Leveraging Large Language Models for Automated Scoring of Open-Ended Questions in Educational Assessment: A Case Study of Obafemi Awolowo University

Abidemi Akanfe
Department of Information Systems, Faculty of Computing Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
akanfeabidemi@pg-student.oauife.edu.ng

Oluwafemi Ojuri
Department of Computer Science, Faculty of Computing Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
ojurifemi132@gmail.com

Rhoda Ikono
Department of Information Systems, Faculty of Computing Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
rudo@oauife.edu.ng

***Abstract*— Assessment is crucial in education as it quantifies learning outcomes and offers feedback that informs instructional practices. The evaluation of open-ended responses remains challenging due to issues of subjectivity, inconsistency, and the extensive time required for manual assessment. The predominant Computer-Based Testing (CBT) systems at Obafemi Awolowo University focus primarily on objective questions and lack mechanisms to assess essay-style responses that evaluate higher-order skills such as critical thinking and reasoning. This study addressed the gap by developing an enhanced CBT system that utilises Large Language Models (LLMs) for the automatic grading of open-ended questions. The system incorporates a Kotlin-based frontend, a NestJS backend, MongoDB and Redis for data management, and GPT-4 for AI-powered grading. Human-assigned scores served as benchmarks for the data derived from undergraduate students' essay responses in CSC507 (Computers in Society). The Root Mean Square Error (RMSE) of 2.23, indicated strong numerical alignment, while the Quadratic Weighted Kappa (QWK) score of 0.71 reflected substantial agreement and fairness across score levels. The findings indicated that integrating GPT-4 into computer-based testing enhanced its utility beyond multiple-choice assessments by enabling precise, effective, and comprehensible grading of open-ended responses. The study demonstrated that LLM-powered systems can enhance the scalability and transparency of educational assessment in higher education, improve grading consistency, and reduce lecturers' workload.**

Keywords*— *Automated Essay Scoring (AES), Large Language Models (LLMs), Educational Assessment, Essay, Artificial Intelligence, Open-Ended Question.

I. INTRODUCTION

Assessment is necessary for regulating the quality of education, its advancement, and outcomes. It offers insights into learning to assess learner strengths and needs and to deliver feedback on teaching and learning. According to [1], assessment is the educational practice of identifying, collecting, and interpreting information about learning and achievement of student. Educational assessment involves the ongoing evaluation of learners at regular intervals, focusing on the cognitive, emotional, and psychomotor domains using a range of instruments, including exams, assignments, observations, interviews, questionnaires, and projects. Assessment informs decisions about educational policies that impact students and provides educators and students with continuous feedback on the teaching and learning process. Examinations are still the most common form of assessment and can include different types of questions, from simple multiple-choice to open-ended questions, such as essays that require students to explain their answers.

Computer-Based Testing (CBT) is an assessment method that delivers, administers, and scores examinations using computer technology rather than traditional paper-based formats [2]. Computer-Based Testing is typically confined to utilising computers for the entire evaluation process, including both the delivery of the assessment and the provision of feedback. Over the years, Computer-Based Testing (CBT), especially for multiple-choice questions, has significantly improved assessment by enhancing efficiency, accuracy, accessibility in examination delivery, and supporting large-scale standardized examinations. Unlike paper-based tests, it provides faster administration and automated grading, reducing human error and ensuring timely feedback for candidates.

Grading is a fundamental aspect of educational assessment that serves as a crucial means of measuring learners' ability levels and providing feedback [3]. Grading essays and other open-ended questions can be tricky due to the fact that essays can measure high- level abilities such as creative thinking, logical thinking, and critical reasoning [4]. Partial or full automation of exam scoring has been made possible by natural language processing (NLP), and, because of its benefits, including reducing human error and the time required to evaluate scripts, automated grading has become increasingly popular.

Artificial Intelligence (AI) refers to machines simulating human intelligence to perform tasks such as learning, reasoning, and problem-solving[5]. Artificial Intelligence has been increasingly integrated into education, particularly in

adaptive learning, grading automation, and intelligent tutoring systems. AI-powered assessment tools, such as intelligent tutoring systems (ITS) and automated essay scoring (AES), have shown promise in reducing educators' workload and improving assessment consistency [6]. Large Language Models (LLMs) are advanced AI systems trained on extensive textual data, utilizing the transformer architecture and attention mechanism to analyze language patterns and generate human-like natural language [7]. The emergence of artificial intelligence (AI), especially large language models (LLMs), holds significant promise for improving the evaluation process in computer-based tests. This is because large language models excel at comprehending text and its context, unlike rule-based methods, thus facilitating the creation of sophisticated assessment systems that accurately evaluate students' understanding and competency [8]. The use of LLM-powered assessment systems represents a major shift in how we evaluate education, offering better accuracy and feedback compared to traditional methods, and the advancement of AI provides significant potential for maximizing LLMs in educational evaluations, which could alter teaching and learning in the digital era.

This research therefore aimed to explore how modern artificial intelligence can be applied to better assess open-ended questions, ensuring fairness and accuracy. In doing so, this aims to reduce the burden on educators, enhance transparency in grading, and ultimately improve the quality of education delivery in higher institutions. The structure of the paper is as follows. Section 2 provides a concise overview of prior research relevant to the topic. Section 3 outlines the methods employed. Implementation and results were discussed in Section 4. In the end, the research conclusion was presented in Section 5.

II. Literature Review

Open-ended questions are crucial for assessing learning outcomes, as they gauge students' comprehension, analytical abilities, and writing proficiency. Evaluating open-ended questions is a multifaceted endeavor that requires assessing descriptive responses and is both time-intensive and subjective. As the ratio of teachers to students escalates, manual assessment gets increasingly intricate. Automated essay scoring (AES) can mitigate these shortcomings by providing consistency, efficiency, and immediate response. AES uses mathematical models to assess essays based on structured attributes such as grammar, coherence, and organisation, guaranteeing consistent scoring across all submissions. AES offers a prompt response, which is essential for extensive evaluations.

