects of Japanese writing, kanji usage, hiragana patterns, discourse structure, might benefit from different evaluation criteria.

Second, the limited training data problem in Japanese AES is not just a quantity issue but a quality issue. Japanese essays exhibit greater linguistic diversity than English essays due to the mixed-script nature of Japanese writing, where kanji, hiragana, and katakana serve different functions. Single-head approaches force this diversity through one evaluation lens, potentially losing nuanced distinctions that are crucial for accurate assessment.

Third, the LoRA fine-tuning success in Japanese AES [12] suggests that efficient adaptation is not just a practical necessity but a methodological advantage. By enabling fine-tuning on smaller, domain-specific datasets, LoRA allows models to capture Japanese-specific patterns that might be lost in larger, more general pre-training. However, current LoRA implementations still use single heads, limiting their ability to capture the multi-faceted nature of Japanese writing quality.

The progression from Jess [8] to modern GPT approaches [12] shows that Japanese AES has evolved in terms of model sophistication but has not addressed the fundamental architectural constraint of single-head classification. This limitation becomes increasingly problematic as models become more sophisticated, as the single-head bottleneck prevents these models from fully leveraging their capacity for Japanese-specific evaluation.

Current MoE implementations in AES lack dynamic guidance mechanisms that could adapt expert selection to essay characteristics. While prompt engineering has shown promise for guiding LLMs in various tasks, its integration with MoE architectures for AES remains unexplored. This integration could enable dynamic activation of experts based on essay content, prompt requirements, or linguistic complexity, addressing the "knowledge hybridity" problem where experts learn overlapping knowledge.

The combination of MoE's parameter efficiency and prompt-based routing offers a promising direction for Japanese AES, where rubric-augmented prompting can be combined with multi-expert routing in settings with limited training data. We use a learned gate over two expert MLPs, together with rubric-augmented prompts, to address both the expressivity limitations of single-head approaches and the unique challenges of Japanese essay assessment.

Our approach addresses this context through multi-expert evaluation, building on the insights from [12] and [11]. We employ two complementary experts whose outputs are mixed by a learned gate.

While LoRA enables efficient fine-tuning, existing implementations like [12] still use single heads, limiting their ability to capture the multi-faceted nature of Japanese writing quality. Building on

the insights from [13] regarding the effectiveness of continual pre-training on domain-specific data, our LoRA-efficient implementation maintains parameter and memory efficiency while enabling sophisticated expert routing. By freezing the backbone and only updating LoRA adapters plus the MoE components, we achieve the benefits of large Japanese pre-trained models without the computational overhead of full fine-tuning. This design choice directly addresses the parameter inefficiency concerns in general MoE architectures while enabling the multi-expert approach needed for nuanced Japanese essay assessment, incorporating the domain-specific knowledge that [13] demonstrated is crucial for Japanese AES performance.

Current MoE implementations like [1] use fixed expert designs that do not adapt to essay characteristics, lacking dynamic guidance mechanisms that could adapt expert selection to essay characteristics and scoring rubrics. Even the most recent AES research continues to face these architectural limitations. [17] proposes unsupervised approaches for essay scoring, while [21] introduces a rubric-based essay dataset. Neither study evaluates a trainable gated MoE scoring head of the kind used here. Our rubric-augmented expert routing addresses this limitation by incorporating explicit scoring rubrics directly into the input prompt. Unlike existing MoE systems with fixed expert designs, we format each essay with its corresponding scoring rubric as a structured prompt, allowing the gating network to dynamically determine the relative importance of each expert based on the rubric-augmented essay content. This enables contextually appropriate evaluation that adapts to different essay types and scoring criteria, making our approach particularly valuable for Japanese AES where the mixed-script nature of Japanese writing and the need for rubric-specific evaluation require more nuanced assessment than single-head approaches can provide.

Our approach addresses this limitation through two key innovations that break through the single-head bottleneck that has constrained AES research for years:

Mixture-of-Experts (**MoE**) on top of a Japanese LLM encoder. Instead of one classifier, we learn a small routing (gate) network that softly selects among multiple lightweight expert MLPs. Their weighted combination feeds the scorer.

Parameter-efficient adaptation at scale. We fine-tune strong Japanese LLMs (Open-CALM family[1], CALM2-7B, and ELYZA Japanese Llama-2 7B[2]) using LoRA (rank 16 for CALM2-7B and ELYZA; rank 8 for Open-CALM) [7] with 4-bit NF4 quantization, updating only adapters plus the new MoE components. This leverages Japanese pretraining while keeping compute modest.

We conduct experiments on Japanese essay scoring to compare our proposed MoE-based model with previous approaches. Our results demonstrate that the proposed approach is effective: CALM2-7B + MoE achieves $\approx$ 92.7% accuracy and QWK $\approx$ 0.915, improving over the same backbone with a single head and the same rubric input. The benefit of MoE is strongest with rubric-augmented prompts and is not uniform across backbones.

Our main contributions are: (1) a multi-expert scoring head that replaces the single classifier with a small gating network over expert subnetworks, used together with rubric-augmented prompts for Japanese AES; (2) parameter-efficient adaptation using LoRA that maintains parameter efficiency while enabling input-dependent expert routing; and (3) empirical validation on the GSK dataset under the same dataset, comprehension trait, per-prompt 5-fold CV, and metrics as prior work, showing that replacing the single head with the proposed MoE head, in the same CALM2-7B + rubric setting, improves automated assessment quality for Japanese essays.

[1] https://huggingface.co/cyberagent/open-calm-large
[2] https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

2 Literature Review

AES has evolved from rule-based systems to sophisticated neural architectures, yet remains fundamentally constrained by the single-head paradigm that forces all quality aspects through one shared representation space. This limitation becomes particularly acute in Japanese AES, where linguistic complexity and limited training data exacerbate the expressivity bottleneck.

2.1 The Single-Head Bottleneck in AES Evolution

Early AES systems relied on handcrafted features, essay length, spelling errors, lexical richness, chosen by psychometricians for interpretability. While these approaches offered transparency, they required substantial feature engineering effort and lacked flexibility for diverse essay types. [3] demonstrated that sophisticated feature engineering, combining string kernels and word embeddings, could achieve performance comparable to deep learning approaches, highlighting the continued relevance of well-designed features. However, these methods still employed single classifier heads that aggregate all quality signals into one decision space.

The shift to deep neural networks promised automatic feature learning, with hierarchical models like [5] combining LSTMs, CNNs, and attention mechanisms to capture both word-level and sentence-level representations. While these models showed superior performance through distributed essay representations, they still employed single classifier heads that force all quality facets through one shared subspace.

Recent **pre-trained language models** (**PLMs**) like BERT and GPT have dominated AES research, yet [10] revealed a critical limitation: fine-tuning BERT for AES often produced performance similar to classical models but at significant computational cost. This suggests that simply scaling model capacity without addressing the fundamental architectural constraint, the single-head bottleneck, yields diminishing returns. The core issue persists: forcing content sufficiency, coherence/logic, and writing style through one shared subspace before classification creates an expressivity bottleneck that we observe in practice.

2.2 Addressing the Bottleneck: Multi-Task and Combined Objectives

Researchers have attempted to address this bottleneck through various strategies. While recent AES research has predominantly focused on cross-prompt generalization [2, 9], single-prompt Japanese AES presents unique challenges that benefit from focused evaluation. The cross-prompt paradigm, exemplified by work on scoring across multiple prompts [2, 9], addresses scalability across diverse writing tasks. However, Japanese AES faces distinct linguistic challenges that make single-prompt specialization more valuable than cross-prompt generalization for this specific language context. [20] introduced R2BERT, a multi-task learning framework that combines regression and ranking losses, achieving the best average **Quadratic Weighted Kappa** (**QWK**) score on the ASAP dataset. [19] extended this approach with Pairwise Contrastive Regression (NPCR), integrating regression with ranking objectives within a contrastive learning framework. While these approaches improve performance by leveraging complementary objectives, they still employ single classifier heads that aggregate all signals into one decision space.

