ESSAY QUESTION EVALUATOR: AI-Powered Essay Evaluation Project

(Essay Question Scores using Large Language Models (LLMs)-Artificial Intelligence Powered Essay Assessment Project)

By

Nannim David Dandam,

THE SCHOOL OF COMPUTING, COLLEGE OF SCIENCES, FEDERAL UNIVERSITY OF EDUCATION PANKSHIN

dandamnannim125@fuep.edu.ng

08141232871

&

Datti Useni Emmanuel

THE SCHOOL OF COMPUTING, COLLEGE OF SCIENCES, FEDERAL UNIVERSITY OF EDUCATION PANKSHIN

datti.useni.emmanuel@fuep.edu.ng

&

Nanbal Jibba Ladan

THE SCHOOL OF COMPUTING, COLLEGE OF SCIENCES, FEDERAL UNIVERSITY OF EDUCATION PANKSHIN

nanballadan@fuep.edu.ng

&

Gokir Justine Ali

THE SCHOOL OF COMPUTING, COLLEGE OF SCIENCES, FEDERAL UNIVERSITY OF EDUCATION PANKSHIN

jgokir1@fuep.edu.ng

Abstract

The increasing enrollment in writing-intensive university courses has intensified the challenge of providing timely, individualized feedback on student writing. This paper presents AI Essay Evaluator, an Automated Essay Scoring (AES) framework powered by large language models (LLMs) to deliver immediate, multidimensional, criterion-referenced evaluations of academic prose. Unlike early AES systems that relied on shallow linguistic features, AI Essay Evaluator leverages transformer-based models (GPT-4) to assess writing along four validated dimensions: Grammar & Mechanics, Coherence & Structure, Lexical Richness & Appropriateness, and Prompt Relevance & Argumentation. A validation study of 250 undergraduate essays demonstrated high agreement (Cohen's $\kappa = 0.86$) between AI Essay

Evaluator and expert human graders, with strong user satisfaction for feedback clarity and utility. The paper concludes by discussing ethical safeguards, limitations, and a sustainable freemium business model to ensure scalability and accessibility.

INTRODUCTION

The demand for scalable and reliable feedback in higher education continues to rise as class sizes increase globally. Writing-intensive disciplines face particular strain, with instructors spending disproportionate time providing individualized feedback (Wilson & Czik, 2016). Yet, as Hattie and Timperley (2007) emphasize, *timely feedback* is among the most powerful influences on learning outcomes. Delays degrade motivation, self-regulation, and revision efficacy.

Automated Essay Scoring (AES) technologies were conceived as a remedy. However, first-generation systems like *Project Essay Grade* (Page, 1994) focused on superficial features—word count, syntactic variety, or lexical frequency—failing to engage with the semantic and rhetorical depth of writing (Shermis & Burstein, 2013). Such tools were useful for large-scale testing but unsuitable for formative feedback.

The emergence of *transformer-based LLMs* (Vaswani et al., 2017) has redefined the possibilities of computational text evaluation. Recent models such as GPT-4 demonstrate not only linguistic precision but also contextual reasoning and rhetorical awareness (Kasneci et al., 2023). AI Essay Evaluator operationalizes these capabilities within a pedagogically structured AES framework, explicitly designed to align with composition theory and evidence-based learning principles.

Automated Essay Scoring (AES) systems are transforming educational assessment by providing rapid, consistent, and scalable feedback on student writing. Traditional manual grading faces limitations such as bias and delayed feedback, which hamper timely student improvements. Recent advances in artificial intelligence, especially natural language processing (NLP), enable the development of AES systems that evaluate essays on dimensions like coherence, grammar, and relevance using machine learning models. Additionally, plagiarism detection and paraphrasing assistance are vital components to uphold academic integrity and foster original writing (Ajit, 2025) (Yakubu, 2024).

This paper details a simplified yet practical AES implementation combining state-of-the-art text classification capabilities with transformer-based paraphrasing and rudimentary plagiarism detection. Presented as a mobile and desktop app-ready solution.

The system is designed to:

i. Assess essay quality on categorical levels (e.g., high, medium, low).
ii. Identify potential plagiarism against a reference text corpus.
iii. Manage essay files and feedback reports, supporting batch processing and scalability.

LITERATURE REVIEW

Evolution of Automated Essay Scoring

AES has evolved through three primary phases. Early statistical systems (e.g., Page, 1994) used linear regression models trained on human-scored essays. Second-generation systems introduced latent semantic analysis (LSA) to assess conceptual similarity (Landauer et al., 2003).

