Artificial Intelligence (AI) - based Driven technique for Marking and Grading of Essay Handwriting Examination Scripts in Tertiary Institution

Blessing Olawale Olasunkanmi
Department of Computer Science, Faculty of Natural and Applied Science, Achievers University,Owo, Ondo State, Nigeria

Sadiq Fatai Idowu
Department of Computer Science, Faculty of Natural and Applied Science, Achievers University, Owo, Ondo State, Nigeria
Department of Computer Science, Ambrose Alli University, Ekpoma

Joshua Tom
Department of Computer Science, Faculty of Natural and Applied Science, Achievers University, Owo, Ondo State, Nigeria

Oyekanmi O. Ezekiel
Department of Computer Science, Faculty of Natural and Applied Science, Achievers University, Owo, Ondo State, Nigeria

Sunday Ebenezer Adepoju
Department of Computer Science, Faculty of Natural and Applied Science, Achievers University, Owo, Ondo State, Nigeria

Abe Oluwaseyi Olugbenga
Department of Computer Science, Faculty of Natural and Applied Science, Achievers University, Owo, Ondo State, Nigeria

ABSTRACT
Manual grading of essay handwritten examination scripts is time-consuming, subject to grader fatigue, and prone to inter-rater inconsistency, particularly when class size is large. This study addressed these challenges by proposing an AI-based technique for marking and grading essay handwritten examination scripts, using a deterministic rule-based engine for multiple choice questions (MCQ) alongside a Gemini 2.5 Flash large language model guided by structured prompts and marking rubrics for theory questions. A two-phase model converted physical handwritten scripts into digital structure for AI evaluation, while a multi-strategy javaScript object notation (JSON) parsing approach extracted and validated structured data using natural language processing (NLP) and deep learning methods. The paper adopted a tri-instance pipeline handling marking-guide extraction and handwritten script transcription for grading, evaluated against human-expert scores using a double-blind design to eliminate anchoring and confirmation bias. The system was tested on forty (40) handwritten scripts, twenty (20) from CSC 307 (Compiler Construction) and twenty (20) from CSC 405 ("Computer Graphics"), collected from the Department of Computer Science, Ambrose Ali University, Ekpoma (AAUE). The digitization pipeline achieved a 100% processing success rate with 300 DPI rasterization. Comparison with human-expert scores yielded a combined Pearson correlation of r = 0.5237 (CSC 405: r = 0.595; CSC 307: r = 0.556), which shows a moderate positive relationship that preserves relative ranking of student performance. The develop AI-driven technique is technically feasible and can complements traditional grading methods for marking and grading essay handwriting of examination scripts.

Keywords
Automated essay grading; hybrid grading model; Vision-Language Model; Gemini 2.5 Flash; handwritten examination scripts; natural language processing

1. INTRODUCTION
As of today, technology has been revolutionizing education in the way we teach and assess students (Mena-Guacas et al., 2025; Gause et al., 2022). Technology continues to transform the way students written works is been thought and accessed. Essay handwriting of examination remains an important method of evaluating the students' ability in their critical thinking and understanding of the subject taught. Manual essay grading, however, encounter challenges such as subjectivity, where different educators that are using the same rubric come up with different scores. Inconsistency, time consumption and the possibility of human error or bias are obvious. Manual marking of essay hand written scripts and its grading is labor intensive, especially in universities or in educational system where there is large number of students to be evaluated thereby, delay feedback and slows student learning. Clear and prompt feedback is crucial for improvement in all institutions in order to strive for fairness, efficiency, and transparency, AI based driven written essay marking and grading approach have received growing attention, this has in turn broader the rise of AI-enabled EdTech across healthcare, finance, autonomous systems, education and other disciplines.

The limitations of Manual marking of written essay and grading affect students' academic progress, morale, inconsistency across graders, enormous manual workload causing delayed feedback and human factors such as grader fatigue and unconscious bias. These challenges motivated the need for the automation of marking essay handwriting and grading technique that not only assigns a score but also gives meaningful, personalized feedback that helps the students in their learning.

Automated Exam Grading System grades easy without the intervention of human experts using as a two-phase pipeline

joining a Vision Language Model (VLM) -based data extraction with rule-based and AI-powered grading. The Hybrid Approach leverage deterministic matching for objective questions with Gemini 2.5 Flash semantic evaluation for the theory questions. The VLM is referred to as the Gemini's multimodal capability for the conversion of handwritten images into structured JSON output. Intelligent Data Extraction covers the six-stage digitization pipeline, from PDF-to-image conversion through JSON parsing, while Multi-Strategy JSON Extraction outlines the four sequential methods used to parse valid JSON from AI responses. The Question Matching Algorithm aligns student answers with the marking guide using normalization and similarity scoring. Objective Question Grading applies deterministic answer matching, whereas Subjective Question AI Grading uses multi-criteria semantic evaluation. The Error Handling and Recovery System describe fallback and retry mechanisms for processing failures, and Handwritten Content specifies the types of student responses digitized and lastly, the Marking Guide defines the structured answer-key data format, and Processing Efficiency quantifies the system's time and memory performance per script.