Reference [18] addressed the long-essay limitation of BERT through multi-scale essay representation, achieving competitive results. However, this work focused on representation learning rather than addressing the fundamental architectural constraint of single-head classification. Similarly, [4] proposed ProTACT for cross-prompt trait scoring using multi-head self-attention and topic-coherence features, but maintained the single-head paradigm for final scoring decisions. This pattern continues in 2025, where [1] demonstrates that even advanced MoE architectures for English AES still employ fixed expert designs that do not adapt to essay characteristics, highlighting the persistent need for more flexible, context-aware evaluation systems.

2.3 Mixture-of-Experts: Beyond the Single-Head Constraint

MoE architectures, originally developed by [15] for large-scale language modeling, offer a principled solution to this bottleneck by enabling conditional computation through specialized expert networks. While MoE has been applied to English AES [1], its application to Japanese AES presents unique challenges that have not been addressed in previous work. Japanese writing exhibits greater linguistic complexity than English due to its mixed-script nature, where kanji provides semantic density, hiragana enables grammatical flexibility, and katakana facilitates foreign word integration. This complexity creates evaluation criteria that can be handled by input-dependent routing rather than the monolithic assessment approaches used in English AES. Unlike single-head approaches that force all quality facets through one shared subspace, a gated MoE head can mix complementary expert transformations in an input-dependent way.

Reference [1] represents the most direct application of MoE to AES, employing three specialized experts for overall quality, relative quality, and prompt relevance assessment. While this approach demonstrates the potential of MoE for AES, it uses a fixed expert design that does not adapt to essay characteristics. The framework models human rater evaluation processes but lacks the dynamic routing that makes MoE powerful for handling diverse essay types. This limitation is particularly evident when compared to the latest 2025 research: [17] and [21] both introduce innovative methodologies but do not combine a trainable gated MoE head with rubric-augmented prompting. This article combines MoE with rubric-augmented prompting for Japanese AES and evaluates on the GSK dataset under the same per-prompt 5-fold protocol as [11], achieving strong performance with limited training data.

2.4 Japanese AES: Limited Progress, Missed Opportunities

Japanese AES research reveals a pattern of adopting English AES methods without addressing Japanese-specific challenges. Early Japanese AES systems like Jess [8] pioneered the use of Japanese-specific linguistic features, including character-type analysis of kanji and hiragana usage, to provide both holistic and multi-trait scores. This work recognized the unique characteristics of Japanese writing but remained constrained by the limitations of handcrafted features and statistical methods.

The field then followed the broader AES community's shift to neural approaches, with researchers exploring various architectures for Japanese essay assessment. [6] found BERT-based models effective for non-native Japanese learners, but this work focused on a narrow subset of Japanese AES, specifically non-native assessment, without addressing the broader challenges of native Japanese essay evaluation. More significantly, these studies simply applied English AES architectures to Japanese data without considering whether the single-head paradigm is appropriate for Japanese linguistic characteristics.

Recent work by [12] represents the most comprehensive exploration of Japanese AES using modern language models. Their study explores Japanese AES using various GPT models with LoRA fine-tuning and soft labels, finding that classification models with soft labels are effective and smaller GPT models can achieve high QWK scores. Crucially, they demonstrate that models pre-trained exclusively on Japanese data (like Open-CALM) outperform multilingual alternatives, suggesting that domain-specific pre-training is crucial for Japanese AES.

Building on this foundation, [14] further advanced Japanese AES by combining generative pre-trained language models with soft labels, demonstrating significant improvements in scoring accuracy and reliability. Their approach showed that the integration of soft label training with Japanese-specific language models can effectively capture the nuanced evaluation criteria required for Japanese essay assessment. This work established that Japanese AES benefits from specialized training approaches that account for the unique characteristics of Japanese writing.

Most recently, [13] investigated the effects of continual pre-training on domain-specific reference data for Japanese AES. Their research revealed that continual pre-training on educational domain data significantly enhances model performance, particularly for specialized contexts like Japanese university entrance examinations. This finding highlights the critical importance of incorporating domain-specific knowledge and demonstrates that Japanese AES requires more sophisticated training strategies than simple fine-tuning of general-purpose models.

However, despite these advances, Japanese AES research reveals a pattern of adopting English AES methods without addressing Japanese-specific challenges. The work by [6] and [12] both employ single classifier heads, missing the opportunity to leverage the linguistic insights gained from Japanese-specific pre-training for input-dependent expert routing. This limitation becomes particularly problematic given the unique characteristics of Japanese writing.

3 Methodology

3.1 From Single-Head Scoring to Routed Experts

The standard approach for Japanese AES follows a conventional pattern: prepend a compact rubric snippet (JP) to the essay, pass the sequence through a decoder-only LLM, pool to a single vector, then apply one classifier head. While this approach has shown promise, it fundamentally constrains the model's ability to capture the nuanced, multi-faceted nature of essay quality assessment.

Template-based Input Formatting: In our approach, we use structured prompt templates that combine the essay with explicit scoring criteria and guidelines. These templates serve as input prompts for the LLM, providing context about what aspects of the essay should be evaluated. Each template includes: (1) specific scoring criteria for the essay topic, (2) evaluation guidelines, (3) the student's essay text, and (4) a request for a score on a 1–5 scale. This template-based approach enables the model to understand both the essay content and the evaluation context, allowing for more targeted and consistent scoring decisions.

Formally, for tokens $x = (x_1, \dots, x_n)$, the encoder yields hidden states $H \in \mathbb{R}^{n \times d}$ and the last-token embedding

$$\mathbf{h} = H_n \in \mathbb{R}^d. \tag{1}$$

A single linear head maps $\mathbf{h}$ to logits and a probability simplex,

$$\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}, \qquad \mathbf{p} = \mathrm{softmax}(\mathbf{z}), \tag{2}$$

and training uses standard cross-entropy on the gold integer score:

$$\mathcal{L}_{\text{single}} = -\log p_{k^\star}, \tag{3}$$

where $k^\star$ is the gold comprehension label. This design is compute-friendly but forces every facet of essay quality (content sufficiency, coherence/logic, writing quality) through one shared subspace before classification, creating an expressivity bottleneck we observe in practice.

3.2 Our Approach: Adaptive Routing to Multiple Experts (Figure 1)

We keep everything upstream of the head identical (rubric conditioning, last-token pooling, and a frozen encoder adapted only via LoRA), and replace the single head with a compact, multi-expert MoE that adapts per essay. This departure from the single-head paradigm addresses the fundamental limitation we identified: the inability to capture the nuanced, multi-faceted nature of essay quality assessment.

Frozen backbone adapted with LoRA (LLM box). The LLM's base weights $\mathbf{W}$ are *frozen* (we store them in 4-bit NF4 for memory efficiency). **Low-Rank Adaptation (LoRA)** [7] updates

