



The Development of an Automated Essay Scoring (AES) Dataset for Indonesian Language Using OCR and Inter-Rater Validation

Aldila Putri Linanzha¹, Daniel Eliazar Latumaerissa^{2*}, Ayu Candra Dewi², Nuzila Putri Al-Bana³, Naimatul Ulumiyah², Yussyfi Akira Arlyn Alzena¹, Kezia Hanna Nugrahningtyas³

¹ Department of Informatics, UPN "Veteran" Yogyakarta, Daerah Istimewa Yogyakarta

² Department of Information System, UPN "Veteran" Yogyakarta, Daerah Istimewa Yogyakarta

³ Department of Industrial Engineering, UPN "Veteran" Yogyakarta, Daerah Istimewa Yogyakarta

Received : September 08, 2026

Revised : September 16, 2026

Accepted : September 17, 2026

Online : October 7, 2026

Abstract

The availability of Automated Essay Scoring (AES) datasets remains limited for low-resource languages such as Indonesian. This study develops and validates a reliable Indonesian AES dataset drawn from 219 authentic handwritten student exams. The methodology uses a multi-stage pipeline that includes image digitization, Optical Character Recognition (OCR) text extraction with meticulous manual verification, and a rigorous cross-validation protocol conducted by three independent expert annotators. For the per-item reliability subset ($n = 59$), the expert panel's consensus achieved good-to-excellent agreement ($ICC = 0.819-0.914$). In contrast, single-evaluator assessments agreed only moderately with the panel ($QWK \leq 0.642$), showing coarse grading tendencies and wider limits of agreement. Consequently, the panel's consensus average was established as the dataset's gold-standard label. Overall, this research contributes a highly reliable, structured educational dataset that provides a robust and ecologically valid foundation for training and evaluating AES models in low-resource contexts.

Keywords: *Automated Essay Scoring, Indonesian Natural Language Processing, Optical Character Recognition, Dataset Construction, Inter-Rater Reliability*

INTRODUCTION

Essay-based educational assessment is an important component of instructor evaluation because it allows teachers to assess students' higher-order thinking, reasoning, conceptual understanding, and communication skills (Liu et al., 2026). However, grading a large number of essay responses can lead to inconsistencies due to fatigue and differences in how graders interpret the criteria (Mi & Yu, 2026; Xu et al., 2024). Consequently, to improve grading efficiency while maintaining consistency and reliability, Automated Essay Scoring (AES) systems have seen rapid development (Gierl et al., 2014; Poonpon et al., 2023).

Advances in natural language processing (NLP) are driving the development of AES systems. Today's AES models no longer rely solely on manually designed linguistic features and conventional machine learning algorithms; instead, they use transformer-based language models, contextual semantic embeddings, and neural network architectures that effectively capture semantic relationships within text (Misgna et al., 2024). According to Li & Ng (2024) and Lin (2026), modern AES research focuses not only on improving prediction accuracy but also on developing assessment systems that are more robust, easier to understand, and widely applicable in education.

Today's AES research still faces challenges regarding the availability of high-quality annotated datasets (Li & Ng, 2024; Misgna et al., 2024). The availability of datasets, the quality of annotations, and the reliability of evaluations can affect the performance of the AES model.

Copyright Holder:

© Aldila, Daniel, Ayu, Nuzila, Naimatul, Yussyfi, & Kezia. (2026)

Corresponding author's email: danieleliazar.latumaerissa@upnyk.ac.id

This Article is Licensed Under:



Currently, benchmark datasets such as the Automated Student Assessment Prize (ASAP) are frequently used in English-language AES research; however, the available datasets remain limited in both quantity and linguistic diversity (Misgna et al., 2024). This limitation also makes it difficult for AES researchers to obtain sufficiently large datasets with reliable human annotations for less commonly used languages.

