

CHAPTER I

INTRODUCTION

1.1 Research Background

Academic assessment is a vital component of the educational process, serving to measure students' abilities and understanding of the material taught. Among the various types of assessment, essay questions hold a central role because they allow students to demonstrate comprehension in a detailed, structured, and reasoned manner. Essays do not only assess cognitive knowledge but also evaluate critical thinking, logical reasoning, and written communication skills. Unlike multiple choice questions, which are limited to binary right or wrong outcomes, essay responses reflect a wide spectrum of comprehension. They are judged on accuracy, depth of analysis, and clarity of explanation, leading to richer and more meaningful insights into student performance.

Despite these advantages, the implementation of essay-based assessment often encounters major challenges that undermine its effectiveness and fairness. One recurring issue is subjectivity and inconsistency in grading. Teachers' perceptions of student behavior, mental and physical fatigue, and limited time for evaluation can influence scores unfairly (Nugraheni et al., 2022). A qualitative study in Indonesian secondary schools reported that some teachers admitted to awarding higher marks to well-behaved students even when academic performance did not fully meet assessment criteria. Time constraints further exacerbate grading problems. (Defianty & Wilson, 2024) report that Indonesian teachers face large class sizes and time pressure, limiting opportunities for providing high-quality

formative assessment. Although technology could support teachers, many still lack adequate training and confidence to apply digital assessment tools effectively (Salzabila et al., 2025).

The impact of unfair assessment is not only theoretical but also measurable. Hasyaty et al., (2023) found that providing structured written feedback on essay tests significantly improved students' learning outcomes. Using N-Gain indicators, the control class achieved only a low improvement rate (0.19), while the experimental class reached a medium improvement rate (0.57). Students in the experimental class also demonstrated stronger self-assessment skills (70%) compared to the control group (60%). These findings confirm that when essay assessments are conducted fairly and systematically, they can substantially enhance both conceptual understanding and metacognitive skills. This aligns with Mertasari et al. (2023), who emphasize that formative performance assessments have a significant positive effect on improving students' metacognitive abilities.

Such problems create a clear gap between the ideal and the reality of school assessment systems. Ideally, evaluations should be objective, consistent, and aligned with national education standards, yet current practices often fall short due to limited human resources, insufficient technological infrastructure, and lack of awareness about the benefits of technology-assisted evaluation. When grading is inconsistent or biased, students' opportunities to develop higher-order thinking skills are hindered, and education risks devolving into rote memorization rather than genuine understanding.

Efforts to address these challenges have begun. (Prasetyo et al., 2022) developed a web-based essay scoring system using association rules and cosine

similarity, achieving up to 62.5% accuracy compared with manual teacher grading. Similarly, Darmayasa et al. (2025) recently developed an automated student learning assessment system within an LMS using RAG, achieving an accuracy of 84.81%, highlighting the effectiveness of combining LLMs and RAG for educational evaluations. However, there is still significant room for improvement in accuracy, scalability, and the ability to provide personalized feedback.

To address those remaining gaps, this study adopts Retrieval Augmented Generation (RAG) as the core AI approach for automated essay scoring. RAG combines an information retrieval component with a generative language model so that each evaluation is explicitly grounded in teacher-curated materials (e.g., ideal answers, rubrics, keywords, and course resources). For long-form essay inputs where students produce paraphrases, distributed arguments across paragraphs, and varying styles, RAG enables chunked retrieval of the most relevant evidence for each argumentative segment and uses that evidence as context when the model generates scores and formative feedback. This design directly tackles three problems that are critical for educational assessment: (1) traceability, every score and comment can be traced to specific source passages used during retrieval, (2) reduction of hallucination, the generator is constrained by retrieved, verifiable evidence instead of relying only on parametric model memory, and (3) scalability and maintainability knowledge updates are performed by updating the retrieval index rather than retraining large models. These properties make RAG particularly well suited to essay scoring where transparency, curriculum alignment, and the ability to handle long, paraphrased student text are essential.