The Emergence of Advanced AEE Systems: Throughout the 1970s and 1980s, significant progress was made in natural language processing and computational linguistics. As computers became more powerful, researchers were able to develop AEE systems capable of analyzing text structure, grammar, and syntax with greater accuracy. These systems could assign preliminary scores based on predefined rules and patterns identified by human experts. (Pirnay-Dummer & Ifenthaler, 2010).

With advances in NLP, neural architectures—RNNs, CNNs, and eventually transformers—enabled models to learn complex semantic representations (Taghipour & Ng, 2016).

Advancements in Neural Networks: In recent years, the application of deep learning and neural networks has revolutionized AEE. These techniques allow AEE systems to understand context and meaning, resulting in more accurate and nuanced evaluations. Neural network-based AEE systems can now analyze large datasets, learn from them, and generate human-like evaluations, marking a significant step forward in essay assessment. (Ifenthaler & Grieff, 2021).

Current LLM-based systems transcend surface-level evaluation by leveraging contextual embeddings to infer argumentation quality, discourse coherence, and rhetorical appropriateness (Kumar et al., 2023).

LLMs in Formative Assessment

Recent research highlights the pedagogical promise of LLMs for formative feedback. Kasneci et al. (2023) and Zawacki-Richter et al. (2023) found that GPT-based models could generate

accurate, rubric-aligned feedback comparable to that of experienced educators when appropriately prompted. Studies by Ding et al. (2024) also emphasize the importance of structured "rubric-aware prompting" to enhance reliability and transparency.

However, most current AES applications remain monolithic—producing a single holistic score rather than diagnostic, multidimensional evaluations. AI Essay Evaluator advances the field by decomposing assessment into discrete pedagogical dimensions, thereby bridging formative assessment theory with modern AI capabilities.

AES systems prominently utilize machine learning on handcrafted or deep-learned features extracted from essay text such as syntactic structure, semantic coherence, and linguistic complexity. FastAI, built on PyTorch, offers efficient transfer learning via AWD-LSTM models for text classification tasks, providing a robust foundation for essay quality scoring. Transformer architectures, such as T5 and BERT, enable contextual understanding and paraphrasing, addressing semantic rewrite needs beyond lexical similarity (Masethe et al., 2024).

Plagiarism detection methods range from string matching techniques (e.g., n-gram or sequence matching) to vector-based semantic similarity approaches leveraging embeddings. While commercial tools like Turnitin integrate extensive databases and advanced algorithms, open-source implementations can initially employ text similarity metrics like 'difflib' to identify high overlap with reference corpora (Amirzhanov et al., 2025).

METHODOLOGY

System Architecture

AI Essay Evaluator is implemented as a Python-based web service integrated with OpenAI's GPT-4 API. A fixed prompt template casts the model as a "Senior Academic Evaluator" with four scoring dimensions. Each essay is processed via secure, anonymized channels, and the output includes:

i. A 0–100 numeric score for each dimension.
ii. Three targeted, criterion-linked bullet points per dimension.
iii. No overall holistic comment, ensuring consistency and modularity.

Evaluation Metrics

i. *Grammar & Mechanics*: Accuracy of grammar, punctuation, and syntax.

ii. *Coherence & Structure*: Logical flow, organization, and transitions.
iii. *Lexical Richness & Appropriateness*: Vocabulary precision and tone.
iv. *Prompt Relevance & Argumentation*: Thesis strength, evidence integration, and reasoning depth.

These metrics were adapted from validated writing assessment frameworks (White, 2007; Shermis & Burstein, 2013).

Validation Study

A dataset of 250 anonymized undergraduate essays was collected from writing courses under IRB approval. Each essay was double-scored by two trained human raters (inter-rater reliability $\kappa = 0.91$). The mean of their scores served as the benchmark for comparison with AI Essay Evaluator's outputs. Cohen's Kappa and Pearson correlation coefficients were computed for agreement analysis. Additionally, 50 students participated in a feedback-utility survey assessing clarity and actionability.

RESULTS AND DISCUSSION

The system's modular design facilitates educational use, allowing students to understand and extend each component. While the classifier performs acceptably on the small sample dataset, real-world deployment requires cloud and edge infrastructure and possibly more sophisticated architectures to capture nuanced quality features (Uyar & Büyükahıska, 2025).

Transformer-based paraphrasing provides a useful substitute for proprietary services, yet lightweight models like T5-small may produce simpler rewrites without deeper contextual creativity. Larger models or fine-tuning specific to paraphrasing could improve suggestions (Malik et al., 2023).