Another challenge is that today's educational institutions still rely on handwritten essay exams. These handwritten responses must undergo a lengthy process before they can be used to develop the model. Handwritten documents must first be digitized using Optical Character Recognition (OCR), followed by text preprocessing, quality verification, and the organization of structured data. If any step in the text extraction process encounters problems, it can affect the quality of the resulting training data. The quality of this data contributes significantly to algorithmic improvements in the overall performance of the automated scoring system in the AES study (Sun et al., 2025).

Another concern in compiling an education dataset is the reliability of essay scores assigned by human evaluators (Pack et al., 2024). According to Shibata et al. (2025), essay grading can inherently involve subjective bias, leading to inconsistencies and inaccuracies among graders and thereby affecting the supervised learning model. Therefore, a rigorous annotation strategy involving multiple expert evaluators and standardized assessment rubrics is necessary to improve the reliability of ground-truth labels (Misgna et al., 2024). Improvements in model performance and the credibility of artificial intelligence (AI) research in the field of education can be achieved through enhanced, reliable data annotation (Mohammed et al., 2025).

Therefore, the objective of this research is to develop and validate a reliable Indonesian Automated Essay Scoring (AES) dataset derived from authentic handwritten student exams, using an OCR-based text extraction pipeline and inter-rater cross-validation. Unlike previous studies that predominantly focus on model architecture, this work establishes a standardized framework for dataset construction to support future educational NLP applications in low-resource settings.

LITERATURE REVIEW

Recent Advances in Automated Essay Scoring

Automated Essay Scoring (AES) systems have been extensively developed to facilitate efficient and consistent grading of essays in the field of education (Elmassry et al., 2025). Research on AES has now evolved from primarily using conventional machine learning approaches to transformer-based architectures, contextual semantic embeddings, and large-scale language models (LLMs). These models can capture semantic meaning, discourse coherence, and contextual relationships more effectively (Ramesh & Sanampudi, 2022; Uto, 2021). Advances in this area improve evaluation performance on educational datasets while expanding the application of AES in the field of education (Huang et al., 2026; Ramnarain-Seetohul et al., 2025; Xiao et al., 2025).

AES research has now shifted its focus from prediction accuracy to improving model robustness, interpretability, and cross-domain generalization to ensure the reliability of the AES system for use in the field of education (Plasencia-Calaña, 2025; Ravindran & Choi, 2025). LLM-based multimodal metrics demonstrate that essay evaluation today is no longer based solely on lexical similarity but also incorporates reasoning and a deep understanding of the response context (Atkinson & Palma, 2025; Zhang et al., 2026).

Dataset Construction and Annotation Quality in AES

The quality of the dataset used has a significant impact on AES performance (Sun et al., 2025). The required dataset quality includes representative essays, a balanced score distribution,

a standardized grading rubric, and reliable human annotations (Latif & Zhai, 2026). Limited availability of datasets also poses a challenge for developing the AES system, particularly for languages with limited NLP resources (Misgna et al., 2024; Švábenský et al., 2026; Yoo et al., 2025).

The importance of dataset construction has become increasingly evident with the introduction of new benchmark datasets designed for educational assessment. Yoo et al. (2025) proposed DREsS, a large-scale, rubric-based dataset comprising nearly 49,000 essays collected from authentic English as a Foreign Language (EFL) learning environments and evaluated by expert assessors. Their study demonstrates that carefully designed datasets, accompanied by standardized annotation procedures, substantially improve the development and evaluation of AES systems.

Another critical aspect concerns the reliability of human-generated labels. Essay scoring is inherently subjective, and rater inconsistencies can introduce noisy annotations that degrade supervised learning performance. Recent studies therefore recommend inter-rater annotation strategies and transparent scoring procedures to improve dataset credibility and reduce annotation bias. Rather than treating annotation merely as a supporting activity, recent AES research increasingly recognizes annotation quality as an integral component of dataset construction (Su et al., 2025; Urrutia et al., 2025).