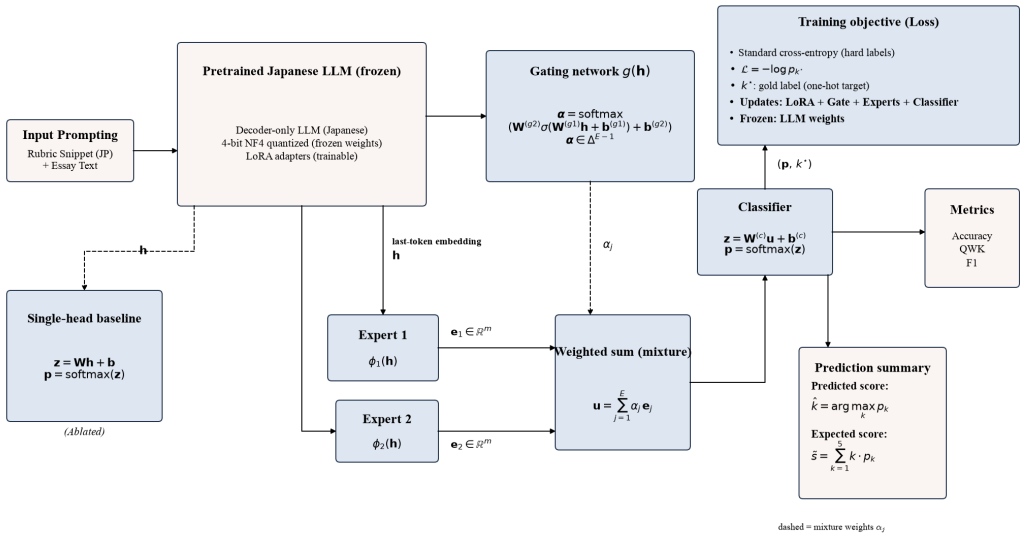

Fig. 1. Methodology overview. Input prompting with template: A rubric-augmented essay (essay + scoring template/prompt) is encoded by a frozen, LoRA-adapted LLM to a last-token vector $\mathbf{h}$. A gate computes α_j and routes $\mathbf{h}$ to expert networks ϕ_j; their weighted sum $\mathbf{u} = \sum_j \alpha_j \mathbf{e}_j$ is classified into a 1–5 score. Only LoRA and the MoE head are updated. The dashed baseline path corresponds to Equation (2).

only small rank-r adapters in selected projections (q, k, v, o for CALM2-7B and ELYZA; the fused `query_key_value` projection for Open-CALM):

$$\mathbf{W}' = \mathbf{W} + \frac{\alpha_{\text{lora}}}{r}\,\mathbf{AB}, \qquad \mathbf{A} \in \mathbb{R}^{d_{\text{out}} \times r},\ \mathbf{B} \in \mathbb{R}^{r \times d_{\text{in}}}, \tag{4}$$

with $\mathbf{W}$ frozen and $\mathbf{A}, \mathbf{B}$ trainable. This design choice maintains parameter and memory efficiency while allowing adaptation to the essay scoring domain. In Figure 1 this is the LLM (frozen + LoRA) box whose output is the *last-token embedding* $\mathbf{h}$.

Gating network $g(\mathbf{h})$ (Gate box; dashed line in the diagram). Instead of feeding $\mathbf{h}$ to one head, a small gate produces mixture weights $\alpha = (\alpha_1, \ldots, \alpha_E) \in \Delta^{E-1}$ that determine how much to trust each expert for the given essay:

$$\alpha = \text{softmax}\Big(\mathbf{W}^{(g2)}\,\sigma\big(\mathbf{W}^{(g1)}\mathbf{h} + \mathbf{b}^{(g1)}\big) + \mathbf{b}^{(g2)}\Big), \qquad \sum_{j=1}^{E} \alpha_j = 1,\ \alpha_j \geq 0. \tag{5}$$

We use a 2-layer MLP: $\mathbf{W}^{(g1)} \in \mathbb{R}^{d_g \times d}$, $\mathbf{b}^{(g1)} \in \mathbb{R}^{d_g}$, $\mathbf{W}^{(g2)} \in \mathbb{R}^{d_g \times E}$, $\mathbf{b}^{(g2)} \in \mathbb{R}^{E}$, with input dimension d (encoder hidden size), hidden dimension $d_g = 128$, and $E = 2$ experts. The activation σ is GELU; the output layer uses standard softmax (temperature $T = 1$). The input dimension d is the encoder's hidden size (e.g. $d = 4096$ for CALM2-7B); $\mathbf{W}^{(g1)}$ maps the last-token embedding to the gate hidden space, and $\mathbf{W}^{(g2)}$ maps from that hidden space to E logits, which are then normalized to mixture weights α. This keeps the gate lightweight while allowing input-dependent routing. If all experts learned similar functions, the mixture would collapse toward a single effective head; in practice we did not observe such collapse (see Discussion). The gate produces a mixture weight for each expert on each essay. In the figure, the *dashed* Gate→Weighted-sum arrow is labeled α_j to emphasize these are scalar weights, not vectors.

Expert networks $\phi_j(\mathbf{h})$ (Expert boxes; solid lines). Each expert receives the same $\mathbf{h}$ and applies a small MLP of width $m = 256$:

$$\mathbf{e}_j = \phi_j(\mathbf{h}) = \mathbf{W}_j^{(2)}\,\mathrm{LN}\Big(\sigma\Big(\mathbf{W}_j^{(1)}\mathbf{h} + \mathbf{b}_j^{(1)}\Big)\Big) + \mathbf{b}_j^{(2)} \in \mathbb{R}^m, \qquad m = 256. \tag{6}$$

Unlike the single-head approach that forces all quality facets through one shared subspace, each expert applies its own MLP to the same embedding. These flow along *solid* arrows to the mixture (Weighted-sum) box.

Dense mixture and classification (Weighted-sum → Classifier). The Weighted-sum box computes a contextually weighted combination of expert outputs:

$$\mathbf{u} = \sum_{j=1}^{E} \alpha_j\,\mathbf{e}_j, \tag{7}$$

which the Classifier maps to logits and probabilities:

$$\mathbf{z} = \mathbf{W}^{(c)}\mathbf{u} + \mathbf{b}^{(c)}, \qquad \mathbf{p} = \mathrm{softmax}(\mathbf{z}). \tag{8}$$

Importantly (as wired in Figure 1), the gate does *not* connect to the classifier; it only modulates the mixture via α_j. The gating weights record how much each expert is used for that essay.

3.3 Why This is Different and Why It Helps

Algebraically, our head *strictly generalizes* the single-head scorer. If E=1 and $\alpha_1 \equiv 1$, then $\mathbf{u} = \phi_1(\mathbf{h})$ and Equations (7)–(8) collapse to a single transformation of $\mathbf{h}$ followed by a linear classifier, the dashed ablated path in Figure 1 and Equation (2). For $E \geq 2$, the gate mixes two transforms of the same embedding, so different essays can use different expert combinations, without touching the frozen encoder and with negligible parameter overhead relative to the LLM.

Parameter and memory efficiency are maintained through careful design choices: experts are lightweight MLPs rather than full transformer layers, the frozen backbone with LoRA reduces memory requirements, and the dense mixture computation has low overhead relative to the LLM encoder.

3.4 Training, Updates, and Outputs (Loss/Prediction/Metrics boxes)

We retain standard cross-entropy, now applied to $\mathbf{p}$ from Equation (8):

$$\mathcal{L} = -\log p_{k^\star}. \tag{9}$$

Here $k^\star$ is the gold comprehension label: equivalently, the training target is one-hot, with probability 1 at $k^\star$ and 0 at the other scores. Backprop *updates only* the LoRA adapters $(\mathbf{A}, \mathbf{B})$, the gate Equation (5), the experts Equation (6), and the classifier Equation (8); the backbone weights remain frozen, matching the locks in the LLM box. At inference, p_k is the model's softmax probability for score $k \in \{1, \ldots, 5\}$ from Equation (8). The Prediction summary box in Figure 1 reports

$$\hat{k} = \arg\max_k p_k, \qquad \tilde{s} = \sum_{k=1}^{5} k \cdot p_k. \tag{10}$$

$\hat{k}$ is the predicted integer score used for accuracy, QWK, and F1. $\tilde{s}$ is the expected score under the same model distribution p (a weighted average of 1–5), not the gold label.

4 Experiments

4.1 Dataset

We evaluate our approach on a comprehensive Japanese essay dataset provided by the Japanese Language Resource Association (GSK) [12]. This dataset spans 12 prompts across four themes, designed to evaluate student understanding across multiple subject areas. The dataset is structured as an Excel workbook containing separate sheets for each essay topic, with essays evaluated on a 5-point scale based on domain-specific criteria.

The essay topics are organized into four major categories:

Global Studies (3 topics): Essays on global income inequality, multinational corporations, and cultural globalization. These topics require students to demonstrate understanding of both positive and negative aspects of global phenomena, with specific emphasis on concrete examples and balanced analysis.

