



Optimizing Automated Essay Scoring with Lightweight Large Language Models and Validated Rubrics

Prayitno^{1*}, Fahima Choirun Nabila², Mohammad Khambali³, Afandi Nur Aziz Thohari⁴, Karisma Trinanda Putra⁵, Viqi Ardaniah⁶, Jurianto⁷

^{1,2,3,4,5}Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang, Indonesia

¹Research Fellow at Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

¹Department of Research and Development, SEAMEO SEAMOLEC, Indonesia

⁵Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Indonesia

^{6,7}Department of English Language and Literature, Universitas Airlangga, Surabaya, Indonesia

*prayitno@polines.ac.id

Abstract

Manual grading of English as a Foreign Language (EFL) essays often leads to inconsistent scores among educators, despite the use of rubrics. While traditional Automated Essay Scoring (AES) systems offer speed, they often fail due to high computational cost, reliance on extensive datasets, and an inability to capture holistic writing qualities such as creativity and humanistic expression. This study addresses these issues by introducing AESCORE, a novel, lightweight, and cost-effective AES framework. Our methodology centers on integrating validated rubric criteria (identified via VOSviewer analysis) with open-source Large Language Models (LLMs), specifically emphasizing a human-centered approach. We evaluated AESCORE across 100 EFL essays using several prompting techniques, including few-shot and multi-trait specialization. The system achieved its most robust performance and high scoring consistency (Quadratic Weighted Kappa QWK = 0.6660) using the DeepSeek-R1 8B LLM with few-shot prompting. AESCORE represents a significant contribution by demonstrating that sophisticated, pedagogically-aligned writing assessment and generative feedback can be achieved with accessible AI, offering a reliable alternative for improving productive writing skills in higher education.

Keywords: automated essay scoring; english as a foreign language essays; large language model; prompting method

How to Cite: Prayitno, "Optimizing Automated Essay Scoring with Lightweight Large Language Models and Validated Rubrics", J. RESTI (Rekayasa Sist. Teknol. Inf.), vol. 10, no. 2, pp. 266 - 374, Apr. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.7012>

Received: July 28, 2025

Accepted: February 23, 2026

Available Online: April 22, 2026

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Manually scoring and providing feedback on student's English essays have been viewed challenging [1] because the rubric assessment still leads to inconsistent score among higher education educators [2-3]. Although such issues are addressed by implementing AES to assess essay writing, educators in higher education have not fully embraced it due to its inability to capture creative, artistic, and humanistic elements [4]. Its application in assessing students' creativity remains unsuccessful [5].

With the rapid advancement of natural language processing and the increasing demand to incorporate creativity and humanistic dimensions in essay scoring systems, technology-based solutions are becoming essential to help students improve their writing more

effectively [6]. One such solution is Automated Essay Scoring (AES) [7], a system that uses natural language processing (NLP) to assess writing based on linguistic features such as sentence structure, grammar, coherence, and argument flow [1]. With AI-based models, AES can provide faster, more objective, and consistent feedback compared to manual grading [8].

In a previous study by Y. Song [9], a few-shot learning method was used on 600 English essays written by children, using various large language models (LLMs), including ChatGLM2-6B, which achieved a Quadratic Weighted Kappa (QWK) score of 0.531. Another study by M. Uto [10] combined Item Response Theory (IRT) and feature engineering with the BERT transformer model on the ASAP Kaggle dataset, reaching a Weighted Kappa score of 0.894. These results show the high potential of LLMs as the foundation for

developing modern AES systems [11]. Moreover, smaller-parameter LLMs are still underexplored, presenting further research opportunities [12].

Despite progress in modern AES, there is still a lack of research utilizing recent open-source large language models (LLMs) [13] available on platforms like Ollama—such as LLaMA 3.1 (8B), Mistral 7B, Qwen 3 (8B), Gemma 3 (4B), and DeepSeek R1 (8B). These low-resource models (4–8B parameters) offer computational efficiency and are well-suited for lightweight AES applications but remain largely underexplored [14]. Building on the work of Mansour et al. [15] which showed that LLaMA 2 (13B) achieved a QWK score of 0.562 through prompting, smaller models like LLaMA 3.1 [16] hold strong potential for developing scalable and cost-effective AES systems, especially in resource-limited settings.

In terms of datasets, most prior AES research has relied on secondary data from platforms like Kaggle [5]. Few studies use primary datasets built through direct data collection. For instance, Y. Song's [9] study used 600 original English essays and performed comparably to studies using Kaggle data. Primary datasets can offer more relevant and authentic context for AES applications [17]. Additionally, many existing AES systems lack automated feedback features and validated rubric assessment taken from experts [18], which are critical in helping students develop productive writing skills, especially when learning English as a foreign language [19].

After reviewing some literatures, this paper highlights several contributions such as the exploration of AES development within the context of low-resource computing through the use of medium-to-mini [20] scale large language models (LLMs), ranging from 4 to 8 billion parameters. These models are considered more computationally efficient [21] yet remain underutilized in AES research, despite prior findings such as those by Mansour et al [15], demonstrating that LLaMA 2, even at a medium scale of 13 billion parameters, can achieve competitive performance based on the Quadratic Weighted Kappa (QWK) metric, the use of a primary dataset of 100 essays, directly curated by English language experts from an autonomous public university in Indonesia, to ensure an authentic representation of the user context and learner proficiency [22], and the integration of automated feedback features that not only assign essay scores but also provide instructional suggestions to support the improvement of students' writing skills.