Numerous AES tools have been created; AES dates back to the 1960s, with Project Essay Grade (PEG) developed by Ellis Page. PEG relied on hand-crafted linguistic features such as word count, sentence length, and syntactic complexity to predict essay quality. This early system was rule-based, meaning it depended heavily on predefined scoring criteria rather than true comprehension of the text [9]. By the 1990s, statistical models such as Latent Semantic Analysis (LSA) and e-rater (developed by ETS) became prominent; e-rater, created by [10], employs a hybrid approach that integrates linguistic features with document structure. This system integrates syntactic, discourse, and topical content analysis using natural language processing (NLP) techniques. Latent Semantic Analysis developed by [11], utilizes Latent Semantic Analysis and a "bag of words" method, disregarding linguistic and structural aspects of the document. Text Categorisation Technique (TCT) designed by [12] applies text classification techniques. Similarly, IntelliMetric by [13] used artificial intelligence and statistical modeling to predict human scores. Another notable system, the Bayesian Essay Test Scoring System (BETSY) [14], applied Bayesian Networks for essay classification, achieving 80% accuracy with relatively small datasets. E-rater V.2 was also created by [15], which adopted a simplified statistical approach, using a fixed, smaller set of intuitive features across all prompts, allowing for a single, standardized scoring model.

The methods for automated scoring can be classified into two broad categories: the hand-engineered feature-based approach and the deep learning model-based approach [16], [17]. A hybrid of both has proven successful in recent years [18]. Feature-based models primarily utilize handcrafted features ranging from basic elements like word count to more complex aspects such as argumentation structure and coherence. These features include linguistic features such as syntactic, semantic, lexical, and readability features, which are used as input to develop automated scoring models [4]. The major modeling approaches used include regression-based models, classification models, and neural networks [19].

The evolution of Automated Essay Scoring (AES) systems has shifted from feature-based methods to deep learning-driven approaches. The modern evolution began with machine learning techniques such as Support Vector Machines (SVMs) and Decision Trees, which classified essays using predefined linguistic features such as lexical diversity, word frequency, and sentence complexity. While effective for structured datasets, these models required extensive manual feature engineering and lacked adaptability to diverse writing styles. The introduction of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks addressed some of these limitations by enabling sequence-based learning, allowing AES systems to retain context across sentences. However, RNNs struggled with long-range dependencies, and LSTMs, despite their gating mechanisms, faced difficulties in fully capturing hierarchical argument structures and were computationally expensive.

Significant progress came with transformer-based models like BERT, which leveraged bidirectional self-attention to analyze words in relation to their entire context, improving the evaluation of coherence, grammar, and meaning. The bidirectional architecture of BERT enables fine-tuning on domain-specific essays, allowing models to adapt to different academic disciplines [20]. Multi-task deep learning further enhanced AES by enabling a single model to assess multiple aspects of writing simultaneously, making scoring more adaptable and generalizable. More recently, Large Language Models (LLMs) such as GPT-3.5, GPT-4, Claude-3, and Mistral-Large have demonstrated remarkable improvements in evaluating open-ended responses.

A. Application of LLMs in AES

Large Language Models (LLMs) and cognitive AI have recently gained popularity, opening new avenues to overcome the shortcomings of Automated Essay Scoring (AES). These methods use retrieval-augmented generation (RAG) techniques, which allow the involvement of other external sources , extensive knowledge bases, and deep contextual understanding to increase evaluation fairness, give richer feedback, and improve grading accuracy [21]. The potential of incorporating large language models in education has

garnered attention due to their versatile applications, and leveraging these models is a promising avenue for improving learning and teaching experiences across various educational levels, including primary, secondary, tertiary, and professional development. Given the diverse learning preferences, abilities, and needs of each individual, large language models present a distinctive opportunity to deliver personalized and impactful learning experiences[22]. The technical foundation of automatic assessment systems has strengthened, and the number of related studies has increased, with excellent prospects for Artificial Intelligence in Education [23].

Automated Essay Scoring (AES) has been enhanced through contextual adaptation by advancements in Large Language Models (LLMs) such as BERT, GPT-4, and Claude-3, which have made fine-tuning for specialised domains easier. Studies show that these models outperform traditional rule-based systems by improving the quality of argument assessment and grading consistency. However, concerns about bias remain, requiring hybrid models that combine AI scoring with human verification to ensure fairness [24]. Reducing grading variability in large language models requires temperature calibration. Temperature changes in LLM-generated scores substantially affect grading variability. According to [25], underscoring the significance of calibrating model outputs to match standardised evaluation metrics.

The ability of LLM-driven AES to provide individualised, helpful feedback that goes beyond grammar and syntax is one of its main advantages; studies have shown that students' writing skills significantly improve as a result. In contrast to students who used conventional AES tools, students who received AI-generated feedback showed a 23% improvement in their argumentation abilities [26]. Similarly, models that used Retrieval-Augmented Generation (RAG) could reference prior student submissions and provide more contextually relevant feedback [25].

With advancements in linguistic analysis and machine learning, AES continues to improve, promising a future where automated grading plays a crucial role in education and assessment. Automated scoring systems can be categorised into hand-engineered feature-based approaches and deep learning model-based approaches. A combination of both has demonstrated success in recent years. Advancements in linguistic analysis and machine learning enhance AES, suggesting a future in which automated grading is integral to education and assessment.

B. Related Works

AutoSAS, a supervised learning model for automatically scoring short answers in essays was developed by [27]. To reduce time, costs, and bias in grading, the system uses features such as word embeddings, POS tags, and word frequency. It was trained on a publicly accessible dataset and provided students with feedback, which might reduce the workload and bias of human graders.

An automated essay-scoring system that accurately assesses and grades student essays, thereby reducing human error and bias, was developed by [28]. Semantic similarity was measured by the system using a three-layer framework and Natural Language Processing tools and algorithms. The system outperformed state-of-the-art systems and showed high efficiency and precision, with higher correlation coefficients with human judgment than other approaches.