This paper aims to design an AI-based driven technique for the marking and grading of written essay scripts to serve as an alternative model in tertiary institutions in Nigeria contest. The technique will employ the use of Natural Language Processing (NLP) and Deep Learning (DL). Meanwhile, the specific objectives are to: (i) design a two-phase method to convert physical handwritten of essay examination scripts into digital structured data for AI evaluation; (ii) develop a multi-strategy JSON-parsing method to extract and validate structured data using the AI-powered driven technique

(iii) Evaluate and implement a robust data-extraction pipeline to extract questions, answers, and marking guides for scoring and grading; and (iv) compare manual scoring and grading methods with the AI-powered results.

2. RELATED WORKS

The first Automated Essay Grading system known as the PEG was developed back in (1966) by Ellis Page, and it leverage on linear regression over surface features like word count, sentence length, and punctuation to predict human scores. PEG system could not even the meaning of the texts it was grading and can easily be manipulated by inflating word count. Foltz, Laham, and Landauer's Intelligent Essay Assessor (1999) was developed using Latent Semantic Analysis (LSA) to capture the conceptual similarity between student essays and reference texts, which also remain specific to a domain and have difficulty with syntactic variation and pragmatics. The advancement of NLP was enabled to achieve deeper analysis of grammar, semantics, discourse structure, and writing style (Dada et al. (2025). E-rater, developed by the Educational Testing Service, combined grammar checking, style analysis to perform and compared human raters, even though its main nature limits transparency and feedbacks while, it focused more on scoring. Integration of Machine learning approaches such as the Support Vector Machines (SVM), Decision Trees, Random Forests, and Naive Bayes classifiers improved on rule-based scoring but its depends on manual feature engineering. Deep learning helps to reduce the dependency with models like Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) addressed the vanishing gradient issues facing ordinary Recurrent Neural Networks and were successfully introduced and applied to Automated Essay System (AES) by Taghipour and Ng (2016) on the Kaggle ASAP dataset. Transformer architectures (Vaswani et al., 2017) and large pre-trained models such as BERT and GPT-3 improved further the contextual understanding of coherence, relevance, and argumentation at a very high computation which is very costly.

Within Nigerian institutions, Adeyanju et al. (2024) built a Java/MySQL rule based grader using NLTK and WordNet, which was tested on only 10 typed short-answer scripts and it achieves a Pearson correlation of 0.93 against a single human rater but have an exact-agreement rate of just 0.40. Obot, Obike, and James (2024) at the University of Uyo in their study make use of Logistic Regression with TF-IDF/Word2Vec semantic similarity on 100 CSC 221 scripts which was manually transcribed also achieving a correlation of 0.888 and mean relative error of 0.59, but this system required to be trsnscribed manually which could not interpret diagrams or tables. Dada, Akinwale, and Tunde-Adeleke (2025) at Covenant University released a benchmark dataset of about 3,000 handwritten, transcribed the responses and compared TF-IDF, BERT, SBERT, RoBERTa, and Longformer models, with Longformer performing best achieving a (Spearman's r = 0.77, MSE = 0.04), but they did not build a deployable grading system.

The gap identified across these Nigerian studies is that no study graded directly from scanned handwritten examination scripts, one of the study manually transcribe test scripts which may not be adequate for written examination scripts and also none of the studies to the best of our knowledge combined a deterministic engine for objective questions with generative AI for theory questions evaluation and available datasets was very small and static. The study with 3000 datasets was not based on essay examination questions nor the scripts.

2.1 Compiler Techniques

Compiler techniques have also been applied to the evaluation of essay assessment. For instance, lexical analysis is deployed to tokenize essays into words, punctuation, and sentences to extract vocabulary and lexical differences, parsing which evaluates grammatical correctness and sentence structure. Hussein et al. (2019) and Ramesh & Sanampudi (2022) opined that combining lexical, syntactic, and semantic features allow automated scoring systems to closely approximate human evaluation. Despite improvement in scores predations, personalized feedbacks still remain un developed area of the AES system. Criterion which is, developed by the Educational Testing Service, provides real-time feedback on grammar, usage, mechanics, and style, but this feedback lacks deeper analysis of structure and argumentation quality also remain at surface level. Grammarly and ProWriting Aid which are commercial tools similarly offer grammar and stylistic corrections without evaluating essay structure or essay critical thinking, since they are not built to perfume academic writing contexts. Kumar et al. (2020), identifies the rhetorical components of an essay statements, which support arguments, counterarguments, and conclusions to generate feedback that is more targeted and meaningful. Most existing tools offer only rule-based or template-driven feedback that does not adapt to an individual students writing level which is a significant challenge to learning and student progression over time.