The goal of this work is to develop a more accurate and context-aware AES system. The system is expected to help students improve their essay writing skills [23], while also offering an efficient and scalable assessment tool for educational institutions [24] which providing quick, objective, and relevant feedback for both academic and international professional contexts.

2. Methods

Manually scoring and providing feedback on student's English essays have been viewed challenging [1] because the rubric assessment still leads to inconsistent score among higher education educators [2], [3]. Although such issues are addressed by implementing AES to assess essay writing, educators in higher education have not fully embraced it due to its inability to capture creative, artistic, and humanistic elements [4]. Its application in assessing students' creativity remains unsuccessful [5].

This study employs an experimental design that integrates prompt engineering with machine learning-based scoring to develop an automated essay scoring system by selecting only one best-performing LLM based on the highest QWK value. The dataset, constructed from manually scored student essays and detailed rubrics, was organized into a structured format and split into balanced training and testing sets.

As shown in Figure 1, four prompting strategies [25] were designed to guide the LLMs: zero-shot [26], few-shot [27] (with 1, 3, or 5 examples), Chain-of-Thought (CoT) [28], and Multi-Trait Specialization Prompts [29] (covering Content Argument, Organization, Language Style, and Theoretical Grounding). These prompts were manually crafted and embedded into the input to enhance the model's understanding of the scoring task. Five pre-trained LLMs—LLaMA 3.1:8b, Mistral 7b, Qwen 3:8b, DeepSeek R1:8b, and Gemma 3:4b—were tested using the proposed prompting methods to generate both predicted scores and textual feedback.

Scoring performance was evaluated using the QWK metric to assess agreement between model predictions and human-assigned scores. Meanwhile, the quality of generated feedback was evaluated using cosine similarity against reference feedback to measure semantic alignment. The LLM–prompt pairing with the highest QWK value was selected and further optimized through prompt tuning, enhancing both scoring accuracy and the relevance of generated feedback.

2.1 Data Source and LLMs Selection

The dataset used in this study consists of 100 academic essays collected from an essay-writing task assigned by lecturers at an autonomous public university. Students were given a set of predetermined topics and asked to write essays in document format. These essays were then submitted to their respective lecturers for evaluation. The scoring process involved a structured rubric-based assessment, where each essay was evaluated against 37 rubrics. These rubrics were developed and validated using keyword network analysis via VOS viewer, based on Scopus-indexed literature to ensure academic relevance, as shown as Figure 2. Additionally, the dataset and code can be accessed at <https://zenodo.org/records/16536693> [30].

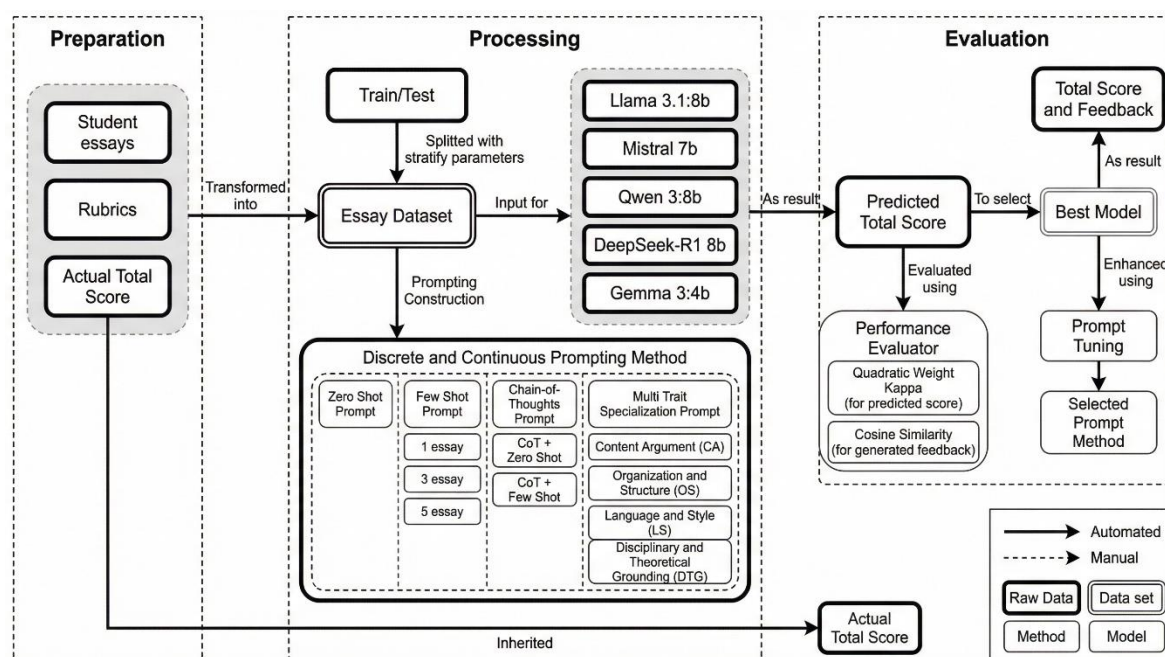


Figure 1. Overview of the Automated Essay Scoring (AES) Methodology Pipeline. This figure illustrates the full workflow for developing the AES system, integrating prompt engineering with machine learning-based scoring.