In addition, a novel AES system was developed to overcome the drawback of only offering a holistic score [29]. To address this, the authors used the GoodWriting dataset—which comprises more than 800 essays written by Japanese students—to develop an AES system. They employed neural network-based techniques using LSTM and BERT models, in addition to feature-based techniques (such as linguistic features specific to Japanese). The results show that the BERT model has the highest holistic and trait scores for both RMSE and QWK, which were used as evaluation metrics. As a result, the BERT model outperformed the others, exhibiting superior overall accuracy and resilience to adversarial distortions. To encourage further research on Japanese learners (AES), the system was made publicly available after its usability and educational applicability were verified.

In addition, [30] evaluated the performance of large language models (LLMs) for non-native Japanese in automated essay scoring. A dataset of 1400 story-writing scripts from learners with first languages across 12 language families was used. The study found that GPT-based methods outperformed linguistic features-based methods and LLM-based approaches, and that LLMs improved performance quality and model interpretability in AES for non-native Japanese language learning.

Again, the accuracy and dependability of the GPT-3 text-davinci-003 AI language model for automated essay scoring (AES) were investigated [31]. By evaluating 12,100 essays from the ETS Corpus of Non-Native Written English (TOEFL11) and comparing them to benchmark levels, the study found that GPT-3 demonstrates some accuracy and reliability, which is further enhanced by the addition of linguistic components. These results suggest that AI models like GPT-3 can effectively support Automated Essay Scoring, which could transform writing evaluation and feedback practices in classrooms.

Furthermore, [25] evaluated students' open-ended written responses with LLMs. The study assessed LLMs (GPT-3.5, GPT-4, Claude-3, Mistral-Large) for grading university students' open-ended responses using the retrieval augmented generation (RAG) framework. Through 4,320 grading iterations with varying temperature settings (0.0 and 0.5), it measured accuracy, consistency, computational efficiency, and fairness by comparing LLM-assigned grades to a benchmark derived from the most frequent scores. The study found significant grading variability across LLMs: GPT-3.5 assigned extreme scores, Claude-3 avoided extremes. Mistral-Large showed the highest consistency (83.3%), GPT-3.5 lowest (18.5%). GPT-3.5 was fastest (3.52s/grade), Claude-3/GPT-4 slowest (>20s). Higher temperature (0.5) increased variability, reducing reliability.

III. METHODOLOGY

This research covered the development of a Computer-Based Testing system capable of assessing open-ended questions using large language models (LLMs). The diagram depicted in Figure 1 presents a conceptual framework for the grading of essay assessments within the university. The procedure commences with the lecturer administering an essay test for students, followed by the marking and grading by the lecturers, the compilation of additional assessment results such as assignments, project work, and attendance, and concludes with result computation and approval by the administrator before the approved result is released.

The system was developed using the Agile SDLC approach, combining requirement gathering, system modeling, and iterative testing. The requirement analysis was determined based on a literature review, Administrative staff interviews, and direct observation of Obafemi Awolowo University's current CBT system, as shown in Figure 2 below, which depicts the conceptual framework of the existing CBT system in Obafemi Awolowo University

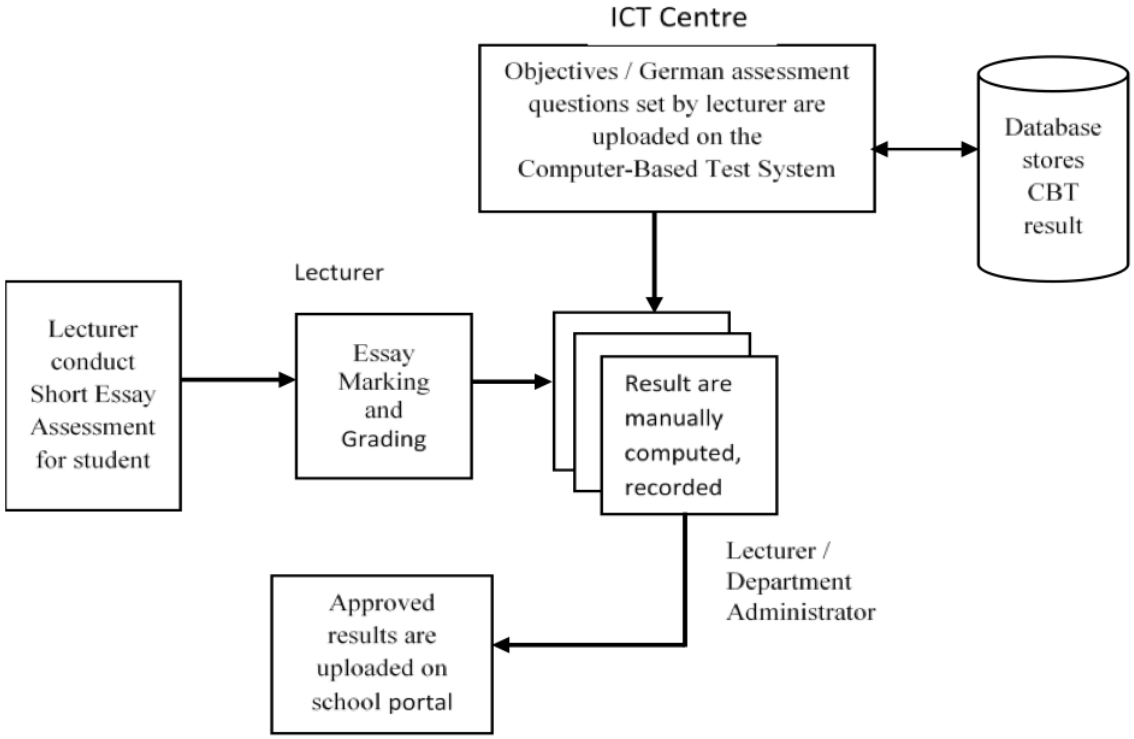


Fig. 1. Existing Conceptual Design Framework of Essay Grading in OAU

A. Inclusion and Exclusion Criteria

Given that this study was tailored towards open-ended assessment, two key factors were considered as the inclusion and exclusion criteria for the data collected during the research. These are:

1. Exclusion of Non-Essay Assessments: Multiple-choice questions (MCQs), true/false, mathematical equations, diagrams, or non-text-based responses and other objective assessment formats are excluded, as the study focused solely on subjective, open-ended textual responses, leaving out highly technical responses that require specialized domain knowledge or formulaic problem-solving. This ensures the scope remains focused on evaluating reasoning-based essay assessments.
2. Exclusion of Non-English Responses: To maintain linguistic consistency in grading, only English-written essays were included, while responses in other languages were excluded.