3. MATERIAL AND METHODS

This section discusses the approach employed to design the AI-based driven technique for marking and grading the hand written essay examination scripts. The study utilizes auto-grading system that grades student handwritten examination scripts using both NLP and DL alongside with rule-based method. The approach was adopted because developing a purely generative AI model for grading can introduce several

challenges, such as hallucination and contextual semantic drift due to limited context window length (Misgna et al., 2026; Bai et al., 2025). The auto grading system output is evaluated against a human expert as the baseline in a controlled condition, the scripts were graded by the system and the human expert, but the grades assigned are controlled through the marking guide and standardized scoring rubric using the marking guide.

The automated grading system can be formally defined as a function $G: \mathcal{D} \times \mathcal{R} \rightarrow \mathcal{S} \times \mathcal{F}$ that maps from the domain of student response documents $\mathcal{D}$ and grading rubrics $\mathcal{R}$ to the codomain of score vectors $\mathcal{S}$ and feedback sets $\mathcal{F}$. For a given document $d \in \mathcal{D}$ containing responses to n questions and a rubric $r \in \mathcal{R}$ containing model answers and scoring criteria. The system computes:

$$G(d, r) = (\langle s_1, s_2, \dots, s_n \rangle, \langle f_1, f_2, \dots, f_n \rangle) \quad (3.1)$$

Where $s_i \in [0, m_i]$ represents the score for question i with maximum mark sm_i, and f_i represents the corresponding feedback. The hybrid approach decomposes this function into two primary operations: extraction $E: \mathcal{D} \rightarrow Q$ that transforms document d into a structured representation $q \in Q$ of questions, answers, and evaluation $Eval: Q \times \mathcal{R} \rightarrow \mathcal{S} \times \mathcal{F}$ that applies both deterministic and probabilistic algorithms to compute scores and feedback.

3.1 Data Collection

The dataset used for this study was obtained from Ambrose Alli University Ekpoma Edo State, computer science department, one course each was selected from two class arms in the department, computer construction (CSC 307) from three hundred level and computer graphics (CSC 405) from four hundred level, the dataset entails handwritten exam scripts for the stated class, examination marking guide along with assigned score to be used has reference for the model baseline..

3.2 Data Processing

Filenames were sanitized using the Werkzeug library to strip paths and prevent path-traversal attacks while the first page of each script containing the students matric number and assigned score was excluded to keep the AI blind to student identity like names and matric numbers and also the human grade recorded at the page. Transcribed answers were filtered with a set-based method to remove duplicates, and regular expressions using regex to validate each courses maximum score. The marking guide, containing computer typed text, was extracted using the Python library pdfplumber which is chosen for its ability to accurately reconstruct tables and columns in which answer keys are structured. Student handwritten scripts were converted to page images using pypdfium2, which is suited to image extraction, unlike pdfplumber. Because PDFs store images at a base resolution of 72 DPI, a scale factor of 4.2 was applied to reach approximately 300 DPI (4.2 x 72 = 302.4 DPI), this improve clarity in the transcription for the vision-language model. All image processing was performed in RAM for faster computation and not on disk. The extracted image for each script was partitioned into batches of three pages so the vision model would not be exceed its context window limit or overwhelmed then the extracted JSON from each batch was merged into a single structured file for each student.

3.3 Proposed AI - Based Driven for Essay Handwriting Scripts Marking and Grading

The system uses a client-server architecture in Fig. 4.1. The architecture is organized into six layers, the first layer is Data collection where handwritten scripts, marking guides, and human expert grade, second layers the Data preprocessing which consist of PDF rasterisation, page batching, and guided text extraction, the third layer is Model building which is a tri-instance pipeline comprising Instance 1 (marking-guide extraction, temperature 0.1), Instance 2 (multi-agent transcription in 3-page batches), and Instance 3 (hybrid rule-based plus AI grading engine) moving to the fourth layer is the Database and persistence , an SQLite database storing marking-guide and grading result tables, fifth layer is the Model evaluation consisting the accuracy metrics (Pearson r, MAE, RMSE), agreement analysis (system bias, Bland-Altman), and paired t-test significance testing and sixth is the System output graded report cards, Excel export, and dashboard visualization..