Building on these advances, this study proposes the development of an AI-powered Learning Management System (LMS) that integrates advanced natural language processing techniques to evaluate essays more effectively. By leveraging artificial intelligence, the system can analyze language variations, process large volumes of responses, and generate individualized feedback, ensuring fairer and more consistent grading. Beyond speeding up the evaluation process, AI technology offers the potential to align school assessments with educational policies that emphasize fairness, transparency, and competency-based learning.

In conclusion, developing an AI powered LMS for essay assessment not only addresses the current shortcomings of manual grading but also sets a pathway toward modern, technology-driven evaluation practices. This innovation is expected to improve assessment efficiency, ensure consistent and objective results, and provide actionable feedback to students. It may also serve as a scalable model for integrating AI in educational assessment at multiple levels.

1.2 Problem Identification

The implementation of essay based assessments in Indonesian schools faces both individual level and systemic level challenges, resulting in inconsistent and sometimes unfair evaluations. These problems are compounded by a lack of reliable data, insufficient teacher training, and the absence of clear and standardized scoring rubrics.

1. Individual Level Problems

The current manual assessment process presents several significant challenges, primarily surrounding teacher fatigue, subjectivity, and limited technological integration. Teachers responsible for grading large numbers of essays often

experience high workloads and time pressure. A 2022 survey by the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) reported that 67% of teachers spend more than 20 hours per week on administrative tasks, including assessment, which drastically reduces the time available for careful, consistent scoring. This time constraint often exacerbates the issue of subjectivity and bias, where teachers' perceptions of student behavior or personal attitudes sometimes influence scores. For instance, (Nugraheni et al., 2022) found cases where "well-behaved" students received higher marks regardless of their actual performance, a practice that undermines fairness and creates distrust among students. Compounding these manual grading issues is the limited training educators receive in technology-assisted assessment. Many teachers have not received adequate guidance on how to utilize digital tools or automated grading systems. Research shows that teachers often lack sufficient training and continuous professional development in educational technology, which limits their confidence and ability to integrate technology effectively in assessment practices (Jiang et al., 2023) Specifically, while digital competence frameworks exist, training programs remain sporadic, superficial, and not always aligned with classroom needs, ultimately causing many educators to rely on manual assessment practices due to low confidence in using digital tools for evaluation.

2. Systemic Level Problems

Beyond individual teacher challenges, assessment practices are further hindered by several systemic-level problems. A primary issue is the lack of standardized rubrics, because schools often do not provide detailed or consistent scoring guidelines, there are significant variations in how different teachers assess the same

type of essay (Mariyati & Putrayasa, 2024). Furthermore, this is compounded by insufficient policy alignment and monitoring. While national education standards emphasize fair and competency-based assessment, practical implementation and institutional oversight remain highly limited (Adiningsih et al., 2024). Consequently, these systemic flaws contribute to a severe workload imbalance that directly affects assessment quality. As the Indonesian Teacher Federation has noted, excessive non-teaching duties have significantly lowered overall instructional quality and consistently slowed the delivery of crucial feedback to students.

3. Long Term Implications

The consequences of these problems extend beyond immediate grading inconsistencies. Students may lose trust in the fairness of the education system, which can reduce motivation and engagement. In the long term, biased or inconsistent assessment weakens the development of higher order thinking skills and can undermine the credibility of academic qualifications. If unaddressed, these issues risk producing graduates who are less prepared for competitive, innovation driven environments.

These issues show the need for an objective and efficient assessment system that reduces teacher workload while ensuring fairness. An AI powered Learning Management System (LMS) can automatically evaluate essay answers using consistent criteria, minimizing bias caused by fatigue, time pressure, or personal perceptions. The system also provides clear, personalized feedback to students, helping them understand their strengths and weaknesses. By standardizing scoring and aligning with competency based education policies, this solution addresses both

individual problems such as teacher subjectivity and lack of training and systemic issues like unclear rubrics and weak policy implementation, creating a more credible and equitable assessment process.

