Automated Essay Grading System with Automated Generated Feedbacks

Alexander Anak Adrian
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan Sarawak, Malaysia
alexxander2620@gmail.com

Tan Ping Ping
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan Sarawak, Malaysia
pptan@unimas.my

2025 International Conference on Electrical Engineering and Informatics (ICEEI) | 979-8-3315-9271-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICEEI68459.2025.11331013

***Abstract*—Providing constructive essay feedback is a crucial yet time-consuming task, particularly in preparation for the Malaysian University English Test (MUET). This study proposes an automated essay grading system (AEGS) that leverages natural language processing (NLP) and artificial intelligence (AI) technologies to assess essays based on MUET rubrics to assist language teachers in grading essays and more importantly, to provide essays feedback efficiently as assist students' learning. The system was developed using generative pre-trained transformers (GPT)-4o OpenAI application programming interface, focusing on key features such as automated grading, detailed feedback generation, and user-friendly user interface design. AEGS extracts content from typed or scanned essays using Optical Character Recognition (OCR) and evaluates submissions via AI-powered analysis. The difference between the grade generated using MUET rubrics and the those marked by the language teachers is +-5 mark (an acceptable range). As proof of concept of this pilot study, an interview with a MUET language teacher and User Acceptance Testing (UAT) was conducted, the functional testing and user feedback indicate improved grading efficiency, consistency, and rubric-aligned feedback delivery. Most AEGS do not provide feedback but our proposed system not only grade but also provide essay writing feedback that the teachers can adjust accordingly. The feedback from the language teacher stated the usefulness of AI-generated essays feedback indicates the potential to reduce teachers' workload, improve overall feedback quality, and the potential to scale for wider academic use.**

Keywords— Artificial intelligence, automated grading, feature-driven development, Malaysian University English Test (MUET), natural language processing, OpenAI

I. Introduction

Essay grading is a fundamental component of academic assessment, particularly for pre-university students preparing for the Malaysian University English Test (MUET). MUET teachers currently evaluate essays manually using predefined MUET rubrics, a process that is both time-consuming and labour-intensive. This manual grading process can lead to variations in scoring among different teachers and delays in providing feedback to students.

Recent advancements in natural language processing (NLP) and artificial intelligence (AI) have opened new possibilities for automating the essay grading process. According to a preliminary survey conducted with the lecturers in Universiti Malaysia Sarawak (UNIMAS), there is a growing need for tools that can streamline the grading process while maintaining assessment quality. The survey revealed that lecturers spend significant time evaluating essays and are concerned about maintaining consistency across different evaluators.

Various automated essay grading systems have been developed, but many lack specific features needed for MUET assessment, such as alignment with standardized rubrics and detailed feedback generation. This gap presents an opportunity to develop a specialized system that addresses these limitations while incorporating the latest developments in AI technology.

This paper presents the development of a pilot AI-powered AEGS with automated generated essay feedback. This web-based solution that automates MUET essay grading using a combination of cutting-edge AI technologies. The system utilizes Azure Document Intelligence [7] to perform Optical Character Recognition (OCR) on scanned or handwritten essays, converting them into recognizable text. The extracted content is then processed using OpenAI GPT-4o mini, a powerful pre-trained large language model, which is guided through carefully engineered prompts to generate a MUET rubric-based scoring and constructive justification. This approach does not involve training a new model, but instead leverages advanced prompt engineering and inference configuration to achieve consistent and accurate evaluation. The primary objectives include reducing the time spent on manual grading, ensuring consistent evaluation standards, and generating comprehensive feedback for student improvement. The system design with scalability in mind to accommodate growing numbers of students and essays. This pilot is designed to address the specific needs of MUET teachers by enhancing the efficiency of the essay grading process. Rather than replacing the teacher's role, it serves as a supportive framework that provides structured guidance to ensure consistent and high-quality assessment in the context of pre-university education. The development follows a systematic approach, incorporating user requirements gathered from UNIMAS lecturers and leveraging modern technologies to create a practical solution for academic assessment needs.