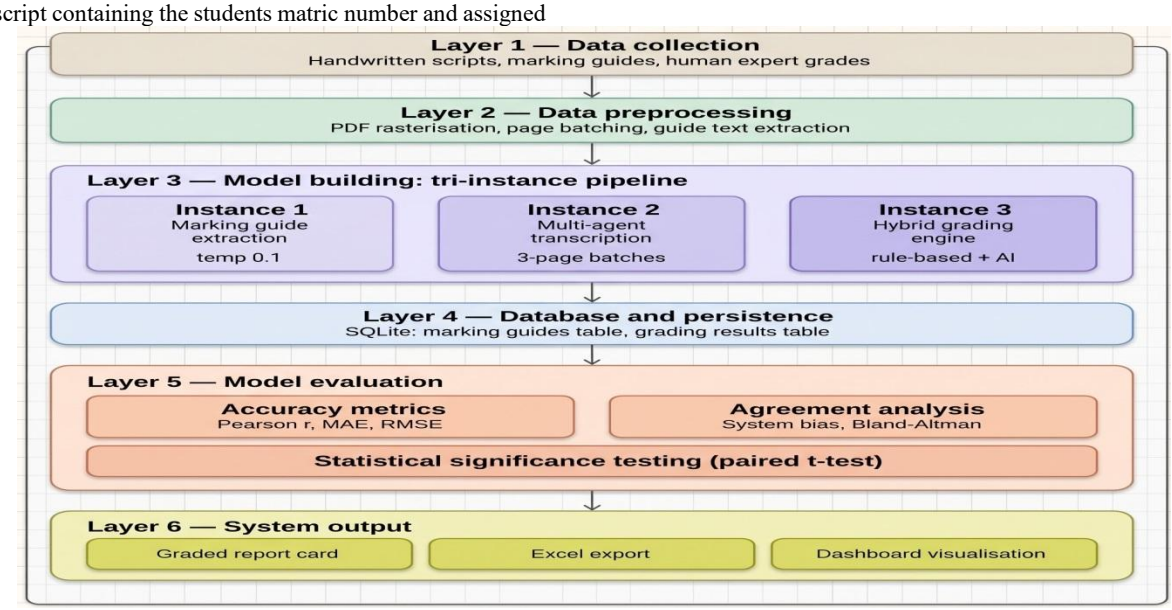

Fig.1 Architecture for Marking Essay Handwritten Examination Scripts

3.4 Model Building (Tri-Instance pipeline)

The development of the model undergoes three pipelines. For this study, Gemini 2.5 flash, one of the families of the Gemini model, was used rather than a localized model. The reason for this is that a sufficiently large language model requires high computational resources. Due to the limited computational resources available, this study utilizes the Gemini-provided

SDK for the development of the Hybrid auto-grading system, specifically the AI part of the framework see Fig 1.

3.4.1 First Pipeline

In Figure 1, the first pipeline of the model development of a system that converts the extracted marking guide into a structured JSON format, assigning a grade to each response as defined in the marking guide. This system employs Gemini 2.5 flash through a prompt-guided approach to structure the marking guide into a well-structured JSON format.

3.4.2 Second Pipeline

The second pipeline of the model entails the development of a system using Gemini 2.5 flash that transcribe student handwritten into a structured json format, the pipeline utilize two separate instance of Gemini 2.5 flash through a system instruction, the first three pages was processed by a model that extract the question and response of student into json and the 4+ pages was processed through another instance of the model, it extract the question into json format. After the extraction, the extracted question and response, now in json was concatenated into a single JSON file. This two-way system employed in this pipeline was implemented to address the issue of limited context window length and hallucination on the part of the model in see Figure 1.

3.4.3 Third Pipeline

In Figure 1, the third pipeline which is the grading system was developed using both the deterministic approach for the objective and the A-I based model for the theory, the deterministic framework employ an objective matching against the marking guide to determine if the student response align with human expert, while the theory was evaluated using Gemini 2.5 flash model with a thinking token of 2048, the

choice for this is to enable to model to take reasonable step before deciding on the grade to award the student. Table 1 illustrates the pipeline components.

Table 1: Pipeline Components

Instance	Role	Temp.	Thinking Budget	Output Schema
Instance 1	Marking guide extraction	0.1	Default	Advanced Marking Guide Extraction
Instance 2	Script transcription (multi-agent)	0.0	Default	Partial Script Extraction
Instance 3	Theory answer grading	0.1	2,048 tokens	Comprehensive Report Card
Deterministic engine	Objective grading	N/A	N/A	JSON score report

3.5 Double-Blind Evaluation Design

The double-blind evaluation was employed in this study to prevent leakage from one party to another in order to ensure the objective performance of the system developed. The blind evaluation comprises three stages.

3.5.1 Blind Stage One

The A-i is blind to Human Grader Score. The student scripts and marking guide were received by the Extracting AI, which is in pipeline 1 and pipeline 2; the AI system does not have access to the student's human experts' awarded mark. The AI grades the student following the first principle against the marking guide.