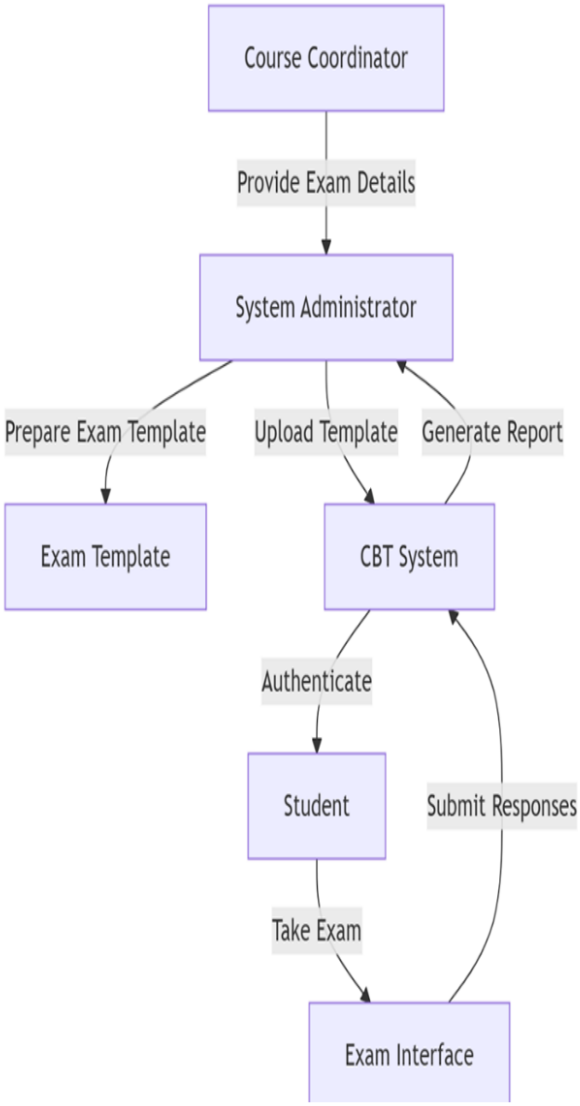


Fig. 2. Conceptual Framework of the current Obafemi Awolowo University's CBT system.

B. Data Collection and Preparation

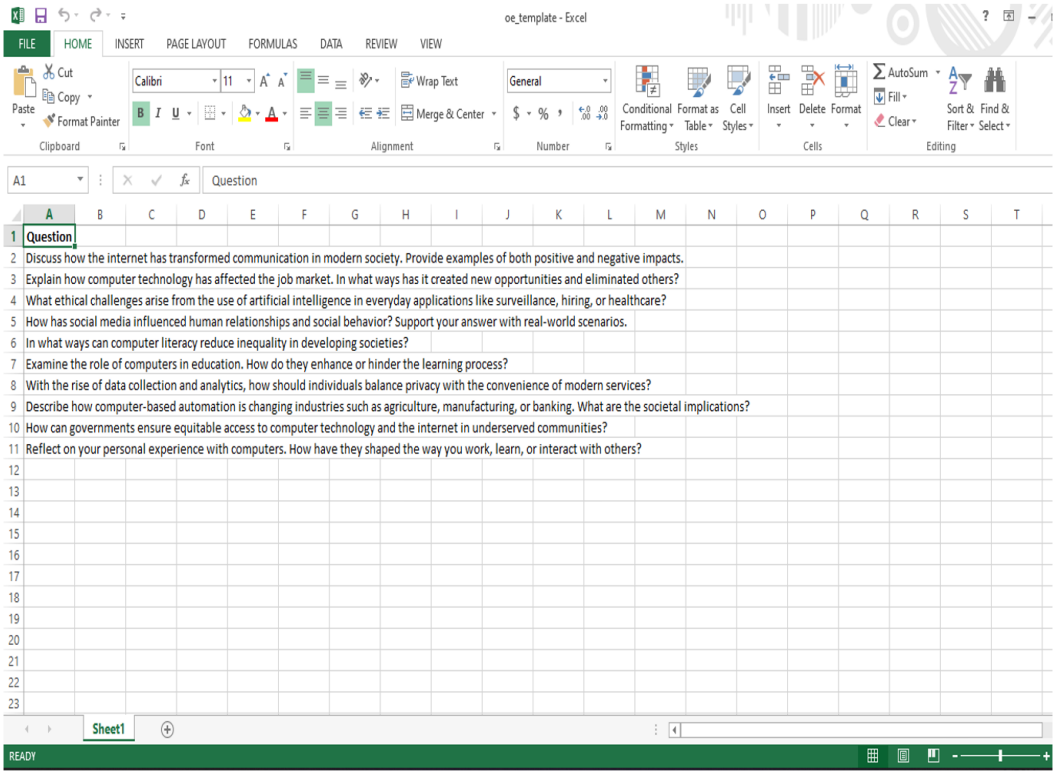
The data employed in this study were an experimental dataset derived from essay responses written by undergraduate students during scheduled test sessions for the Course CSC507 Computers in the Society, conducted in the Department of Computer Science and Engineering at Obafemi Awolowo University, Ile Ife, Nigeria. The question set, prepared by the course instructors, consisted of essay-style prompts that required critical reasoning and argumentation. These questions were paired with marking guides that served as the grading rubric, forming the input framework for both human and AI grading." A sample of the open-ended questions and marking guide is presented in Figures 3 and 4 below, respectively. Responses of five students were collected within the CBT system and graded by both the AI. The human-assigned scores on the written essay served as the benchmark for evaluating the AI's grading accuracy and consistency.