Science (3 topics): Essays covering scientific methodology, the dual nature of natural sciences, and future directions in science education (STL). These topics assess students' understanding of scientific principles, their ability to connect theoretical concepts to practical applications, and their capacity for forward-thinking analysis.

East Asian Studies (3 topics): Essays on economic interdependence among Japan, China, and South Korea; the middle-income trap; and common economic challenges in East Asia. These topics require regional knowledge, data-driven analysis, and understanding of complex economic relationships.

Critical Thinking (3 topics): Essays on critical thinking perspectives, correlation vs causation analysis, and pseudoscience identification. These topics assess students' analytical skills, logical reasoning, and ability to evaluate evidence critically.

Each essay is evaluated using domain-specific rubrics that emphasize different aspects of student performance, from factual accuracy and example usage to analytical depth and logical coherence. The scoring criteria are designed to reflect the complexity and nuance required in Japanese academic writing, making this dataset particularly valuable for testing the capabilities of automated scoring systems.

All essays in the dataset were double-scored by human raters on several criteria and assigned integer scores from 1 to 5 for each criterion [12]. The criteria include comprehension, logic, validity of arguments, and grammar. As in prior work on this dataset [12], we use the comprehension score as the prediction target; the dataset also provides scores for logic, validity, and grammar, which could support multi-trait scoring in future work. The two human raters' scores were averaged (and in practice often agreed or differed by at most 1 point) to produce a final gold score for each essay. In total, the dataset comprises approximately 3,700 essays (roughly 300 essays per prompt). The essay length varies widely from about 100 to 800 Japanese characters depending on the prompt [12], providing a diverse range of responses from very short to fairly long essays.

For each prompt we run 5-fold stratified CV on that prompt's essays (about 300 essays; score labels 1–5 kept balanced across folds), then we repeat the same procedure for the next prompt, through all 12 prompts. A theme score such as Science is the average of that theme's three prompts; the overall score is the average over the 12 prompts.

4.2 Experimental Setup

4.2.1 Training Protocol. We fine-tuned each model (and our MoE framework) on the above dataset using 5-fold stratified cross-validation on each prompt separately. Our MoE architecture employs exactly 2 expert networks ($E = 2$), each implemented as lightweight MLPs. Two experts keep the head lightweight and training tractable on a single GPU; with roughly 300 essays per prompt, more experts would be harder to fit and would risk overfitting. The choice of expert count on larger

datasets, or learning it from data, remains an open direction. For a given prompt, that prompt's essays were split into 5 folds ensuring each fold had a similar score distribution (since scores 1–5 are ordinal, stratification helps maintain class balance in each fold). We trained for 5 epochs on each fold, using early stopping if the validation loss did not improve. All training was done on a single NVIDIA RTX-series GPU with 24 GB memory; thanks to our 4-bit compression and LoRA, even the 7B models could be trained in this environment. We set a batch size of 4 essays (effective batch 8 with gradient accumulation) given the input length and memory limits. For each fold, we saved the model with the best validation loss and evaluated it on that fold's hold-out set. We report the 5-fold mean for each prompt, then the mean across the 12 prompts.

Training used AdamW with seed 42. Learning rates were 1×10^{-4} (LoRA), 3×10^{-4} (experts), 1×10^{-3} (gate), and 2×10^{-4} (classifier). LoRA used rank 16, $\alpha = 32$, dropout 0.1, and target modules q, k, v, o for CALM2-7B and ELYZA; Open-CALM used rank 8, $\alpha = 16$, dropout 0.05, and the fused `query_key_value` projection. Each expert is an MLP of width $m = 256$ with GELU, dropout 0.2, and LayerNorm; the gate hidden size is $d_g = 128$ and $E = 2$. Sequences were truncated at 512 tokens; gradients were clipped at norm 1.0.

4.2.2 Evaluation Metrics. Following common practice in AES, we evaluate our models using Accuracy, QWK, and F1-score. Accuracy is the percentage of essays for which the predicted score equals the gold comprehension label (the averaged score of the two raters). QWK measures ordinal agreement between the predicted score and that gold label. We compute QWK for each fold, then average. We also report the weighted F1-score (which balances precision and recall across the 5 classes) to give insight into performance across all score levels, especially since the class distribution is not uniform (middle scores like 3 are more frequent than extreme scores 1 or 5). These metrics together provide a comprehensive evaluation of the scoring model's quality.

4.2.3 Baselines and Comparisons. We compare our MoE-enhanced model against several baseline variants. Firstly, we consider each pre-trained model fine-tuned without the MoE mechanism, i.e., using a single linear classifier on the final embedding (this corresponds to a standard fine-tuning approach similar to Hirao et al. [6]). This lets us isolate the impact of the MoE component. Secondly, we compare across different base model architectures and sizes: Open-CALM models of small, medium, large, and 7B parameters; the CALM2-7B model; and the ELYZA Llama-2 7B model. All models are fine-tuned under the same conditions for fairness. For direct comparison with prior work, we use the same GSK dataset, comprehension trait, per-prompt 5-fold stratified CV, and QWK/accuracy metrics as Okgetheng and Takeuchi [11], who report results with Open-CALM and single-head classification on this dataset.

5 Results

Our proposed MoE-based approach achieved strong results on the Japanese essay dataset. The best-performing configuration was with the CALM2-7B model as the base, obtaining an average accuracy of 92.7% and a QWK of 0.915 across the 12 prompts. The controlled comparison is the same CALM2-7B backbone with rubric templates, with vs. without MoE (Table 5): accuracy 91.0% / QWK 0.907 without MoE vs. 92.7% / 0.915 with MoE. Prior Open-CALM numbers from [11] use the same dataset, trait, per-prompt 5-fold CV, and metrics, but a different backbone and no rubric templates (Table 2). These results demonstrate that high-performance Japanese AES is achievable with limited training data. While English AES research typically relies on large datasets like ASAP (12,976 essays) or TOEFL (TOEFL11, 12,100 essays), our Japanese dataset contains approximately 3,700 essays across 12 prompts (roughly 300 per prompt). Despite this smaller labeled set, our CALM2-7B MoE model achieves 92.7% accuracy and 0.915 QWK. This success demonstrates that parameter-efficient adaptation combined with input-dependent routing can help overcome data

Table 1. Average Results Across all Models (5-fold)

Model Name	Params	Train Loss	Val Loss	Accuracy	QWK	F1
cyberagent/open-calm-small	123,529,351	0.3742	0.5222	0.8272	0.7556	0.8036
cyberagent/open-calm-medium	259,585,671	0.6232	0.7595	0.7318	0.5921	0.6876
cyberagent/open-calm-large	502,557,319	0.3868	0.6016	0.8268	0.7696	0.8012
cyberagent/open-calm-7b	3,657,707,143	0.3731	0.5188	0.8341	0.7798	0.8135
cyberagent/calm2-7b	3,790,478,983	**0.1495**	**0.3716**	**0.9272**	**0.9153**	**0.9217**
cyberagent/calm2-7b-chat	3,790,478,983	0.1839	0.3771	0.8992	0.9021	0.8959
elyza/ELYZA-japanese-Llama-2-7b	3,519,946,375	0.2770	0.5533	0.8512	0.8173	0.8396
elyza/ELYZA-japanese-Llama-2-7b-instruct	3,519,946,375	0.2629	0.5733	0.8494	0.8108	0.8367

Best values in bold. Params are the counts returned after 4-bit NF4 loading (quantized tensors are packed); a catalog "7B" model still has about 7B weights.

Table 2. Comparison on the GSK Comprehension Task

Method	Accuracy	QWK
This article: CALM2-7B + MoE + rubric templates	92.7%	0.915
This article: CALM2-7B + rubric templates, single-head	91.0%	0.907
Okgetheng and Takeuchi [11] (Open-CALM-Large, with soft labels)	59.8%	0.490

The first two rows use the same backbone and rubric input, with vs. without MoE. The ANLP 2024 row uses the same dataset, trait, per-prompt 5-fold CV, and metrics, but a different backbone and no rubric templates.

scarcity challenges, making the approach particularly valuable for languages with limited annotated resources.