For each rubric, lecturers assigned a binary score of either 0 (not fulfilled) or 1 (fulfilled). Each rubric was also assigned a specific weight, and the final essay score was calculated by accumulating the weighted rubric scores, providing a comprehensive assessment of the student's writing ability, as shown in Table 1 (Essay Scoring Rubrics). After manual scoring by several English lecturers, we obtained a curated dataset of 100 academic essays written by students at an autonomous public university. Each essay was evaluated using 37 validated rubrics, scored with binary values (0 or 1), where each rubric carried a specific weight contributing to the final score. The resulting dataset includes both overall essay scores and fine-grained rubric-level

annotations, making it well-suited for automated scoring model development.

An overview of the dataset characteristics is provided in Table 2. Table 2 summarizes the dataset by categorizing essays into score ranges. No essays fall within the 0–20 range, while the majority are clustered between 41–80. Interestingly, higher-scoring essays tend to have lower average word and sentence counts. For example, essays in the 21–40 range average 687.75 words and 35.50 sentences, compared to 574.73 words and 28.82 sentences in the 81–100 range. Despite variations in length, syntactic depth—a proxy for sentence complexity—remains relatively stable across all score groups, suggesting that essay quality is not solely determined by length or structural complexity.

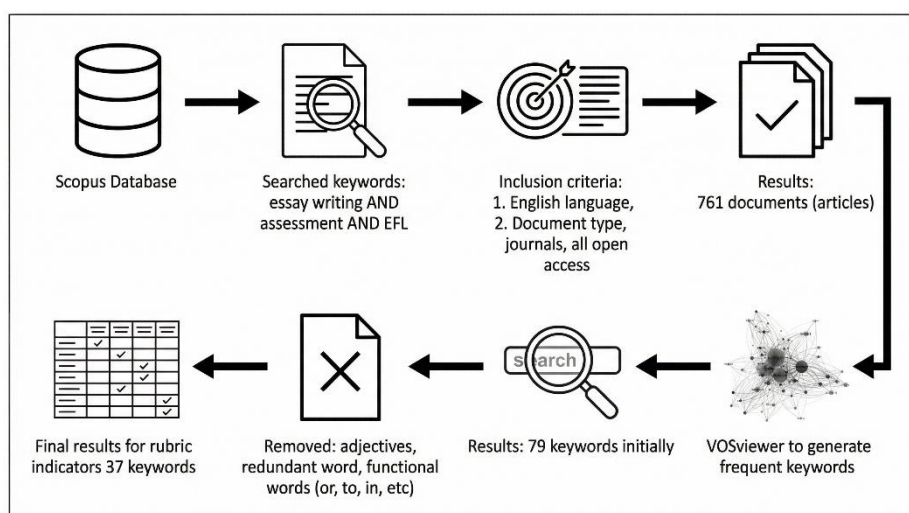


Figure 2. Essays were written by university students and graded using 37 rubric items derived from Scopus-indexed literature via VOSviewer. Lecturers assigned binary scores (0 or 1) to each item, weighted to calculate a total score.

Table 1. Essay Scoring Rubrics

w	Database Keyword	Indicator Description
4	argumentative	The essay has a clear Thesis statement to discuss; topic and controlling idea.
2	comparison	The essay compares two objects/two subjects as examples
3	syntactic complexity	The essay has a various type of sentences; simple, complex, compound, compound-complex
3	condition	The essay presents the condition/context where the problem arises
4	current study	the essay uses data/example/statistic that is current/up-to-date
2	difference	The essay indicates the essay is different from other essays.
2	engagement	The essay can attract readers to be engaged in reading the whole essay despite its introduction.
2	genre	The essay has a clear category /type of essay
4	grammar	The essay uses proper tenses in 80% of the whole sentences.
4	lexical sophistication	The essay uses proper dictions that do not reflect 'AI generated' words/dictions
2	motivation	The essay presents the writer's motivation in writing the essay /goals/aims
4	organization	The essay has a clear essay organization; introduction, thesis statement, body paragraphs, topic sentence, conclusion
2	pedagogical implication	the essay simply indicates impacts in practice.
3	relation	The essay presents a good relation from one sentence to another sentence
4	topic	The body paragraph has Topic Sentence that relates to the thesis statement
3	total	The essay meets the number of required words in total
2	vocabulary	The essay presents varied vocabulary despite sophisticated
2	combination	The essay combines definition, examples, details
2	education	The essay addresses issues relating or being implemented to Higher Education in practices.
2	effectiveness	The sentences are straight forward
2	field	The essay can be categorized in a precise field or multidisciplinary field.
4	form	the essay has a five-paragraph form without subtitles
2	framework	The essay employs a theoretical perspective
3	future research	The essay gives recommendation, suggestions, predictions in its conclusion
2	limitation	The essay presents what is not covered in the essay, or limitations
2	need	The essay presents there is a need to raise the issue written in the essay
4	review	The essay reviews the problems/ the context at the introduction
2	information	Information provided is concise, clear, chronologically
2	academic	Its vocabulary is academic
4	article	The essay cites references from the last 5-year articles.
2	discourse	The essay presents author's individual perspective in the Indonesian context
2	issue	The issue raised in the essay is happening now
		The essay addressess clear readers , to whom the essay is addressed bringing more impacts to readers.
2	reader	
4	reason	Body paragraphs present some reasons to support arguments.
2	strategy	The essay employes pros and cons in its structure
2	theory	The essay inserts some citations from a theoretical perspective

Each rubric has a weight (w) and is scored 0 or 1 by the lecturer. The final score per rubric is calculated as $w \times \text{binary score}$ (e.g., $4 \times 0 = 0$).