3.5.2 Blind Stage Two

The AI is blind to student identity and assigned grade. The student's handwritten script, before feeding the student script to the Gemini model, the first page of the student script, which contains the student's matric number and the assigned grade by a human expert, was dropped; the system therefore graded the student objectively without knowing whose script it was reading.

3.5.3 Blind Stage Three

Human grader is blind to the AI scores. The lecturer graded and completed the grading of the student before the development of the Hybrid grading system; they did not know AI-awarded marks at the time of grading. This implies that the Human expert grading cannot be influenced by the hybrid model developed in this study.

3.5.4 System Performance

The system performance in grading students' handwritten work will be evaluated against the human expert score using various metrics, such as: Pearson Product-Moment Correlation Coefficient (r)

The Pearson correlation coefficient measures the strength and direction of the linear relationship between AI-awarded scores (x) and corresponding human expert scores (y). It is computed as: $r = \frac{\sum_{i=1}^{n} (X-\underline{X})(Y-\underline{Y})}{\sqrt{\sum_{i=1}^{n} (X-\underline{X})^2 \sum_{i=1}^{n} (Y-\underline{Y})^2}}$ (3.2)

A positive r implies the hybrid system and the human expert are linearly correlated and going in the same direction, and a negative r implies the relationship between the system and the human expert is negatively linearly related.

3.5.4.1 Mean Absolute Error (MAE)

The MAE is the average magnitude of absolute differences between AI-grading system and human-expert scores per student, providing an intuitive measure of typical scoring deviation

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \quad (3.3)$$

A high MAE implies there is a huge variation between the AI grading system and the human expert.

3.5.4.2 Root Mean Square Error (RMSE)

The RMSE measures the average distance between the system grade and the human expert; the RMSE penalizes large individual errors more heavily than MAE by squaring errors before averaging:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (3.4)$$

An RMSE substantially higher than the MAE for the same dataset indicates the presence of a small number of large individual discrepancies in scripts where AI and human scores diverged markedly. RMSE, therefore, provides a conservative estimate of grading deviation and is particularly sensitive to outlier scripts.

3.5.4.3 System Bias

System bias is defined as the mean signed difference between AI-awarded scores and human-expert scores:

$$\underline{d} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - X_{1,i}\right) \quad (3.5)$$

A positive bias indicates systematic AI over-scoring relative to the human grader; a negative bias indicates systematic AI under-scoring. Unlike MAE and RMSE, which measure the magnitude of deviation, bias measures its direction. A system with zero bias but high MAE applies large but random deviations; a system with large negative bias applies consistent, directional deviations that require calibration.

4. RESULTS AND DISCUSSION

The study evaluated 40 students' essay handwritten examination scripts in total for first semester 2024/2025 academic session from Ambrose Alli University, Ekpoma (AAUE) in Table 2 Twenty (20) each from CSC 307 (Compiler Construction, 300 level) and CSC 405 (Computer Graphics, 400 level), each containing compulsory and optional objective and theory questions. All the 40 handwritten scripts were processed successfully at 300 DPI rasterisation setting to improve clarity which has a 100% processing success rate across both courses.

Table 2: Dataset Composition for Experimental Evaluation

Course	Course Title	Scripts Graded	Max Score	Question Structure
CSC 307	Compiler Construction	20	70	Compulsory + Optional; Objective & Theory
CSC 405	Computer Graphics	20	71	Compulsory + Optional; Objective & Theory

For CSC 405, because the extracted marking guide did not assign a particular mark to the objective section, so by default the model was set to assign two marks per objective question containing 10 compulsory objective questions, this gave 20 marks, while the theory section allowed three of several optional questions worth 17 marks each (51 marks in total), for a combined maximum of 71 marks. For CSC 307, the objective section was set to two marks per question except question 10 which is one mark, giving 19 marks for the compulsory objective section, with the same 51-mark optional theory structure, for a combined maximum of 70 marks all together in total with the descriptive statistics shown in Table 3 and table 4 for CSC 405 and CSC 307, while Table 5 and table 6 presents the comparative performance of AI-based driven technique against human expert evaluation.