Table 1 reports the full method (rubric templates + MoE). Under that setting, Open-CALM-7B reaches 83.4% accuracy and QWK 0.78; the same backbone with a single head and templates is stronger (86.3% / 0.835; Table 8). The Japanese ELYZA-Llama2-7B model achieved about 85% accuracy and QWK 0.82 under the full method. Within the Open-CALM family, larger size does not strictly imply higher accuracy: Small reaches ~82.7% while Medium reaches ~73.2%, with Large and 7B around 83%.

MoE does not uniformly improve over a single head. Table 3 summarizes the overall with/without-MoE comparison for CALM2-7B and Open-CALM Medium. For the CALM2-7B model with rubric templates, adding MoE increased QWK from 0.9068 to 0.9153 and accuracy from 91.0% to 92.7%. Without templates, MoE did not improve CALM2-7B (QWK 0.903 to 0.900; Table 5). The improvement was more pronounced on weaker base models, with the Open-CALM-Medium model's QWK jumping from 0.5523 to 0.5921 with MoE enhancement. On some prompts the picture is mixed: Table 4 shows Q1 QWK decreasing with MoE (0.9286 to 0.9096) even as accuracy rises.

Performance across essay themes showed that Globalization prompts were easiest for the model (94.3% accuracy), while Critical Thinking prompts were slightly more challenging (90.9% accuracy). Nonetheless, our model maintained over ~0.90 QWK on every theme, indicating strong performance across the 12 prompts and four thematic domains.

Error analysis revealed that most prediction errors were off by only one score level, with very few essays (under 2%) mis-predicted by more than one point. The QWK of ~0.91 is agreement between the predicted score and the gold (adjudicated) comprehension label.

5.1 Statistical Significance

To assess whether the observed improvements are statistically significant, we conducted paired t-tests on fold-level QWK and accuracy metrics. For each model, we compare configurations on 60 cells: 12 prompts × 5 folds, from a separate 5-fold CV on each prompt. Within a prompt the training

Table 3. Overall Comparison with and Without MoE (Second Experiment Settings)

Model	Train Loss	Val Loss	Accuracy	QWK	F1
CALM2-7B (without MoE)	0.1473	0.4250	0.9102	0.9068	0.9059
CALM2-7B (with MoE)	0.1495	**0.3716**	**0.9272**	**0.9153**	**0.9217**
Open-CALM Medium (without MoE)	0.6637	0.7019	0.7128	0.5523	0.6623
Open-CALM Medium (with MoE)	0.6232	0.7595	0.7318	0.5921	0.6876

Best values in bold.

Table 4. Per-Prompt Results for the Best Model (CALM2-7B) with and Without MoE

Prompt Group	Train Loss	Val Loss	Accuracy	QWK	F1
Q1 (Global1, Science1, EastAsia1, Criticize1), without MoE	0.1482	0.4299	0.9158	0.9286	0.9107
Q1 (Global1, Science1, EastAsia1, Criticize1), with MoE	**0.1432**	0.3598	**0.9313**	0.9096	0.9172
Q2 (Global2, Science2, EastAsia2, Criticize2), without MoE	0.1460	0.4236	0.9095	0.9215	0.9069
Q2 (Global2, Science2, EastAsia2, Criticize2), with MoE	0.1494	**0.3349**	0.9237	**0.9366**	**0.9306**
Q3 (Global3, Science3, EastAsia3, Criticize3), without MoE	0.1477	0.4216	0.9054	0.8703	0.9001
Q3 (Global3, Science3, EastAsia3, Criticize3), with MoE	0.1565	0.3376	0.9298	0.8847	0.9099

Best values in bold.

Table 5. Ablation Study: Effect of Rubric Templates and MoE (CALM2-7B, Mean Over 12 Prompts, and 5-fold CV)

Condition	Templates	MoE	Acc (%)	QWK	F1
Full approach	Yes	Yes	**92.7**	**0.915**	**0.922**
Rubric only	Yes	No	91.0	0.907	0.906
Essay-only + MoE	No	Yes	91.5	0.900	0.910
Essay-only only	No	No	91.5	0.903	0.907

Best values in bold.

folds overlap, so the cells are not independent; we treat the p-values as approximate. Tables 9 and 10 present comprehensive statistical test results for all models. The 95% bootstrap CIs resample the 60 paired cells with replacement (5,000 replicates) and report the 2.5th–97.5th percentiles of the mean paired difference. Those unweighted 60-cell means can differ from the difference of the prompt-level or essay-level averages shown in Tables 1 and 5.

For the main model (CALM2-7B), the full method (Templates + MoE) significantly outperformed the rubric-only configuration (Templates + No MoE): QWK difference +0.013 (95% bootstrap CI [0.003, 0.024], paired t-test $p = 0.014$), and accuracy difference +0.010 (95% CI [0.001, 0.021], $p = 0.048$). In the no-templates condition, MoE and single-head showed no significant difference for CALM2-7B ($p = 0.67$ for QWK, $p = 0.97$ for accuracy), consistent with our finding that MoE benefits most when rubric context is present.

For Open-CALM small, MoE yielded significant QWK gains in both templates and no-templates conditions ($p < 0.001$). For Open-CALM medium, the no-templates QWK gain is $p < 0.001$, while the templates QWK gain is $p = 0.028$ (accuracy $p = 0.09$). However, for Open-CALM 7B and ELYZA-7B with templates, single-head was significantly better ($p < 0.05$), indicating that MoE benefit is model-dependent and most effective when combined with rubric-augmented prompting for certain model architectures.

6 Discussion

Impact of the MoE Layer: On CALM2-7B, the MoE head yields a modest gain when combined with rubric templates (Table 5: +1.7 percentage points accuracy, +0.008 QWK at the table-level

Table 6. Ablation Study: No Templates + MoE — All Models (Mean Across 12 Prompts, 5-fold CV)

Model	Train Loss	Val Loss	Accuracy	QWK	F1
open-calm-small	0.3203	0.4641	0.838	0.812	0.830
open-calm-medium	0.6091	0.7065	0.747	0.643	0.711
open-calm-large	0.3943	0.5373	0.828	0.790	0.809
open-calm-7b	0.3321	0.4045	0.861	0.822	0.845
calm2-7b	**0.1305**	0.4383	0.915	0.900	0.910
calm2-7b-chat	0.1484	**0.3796**	**0.916**	**0.903**	**0.911**
ELYZA-7b	0.3091	0.4870	0.873	0.868	0.863
ELYZA-7b-instruct	0.2355	0.4361	0.893	0.888	0.887

Best values in bold.

Table 7. Ablation Study: No Templates + No MoE - All Models (Mean Across 12 Prompts, 5-fold CV)

Model	Train Loss	Val Loss	Accuracy	QWK	F1
open-calm-small	0.4697	0.5531	0.782	0.702	0.765
open-calm-medium	0.6602	0.7021	0.708	0.537	0.662
open-calm-large	0.3932	0.4762	0.818	0.721	0.788
open-calm-7b	0.2702	**0.3501**	0.876	0.848	0.863
calm2-7b	**0.1276**	0.4641	**0.915**	**0.903**	**0.907**
calm2-7b-chat	0.1490	0.4703	0.912	0.901	0.905
ELYZA-7b	0.2566	0.4770	0.880	0.872	0.871
ELYZA-7b-instruct	0.2023	0.3998	0.895	0.882	0.886

Best values in bold.