Table 2. Overview of Dataset

score range	sample size	word count mean	sentence count mean	syntactic depth mean
0-20	0	null	null	null
21-40	4	687.75	35.50	20.82
41-60	29	639.03	30.90	20.84
61-80	45	590.96	31.16	19.66
81-100	22	574.73	28.82	20.14

The selected models—LLaMA 3.1:8b, Mistral 7b, DeepSeek R1:8b, Qwen 3:8b, and Gemma 3:4b—were chosen for their structural similarity, modern optimization techniques, and suitability for low-resource environments. As shown in Table 3, most operate within the 7–8B parameter range and share consistent architectural traits such as 32 layers, 4,096-dimensional embeddings, and 32 attention heads, enabling fair benchmarking. One 4B model was intentionally included to explore whether a smaller-scale model can outperform larger ones, enabling comparison between mini and medium-sized models. Additionally, features like extended context lengths (e.g., Qwen 3: 128K, DeepSeek R1: 32K) and tuning methods (e.g., GQA, SFT, DPO) enhance their relevance for essay scoring tasks requiring nuanced semantic understanding over long sequences. Their

open-source availability on Ollama further supports reproducibility and accessibility.

Table 3. Selection of Open Source Lightweighted LLMs

Name	Publisher	Parameter Category	Release Time
Llama 3.1:8b	Meta	Medium	July 23, 2024
Mistral 7b	Mistral AI	Medium	September 27, 2023
Qwen 3:8b	Alibaba Cloud	Medium	April 29, 2025
Deepseek R1:8b	Deepseek	Medium	January 2025
Gemma 3:4b	Google	Mini	March 10, 2025

2.2 Performance Metric

This research is using Quadratic Weighted Kappa (QWK) as performance metric calculation where is formulated as shown in Equation 1. First, a histogram matrix $O \times N$ is constructed, where O, j corresponds to the number of installation_id i (actual) that receive the predicted scores. A weight matrix $N \times N$, ω is then calculate based on the difference between the actual and predicted values:

$$\omega_{i,j} = \frac{(i-j)^2}{(N-i)^2} \quad (1)$$

The histogram matrix $N \times N$ of the expected result, E is calculated under the assumption that there is no correlation between the scores. It is computed as the outer product of the actual outcome histogram vector and the predicted outcome histogram vector, normalized so that E and O have the same total sum. Thus, based on these three matrices, the Quadratic Weighted Kappa (QWK) is calculated using Equation 2.

$$\kappa = 1 - \frac{\sum_i \omega_{i,j} O_{i,j}}{\sum_i \omega_{i,j} E_{i,j}} \quad (2)$$

As shown in Table 4. It is important to note that the Quadratic Weighted Kappa score is a ratio that can range from -1 to 1. A negative QWK score implies that the model performs worse than a random model. A random model is expected to produce a score close to 0. Finally, a perfect prediction will yield a score of 1.

Table 4. Value Range of QWK

Value Range	Level of Consistency
< 0.2	Very Low
0.2–0.4	Low
0.4–0.6	Moderate
0.6–0.8	High
0.8–1.0	Almost Perfect

Furthermore, Figure 3 illustrates the performance of five Large Language Models (LLMs) in AES using the QWK metric across nine prompting techniques. It reveals significant variability in AES performance based on both the LLM and the prompting strategy, with few-shot learning (e.g., FS-1e) often proving beneficial for models like Llama 3.1:8b and Mistral 7b, while Chain-of-Thought (CoT) prompting can yield mixed results, even detrimental for some like Gemma 3:4b. The data emphasizes that selecting the optimal LLM and prompting technique is crucial for robust AES, as there is no single best approach for all scenarios.