C. System Implementation

The CBT system was improved to include open-ended essay questions by adding both a question set and a marking guide during test creation. The question set *included prompts for analysis, and the marking guide outlined the expected main points and scoring rubrics.* A GPT-4 pipeline graded the students' answers by using the marking guide to give them consistent scores and feedback on the quality of their work. By aligning AI grading with academic standards, this method made things fairer, less subjective, and more open. The diagram in Figure 5 below gives a clear view of the workflow on the improved CBT system.

The system is divided into two core components: the frontend, which provides the Graphical User Interface (GUI) for user interaction, and the backend, which handles the system's logic, data processing, and AI-powered grading operations. The frontend is implemented using Kotlin Multiplatform, enabling a responsive and consistent user experience across platforms, while supporting dynamic interaction with the backend. The backend is built with NestJS, a scalable Node.js framework that manages business logic, API endpoints,

authentication, and integration with the OpenAI GPT-4 API for automated grading.



	A
1	Question
2	Discuss how the internet has transformed communication in modern society. Provide examples of both positive and negative impacts.
3	Explain how computer technology has affected the job market. In what ways has it created new opportunities and eliminated others?
4	What ethical challenges arise from the use of artificial intelligence in everyday applications like surveillance, hiring, or healthcare?
5	How has social media influenced human relationships and social behavior? Support your answer with real-world scenarios.
6	In what ways can computer literacy reduce inequality in developing societies?
7	Examine the role of computers in education. How do they enhance or hinder the learning process?
8	With the rise of data collection and analytics, how should individuals balance privacy with the convenience of modern services?
9	Describe how computer-based automation is changing industries such as agriculture, manufacturing, or banking. What are the societal implications?
10	How can governments ensure equitable access to computer technology and the internet in underserved communities?
11	Reflect on your personal experience with computers. How have they shaped the way you work, learn, or interact with others?

Fig. 3. Open-Ended Question Sample Set

1. Discuss how the internet has transformed communication in modern society. Provide examples of both positive and negative impacts.
Marking Guide (20 marks total):

- Clear understanding of internet's role in modern communication (4 marks)
- Examples of **positive impacts** (e.g., instant messaging, global connectivity, online collaboration) – at least 2 (6 marks)
- Examples of **negative impacts** (e.g., misinformation, digital divide, cyberbullying) – at least 2 (6 marks)
- Coherence, depth, and balanced analysis (4 marks)

Look out for: Insightful connections to everyday life and mention of specific technologies like email, WhatsApp, Zoom, etc.

2. Explain how computer technology has affected the job market. In what ways has it created new opportunities and eliminated others?
Marking Guide:

- Description of technological transformation in the workplace (3 marks)
- **New job opportunities** (e.g., IT roles, data analysis, remote work) – at least 2 examples (6 marks)
- **Job displacement** or automation of manual jobs – at least 2 examples (6 marks)
- Consideration of upskilling/reskilling or the gig economy (3 marks)
- Structure and clarity (2 marks)

Fig 4. Sample Marking Guide

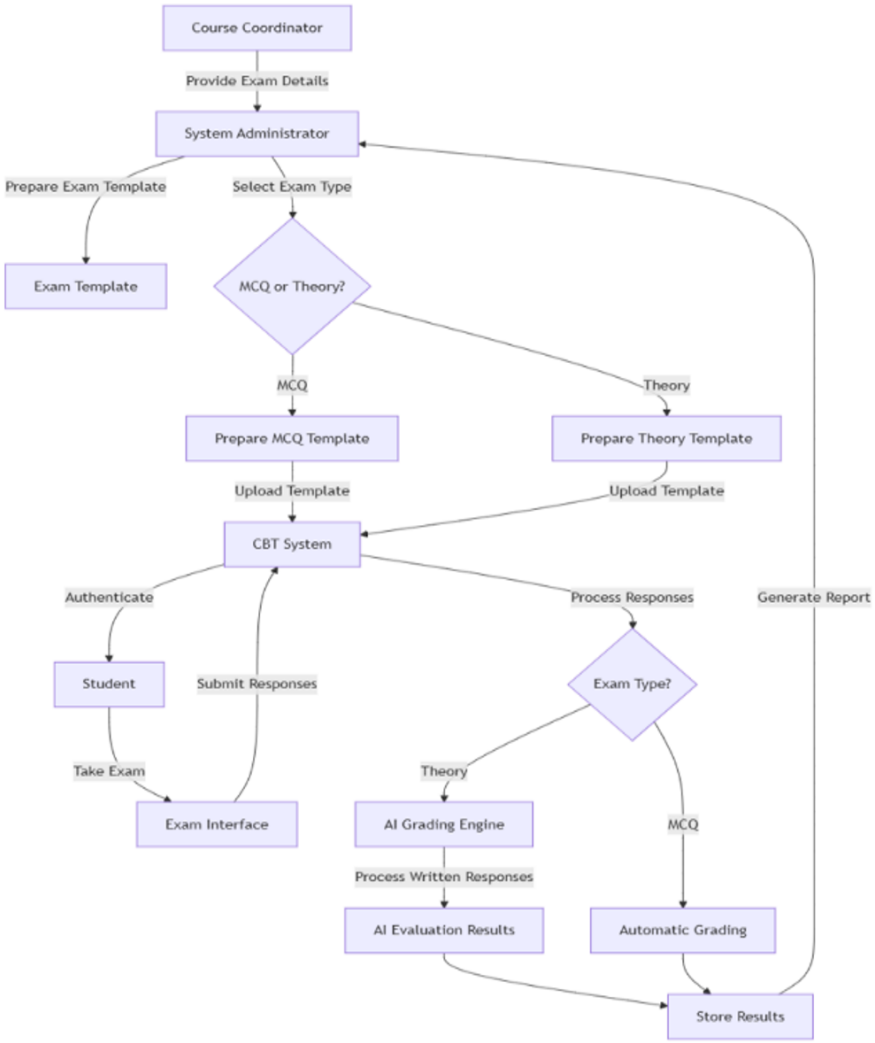


Fig 5. Workflow of the Enhanced CBT System

MongoDB served as the primary database for storing user profiles, exam data, essay responses, and grading reports, while Redis was used for caching, performance optimization, and to ensure questions generated for a student remain the same for a period of time. Security was enforced through JWT authentication and password hashing with Bcrypt, ensuring data confidentiality and integrity.