II. Literature Review

Four systems were compared with, that are relevant to the proposed Automated Essay Grading System, including the conventional manual grading approach and three AEGS sourced from GitHub repositories. The four systems compared include the conventional manual-based approach, EssayScore_FYP by tingwei3931 [9], Automatic-Essay-Scoring by sankalpjain99 [4], and Automated-Essay-Grading-with-NLP by AlexEBall [2]. The chosen reviewed systems were analyzed for its features, strengths, and limitations to identify gaps that the proposed system aims to address. These systems are also compared in aspect of the main functions that

are grading process using AI and feedback mechanisms. Table 1 shows the summary of existing system and the comparison against the proposed system.

TABLE I. SUMMARY COMPARISON BETWEEN PROPOSED SYSTEM AND EXISTING SYSTEMS

Criteria	Conventional Manual-Based System	EssayScore_FYP	Automatic-Essay-Scoring	Automated-Essay-Grading-with-NLP (Proposed System)
Target Users	Lecturers	Students	Students and educators	Educators
Essay Submission	Physical or digital upload	Digital upload or textbox	Textbox	None
Grading Process	Manual (rubric-based)	Automated using neural network	Automated using machine learning and neural network	Automated using natural language processing
Feedback Mechanism	Limited, manual feedback	Numeric score only	Numeric score only	None
Customizability	Flexible	Fixed essay topic	Limited	Supports multiple trained models
Scalability	Low	Moderate	High	Moderate
Rubric Alignment	MUET's rubrics	Generic	Generic	Undefine
Usability	No interface	Simple user interface	Basic user interface	Lack user interface
Strengths	Personalized evalution	Flexible submission format, fast scoring, provide interface for spelling error detection and correction recommendation	Simple interface, fast scoring	Custom trained model
Limitations	Labor intensive, inconsistent grading	No submission history list, no feedback justification	No submission history list, no feedback justification, no interface for showing essay mistake	No user interface, Focus on experimentation

III, METHODOLOGY

A. Dataset

The dataset used is essays marked by a language teacher during the MUET classes conducted in a Pre University in Universiti Malaysia Sarawak. There is a total of 20 essays used as test set because the proposed model uses pre-trained model. For this pilot study, the main and only user is the MUET teachers. Table II shows the functional requirements for the main user.

TABLE II. SUMMARY COMPARISON BETWEEN PROPOSED SYSTEM AND EXISTING SYSTEMS

Criteria	**Priority**
Login to the system.	High
Register an account with the system.	High
Upload essays to the system.	High
View a list of submitted essays.	High
View individual essay scores, score breakdowns and score justification.	High
Manage account	Medium

B. Tools and Technologies

Table III provides an overview of the development tools and technologies utilized for development of the proposed system.

TABLE III. TOOLS USED TO DEVELOP THE PROPOSED AUTOMATED ESSAY GRADING SYSTEM

Tools	**Applications**
Azure AI Document Intelligence	For text analysis, performing tasks like tokenization, POS tagging, dependency parsing, NER, and extracting text, key-value pairs, tables, and structures from documents.
OpenAI API (GPT-4o mini)	For evaluating essays based on rubric criteria, assigning scores, and providing insights into semantic and contextual coherence.
Django	For developing a robust web application, creating APIs, managing databases, and building user interfaces.
SQLite	For storing essay data, user information, and grading results

C. Prompt Design & Model Configuration

The essay evaluation module of the proposed system utilizes OpenAI GPT-4o mini, a state-of-the-art large language model known for its multi-modal processing capabilities and efficiency [8]. Rather than training a model from scratch, this approach leverages the pre-trained capabilities by employing prompt engineering, where task-specific instructions and context are embedded directly in the prompt submitted to the model.

The prompts were carefully crafted to guide the AI in evaluating essays based on predefined MUET rubric components, including Task Fulfilment and Content, and Language and Organization which are detailed further in [5] and Table IV. A breakdown of rubric-based scoring instructions is embedded into the request payload of the system to ensure consistency and interpretability.