Table 3: Descriptive Statistics CSC 405 (n = 20, Max = 71 marks)

Statistic	AI Score	Human Score
Count	20	20
Mean	18.90	28.40
Std. Dev.	8.39	4.26
Min	2.00	21.00
Median (Q2)	20.00	28.00
Max	32.00	40.00

Table 4: Descriptive Statistics CSC 307 (n = 20, Max = 70 marks)

Statistic	AI Score	Human Score
Count	20	20
Mean	20.50	34.95
Std. Dev.	7.78	6.11
Min	6.50	25.00
Median (Q2)	23.00	35.50
Max	35.50	43.00

Table 5: Descriptive Statistics: AI vs. Human Scores by Course

Statistic	CSC 405 AI	CSC 405 Human	CSC 307 AI	CSC 307 Human
Number of students	20	20	20	20
Max possible score	71	70	70	70
Mean score	18.9	28.4	20.5	35.0
Mean % of maximum	26.6%	40.0%	29.3%	50.0%

From both courses the AI graded students lower than the human expert. For CSC 405, the AI mean score was 18.9 versus a human mean of 28.4, which now making a difference of 9.5 marks, and for CSC 307, the AI mean was 20.5 versus a human mean of 35.0 making a difference of 14.45 marks. The overall system bias was -11.975 marks, confirmed as statistically significant by a paired t-test ($t = -10.644$, $p = 4.2451 \times 10^{-13}$), well below the alpha = 0.05 threshold.

Table 6: Global Comparative Performance Metrics AI vs. Human Grader

Metric	CSC 405 (n=20)	CSC 307 (n=20)	Combined (n=40)
Pearson r	0.595	0.556	0.5237
MAE (marks)	9.75	14.45	—
System bias (AI - Human)	-9.50	-14.45	-11.975
Paired t p-value	5.13 x 10^-6	9.67 x 10^-9	4.2451 x 10^-13

The Pearson correlations of 0.595 (CSC 405), 0.556 (CSC 307), and 0.5237 combined shows that a moderate positive linear relationship exist between the AI score and human scores in both courses in Table 6, because both graders shows a similar student performance. In addition, the grading dashboard for the result is in Fig 2, Fig 3 and Fig.4 respectively.

Fig 2 Students batch grading

Fig 3 Results dashboard

	S/N	Matric No	Course Code	Total Score	Max Score	Percentage (%)	Graded At	Expert Grade
0	1	student_1	CSC 405	25.5	71	35.9	2026-06-25 10:59:28.686	34
1	2	student_2	CSC 405	14.0	71	19.7	2026-06-25 10:56:14.145	21
2	3	student_3	CSC 405	20.0	71	28.2	2026-06-25 10:54:17.751	30
3	4	student_4	CSC 405	23.5	71	33.1	2026-06-25 10:53:06.874	27
4	5	student_5	CSC 405	6.5	71	9.2	2026-06-24 10:21:53.579	24

Fig 4 Final graded table by the model, along with the Human Expert grade

4.1 Score Alignment - AI Auto grader vs. Human Expert

The scatter plot in fig 5 shows an upward trend towards the right for the AI grading system and the Human expert for both courses. Although not a perfect straight line, the direction of the AI grading score moderately aligns with that of a human expert, this however shows both the AI and human expert identify the diversity and brilliance in student performance.

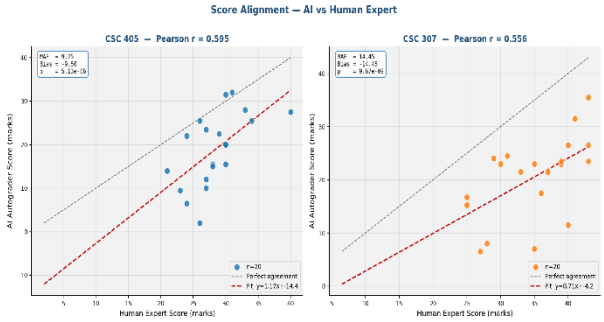

Fig 5 Score Alignment - AI Auto grader vs. Human Expert (CSC 405 and CSC 307)

4.2 Student Score Trend

The dynamic trend of the AI score and the Human expert for both courses shows that the AI grading system grades students lower as compared to the human expert. Despite this, both graders score a strong positive trend, indicating a high level of agreement between the underlying intellectual behaviors of student performance in their exam. See fig 6.

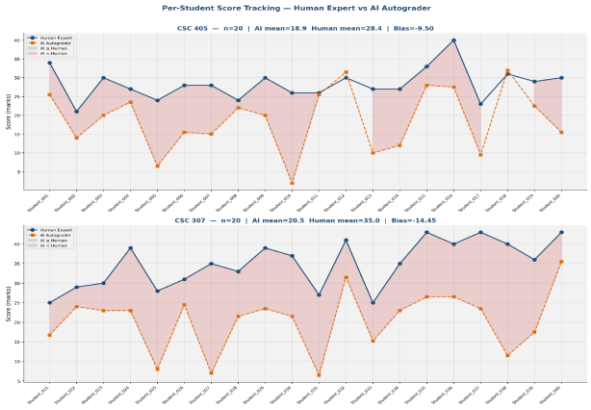

Fig 6 Per-Student Score Tracking - Human Expert vs. AI Auto grader

4.3 Percentage Score Boxplots

Shown in fig 7, boxplots distribution of the auto-grading system and the Human expert score shows that for CSC 405 the AI model graded 50% of the student below 30 marks and 25% of the student below 20 marks while the human expert grader 50% of the student close to 40 marks and 25% of the student also above 30 marks a little bit farther from 40 marks.