Table 8. Ablation Study: Templates + No MoE - All Models (Mean Across 12 Prompts, 5-fold CV)

Model	Train Loss	Val Loss	Accuracy	QWK	F1
open-calm-small	0.517	0.579	0.772	0.671	0.748
open-calm-medium	0.664	0.702	0.713	0.552	0.662
open-calm-large	0.456	0.514	0.796	0.692	0.759
open-calm-7b	0.319	**0.419**	0.863	0.835	0.852
calm2-7b	**0.147**	0.425	**0.910**	**0.907**	**0.906**
calm2-7b-chat	0.188	0.452	0.901	0.895	0.896
ELYZA-7b	0.243	0.451	0.871	0.825	0.864
ELYZA-7b-instruct	0.240	0.495	0.857	0.804	0.850

Best values in bold.

average; +0.010 accuracy and +0.013 QWK on the unweighted 60-cell test). Without templates, MoE does not improve this backbone. The gain was more pronounced on weaker base models, suggesting that the MoE mechanism can help when the base model's capacity is more limited. The gate typically assigned non-trivial weight to both experts rather than collapsing to one. The routing weights varied across essays, suggesting input-dependent use of the experts. Stronger regularization to encourage more distinct expert roles could be explored in future work.

Performance Across Essay Themes: The variation in performance across essay themes provides insights into the model's capabilities and limitations. Globalization prompts achieving the highest

Table 9. Statistical Significance: No Templates Condition—MoE vs No MoE (Paired t-test, 60 Observations Per Model: 5-fold CV × 12 Prompts)

Model	QWK: MoE vs No MoE	QWK p-value	Acc: MoE vs No MoE	Acc p-value
open-calm-small	+0.110 (95% CI [0.075, 0.149])	$p < 0.001$	+0.056 (95% CI [0.035, 0.077])	$p < 0.001$
open-calm-medium	+0.106 (95% CI [0.062, 0.156])	$p < 0.001$	+0.039 (95% CI [0.020, 0.061])	$p < 0.001$
open-calm-large	+0.069 (95% CI [0.032, 0.113])	$p = 0.001$	+0.011 (95% CI [−0.005, +0.027])	$p = 0.19$
open-calm-7b	−0.026 (95% CI [−0.052, −0.001])	$p = 0.050$	−0.015 (95% CI [−0.031, −0.002])	$p = 0.047$
calm2-7b	−0.003 (95% CI [−0.017, +0.011])	$p = 0.67$	≈ 0 (95% CI [−0.007, +0.007])	$p = 0.97$
calm2-7b-chat	+0.002 (95% CI [−0.019, +0.019])	$p = 0.86$	+0.004 (95% CI [−0.009, +0.015])	$p = 0.55$
ELYZA-7b	−0.004 (95% CI [−0.021, +0.015])	$p = 0.67$	−0.007 (95% CI [−0.021, +0.009])	$p = 0.41$
ELYZA-7b-instruct	+0.006 (95% CI [−0.009, +0.023])	$p = 0.46$	−0.001 (95% CI [−0.016, +0.014])	$p = 0.87$

Table 10. Statistical Significance: Templates Condition - Full Method (Templates+MoE) vs Rubric-Only (Templates+NoMoE) (Paired t-test, 60 Observations Per Model)

Model	QWK: Full vs Rubric-only	QWK p-value	Acc: Full vs Rubric-only	Acc p-value
calm2-7b	+0.013 (95% CI [0.003, 0.024])	$p = 0.014$	+0.010 (95% CI [0.001, 0.021])	$p = 0.048$
open-calm-small	+0.116 (95% CI [0.076, 0.161])	$p < 0.001$	+0.057 (95% CI [0.037, 0.078])	$p < 0.001$
open-calm-medium	+0.039 (95% CI [0.005, 0.074])	$p = 0.028$	+0.016 (95% CI [−0.003, +0.035])	$p = 0.09$
open-calm-large	+0.058 (95% CI [0.028, 0.092])	$p < 0.001$	+0.027 (95% CI [0.011, 0.045])	$p = 0.001$
open-calm-7b	−0.042 (95% CI [−0.078, −0.013])	$p = 0.011$	−0.020 (95% CI [−0.033, −0.007])	$p = 0.003$
ELYZA-7b	−0.026 (95% CI [−0.051, −0.004])	$p = 0.032$	−0.020 (95% CI [−0.036, −0.006])	$p = 0.009$
calm2-7b-chat	+0.009 (95% CI [−0.007, +0.026])	$p = 0.30$	+0.005 (95% CI [−0.008, +0.019])	$p = 0.50$
ELYZA-7b-instruct	−0.006 (95% CI [−0.030, +0.015])	$p = 0.60$	−0.009 (95% CI [−0.026, +0.005])	$p = 0.23$

scores (94.3% accuracy) aligns with prior observations that the Global theme had the highest human agreement and model scores in earlier studies [11]. This suggests that factual, content-rich topics may be easier for the model to assess consistently. In contrast, the Critical Thinking prompts were slightly more challenging (90.9% accuracy), involving identifying logical fallacies or pseudo-science. This trend mirrors human scoring difficulty: some prompts naturally lead to more variability in answers due to the subjective nature of critical thinking assessment. Nonetheless, our model maintains over ∼0.90 QWK on every theme, indicating strong performance across the 12 prompts and four thematic domains. The high agreement (QWK in the 0.90+ range) is agreement with the gold labels on those themes.

Error Analysis and Model Behavior: Our error analysis revealed important patterns in the model's decision-making process. Most prediction errors were off by only one score level, with very few essays (under 2%) mis-predicted by more than one point. This suggests the model has learned a reasonable ordinal understanding of essay quality. Many of the larger errors occurred on essays that were borderline in quality or had unusual writing styles. For instance, a few essays in the Criticize theme used unconventional reasoning that confused the model's assessment. In some cases, the model tended to slightly over-predict the score for very short essays, possibly because the rubric prompt text dominated the input and the model expected a certain level of content. This indicates a potential area for improvement: using length-aware or content-aware gating could mitigate this bias in future work.

Agreement with the Gold Label: Achieving a QWK of ∼0.91 is a strong result for automated scoring on this dataset: it is the agreement between the predicted score and the gold comprehension label (the averaged score of the two raters), which is the target used for training and evaluation. This level of agreement with the official labels, together with most errors being off by only one point, supports use as an assistive scoring tool in educational settings. The same QWK remains

above ∼0.90 on every theme, which further supports practical utility across the four assessment topics.

Ablation and Component Analysis: The 2×2 ablation shows that for CALM2-7B the best result uses templates and MoE together; templates alone do not raise accuracy over essay-only. We conducted a full 2×2 ablation design across all 8 models, comparing all four combinations: (1) Templates + MoE (full method), (2) Templates + No MoE (rubric-only), (3) No Templates + MoE, and (4) No Templates + No MoE. Tables 6, 7, and 8 present detailed results for each condition across all models, while Table 5 summarizes the 2×2 comparison for CALM2-7B.

The ablation results demonstrate several key findings: (1) The full approach (Templates + MoE) achieves the best performance for CALM2-7B (92.7% accuracy, 0.915 QWK), outperforming all other configurations. (2) On CALM2-7B, rubric templates help mainly together with MoE: rubric-only accuracy (91.0%) is slightly below the essay-only baseline (91.5%), while templates+MoE reaches 92.7%. (3) MoE provides additional gains when used with templates for CALM2-7B and smaller Open-CALM models, but the benefit is model-dependent (see Tables 9 and 10 for statistical significance). (4) For smaller models (Open-CALM small and medium), MoE yields significant improvements even without templates, suggesting that extra head capacity helps when base model capacity is limited.

Removing the MoE experts (i.e., using a single linear head) reduced accuracy and QWK for CALM2-7B and smaller models when rubric templates are used. Training without the prompt templates (feeding the essay alone to the model) led to a small drop in performance for the full CALM2-7B method (∼1–2% lower accuracy), confirming that including the prompt/rubric helps when combined with MoE. Among the LoRA-tuned models in this article, CALM2-7B was substantially stronger than the smaller Open-CALM variants. These findings indicate that both the scale of the base model and our methodological innovations (prompting and MoE fine-tuning) are crucial to the system's success.