Handwritten Essay Digitization and Research Gap

Studies focused on developing essay datasets remain relatively limited. Existing datasets are still in digital text format; none are sourced from handwritten exam papers, which must first undergo a digitization process. Therefore, Optical Character Recognition (OCR) must be used to convert educational datasets into machine-readable formats suitable for NLP applications (Ghadekar et al., 2025; Pawar et al., 2025). AES datasets are currently more widely available in English and other commonly used languages, whereas datasets in Indonesian are still very limited (Wilie et al., 2020; Winata et al., 2023). The lack of validated Indonesian essay datasets limits opportunities for reproducible research and objective comparison of AES models developed for the Indonesian language.

Based on the reviewed literature, previous studies have predominantly emphasized algorithm development, neural architectures, and scoring optimization, whereas considerably less attention has been devoted to the systematic development of validated handwritten essay datasets. Therefore, this study addresses this research gap by proposing a reproducible workflow for constructing an Indonesian handwritten essay dataset, covering handwritten essay acquisition, OCR-based text extraction, preprocessing, structured dataset organization, and independent annotation by three expert raters. The resulting dataset is expected to support future AES research and provide a reliable benchmark for Indonesian educational NLP applications.

RESEARCH METHOD

This study proposes a multi-stage methodology for constructing and validating a reliable Indonesian Automated Essay Scoring (AES) dataset from physical student answer sheets. The framework comprises four primary phases: (1) data collection, (2) Optical Character Recognition (OCR) processing, (3) data storage and structuring, and (4) inter-rater cross-validation.

Data Collection and Digitization

The primary data consisted of 438 authentic handwritten essay responses collected from 219 students during real classroom assessments. The sample was selected using a purposive sampling method (Campbell et al., 2020) to meet specific coursework criteria, thereby ensuring the validity of the dataset. To maintain ecological validity while facilitating text extraction, each

physical answer sheet was digitized into an image dataset. The image capture protocol was strictly controlled to ensure even lighting, an upright orientation, and adequate resolution to preserve the quality of the raw handwritten artifacts. Furthermore, ethical research procedures were rigorously implemented, including securing informed consent and replacing all student names with unique anonymized identifiers to protect participant privacy.

Optical Character Recognition (OCR) and Text Extraction

Text extraction from the digitized images was performed using a Handwritten Text Recognition (HTR/OCR) engine. Given that OCR outputs for the Indonesian language are highly prone to errors, a comprehensive manual verification process was conducted across all collected responses. The unedited OCR text was preserved in the database, while the manually corrected text served as the definitive ground truth, enabling calculation of the Character Error Rate (CER) and Word Error Rate (WER). Crucially, the essay scoring procedure was performed by viewing the original handwritten images rather than the OCR-generated text, ensuring that extraction errors did not compromise the validity of the assigned scores.

Data Storage and Structuring

After the extraction and verification phases, all data were systematically organized into a Comma-Separated Values (CSV) format. The database schema includes a unique anonymized identifier, generalized course and class codes, specific question item numbers, the manually verified OCR text, and the evaluation scores. To support comprehensive analysis of annotation quality, the dataset structure explicitly links the original score assigned by the initial instructor with the independent scores provided by the three expert validators. In the final dataset release, the established consensus score serves as the primary gold-standard label, while the individual evaluator scores are documented transparently as supplementary properties.

Inter-Rater Cross-Validation and Reliability Analysis

To ensure high-quality ground-truth annotations, an inter-rater cross-validation protocol was implemented in which three independent expert validators blindly rescored every answer using a standardized rubric. Inter-rater reliability was quantitatively assessed per question ($n = 59$ for a specific question item) using the Intraclass Correlation Coefficient (ICC) ([Koo & Li, 2016](#)) and Krippendorff's alpha ([Hayes & Krippendorff, 2007](#)), with 95% bootstrap confidence intervals to assess sample-size stability. Additionally, agreement between the initial single-rater score and the panel's consensus was evaluated using Quadratic Weighted Kappa (QWK), Spearman correlation, systematic bias, and Bland-Altman limits of agreement. Based on these rigorous reliability metrics, the panel's consensus average was adopted as the definitive gold-standard label for training Automated Essay Scoring (AES) models, providing a significantly more stable and reliable target than a single idiosyncratic evaluator's score.