TABLE IV. MUET RUBRICS TEST SPECIFICATION

Component	**Test Specifications**
Writing	Candidates are assessed on their ability to write various types of text covering a range of rhetorical styles. Assessment will cover the following: **(i) accuracy** • using correct spelling and mechanics • using correct grammar • using correct sentence structures **(ii) appropriacy** • using varied vocabulary and expressions

Component	Test Specifications
	• using clear varied sentences • using language appropriate for the intended purpose and audience • observing conventions appropriate to a specific situation or text type **(iii) coherence and cohesion** • developing and organising ideas • using appropriate markers and linking devices • using anaphora appropriately together with other cohesive devices **(iv) use of language functions** • defining, describing, explaining • comparing and contrasting • classifying • giving reasons • giving opinions • expressing relationships • making suggestions and recommendations • expressing agreement and disagreement • persuading • interpreting information from non-linear texts • drawing conclusions • stating and justifying points of view • presenting an argument **(v) task fulfilment** • presenting relevant ideas • providing adequate content • showing a mature treatment of topic

The model is configured with the following inference parameters:

- Temperature: 0.2 (to reduce randomness and maintain response stability)
- Max Tokens: 1024
- Model: gpt-4o-mini via OpenAI API [7] (June 2025 release)
- Input Format: Structured prompt containing either OCR-extracted or user-submitted essay text with MUET rubric.
- Output Format: JSON-formatted response including two main scores and a detailed justification

The AI-generated scores are then mapped to the internal database and compared against teacher-assigned scores for evaluation and visualization. This pilot study does not involve fine-tuning of the GPT-4o mini model. Instead, it uses prompt engineering to adapt the model for MUET essay scoring. This approach was selected for its potential for scalability, practical integration, and resource efficiency that make it ideal for embedding NLP and AI into a production-ready academic assessment tool. This method ensures standardized and scalable essay assessment without the need for fine-tuning a custom model. Instead, it offers an efficient way to embed domain knowledge into a general-purpose LLM, balancing performance with simplicity of integration.

D. System Overview

The system has account management where the user can manage, edit or delete the account. On the essay submission, users have two types of essay input option which are typed or image based. It extracts essay content using OCR, grades the essay using GPT models based on predefine MUET rubrics, and generates score justification following by criterion topic sentence, elaboration, example, explanation, linking sentence, language & organization, task fulfilment & content. The key features of the system are essay upload (text/image), AI-powered scoring, rubric-based justification, score breakdown, dashboard and essay list.

III. Result & Discussion

The proposed pilot study undergone a comprehensive evaluation based on two core aspects which are testing the accuracy of AI-generated scores compared to human scoring and getting user experience feedback gathered through User Acceptance Testing (UAT). The accuracy evaluation was performed by comparing AI-assigned scores to scores given by MUET-trained lecturers, using statistical metrics to quantify the alignment. Meanwhile, the UAT involved hands-on testing by target users to validate system functionality, usability, and practical usefulness in real-world academic settings.

A. Accuracy Analysis: AI vs Human Scoring

To evaluate the accuracy and reliability of the AI-generated scores, a dataset of student essays was compared against human-assigned scores from MUET-trained teachers. The analysis focused on two rubric components: Task Fulfilment and Content (TFC) and Language and Organization (LO).

The following metrics were calculated:

- Mean Absolute Error (MAE): The average absolute difference between the AI-generated scores and lecturer-assigned scores was 2.25 points. This relatively low MAE indicates that the AI model consistently assigns scores that are close to human evaluations, suggesting reliable performance in automated grading.
- Within ±5 Point Range: 100% of the AI-generated scores fell within a ±5 point range of the lecturer's scores. This perfect alignment suggests that the AI model demonstrates strong agreement with human judgment in terms of general scoring consistency, even if not matching exactly on every essay.
- Standard Deviation: The standard deviation of the score differences was 1.69, reflecting a reasonably tight clustering of discrepancies. A low standard deviation means that most differences between AI and human scores were small and consistent, rather than erratic or widely spread.

As illustrated in Figure 1, the line chart shows that AI-assigned scores closely follow the trend of lecturer-assigned scores, indicating strong alignment across multiple essays. Meanwhile, Figure 2 visualizes the score difference per essay, where all differences fall within the acceptable ±5 point tolerance range. This reinforces the reliability of the AI model for practical academic use, especially where slight variation is tolerable. These results suggest that the system performs with high reliability, particularly within practical grading tolerance margins accepted by MUET examiner.