Limitations: Our study has several limitations that should be acknowledged. First, the evaluation is conducted on a single dataset (GSK) with limited size (approximately 300 essays per prompt), which raises questions about generalizability to other Japanese essay datasets or domains. Second, the benefit of MoE is model-dependent: while it significantly improves performance for CALM2-7B and smaller Open-CALM models when combined with rubric templates, for some models (e.g., Open-CALM 7B and ELYZA-7B) single-head can perform better, suggesting that MoE is not universally beneficial. Third, performance varies across prompts and themes, with Critical Thinking prompts showing higher variance and lower scores than Global or Science prompts, indicating that some essay types remain more challenging. Fourth, our evaluation focuses on a single language (Japanese) and a single scoring trait (comprehension), limiting conclusions about cross-lingual transfer or multi-trait prediction. Finally, potential biases may exist, such as the model's tendency to over-predict scores for very short essays when rubric text dominates the input. Future work should evaluate on larger and more diverse datasets, explore cross-lingual transfer, and investigate methods to reduce prompt-level variance and scoring biases.

Future Directions: The success of our approach opens several promising avenues for future research. Explicit expert regularization could help enforce more interpretable specialization among the expert networks. Length-aware gating could address the observed bias toward over-predicting scores for very short essays. Adding trait-level supervision could increase interpretability by providing more granular feedback on different aspects of essay quality. Furthermore, exploring the scalability of this approach to even larger models and investigating the transferability of the learned expert specializations across different essay domains could lead to more robust and generalizable automated scoring systems.

In summary, our experiments demonstrate that rubric-augmented prompting with a two-expert MoE head improves Japanese AES over a controlled single-head CALM2-7B baseline, while the benefit of MoE is model-dependent and context-dependent. This underscores the benefit of combining large language models with task-specific adaptations like LoRA and MoE for complex NLP tasks, providing a robust foundation for advanced AES systems.

7 Conclusion

We have presented an approach to Japanese AES that combines multi-expert MoE routing with parameter-efficient LoRA fine-tuning. Our method achieves strong performance on a comprehensive Japanese essay dataset, demonstrating the effectiveness of combining large language models with task-specific architectural innovations. Our primary contribution is the introduction of a multi-expert MoE head that replaces the standard single linear classifier in Japanese AES systems, enabling input-dependent routing over two lightweight experts. The CALM2-7B MoE model achieved 92.7% accuracy and 0.915 QWK across 12 diverse essay prompts spanning four academic themes, improving over the same backbone with a single head and the same rubric input.

The combination of LoRA-based parameter-efficient fine-tuning with the MoE mechanism provides a scalable solution that maintains parameter efficiency while achieving strong performance. The QWK of ~0.91 is agreement between the predicted score and the gold comprehension label. Future research directions include explicit expert regularization for more interpretable specialization, length-aware gating to address scoring biases, and trait-level supervision for more granular feedback. Our approach demonstrates that combining the representational power of large language models with task-specific architectural innovations can significantly improve upon previous methods for AES in educational applications.

References

[1] Po-Kai Chen, Bo-Wei Tsai, Shao Kuan Wei, Chien-Yao Wang, Jia-Ching Wang, and Yi-Ting Huang. 2025. Mixture of ordered scoring experts for cross-prompt essay trait scoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 18071–18084. DOI: https://doi.org/10.18653/v1/2025.acl-long.884

[2] Yuan Chen and Xia Li. 2024. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, 12775–12786. Retrieved from https://aclanthology.org/2024.lrec-main.1118/

[3] Mădălina Cozma, Andrei M. Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 503–509. DOI: https://doi.org/10.18653/v1/P18-2080

[4] Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 1538–1551. DOI: https://doi.org/10.18653/v1/2023.findings-acl.98

[5] 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)*. Association for Computational Linguistics, Vancouver, Canada, 153–162. DOI: https://doi.org/10.18653/v1/K17-1017

[6] Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. 2020. Automated essay scoring system for nonnative japanese learners. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1250–1257. Retrieved from https://aclanthology.org/2020.lrec-1.157/

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=nZeVKeeFYf9

[8] 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 Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, 233–240. DOI: https://doi.org/10.3115/1220175.1220205

[9] Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 7661–7681. DOI: https://doi.org/10.18653/v1/2024.acl-long.414

[10] Elijah Mayfield and Alan W. Black. 2020. Should you fine-tune bert for automated essay scoring?. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Seattle, WA, USA, 151–162. DOI: https://doi.org/10.18653/v1/2020.bea-1.15

[11] Boago Okgetheng and Koichi Takeuchi. 2024. Estimating japanese essay grading scores with large language models. In *Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing*. The Association for Natural Language Processing, 643–647. Retrieved from https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/B3-2.pdf

[12] Boago Okgetheng and Koichi Takeuchi. 2024. Modeling score estimation for japanese essays with generative pre-trained transformers. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*. Association for Computational Linguistics, Trento, 64–73. Retrieved from https://aclanthology.org/2024.icnlsp-1.8/

[13] Boago Okgetheng and Koichi Takeuchi. 2025. Automated essay scoring for japanese: The effects of continual pre-training on domain-specific reference data. In *Proceedings of the 2025 8th International Conference on Information and Computer Technologies (ICICT)*. IEEE, 265–272. DOI: https://doi.org/10.1109/ICICT64582.2025.00048

[14] Boago Okgetheng and Koichi Takeuchi. 2025. Japanese essay scoring with generative pre-trained language models combined with soft labels. *IEICE Transactions on Information and Systems* E108-D, 9 (2025), 1095–1107. DOI: https://doi.org/10.1587/transinf.2024EDP7189

[15] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=B1ckMDqlg

[16] 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*. Association for Computational Linguistics, Austin, Texas, 1882–1891. DOI: https://doi.org/10.18653/v1/D16-1193

[17] Felipe Urrutia, Cristian Buc, Roberto Araya, and Valentin Barriere. 2025. Unsupervised automatic short answer grading and essay scoring: A weakly supervised explainable approach. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*. Association for Computational Linguistics, Vienna, Austria, 38–54. DOI: https://doi.org/10.18653/v1/2025.bea-1.4

[18] 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*. Association for Computational Linguistics, Seattle, United States, 3416–3425. DOI: https://doi.org/10.18653/v1/2022.naacl-main.249

[19] Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2724–2733. Retrieved from https://aclanthology.org/2022.coling-1.240/

[20] 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*. Association for Computational Linguistics, Online, 1560–1569. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.141

[21] Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. DREsS: Dataset for rubric-based essay scoring on efl writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 13439–13454. DOI: https://doi.org/10.18653/v1/2025.acl-long.659

A Prompt Templates

Tables 11, 12, 13, and 14 list the Japanese rubric-augmented prompt templates used for each theme (Global, Science, East Asia, and Critical Thinking).