Second boxplot for CSC 307 shows a similar distribution as the AI grade, 50% of the students between 30 marks and 40 marks, while the human expert grade 50% of the students a little bit above 50 marks. This implies that the AI model is stricter in grading students compared to the Human expert.

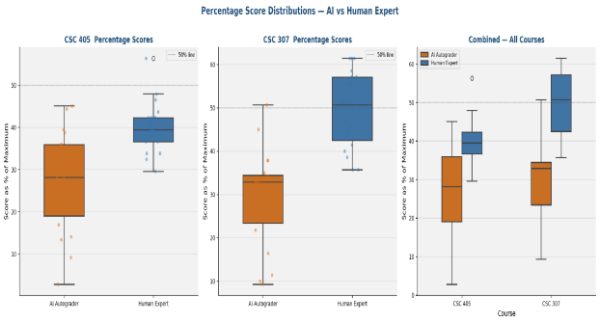

Fig 7 Percentage Score Distributions AI vs. Human Expert

4.4 Error Bias Histogram

Fig 8, illustrates the distribution of bias across each student. As regard both courses. CSC 405, the bias exhibits a standard deviation of 6.61 marks, and the error or difference falls heavily within the negative side; the vast majority of the bias is between -15 and -5 marks, while for CSC 307, similarly, the standard deviation of the bias, which is the difference between the AI grader and the human expert, falls towards the negative side.

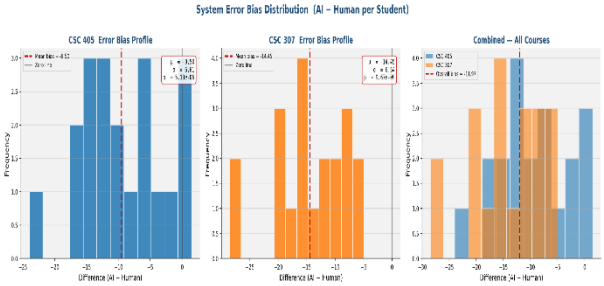

Fig 8 System Error Bias Distribution AI minus Human per Student

4.5 Bland-Altman Agreement Analysis

Bland-Altman residual was employed to assess the level of agreement between the AI model and the human expert, shown in figure 9. CSC 405 exhibits a mean bias of -9.5 marks; the 95% limit of agreement falls between -22.45 and +3.45 marks, while for CSC 307, the system bias from the human expert is -14.45 marks, and the limit of agreement between the two graders falls between -27.27 and -1.63 marks. This implies the auto-grading model is stricter as compared to the human grader.

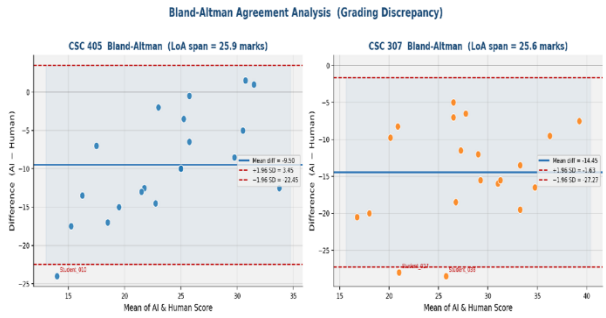

Fig. 9 Bland-Altman Agreement Analysis

4.6 Non-mutually Exclusive Mechanisms Explain this Finding

Three non-mutually exclusive mechanisms are proposed to explain the negative bias observed in Fig. 4.10. First is the strict criterion-matching, the AI grader requires student answers to align closely with specific keywords and concepts in the marking guide, if not it partially penalize correct answers more severely than a human grader applying holistic judgment, in this case human experts naturally recognize equivalent or close formulations and contextually correct student answer not in resemblance with the marking guide literal language. Secondly, transcription error propagation where handwriting recognition errors in instance 2 output can produce missing or unclear answer text causing instance 3 to award zero marks for questions the students answered, this happened mostly for scripts with dense or nonstandard handwriting and is reflected in the wider Bland-Altman limits observed for CSC 307. Thirdly the marking guides used in this study was designed for human graders and it may lack sufficiently explicit partial credit thresholds for the AI to calculate between no mark and full mark for quality answers but with a marking guide redesigned with explicit multi-level partial credit criteria would likely reduce the negative bias with patterns depict in Fig. 4.10.