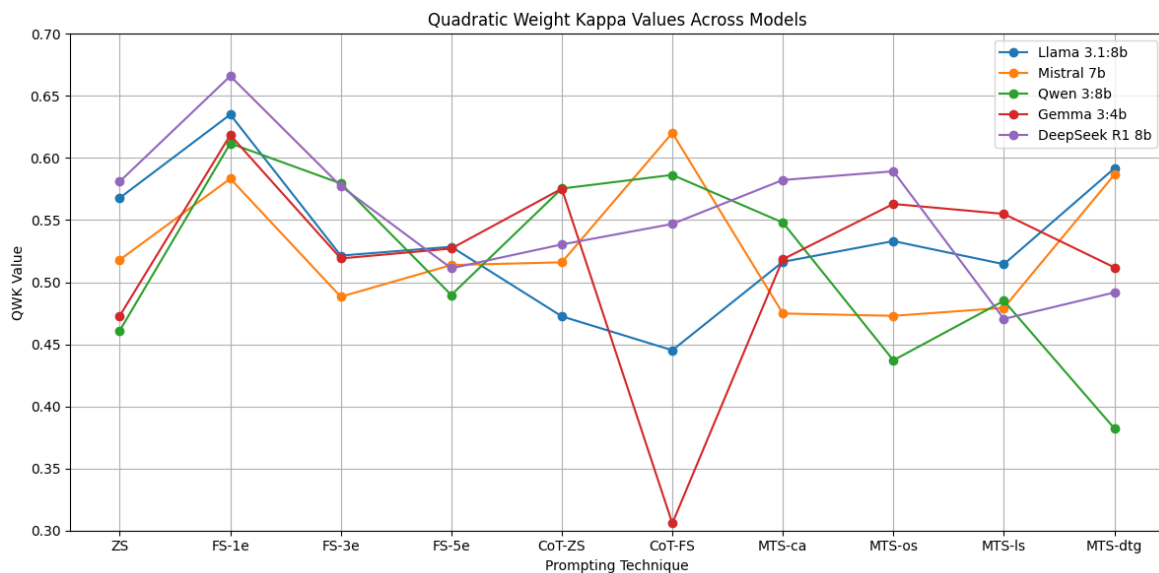


Figure 3. QWK performance comparison across five large language models (LLMs) under nine different prompting techniques for Automated Essay Scoring (AES).

3. Results and Discussions

The result highlights the comparative performance of different large language models (LLMs) based on their QWK scores across various prompting techniques. As shown in Table 5 (Summary of AES Performance Result) among the evaluated models, DeepSeek R1 8b emerges most frequently as the best-performing model, securing the highest QWK in four different techniques: ZS, FS-1e, FS-3e, and MTS-os. This indicates its strong reliability in both zero-shot and few-shot prompting scenarios. Qwen 3:8b excelling in four other techniques including FS-5e, CoT-ZS, MTS, and MTS-ca, showcasing its robustness particularly in multi-turn and few-shot settings. Importantly, Gemma 3:4b achieves the overall highest QWK score of 0.6660 under the FS-1e setup, indicating its peak performance when provided with a single example. Conversely, Gemma 3:4b and Llama 3:8b appear less frequently, suggesting

more limited generalizability across methods compared to DeepSeek R1 8b and Qwen 3:8b.

Table 5. Summary of AES Performance Result

Method	Best Model	Best QWK
ZS	DeepSeek R1 8b	0.5811
FS-1e	DeepSeek R1 8b	0.666
FS-3e	DeepSeek R1 8b	0.5774
FS-5e	Qwen 3:8b	0.5285
CoT	Gemma 3:4b	0.6347
CoT-ZS	Qwen 3:8b	0.5754
CoT-FS	Mistral 7b	0.6202
MTS	Qwen 3:8b	0.6341
MTS-ca	Qwen 3:8b	0.548
MTS-os	DeepSeek R1 8b	0.5893
MTS-ls	DeepSeek R1 8b	0.5944
MTS-dtg	Llama 3:8b	0.5917

Violin plot as shown in Figure 4 visualizes the density and dispersion of essay scores produced by the model versus the actual ground truth.

The left panel ("Actual") represents the score distribution as assigned by human raters, while the right panel ("Predicted") reflects the scores generated by DeepSeek R1 8b. Both distributions exhibit similar central tendencies and spread, suggesting that the model effectively captures the underlying scoring patterns. The interquartile ranges and median positions indicate that the model tends to approximate the human-labeled scores with minimal systematic bias, although slight overestimation is observed in the upper quartile. This alignment reflects DeepSeek R1 8b's robustness and sensitivity to the few-shot example provided, further supporting its strong performance in earlier QWK evaluations.



Figure 4. Distribution comparison between actual and predicted essay scores using best achieved QWK of 0.666 DeepSeek R1 8b with a few-shot (1-example) prompt.

3.1 Zero Shot Method

Zero-shot is a prompting technique that is natural and intuitive, as it involves interacting directly with the LLM without providing any examples or demonstration flows. As previously explained in methodology this research utilizes five (5) types of medium and mini LLMs, with parameter variations ranging from 8 billion (medium-size) to 4 billion (mini-size), for performance comparison. Table 6 presents the performance results of these five LLM models after applying the zero-shot prompting technique.

Table 6. AES Performance Result on Zero Shot Method

Model	QWK
Llama 3.1:8b	0.567
Mistral 7b	0.518
Qwen 3:8b	0.461
DeepSeek R1:8b	0.581
Gemma 3:4b	0.472

The results presented in the Table 6, DeepSeek R1:8B achieved the highest QWK score of 0.581, indicating the strongest agreement with human-assigned essay scores. LLaMA 3.1:8B with a QWK of 0.567, also demonstrating a high level of predictive reliability. In contrast, Qwen 3:8B and Gemma 3:4B performed less effectively, with QWK scores of 0.461 and 0.472 respectively, suggesting weaker alignment with human

evaluations. Mistral 7B placed in the middle with a score of 0.518.

3.2 Few-Shot Method

In contrast, few-shot prompting leverages the in-context learning capabilities of LLMs by including a small number of relevant input-output examples directly within the prompt. This research specifically explores three few-shot configurations: 1 essay, 3 essays, and 5 essays as illustrative examples. The selection of these quantities is aimed at identifying how the model adapts its understanding of the task with varying levels of contextual information—minimal (1 essay), moderate (3 essays), and slightly richer (5 essays). These configurations allow for the analysis of the point at which adding more examples either yields diminishing returns or leads to optimal performance on the given task, while also testing the model's ability to generalize from a limited set of demonstrations.