IV. Results and Discussion

The evaluation process examined how well the system met its functional goals of improving grading consistency, reducing human subjectivity, and providing transparent feedback. The system was tested using scheduled examinations for a group of students. Multiple-choice questions were graded instantly by the backend, while open-ended responses were saved, queued, and later graded by the AI engine. To assess performance, results from the AI grader were compared against those from seven human graders as shown in Table 1.

Statistical analysis, as presented in Table 2, was performed to assess grading subjectivity among humans and the consistency of the AI system. The standard deviation and range of human scores indicated variability, thereby affirming the personal nature of manual grading. The AI generated consistent outcomes, providing a numerical score alongside qualitative feedback. Correlation analysis of AI scores and average human scores revealed a strong alignment, indicating that although human judgments varied, the AI exhibited consistent performance without significant deviation from human consensus. Figure 6 shows *a grouped bar chart of the AI score compared with the average human score and the human scores for Q1–Q5.*

The grading results presented in Figure 7 indicate that the AI effectively guided coordinators through its reasoning process when assigning scores, facilitating reevaluation in instances of student appeals. The interpretability enhanced

the system's fairness and reliability. To thoroughly assess the precision and dependability of the AI-driven grading system in comparison to human evaluators, two prevalent metrics in Automated Essay Scoring (AES) research were utilised: Root Mean Square Error (RMSE) and Quadratic Weighted Kappa (QWK). The average difference between AI scores and human average scores is called the Root Mean Square Error (RMSE).

Figure 8 below shows the comparison of AI and Human Average Scores with RMSE and QWK Metrics. The RMSE that was calculated was 2.23, which means that the AI grader's scores were about two points off from the human averages on the grading scale. This relatively low error shows that AI grading and human consensus are very close to each other. Quadratic Weighted Kappa (QWK) measures how much two raters agree with each other and punishes bigger disagreements more harshly. The QWK result of 0.71 between the AI grader and human averages shows a lot of agreement. This shows that the AI system not only gives scores that are close to human averages, but it also keeps fairness and consistency across different score ranges. RMSE shows how close AI grading is to human averages, while QWK shows how fair the agreement is by taking into accoun t how big the score differences are. The results (RMSE = 2.23; QWK = 0.71) show that the AI grading system is both accurate and reliable. This means that it could be used in educational assessments in a way that is clear and can be used by a lot of people.

Overall, the evaluation results confirmed that the developed system successfully extends the traditional CBT platform at Obafemi Awolowo University by enabling automated grading of open-ended questions with accuracy, transparency, and reduced subjectivity.

TABLE 1. AI Scores versus Human Grader (HG) Scores

Question	AI Score	Human Grader1	Human Grader 2	Human Grader3	Human Grader4	Human Grader5	Human Grader6	Human Grader7
Q1	6	7	10	12	3	6	7	5
Q2	111	13	12	13	8	6	14	10
Q3	14	14	8	10	12	8	10	13
13Q4	9	16	12	13	10	7	10	10
Q5	8	13	10	12	12	5	12	12

TABLE 2. Statistical Analysis on the scores

Question	AI Score	Human Grader Average	Human Grader Standard Deviation	Human Grader Range	AI Vs HG Average
Q1	6	7.14	3.02	9	1.14
Q2	11	10.86	2.97	8	0,14
Q3	14	10.71	2.36	6	3.29
13Q4	9	11.14	2.85	9	2.14
Q5	8	10.86	2.73	8	2.86