Overall, the moderate positive correlation confirms that there is relationship between AI auto grader and the student scores by human rater, while the statistically significant negative bias shows that it is not a replacement for human grading at the individual score level but an assistant tools. The system is therefore best positioned as a grading assistance, with human supervision needed for final mark assignment pending further calibration and marking guide redesign.

5. CONTRIBUTIONS OF THE PAPER

The contributions are as follow; a multi-strategy using JSON phrasing method to extract and validate structured data using the AI–based driven technique powered with Natural NLP and Deep Learning (DL) is designed. The AI-based driven technique using NLP and DL for marking essay handwritten examination scripts is evaluated based on statistics metrics. The technique can complements the traditional method currently being used in tertiary institutions through the achieved promising result. Comparison of manual grading scoring method with the develop AI powered model is conducted

5.1 Conclusion and Recommendation

The paper shows that an AI-based driven technique for marking and grading essay handwritten examination scripts is feasible, with the model achieving a 100% processing success rate and exhibiting a moderate correlation with human expert scores, confirming that it can capture students' underlying performance and is suitable for deployment.

It is recommended that marking and grading technique should not replace human experts entirely but rather serve as an assistance tool for grading handwritten essay examination scripts in tertiary institutions. The technique if embraces will reduce the volume of scripts requiring full manual review and provides rapid preliminary scores, and flags low-confidence cases for lecturer attention, making the overall grading process faster, more consistent, and less prone to fatigue related errors. Examination bodies in Nigeria, including WAEC, NECO, JAMB, and tertiary institutions, could consider the adoption of suggested technique. The model can process physical handwritten scripts without needing a transition to computer-based examinations, making it particularly suited to the Nigerian academic environment, where pen-and-paper assessments remain the dominant mode.

5.2 Future research

should expand the dataset across disciplines, levels, and institutions beyond the two Computer Science courses evaluated that have been evaluated here, to assess and cover more disciplines and observe maybe the observed negative bias is happening consistently across other subject areas. Researchers should also endeavor to compare prompt guided grading against models that are fine tuned in variants, trained on individual lecturers grading patterns, and conduct longitudinal studies on an AI-generated feedback to reaffirm if truly it will relief lecturer method of students essay handwritten examination scripts marking assessment and grading scores.

6. REFERENCES

[1] Adeyanju, I. A., Oderinde, K. R., Adedeji, O. T., Gbadamosi, O. A., Makinde, B. O., & Falohun, A. S. (2024). Artificial Intelligence Based Essay Grading System. Engineering and Technology Journal, 9(7), 4382-4388.Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment, 4(3).

[2] Bai, Y., Lv, X., Zhang, J., Xie, Peng., & Li, J. (2025). LongBench v2: Towards multi-task and long-context evaluation for large language models. *arXiv preprint arXiv:2412.15204*.

[3] Dada, I. D., Akinwale, A. T., & Tunde-Adeleke, T.-J. (2025). A Structured Dataset for Automated Grading: From Raw Data to Processed Dataset. Data, 10(6), 87.

[4] Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y.-I. (2024). Hybrid Approach to Automated Essay Scoring. Mathematics, 12(21), 3416.

[5] Foltz, P., Laham, D., & Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning.

[6] Gause G, Mokgaola IO, Rakhudu MA. (2022). Technology usage for teaching and learning in nursing education: An integrative review. Curationis. 2022 Jun

15;45(1):e1 e9. doi: 10.4102/curationis.v45i1.2261.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

[8] Lewis Sevcikova, B. (2018). Human versus Automated Essay Scoring: A Critical Review. Arab World English Journal, 9(2).

[9] Mena-Guacas, A. F., López-Catalán, L., Bernal-Bravo, C., & Ballesteros-Regaña, C. (2025). Educational Transformation Through Emerging Technologies: Critical Review of Scientific Impact on Learning. Education Sciences, 15(3), 368. https://doi.org/10.3390/educsci15030368 Misgna, H., & Haile, A. (2026). LEAF++: Trait-annotated essay feedback dataset with validation of trait scores via LLMs. IEEE Access, 14, 11303–11315. https://doi.org/10.1109/ACCESS.2026.11303736.

[10] Misgna, H., & Haile, A. (2026). LEAF++: Trait-annotated essay feedback dataset with validation of trait scores via LLMs. IEEE Access, 14, 11303–11315. https://doi.org/10.1109/ACCESS.2026.11303736

[11] Obot, O. U., Obike, P., & James, I. (2024). Automated Marking System for Essay Questions. Journal of Engineering Research and Reports, 26(5), 107-126.

[12] Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 47, 238-243.

[13] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[14] Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays. In Handbook of Automated Essay Evaluation.

[15] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of EMNLP.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.