Alignment with Educational Goals

This project exemplifies core competencies from a file processing course, including:

i. Integration of modern and scalable AI libraries and application for real-world application development.
ii. Automation and report generation, simulating workplace feedback pipelines.
iii. Error handling and modular code structure aligned with software engineering principles.

Its hands-on nature supports learners in bridging theoretical AI concepts with tangible programming skills in cloud environments.

Agreement with Human Graders

AI Essay Evaluator achieved strong concordance with human ratings (κ = 0.86 overall). Subscale reliability was highest for Grammar (κ = 0.92) and lowest for Lexical Richness (κ = 0.80), reflecting the inherently subjective nature of lexical evaluation.

Metric	Mean (AI)	Mean (Human)	Pearson r	Cohen's κ
Grammar & Mechanics	84.2	85.1	.93	.92
Coherence & Structure	79.0	79.4	.89	.84
Lexical Richness	77.3	78.0	.86	.80
Prompt Relevance	81.0	82.2	.90	.85
Composite	**80.4**	**81.2**	**.91**	**.86**

Where $p < .001$

Student Perceptions of Feedback

Eighty-eight percent of surveyed students rated AI Essay Evaluator feedback as "clear and specific," and 74% found it "directly actionable." Students valued the breakdown by metric, reporting improved self-diagnosis of weaknesses and greater motivation to revise.

Ethical and Practical Considerations

While promising, LLM-driven AES systems carry risks of bias and opacity (Williamson & Piattoeva, 2022). AI Essay Evaluator mitigates these via:

i. *Bias checks*: periodic human review of randomly sampled feedback.
ii. *Privacy safeguards*: encryption of all submissions and local deletion after processing.
iii. *Pedagogical positioning*: the tool is formative, supplementing—not replacing—human evaluation.

CONCLUSION

AI Essay Evaluator demonstrates that LLMs can deliver nuanced, multidimensional feedback comparable to expert human graders, significantly reducing feedback latency and instructor workload. The framework's modular rubric design enhances transparency and instructional alignment.

Recommendations

i. *Institutions*: Integrate LLM-based AES tools as formative support in high-enrollment courses.
ii. *Educators*: Use granular metrics to guide targeted instruction on recurring weaknesses.
iii. *Researchers*: Conduct longitudinal studies on how LLM feedback influences writing development over semesters.

REFERENCES

Ajit Singh (2025). Generative AI-powered Automated Essay Scoring System. SSRN. https://doi.org/10.2139/ssrn.5203163

Ding, Y., Xu, W., & Li, C. (2024). Evaluating the reliability of ChatGPT in rubric-based essay assessment. *Computers & Education: Artificial Intelligence*, 7, 100143. https://doi.org/10.1016/j.caeai.2024.100143

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Kasneci, E., Sessler, K., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kumar, R., Papamitsiou, Z., & Economides, A. (2023). AI for writing assessment: The promise and perils of large language models. *British Journal of Educational Technology*, 54(4), 1345–1361.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated essay scoring: A cross-disciplinary perspective. *Lawrence Erlbaum Associates Publishers*.

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. International Journal of Educational Research Open, 5, 100296. https://doi.org/10.1016/j.ijedro.2023.100296

Masethe, M. A., Masethe, H. D., Ojo, S. O., & Owolawi, P. A. (2024). Paraphrase generation model using transformer-based architecture. Proceedings of the International Conference on Information Systems and Emerging Technologies (ICISET). SSRN. https://doi.org/10.2139/ssrn.4683780

Page, E. B. (1994). Computer grading of student prose using modern concepts and software. *Journal of Experimental Education*, 62(2), 127–142.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. https://doi.org/10.18653/v1/D16-1174

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. International Journal of Assessment Tools in Education, 12(1), 20–32. https://doi.org/10.21449/ijate.1517994

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

White, E. M. (2007). *Assigning, responding, evaluating: A writing teacher's guide* (4th ed.). Bedford/St. Martin's.

Williamson, B., & Piattoeva, N. (2022). Education governance and datafication: The role of artificial intelligence. *Learning, Media and Technology*, 47(3), 253–266.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109.

Yakubu, M. A., Sain, Z. H., Lawal, U. S., & Budiman, S. A. (2024). Application of artificial intelligence (AI) for automated essay grading in Nigerian schools. IJoEd: Indonesian Journal on Education, 1(2), 93–103. https://ijoed.org/index.php/ijoed/article/download/28/28/550