Table 7 presents the QWK scores achieved by five LLMs across three different few-shot prompting settings, using 1, 3, and 5 examples (fws-1e, fws-3e, fws-5e). Among all models, DeepSeek R1:8B consistently outperformed others in the 1-example setting with the highest QWK score of 0.666, indicating strong agreement with human raters even with minimal guidance. LLaMA 3.1:8B and Gemma 3:4B also demonstrated relatively stable performance in the 1-shot setup with QWK values of 0.635 and 0.618, respectively. However, performance tended to decline or fluctuate as the number of examples increased, suggesting that not all models benefitted from more demonstrations. Interestingly, Qwen 3:8B achieved its highest score in the 3-example setting (0.579), while most other models peaked with only one example, highlighting the variable impact of example count on model scoring consistency.

Table 7. AES Performance Result on Few Shots Method

Model	QWK fws-1e	QWK fws-3e	QWK fws-5e
Llama 3.1:8b	0.635	0.521	0.528
Mistral 7b	0.583	0.488	0.513
Qwen 3:8b	0.611	0.579	0.489
DeepSeek R1:8b	0.666	0.577	0.510
Gemma 3:4b	0.618	0.519	0.527

3.3 Chain-of-Thought Method

This research also explores the integration of Chain-of-Thought (CoT) prompting with both zero-shot and few-shot approaches for interaction with LLMs, specifically in the context of essay scoring tasks. The zero-shot CoT approach is applied by instructing the model to generate reasoning steps and responses without explicit examples, relying solely on the model's pre-trained knowledge. In contrast, the few-shot CoT method leverages the in-context learning capability of LLMs by providing a single essay as a reference example, enabling the model to mimic the desired reasoning structure. The use of only one example in the few-shot

setup is intended to test the model's ability to adapt its understanding of the task with minimal contextual information, while also evaluating its capacity to generalize and construct a coherent chain of thought from a highly limited demonstration in performing essay assessment.

Table 8 presents result using Chain-of-Thought (CoT) prompting strategies, combined with either zero shot (cot+zs) or few-shot (cot+fws) configurations. The results highlight that incorporating few-shot examples into CoT prompting generally leads to improved scoring consistency with human evaluators, particularly evident in Mistral 7B and Gemma 3:4B, which achieved QWK scores of 0.620 and 0.546, respectively—both higher than their zero-shot counterparts. Qwen 3:8B exhibits strong performance under both strategies, showing a slight increase from 0.575 (cot+zs) to 0.586 (cot+fws). In contrast, models such as LLaMA 3.1:8B and DeepSeek R1:8B perform better with zero-shot CoT, suggesting that the effect of additional few-shot examples may vary depending on model-specific tuning and architecture.

Table 8. AES Performance Result on Chain-of-Thought Method

Model	QWK cot+zs	QWK cot+fws
Llama 3.1:8b	0.472	0.445
Mistral 7b	0.516	0.620
Qwen 3:8b	0.575	0.586
DeepSeek R1:8b	0.530	0.456
Gemma 3:4b	0.505	0.546

3.4 Multi-Trait Specialization

The Multi-Trait Specialization technique decomposes the evaluation process into four distinct dimensions, each aligned with a specific scoring rubric: Content Argumentation (CA), Language and Style (LS), Organization and Structure (OS), and Disciplinary and Theoretical Grounding (DTG). For each rubric, the LLM is instructed to analyze and assign scores independently, enabling a more granular and focused assessment of each essay component. This approach aims to enhance the accuracy and interpretability of model evaluations by leveraging the LLM's ability to process complex and targeted instructions for each scoring dimension, in contrast to traditional holistic scoring methods.

As described in Table 9, Qwen 3 model achieved the highest QWK in the CA and DTG dimensions, indicating strong capabilities in content relevance and theoretical articulation. Meanwhile, DeepSeek R1 demonstrated superior performance in OS, suggesting its strength in understanding structural coherence. Gemma 3 performed consistently well across all traits despite its smaller parameter size (4B), particularly in OS and CA. These results highlight that while no single model dominates in all traits, specific models excel in certain scoring dimensions, offering valuable insights for targeted AES implementations.

Table 9. AES Performance Result on Multi-Trait Specialization Method

Model	mts-ca	mts-os	mts-ls	mts-dtg
Llama 3.1:8b	0.5164	0.5318	0.5146	0.472
Mistral 7b	0.4748	0.4729	0.4793	0.516
Qwen 3:8b	0.5480	0.4371	0.4850	0.575
DeepSeek R1:8b	0.4970	0.5893	0.4881	0.530
Gemma 3:4b	0.5185	0.5667	0.5089	0.505

The future trajectory of Automated Essay Scoring (AES) lies in developing Explainable Artificial Intelligence (XAI) systems that can match the rigor and interpretability demonstrated in advanced computational modeling. To truly realize the potential of AES, future work must focus on integrating deep learning architectures with explicit pedagogical insights, moving beyond surface-level feature detection. This necessitates building transparent and high-confidence assessment systems that utilize deep learning for performance prediction [31], while simultaneously tackling critical challenges such as data heterogeneity and bias inherent in educational datasets [31]. Furthermore, enhancing the core AI models requires continuous hyperparameter tuning and optimization techniques [32] to improve scoring accuracy for nuanced, qualitative constructs—the current weakness of AES. The field must also embrace advanced system architectures, such as Federated Learning and Edge Computing, to manage privacy and scale the deployment of these complex models across diverse institutional networks[33-34]. Finally, future AES development must leverage robust data methodologies, ensuring effective distributed data management and processing to underpin the efficiency of next-generation, integrated AI assessment platforms in a friendly elearning platform [35-36].