FINDINGS AND DISCUSSION

OCR Performance

The digitization process successfully converted the physical handwritten answer sheets into an image dataset, establishing a traceable source for text extraction. Recognizing that Indonesian handwriting is prone to recognition errors, the raw text generated by the OCR engine was meticulously verified and manually corrected against the original images to ensure high data fidelity. The final structured dataset systematically maps each anonymized student response to specific question items, preserving both the raw OCR output and the manually cleaned text.

Crucially, to prevent error propagation, the scoring process was conducted directly on the original images, so OCR inaccuracies affect text features but do not compromise score validity.

Inter-Rater Reliability

To evaluate annotation quality, two complementary analyses were conducted on the responses to questions 1a and 1b ($n = 59$ per question): (i) inter-rater reliability among the three independent expert validators, and (ii) agreement between the original lecturer score and the panel consensus (the mean of the three validators). Because these analyses address different questions, they are interpreted separately. The results are summarized in Table 1.

Table 1. Reliability Test Results

	Question 1a	Question 1b
ICC (2,k)	0.819	0.914
ICC (2,1)	0.780	0.601
Krippendorff's- α	0.596	0.778
Quadratic Weighted Kappa (QWK)	0.199	0.642
Spearman rho	0.001	0.465
Systematic Bias	+4.44 (Significant)	-0.83 (Non-significant)
Bland-Altman	-2.06	-9.55

Agreement among the three validators was good to excellent. For the averaged panel score, ICC(2,k) was 0.914 for question 1a and 0.819 for question 1b, indicating that the mean of the three validators is a highly reliable measure in both cases (Koo & Li, 2016). Krippendorff's- α , which reflects agreement closer to the single-rating level, was 0.778 for 1b (acceptable) but lower for 1a at 0.596, below the 0.667 tentative threshold (Hayes & Krippendorff, 2007).

The gap between a high ICC(2,k) and a more modest α for question 1a is expected rather than contradictory: ICC(2,k) is an average-measures coefficient and is therefore higher when ratings are aggregated across the three raters, whereas α approximates the reliability of a single rating. In practical terms, the three-validator consensus is reliable for both questions, whereas agreement at the level of an individual validator is weaker for 1a than for 1b.

In this dimension, the two questions diverge sharply. For question 1b, the lecturer's scores aligned substantially with the panel consensus (QWK = 0.642, "substantial" on the Landis & Koch scale; Spearman $\rho = 0.465$), with no significant systematic bias (-0.83 , non-significant). The lecturer and the panel therefore ranked and scored responses to 1b in broadly consistent ways.

For question 1a, agreement was much weaker. QWK dropped to 0.199 ("slight"), and Spearman ρ was 0.001, indicating virtually no rank correlation between the lecturer's ordering of students and the panel's. This was accompanied by a significant positive bias of +4.44 points, meaning the lecturer scored responses to 1a markedly higher than the panel on average. In other words, for 1a, the lecturer's single scores deviated from the consensus both in ordering and in level.

It should be noted that this analysis measures agreement, not construct validity; the panel's strong internal reliability indicates consistency, not proof that the scores capture essay quality in an absolute sense. Content and construct validity are left to future work.

However, a critical divergence was observed: the QWK metric (0.642) outperformed the Spearman correlation (0.465), indicating that although the assigned scores were numerically similar, the relative ranking of students differed between the instructor and the panel. Furthermore, the initial instructor's tendency to assign scores in coarse intervals (e.g., 0, 5, 10, 15) produced numerous ties, which mechanically depressed the rank correlation and widened the Bland-Altman limits of agreement (question 1a: -9.55 to $+7.90$; question 1b: -2.06 to $+10.95$). This

finding empirically demonstrates the vulnerability of single-rater annotations, in which individual grading idiosyncrasies, such as coarse granularity, become embedded in the data.