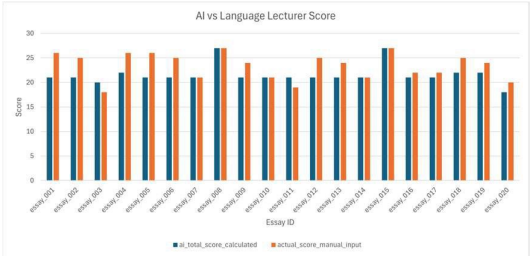

Fig. 1. Comparison of AI vs Human Scores

Fig. 2. Score Difference between Essay

B. User Acceptance Testing (UAT)

As a proof of concept, the User Acceptance Testing (UAT) session was conducted involving a MUET teacher and three student testers. Participants were asked to complete a set of related tasks using the proposed AEGS platform, including registering an account, uploading essays in both typed and image-based formats, viewing AI-generated feedback, editing content, navigating the dashboard, and managing account information. The primary goal was to determine whether the system met the functional, usability, and workflow expectations of its target users.

Feedback was collected through Google Forms and organized into four categories: tester information, core module satisfaction, system usability and accuracy, and open-ended comments. Testers reported that the system was easy to navigate, with an average usability rating of 4.5 out of 5 as shown in Figure 3. Essay uploads, both typed and image-based, were deemed smooth and straightforward, scoring an average of 4 out of 5 as shown in Figure 4. Respondents highlighted that the AI-generated scores and rubric-aligned justifications were informative and helpful as shown in Figure 5. Visual tools such as dashboards and graphs were praised for summarizing student performance effectively. While no significant bugs were encountered, a suggestion was made to support multiple essay uploads and provide post-editing analysis capabilities.

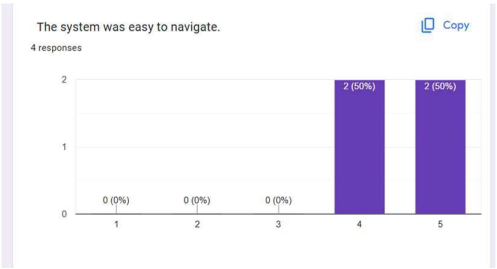

Fig. 3. System Usability UAT Result

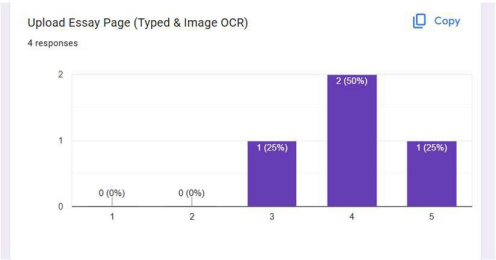

Fig. 4. Essay Upload UAT Result

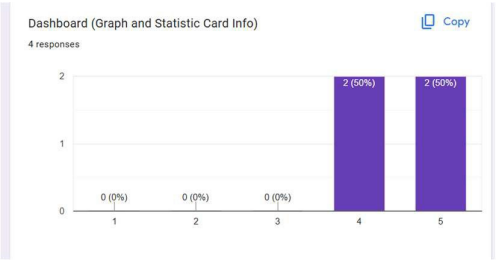

Fig. 5. Dashboard UAT Result

In the broader discussion of the effectiveness of the proposed AEGS, testing confirmed that AEGS fulfils its core functional requirements. All critical modules, including user registration, essay processing, AI scoring, feedback generation, and account management, operated as intended. Non-functional aspects, particularly interface design and overall usability, were also validated during UAT. FELC lecturers noted that the platform was intuitive and easy to use with minimal instruction, an important factor for adoption in academic environments.

The OCR module, powered by Azure Document Intelligence, effectively extracted text from handwritten essays, although its accuracy was influenced by image clarity and handwriting legibility. The integration of Azure OpenAI for automatic scoring and justification produced fast, consistent results that testers described as a valuable time-saving feature.

Most feedback focused on future improvements rather than critical issues. Suggestions included supporting a wider range of essay types beyond MUET and enhancing system flexibility. These insights point to the platform's readiness for further expansion or deployment in classroom settings. Overall, the UAT results and discussion confirm that AEGS is a stable, functional, and user-friendly system that aligns well with the practical needs of language lecturers.