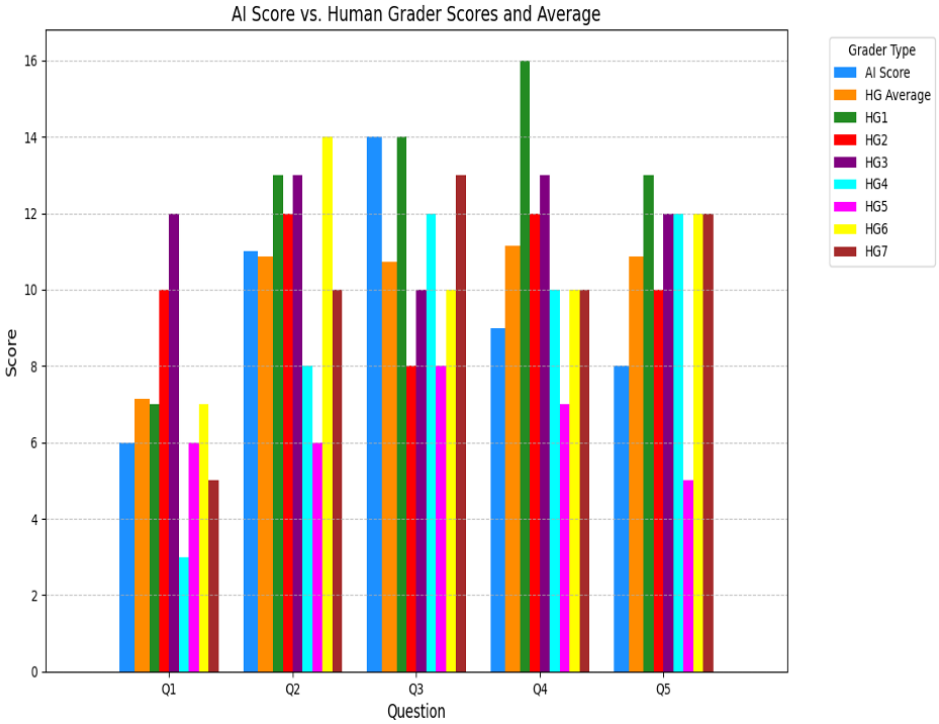


Fig 6. Grouped Bar Chart showing the AI Score, Average Human Score and Huma

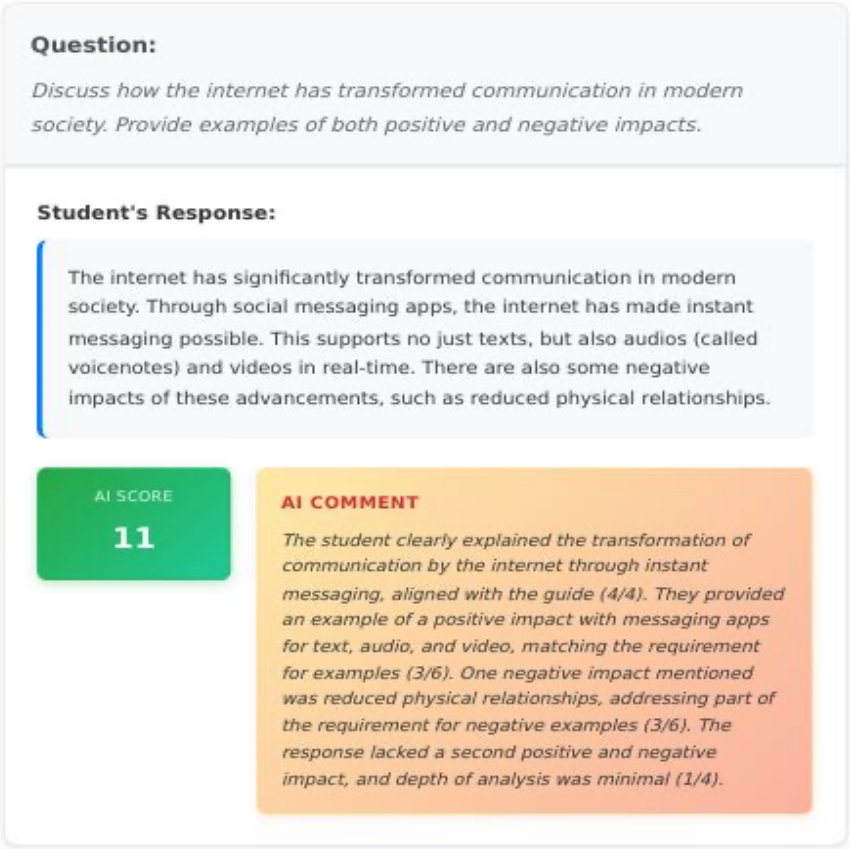


Fig 7. Sample grading result

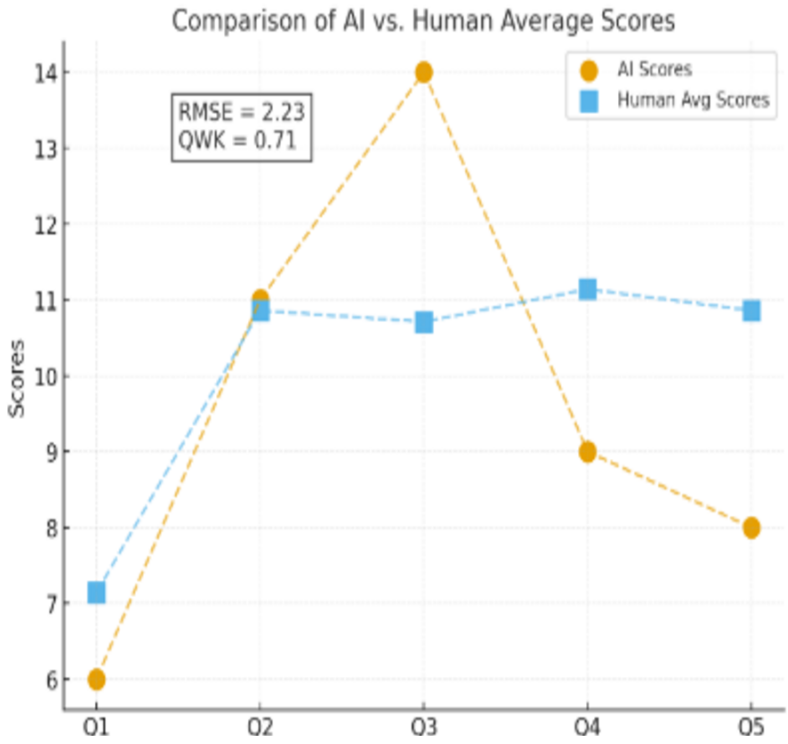


Fig 8. Comparison of AI and Human Average Scores with RMSE and QWK Metrics

V. Conclusion

The research set out to overcome the limitation of existing computer-based testing systems that largely focus on objective questions and lack the ability to evaluate essay-type responses. To achieve this, a CBT platform was designed and implemented with two main user roles: students and administrators. The backend was developed with NestJS, MongoDB, and Redis, while the frontend relied on Kotlin Multiplatform for a consistent cross-platform experience. The central innovation of the system was the integration of OpenAI's GPT-4 model, which enabled automated grading of open-ended responses.

Through testing, the system was shown to successfully handle exam creation, student participation, multiple-choice grading, and essay evaluation. More importantly, the AI grading engine did not merely produce scores but also provided explanations of its decisions. This made the grading process more transparent and easier for course coordinators to review, especially in cases where a student's script might need to be re-evaluated. This study demonstrated the practical application of Large Language Models in educational assessment with, its result showing RMSE = 2.23 and QWK = 0.71. It extends the traditional scope of CBT systems by incorporating the grading of open-ended questions, an area that has typically been left to manual evaluation. In addition, it highlights how AI-generated feedback can enhance transparency, enabling both students and coordinators to understand the reasoning behind each score rather than seeing results as arbitrary numbers.