Implications for Dataset Labeling and AES Application

The primary implication of the reliability analysis is establishing the panel's consensus average as the dataset's gold-standard label, rather than relying on the idiosyncratic single-rater score. An AES model is only as robust as the labels it learns from; thus, using a highly reliable consensus score (ICC = 0.914) provides a significantly more stable training target. Although the constructed dataset is limited to a specific course and cohort and relies on human consensus, which inherently carries subjectivity, it still contributes a rare, validated Indonesian educational resource. By transparently documenting the pipeline from authentic handwritten exams to a structured dataset with measured reliability, this research provides a vital foundation for advancing multilingual AES models in low-resource contexts.

CONCLUSIONS

This study developed and validated a highly reliable Indonesian Automated Essay Scoring (AES) dataset from authentic handwritten exams, providing a crucial educational resource with a stable panel-consensus gold standard (ICC = 0.914) for low-resource language NLP. The analysis shows that inter-rater consensus scores serve as a significantly more robust training target than single-evaluator assessments, which often display coarse grading idiosyncrasies and wide limits of agreement.

LIMITATIONS & FURTHER RESEARCH

The current dataset is fundamentally limited by its single-cohort scope and its primary focus on measuring annotation reliability rather than on full construct or content validity. Future research should expand data collection across diverse academic disciplines, establish formal validity evidence, apply Generalizability Theory for more detailed rater analysis, and use this validated dataset to train and evaluate hybrid AES models.

REFERENCES

- Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15, 14551. <https://doi.org/10.1038/s41598-025-87862-3>
- Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., & Walker, K. (2020). Purposive sampling: Complex or simple? Research case examples. *Journal of Research in Nursing*, 25(8), 652–661. <https://doi.org/10.1177/1744987120927206>
- Elmassry, A., Zaki, N., Alsheikh, N., & Mediani, M. (2025). A systematic review of pretrained models in automated essay scoring. *IEEE Access*, 13, 121902–121917. <https://doi.org/10.1109/ACCESS.2025.3584784>
- Ghadekar, P., Khanvilkar, O., Kharat, J., Joshi, K., Kulkarni, T., & Kulkarni, Y. (2025). Automating the grading of handwritten examinations through the integration of optical character recognition and machine learning algorithms. *Journal of Integrated Science and Technology*, 13(6), 1128. <https://doi.org/10.62110/sciencein.jist.2025.v13.1128>
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48(10), 950–962. <https://doi.org/10.1111/medu.12517>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>

- Huang, Y., Palermo, C., & Wilson, J. (2026). Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing*, 69, 101047. <https://doi.org/10.1016/j.asw.2026.101047>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Latif, E., & Zhai, X. (2026). Using artificial intelligence for automatic scoring. In X. Zhai & G. Lee (Eds.), *Artificial intelligence for STEM education research: Advanced methods and applications* (pp. 17–37). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-06565-0_2
- Li, S., & Ng, V. (2024). Automated essay scoring: A reflection on the state of the art. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17876–17888). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.991>
- Lin, F. (2026). CNN–transformer-based trait-aware scoring and student-aware feedback generation for English writing. *Discover Artificial Intelligence*, 6(1), 405. <https://doi.org/10.1007/s44163-026-01132-3>
- Liu, T., Ye, L., & Yan, W. (2026). A framework for evaluation of large language models in essay assessment: Reliability, alignment, and causal reasoning. *Computers and Education: Artificial Intelligence*, 10, 100565. <https://doi.org/10.1016/j.caeai.2026.100565>
- Mi, T., & Yu, T. (2026). GDBERT-score: Semantic graph-enhanced DeBERTa for automated essay scoring in higher education. *Alexandria Engineering Journal*, 145, 154–163. <https://doi.org/10.1016/j.aej.2026.04.019>
- Misgna, H., On, B.-W., Lee, I., & Choi, G. S. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2), 36. <https://doi.org/10.1007/s10462-024-11017-5>
- Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, 132, 102549. <https://doi.org/10.1016/j.is.2025.102549>
- Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
- Pawar, P. A. H., Chavan, V., Pol, V., Shinde, H., & Thorat, S. (2025). Automating handwritten answer evaluation: A deep learning and OCR integrated approach. *International Research Journal of Education and Technology*, 8(4), 1120–1123. <https://www.irjweb.com/Automating%20Handwritten%20Answer%20Evaluation%20A%20Deep%20Learning%20and%20OCR%20Integrated%20Approach.pdf>
- Plasencia-Calaña, Y. (2025). *Operationalizing automated essay scoring: A human-aware approach* (arXiv:2506.21603). arXiv. <https://doi.org/10.48550/arXiv.2506.21603>
- Poonpon, K., Manorom, P., & Chansanam, W. (2023). Exploring effective methods for automated essay scoring of non-native speakers. *Contemporary Educational Technology*, 15(4), ep475. <https://doi.org/10.30935/cedtech/13740>
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Ramnarain-Seetohul, V., Rosunally, Y., & Bassoo, V. (2025). A unified conceptual hybrid framework for the automated assessment of short answers. *International Journal of Artificial Intelligence in Education*, 35(5), 2923–2974. <https://doi.org/10.1007/s40593-025-00487->