IV. Conclusion & Future Work

The development and evaluation of the proposed Automated Essay Grading System (AEGS) demonstrate the practical viability of using Artificial Intelligence (AI) and Optical Character Recognition (OCR) technologies to automate the grading and feedback generation of MUET essays. AEGS successfully met its original objectives, including the integration of Azure Document Intelligence for OCR, the application of OpenAI models for rubric-based essay evaluation, and the validation of system usability through structured user testing.

Feedback collected during User Acceptance Testing (UAT) confirmed that teachers found the system intuitive, practical, and well-aligned with their existing grading workflows. Participants praised the clarity of AI-generated score justifications, the accuracy of OCR for handwritten content, and the overall ease of navigating the platform. These responses indicate strong user satisfaction and affirm the relevance of the system in academic environments. No major bugs were encountered, and suggestions focused primarily on feature expansion and customization, rather than core functionality.

Despite its success, the current version of AEGS has several limitations including ethical considerations, such as bias, fairness, and student data privacy. The system only supports MUET writing tasks and may not generalize to other essay types without additional AI model tuning. Its user interface, while functional, could benefit from improved consistency, responsiveness, and accessibility features. Moreover, the system lacks advanced security implementations, such as multi-factor authentication and encrypted API communication, which are essential for real-world deployment.

Future development should focus on enhancing AI scoring accuracy by training models on a larger, more diverse dataset of graded MUET essays. Expanding the system to cover additional MUET components, such as grammar assessments, listening tasks, and oral evaluations, could also increase its utility in classroom settings. Improvements to the user interface should include mobile responsiveness, accessibility standards, and a more modern visual design. Security upgrades like HTTPS, user activity logs, and role-based access control are also crucial for protecting sensitive data. Additionally, integrating a feedback mechanism for lecturers to validate or comment on AI-generated results would foster continuous improvement and build user trust.

In conclusion, AEGS has laid a solid foundation for AI-assisted grading in educational contexts. Its successful implementation highlights the potential of NLP and OCR technologies in reducing teachers' workload, improving grading consistency, and delivering timely feedback. With the planned enhancements, AEGS can evolve into a comprehensive, secure, and scalable tool for supporting English language instruction at the Pre-University level and beyond. With continued development, AEGS has the potential to be expanded, and secure the system for larger-scale adoption as a practical academic tool in educational institutions across Malaysia and beyond MUET assessment.

ACKNOWLEDGMENT

We would like to express sincere gratitude to Miss Feona Anak Albert, language teacher at the Faculty of Education, Language and Communication (FELC), for her valuable insight and expertise in MUET essay assessment. I also extend the appreciation to all survey respondents who provided valuable feedback for system improvement and User Acceptance Testing (UAT).

REFERENCES

[1] Ambler, S. (2023, November 26). Feature Driven Development (FDD) and Agile Modeling. The Agile Modeling (AM) Method - Effective Strategies for Modeling and Documentation. https://agilemodeling.com/essays/fdd.htm

[2] Ball, A. E. (2019, September 1). Automated-Essay-Grading-with-NLP. GitHub. https://github.com/AlexEBall/Automated-Essay-Grading-with-NLP

[3] Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Thomas, D. (2001, February). The agile manifesto.

[4] Jain, S., & Thakkar, K. (2020, March 9). Automatic-Essay-Scoring (AES). GitHub. https://github.com/sankalpjain99/Automatic-Essay-Scoring

[5] Majlis Peperiksaan Malaysia. (2015). *MUET Test Specification | MPM*. Majlis Peperiksaan Malaysia (MPM). http://webmpm1.mpm.edu.my/download_MUET/MUET_Test_Specification_2015VersiPortal.pdf

[6] Microsoft Azure. (2025). *Azure Ai Document Intelligence*. https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence?msockid=0f0130df5d4c651d14a526fa5c856412

[7] Microsoft Learn. (2025, February 7). *Azure OpenAi in Azure AI Foundry Models - Azure OpenAI | Microsoft Learn*. Azure OpenAI | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/models?tabs=global-standard%2Cstandard-chat-completions#gpt-4o-and-gpt-4-turbo

[8] OpenAI. (2024, July 18). *GPT-4O Mini: Advancing cost-efficient intelligence | OpenAi*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[9] Wei, L. T. (2019, December 24). EssayScore: Automated Essay Scoring with Deep Learning (Final Year Project). GitHub. https://github.com/tingwei3931/EssayScore_FYP