The research suggests future work should extend beyond essay-based questions to include systems capable of assessing mathematical problems at scale, where solutions may involve both symbolic expressions and procedural steps. Similarly, more niche-specific examinations, such as oral tests or interactive assessments, could be explored, allowing the grading engine to adapt to a wider range of academic disciplines. Another important direction for future research lies in the refinement of prompt engineering, as the quality and reliability of AI grading are highly dependent on how grading instructions are structured for the model. Developing optimal prompt strategies that can consistently produce fair and accurate evaluations across different question types would improve both transparency and scalability.

References

[1] J. W. Anyor and A. A. Joshua, Mathematics curriculum change and assessment models: The quest for an integrated approach, 2024.

[2] C. E. Iderima, "Implications of computer based testing on open and distance learning," International Journal of Innovative Information Systems & Technology Research, vol. 11, no. 3, pp. 122–134, 2023.

[3] O. Fagbohun, N. P. Iduwe, M. Abdullahi, A. Ifaturoti, and O. M. Nwanna, "Beyond traditional assessment: Exploring the impact of large language models on grading practices," J. Artif. Intell. Mach. Learn. Data Sci., vol. 2, no. 1, pp. 1–8, 2024.

[4] M. Uto, "A review of deep-neural automated essay scoring models," Behaviormetrika, vol. 48, no. 2, pp. 459–484, 2021, doi: 10.1007/s41237-021-00142-y.

[5] D. S. Grewal, "A critical conceptual analysis of definitions of artificial intelligence as applicable to computer engineering," IOSR Journal of Computer Engineering, vol. 16, no. 2, pp. 9–13, 2014.

[6] Y. E. Yesilyurt, "AI-enabled assessment and feedback mechanisms for language learning: Transforming pedagogy and learner experience," in Transforming the Language Teaching Experience in the Age of AI, IGI Global, 2023, pp. 25–43.

[7] P. Kumar, "Large language models (LLMs): survey, technical frameworks, and future challenges," Artificial Intelligence Review, vol. 57, no. 10, p. 260, 2024.

[8] I. Kondurkar, A. Raj, and D. Lakshmi, "Modern applications with a focus on training ChatGPT and GPT models: Exploring generative AI and NLP," in Advanced Applications of Generative AI and Natural Language Processing Models, IGI Global, 2024, pp. 186–227.

[9] E. B. Page, "The imminence of grading essays by computer," Phi Delta Kappan, pp. 238–243, Jan. 1966.

[10] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, "Enriching automated essay scoring using discourse marking," in Proc. Workshop on Discourse Relations and Discourse Markers, Annual Meeting of the Association of Computational Linguistics, August, Montreal, Canada., Aug. 1998.

[11] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse Processes, vol. 25, pp. 259–284, 1998.

[12] L. S. Larkey, "Automatic essay grading using text categorization techniques," in Proc. 21st Annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 90–95.

[13] S. M. Elliot, "IntelliMetric: From here to validity," in M. D. Shermis and J. C. Burstein, Eds., Automated Essay Scoring: A Cross-Disciplinary Perspective. Routledge, 2003, pp. 71–86.

[14] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," J. Technol., Learn. Assess., vol. 1, no. 2, 2002.

[15] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," J. Technol., Learn. Assess., vol. 4, no. 3, 2006.

[16] S. Haller, A. Aldea, C. Seifert, and N. Strisciuglio, "Survey on automated short answer grading with deep learning: From word embeddings to transformers," *arXiv*, doi: 10.48550/arXiv.2204.03503, 2022.

[17] . Lottridge, C. Ormerod, and A. Jafari, "Psychometric considerations when using deep learning for automated scoring," in V. Yaneva and M. von Davier, Eds., Advancing Natural Language Processing in Educational Assessment. Routledge, 2023, pp. 15–30, doi: 10.4324/9781003278658-3.

[18] J. Whitmer, E. Y. Deng, C. Blankenship, M. Beiting-Parrish, T. Zhang, and P. Bailey, Results of NAEP reading item automated scoring data challenge (Fall 2021), 2023.

[19] D. Ramesh and S. K. Sanampudi, "An automated essay scoring system: A systematic literature review," Artificial Intelligence Review., vol. 55, no. 3, pp. 2495–2527, 2022.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[21] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.

[22] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and individual differences., vol. 103, p. 102274, 2023.

[23] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. Srinivasa, "Automatic assessment of text-based responses in post-secondary education: A systematic review," Computers and Education: Artificial Intelligence., p. 100206, 2024.

[24] D. Perin and M. Lauterbach, "Assessing text-based writing of low-skilled college students," International Journal of Artificial Intelligence in Education., vol. 28, pp. 56–78, 2018, doi: 10.1007/s40593-016-0122-z.

[25] J. S. Jauhiainen and A. G. Garagorry, Evaluating students' open-ended written responses with LLMs.

[26] J. Wilson, Y. Huang, C. Palermo, G. Beard, and C. A. MacArthur, "Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI Write," International Journal of Artificial Intelligence in Education, vol. 31, pp. 234–276, 2021, doi: 10.1007/s40593-020-00236-w.

[27] Y. Kumar, S. Aggarwal, D. Mahata, R. R. Shah, P. Kumaraguru, and R. Zimmermann, "Get it scored using AutoSAS—an automated system for scoring short answers," in *Proc. AAAI Conf. Artificial. Intelligenece.*, vol. 33, no. 1, pp. 9662–9669, Jul. 2019.

[28] M. Joundy Hazar, Z. Hussein Toman, and S. Hussein Toman, "Automated scoring for essay questions in E-learning," in Journal of Physics: Conference Series, vol. 1294, no. 4, p. 042014, Sep. 2019, IOP Publishing.

[29] R. Hirao, M. Arai, H. Shimanaka, S. Katsumata, and M. Komachi, "Automated essay scoring system for nonnative Japanese learners," in Proc. 12th Lang. Resource. Evaluation. Conference., 2020, pp. 1250–1257.

[30] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," Humanities and Social Sciences Communications., vol. 11, no. 1, pp. 1–15, 2024.

[31] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," Research Methods in Applied Linguistics, vol. 2, no. 2, article no. 100050, Aug. 2023, doi: 10.1016/j.rmal.2023.10005