- Ravindran, R., & Choi, I. (2025). Investigating adversarial robustness in LLM-based AES. In J. Wilson, C. Ormerod, & M. Beiting Parrish (Eds.), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated session papers* (pp. 86–91). National Council on Measurement in Education. <https://aclanthology.org/2025.aimecon-sessions.10/>
- Shibata, T., Tomikawa, Y., Ito, Y., & Uto, M. (2025). Enhancing neural automated essay scoring accuracy by removing noisy data through data valuation. In *Proceedings of the Second Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC 2025)*. CEUR Workshop Proceedings. <https://ceur-ws.org/Vol-4006/paper6.pdf>
- Su, J., Yan, Y., Fu, F., Han, Z., Ye, J., Liu, X., Huo, J., Zhou, H., & Hu, X. (2025). EssayJudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 6363–6389). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.329>
- Sun, J., Song, T., Peng, W., & Song, J. (2025). A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing*, 650, 130916. <https://doi.org/10.1016/j.neucom.2025.130916>
- Švábenský, V., Flanagan, B., López Zapata, E. D., & Shimada, A. (2026). Open datasets in learning analytics: Trends, challenges, and best practice. *ACM Transactions on Knowledge Discovery from Data*, 20(4), Article 56. <https://doi.org/10.1145/3798096>
- Urrutia, F., Buc, C., Araya, R., & Barriere, V. (2025). Unsupervised automatic short answer grading and essay scoring: A weakly supervised explainable approach. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)* (pp. 38–54). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.bea-1.4>
- Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2), 459–484. <https://doi.org/10.1007/s41237-021-00142-y>
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In K.-F. Wong, K. Knight, & H. Wu (Eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 843–857). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.aacl-main.85>
- Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Fung, P., Baldwin, T., Lau, J. H., Sennrich, R., & Ruder, S. (2023). NusaX: Multilingual parallel sentiment datas
- Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025). Human-AI collaborative essay scoring: A dual-process framework with LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 293–305). Association for Computing Machinery. <https://doi.org/10.1145/3706468.3706507>
- Xu, W., Mahmud, R., & Hoo, W. L. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, 12, 77639–77657. <https://doi.org/10.1109/ACCESS.2024.3399163>
- Yoo, H., Han, J., Ahn, S.-Y., & Oh, A. (2025). DREsS: Dataset for rubric-based essay scoring on EFL writing. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
(pp. 13439–13454). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2025.acl-long.659>

Zhang, H., Su, J., & Liu, L. (2026). *Decision-level ordinal modeling for multimodal essay scoring with large language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2603.14891>