Table 11. Japanese Prompt Templates for Scoring Essays on Global Theme

Theme and Prompt	Template
Global Q1 (Original)	【採点基準】 ① 世界の所得格差：光の側面（縮小） ② 各国の所得格差：影の側面（拡大） 【評価指針】 3 具体例の存在 3 データ引用 3 要因分析 【小論文】 {essay} 【採点結果（1–5の整数）】:
Global Q1 (Translated)	【Grading Criteria】 ① Global Income Inequality: Positive Aspect (Decreasing) ② National Income Inequality: Negative Aspect (Increasing) 【Evaluation Guidelines】 3 Presence of Specific Examples 3 Data Citation 3 Analysis of Causes 【Essay】 {essay} 【Grading Result (Integer 1-5)】:
Global Q2 (Original)	【採点基準】 ① 多国籍企業の光の側面 ② 多国籍企業の影の側面 【評価指針】 3 光と影の両側面への言及（どちらか一方のみなら2点、両方に言及すれば3点以上） 3 光・影それぞれの具体例を用いた説明（複数組み合わせ、詳細に説明すれば4点） 3 客観的評価（光と影を踏まえた総合的・批判的視点）があれば5点 【例示】 光の例：市場拡大、世界経済の成長促進、直接投資による新興市場の発展、現地雇用の創出、 地域活性化、納税による財政貢献、自由市場経済の効率化、製品・サービスの低価格化 など 影の例：熾烈な競争による現地企業の衰退、弱者の貧困化、開発途上国経済の支配、地域格差の拡大、 低賃金の固定化、不法移民の増大、伝統的生産様式の崩壊 など 【小論文】 {essay} 【採点結果（1–5の整数）】:
Global Q2 (Translated)	【Grading Criteria】 ① Positive aspects of multinational corporations ② Negative aspects of multinational corporations 【Evaluation Guidelines】 3 Mention of both positive and negative aspects (2 points for one aspect only; 3 points or more for both) 3 Explanation using specific examples for each positive/negative aspect (4 points for combining multiple examples and providing detailed explanations) 3 5 points if an objective evaluation is provided (comprehensive and critical perspective considering both light and shadow) 【Examples】 Light Examples: Market expansion, promotion of global economic growth, development of emerging markets through direct investment, creation of local employment, regional revitalization, fiscal contributions through taxation, efficiency of free market economies, lower prices for products and services, and so on. Shadow examples: Decline of local businesses due to fierce competition, impoverishment of vulnerable groups, domination of developing country economies, widening regional disparities, entrenchment of low wages, increase in illegal immigration, collapse of traditional production methods, and so on. 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Global Q3 (Original)	【採点基準】 ① 文化のグローバリゼーションに関する記述（光または影） ② 具体例の有無（アニメ、ファーストフード、ひらがななど） ③ 光と影の両方への言及、および自分の見解 【評価指針】 3 光または影のいずれかに言及（3点） 3 具体例も含む（4点） 3 光と影の両方に加え、具体例と自分の見解を含む（5点） 【小論文】 {essay} 【採点結果（1–5の整数）】:
Global Q3 (Translated)	【Grading Criteria】 ① Description of cultural globalization (positive or negative aspects) ② Presence of specific examples (e.g., anime, fast food, hiragana) ③ Mention of both positive and negative aspects, plus personal opinion 【Evaluation Guidelines】 3 Mention of either positive or negative aspect (3 points) 3 Includes specific examples (4 points) 3 Includes both positive and negative aspects, specific examples, and personal opinion (5 points) 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Table 12. Japanese Prompt Templates for Scoring Essays on Science Theme

Theme and Prompt	Template
Science Q1 (Original)	【採点基準】 ①「科学的」であるための条件（再現性、客観性、検証可能性など） 【評価指針】 3 明確で正確な条件を簡潔に述べているか（100字以内） 【小論文】 {essay} 【採点結果（1–5の整数）】:
Science Q1 (Translated)	【Grading Criteria】 ① Conditions for being "scientific" (reproducibility, objectivity, verifiability, and so on.) 【Evaluation Guidelines】 3 Does it concisely state clear and precise conditions (within 100 characters)? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:
Science Q2 (Original)	【採点基準】 ① 自然科学の二つの側面（講義より） ② 自然科学の社会的または実用的な役割に関する記述 【評価指針】 3 講義で学んだ側面に触れ、かつその意義を説明しているか 3 一貫性のある論理と例を含んでいるか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Science Q2 (Translated)	【Grading Criteria】 ① Two aspects of natural science (from the lecture) ② Description of the social or practical role of natural science 【Evaluation Guidelines】 3 Does it address the aspects learned in the lecture and explain their significance? 3 Does it contain consistent logic and examples? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Science Q3 (Original)	【採点基準】 ① STL（Scientific and Technological Literacy for All）の狙いを理解 ② 未来の科学教育の方向性に対する考察 【評価指針】 3 STLの要点を踏まえているか 3 科学教育の具体的な提案がされているか（500〜800字） 3 論理的かつ明快に構成されているか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Science Q3 (Translated)	【Grading Criteria】 ① Understanding the objectives of STL (Scientific and Technological Literacy for All) ② Reflection on the future direction of science education 【Evaluation Guidelines】 3 Does it incorporate the key points of STL? 3 Does it include specific proposals for science education? (500800 characters) 3 Is the essay logically and clearly structured? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Table 13. Japanese Prompt Templates for Scoring Essays on Easia Theme

Theme and Prompt	Template
Easia Q1 (Original)	【採点基準】 ① 日中韓の経済的相互依存についてデータを用いた説明 ② 協力・協業の具体例の提示（300字以内） 【評価指針】 3 データに基づいた説明があるか 3 協力例が具体的であるか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Easia Q1 (Translated)	【Grading Criteria】 ① Explanation of economic interdependence among Japan, China, and Korea using data ② Presentation of specific examples of cooperation/collaboration (within 300 characters) 【Evaluation Guidelines】 3 Does the explanation include data? 3 Are the cooperation examples specific? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:
Easia Q2 (Original)	【採点基準】 ① 中所得国の罠の概要 ② それを克服するための提案（250字以内） 【評価指針】 3 中所得国の罠について簡潔かつ正確に説明しているか 3 克服手段が具体的かつ現実的であるか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Easia Q2 (Translated)	【Grading Criteria】 ① Overview of the Middle-Income Trap ② Proposals for overcoming it (within 250 characters) 【Evaluation Guidelines】 3 Does it explain the Middle-Income Trap concisely and accurately? 3 Are the proposed solutions concrete and realistic? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Easia Q3 (Original)	【採点基準】 ① 日中韓共通の経済問題（少子化・環境問題など） ② 各国特有の経済課題 ③ 協調と対立の構造の説明（300字以内） 【評価指針】 3 共通課題と国別課題を的確に説明 3 協調・対立の構造を論理的に示しているか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Easia Q3 (Translated)	【Grading Criteria】 ① Common economic issues among Japan, China, and Korea (e.g., declining birth rates and environmental problems) ② Unique economic challenges for each country ③ Explanation of the structure of cooperation and conflict (within 300 characters) 【Evaluation Guidelines】 3 Accurate explanation of common and country-specific issues 3 Logical presentation of the structure of cooperation and conflict 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Table 14. Japanese Prompt Templates for Scoring Essays on Criticize Theme

Theme and Prompt	Template
Criticize Q1 (Original)	【採点基準】 ① 批判的思考の3つの共通観点（100字以内） 【評価指針】 3 講義・研究で共通に見られる観点（分析、評価、推論など）を含んでいるか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Criticize Q1 (Translated)	【Grading Criteria】 ① Three Common Perspectives of Critical Thinking (within 100 characters) 【Evaluation Guidelines】 3 Does it include perspectives commonly seen in lectures and research (e.g., analysis, evaluation and reasoning)? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:
Criticize Q2 (Original)	【採点基準】 ① カラーテレビの所有数と長寿の因果関係についての誤りを説明 ② グラフの形成理由に触れる（400字以内） 【評価指針】 3 因果関係と相関関係の違いを理解しているか 3 グラフ形成の背景（第三の変数など）に触れているか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Criticize Q2 (Translated)	【Grading Criteria】 ① Explain the error regarding the causal relationship between the number of color TVs owned and longevity ② Address the reasons for the graph's formation (within 400 characters) 【Evaluation Guidelines】 3 Understanding the difference between causation and correlation 3 Addressing the background of the graph formation (e.g., third variables) 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Criticize Q3 (Original)	【採点基準】 ① ニセ科学の実例（講義で扱っていないもの） ② それがニセ科学であると示す方法・証拠・説明 ③ 信じてしまいやすい要因の分析（500〜800字） 【評価指針】 3 実例が明確で適切か 3 科学的な検証方法が正しく使われているか 3 信念形成に関する心理的・社会的要因を説明しているか 【小論文】 {essay} 【採点結果（1–5の整数）】:
Criticize Q3 (Translated)	【Grading Criteria】 ① Examples of pseudoscience (not covered in lectures) ② Methods, evidence, and explanations demonstrating it is pseudoscience ③ Analysis of factors making it easy to believe (500800 characters) 【Evaluation Guidelines】 3 Are the examples clear and appropriate? 3 Are scientific verification methods used correctly? 3 Does it explain psychological and social factors related to belief formation? 【Essay】 {essay} 【Grading Result (Integer 1-5)】:

Received 1 October 2025; revised 17 August 2026; accepted 13 September 2026