4. Conclusions

This study presented AESCORE, an automated essay scoring framework that integrates validated rubric-based assessment with lightweight open-source large language models (LLMs) to support reliable and pedagogically aligned writing evaluation. Through extensive experimentation on 100 manually scored EFL essays, the results demonstrate that medium-to-mini scale LLMs (4–8B parameters) can achieve strong agreement with human raters, particularly when guided by carefully designed prompting strategies.

The findings highlight that few-shot prompting with a single example offers an effective balance between scoring accuracy and computational efficiency, enabling cost-effective deployment without reliance on large proprietary models. This is especially significant for higher education institutions in resource-constrained environments, where access to scalable and transparent assessment tools remains limited.

Beyond scoring accuracy, AESCORE provides instructional feedback aligned with expert-validated rubrics, supporting formative assessment and the

development of students' academic writing skills. This positions the proposed framework not only as an evaluation tool, but also as a practical learning support system that complements human instruction rather than replacing it.

Overall, this work contributes to the advancement of human-centered, explainable, and scalable AES systems, demonstrating that accessible AI technologies can meaningfully support writing assessment in real educational contexts. Future work will extend this framework through explainable AI techniques, larger multi-institutional datasets, and privacy-preserving deployment strategies to further enhance robustness and adoption.

Acknowledgements

This work was supported by a research grant from the Politeknik Negeri Semarang, Universitas Airlangga, and the Universitas Muhammadiyah Yogyakarta (UMY) Research Fellowship Program.

References

- [1] M. Faseeh et al., "Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy," *Mathematics*, vol. 12, no. 21, 2024, <https://doi.org/10.3390/math12213416>
- [2] S. Mumford and D. Atay, "Teachers' perspectives on the causes of rater discrepancy in an English for Academic Purposes context," *Assess. Writ.*, vol. 48, p. 100527, Apr. 2021, <https://doi.org/10.1016/j.asw.2021.100527>
- [3] N. B. Uzun, D. Alici, and M. Aktas, "Reliability of the Analytic Rubric and Checklist for the Assessment of Story Writing Skills: G and Decision Study in Generalizability Theory," *Eur. J. Educ. Res.*, vol. 8, no. 1, pp. 169–180, Jan. 2019, <https://doi.org/10.12973/eu-jer.8.1.169>
- [4] R. Conijn, P. Kahr, and C. Snijders, "The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation," *J. Learn. Anal.*, vol. 10, no. 1, pp. 37–53, Mar. 2023, <https://doi.org/10.18608/jla.2023.7801>
- [5] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," *J. Comput. High. Educ.*, vol. 33, no. 3, pp. 727–746, Dec. 2021, <https://doi.org/10.1007/s12528-021-09283-1>
- [6] A. V. Y. Lee, A. C. Luco, and S. C. Tan, "A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning," *Educ. Technol. Soc.*, vol. 26, no. 1, pp. 147–159, 2023, [https://doi.org/10.30191/ETS.202301_26\(1\).0011](https://doi.org/10.30191/ETS.202301_26(1).0011)
- [7] X. Li, L. Qu, M. Tan, and Y. Jia, "Automated essay scoring based on the enhanced chimp optimization algorithm-back propagation (ENChOA-BP) and K-means," *Multimed. Tools Appl.*, vol. 1–32, no. 2024, 2024, <https://doi.org/10.1007/s11042-024-20313-9>
- [8] X. Tang, H. Chen, D. Lin, and K. Li, "Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments," *Heliyon*, vol. 10, no. 14, 2024, <https://doi.org/10.1016/j.heliyon.2024.e34262>
- [9] Y. Song, Q. Zhu, H. Wang, and Q. Zheng, "Automated Essay Scoring and Revising Based on Open-Source Large Language Models," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1880–1890, 2024, <https://doi.org/10.1109/TLT.2024.3396873>
- [10] M. Uto and M. Okano, "Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases," *IEEE Trans. Learn. Technol.*, vol. 14, no. 6, pp. 763–776, Dec. 2021, <https://doi.org/10.1109/TLT.2022.3145352>
- [11] J. Ruan et al., "Applying Large Language Models to Power Systems: Potential Security Threats," 2023, *arXiv*, <https://doi.org/10.48550/ARXIV.2311.13361>
- [12] A. Kundu and D. Barbosa, "Are Large Language Models Good Essay Graders?," 2024, *arXiv*, <https://doi.org/10.48550/ARXIV.2409.13120>
- [13] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha, "A Literature Survey on Open Source Large Language Models," in *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, Singapore: ACM, Jan. 2024, pp. 133–143, <https://doi.org/10.1145/3647782.3647803>
- [14] S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope, and T. Jhamtani, "Fine Tuning LLMs for Low Resource Languages," in *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, Dhulikhel, Nepal: IEEE, Jul. 2024, pp. 511–519, <https://doi.org/10.1109/ICIPCN63822.2024.00090>
- [15] W. Mansour, S. Albatarni, S. Eltanbouly, and T. Elsayed, "Can Large Language Models Automatically Score Proficiency of Written Essays?," Apr. 16, 2024, *arXiv*: arXiv:2403.06149, <https://doi.org/10.48550/arXiv.2403.06149>
- [16] R. Zhang et al., "LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention," Sep. 18, 2024, *arXiv*: arXiv:2303.16199, <https://doi.org/10.48550/arXiv.2303.16199>
- [17] H. Yoo, J. Han, S.-Y. Ahn, and A. Oh, "DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing," 2024, *arXiv*, <https://doi.org/10.48550/ARXIV.2402.16733>
- [18] J. Wilson and R. D. Roscoe, "Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy," *J. Educ. Comput. Res.*, vol. 58, no. 1, pp. 87–125, Mar. 2020, <https://doi.org/10.1177/0735633119830764>
- [19] Y. Karatay and L. Karatay, "Automated writing evaluation use in second language classrooms: A research synthesis," *System*, vol. 123, 2024, <https://doi.org/10.1016/j.system.2024.103332>
- [20] Y.-S. Lee, "Analysis of Small Large Language Models (LLMs)," *Int. J. Adv. Smart Conver.*, vol. 13, no. 4, pp. 155–160, Dec. 2024, <https://doi.org/10.7236/IJASC.2024.13.4.155>
- [21] S. Cahyawijaya, H. Lovenia, and P. Fung, "LLMs Are Few-Shot In-Context Low-Resource Language Learners," 2024, *arXiv*, <https://doi.org/10.48550/ARXIV.2403.16512>
- [22] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022, <https://doi.org/10.1007/s10462-021-10068-2>
- [23] J. Fleckenstein, L. W. Liebenow, and J. Meyer, "Automated feedback and writing: a multi-level meta-analysis of effects on students' performance," *Front. Artif. Intell.*, vol. 6, 2023, <https://doi.org/10.3389/frai.2023.1162454>
- [24] A. Alsharif, "Artificial Intelligence and the Future of Assessment: Opportunities for Scalable, Fair, and Real-Time Evaluation," *Libyan J. Educ. Res. E-Learn. LJERE*, pp. 42–52, Feb. 2025, <https://doi.org/10.65417/ljere.v1i1.5>
- [25] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, "Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation," 2024, *arXiv*, <https://doi.org/10.48550/ARXIV.2404.15845>
- [26] Q. Wang and J. M. Gayed, "Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting," *Comput. Assist. Lang. Learn.*, pp. 1–29, Jul. 2024, <https://doi.org/10.1080/09588221.2024.2371395>
- [27] L. Yoshida, "The Impact of Example Selection in Few-Shot Prompting on Automated Essay Scoring Using GPT Models," in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, vol. 2150, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds., in *Communications in Computer and Information Science*, vol. 2150, Cham:

- Springer Nature Switzerland, 2024, pp. 61–73. https://doi.org/10.1007/978-3-031-64315-6_5
- [28] G.-G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai, “Applying large language models and chain-of-thought for automatic scoring,” *Comput. Educ. Artif. Intell.*, vol. 6, p. 100213, Jun. 2024, <https://doi.org/10.1016/j.caeai.2024.100213>
- [29] T. Shibata and Y. Miyamura, “LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models,” 2025, *arXiv*. <https://doi.org/10.48550/ARXIV.2505.08498>
- [30] Prayitno, “Automatic AES dataset and Code.” Politeknik Negeri Semarang, Jul. 28, 2025. Accessed: Jul. 28, 2025. [Online]. Available: <https://zenodo.org/records/16536693>
- [31] H.-C. Chen et al., “Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence,” *Appl. Sci.*, vol. 12, no. 4, p. 1885, Feb. 2022, <https://doi.org/10.3390/app12041885>
- [32] K. Putra et al., “Optimizing the Performance of AI Model for Non-Invasive Continuous Glucose Monitoring: Hyperparameter Tuning and Random Oversampling Approach,” *JOIV Int. J. Inform. Vis.*, vol. 8, no. 2, p. 991, May 2024, <https://doi.org/10.62527/joiv.8.2.2047>
- [33] K. T. Putra et al., “Federated Compressed Learning Edge Computing Framework with Ensuring Data Privacy for PM2.5 Prediction in Smart City Sensing Applications,” *Sensors*, vol. 21, no. 13, p. 4586, Jul. 2021, <https://doi.org/10.3390/s21134586>
- [34] Prayitno et al., “A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications,” *Appl. Sci.*, vol. 11, no. 23, Art. no. 23, Nov. 2021, <https://doi.org/10.3390/app112311191>
- [35] D. Purbohadi and J. Santoso, “Child-friendly e-learning for artificial intelligence education in Indonesia: conceptual design,” *IAES Int. J. Artif. Intell.*, vol. 14, no. 4, pp. 2622–2633, 2025, <https://doi.org/10.11591/ijai.v14.i4.pp2622-2633>
- [36] Sunardi, Prayitno, B. P. Kamiel, A. D. Saputri, Z. H. Muizza, and A. Yobioktabera, “Smart Harvest: Web-Integrated Ripeness Detection for Apples with CNN Algorithm,” *Ingénierie Systèmes Inf.*, vol. 29, no. 6, pp. 2181–2190, Dec. 2024, <https://doi.org/10.18280/isi.290608>