1.3 Research Questions

To address the issues highlighted in the background, this study formulates the core research problems as follows:

1. How's the design and the implementation of an automatic essay assessment system based on LMS using RAG to ensure it functions according to the established functional and non functional requirements?
2. How accurate is the level of the accuracy in text based subject (declarative knowledge, historical facts, narratives, and qualitative concepts) compared to the result of manual grading by teachers?
3. What is the response to the implementation of this automatic assessment system?

1.4 Research Objectives

To provide a clear direction for this study, the following research purposes are formulated in accordance with the identified problems:

1. Develop a Learning Management System powered by AI to reduce subjectivity in academic assessment.
2. Ensure assessment consistency by utilizing technology to reduce the impact of tired teachers' physical and mental conditions and limited time pressure.
3. Reduce teachers administrative workload in the assessment process, allowing them to focus on providing providing in depth evaluations.

4. Implement automated assessment technology to provide faster, more accurate, and fair assessments, as well as provide more personalized feedback to students.

1.5 Research Limitations

To ensure the feasibility and effectiveness of this research project, the following scope and limitations are defined:

1. Essay Only Assessment

This system will only assess short essay type questions that require descriptive answers, not mathematical, programming, or formula based problem solving tasks.

2. Text Based Subjects Only

The system will be tested and implemented in text based subjects such as history, and does not process images, diagrams or mathematical equation in questions or answers where conceptual understanding and reasoning are evaluated through written responses.

3. Only in Bahasa Indonesia Input

For this prototype phase, the system will only support essay responses written in Indonesia to simplify language processing and avoid the complexity of multi language NLP support.

4. Use of Static Reference Materials

The retrieval system will rely on a predefined and curated set of reference materials (e.g., textbooks, lesson modules) rather than open web sources, to control the quality and relevance of retrieved information.

5. No Mathematical Expression Processing

The system will not interpret or assess mathematical expressions, equations, or

symbolic logic, which require specialized parsing and evaluation methods beyond the scope of this study.

6. Single Subject Dataset for Evaluation

During development and testing, only one academic subject (history) will be used to train and evaluate the system's performance.

7. Cognitive Limitation

This system is limited to assessing cognitive levels from C1 to C4 (Remembering, Understanding, Applying, and Analyzing). Higher order cognitive processes such as C5 (Evaluating) and C6 (Creating) are not included in this study due to the complexity of subjective judgment required beyond rubric based AI evaluation.

8. Teacher Curated Answer Keys

The grading reference will be based on answer keys and exemplar answers provided manually by subject teachers.

9. Not Including a Comprehensive UI & UX Testing

This research focuses primarily on the computational accuracy and technical implementation of the grading algorithm using the Retrieval-Augmented Generation (RAG) method and does not include comprehensive User Interface (UI) and User Experience (UX) testing from the perspective of end user design or pedagogical ergonomics.

1.6 Research Significances

A. For Teachers

1. **Reduce Workload:** Automated assessment applications can reduce teachers' workload in correcting essay answers, leaving them with more time to focus on other teaching tasks.
2. **Increase Objectivity:** By using Learning Management System powered by AI, assessment can be done objectively without being influenced by personal views on student behavior.
3. **Assessment Consistency:** Automated assessment technology can help maintain consistency in assessment, reducing errors that may occur due to physical or mental fatigue.
4. **Administrative Workload Reduction:** The system technically accelerates the grading process and is perceived to reduce teacher workload, allowing them to provide faster feedback to students

B. For Students

1. **Fair Assessment:** Students will receive a fairer and more objective assessment, based on their academic ability without being influenced by the teacher's personal views.
2. **Quick and Personalized Feedback:** The app can provide faster and personalized feedback, helping students understand their strengths and weaknesses better.
3. **Increases Learning Motivation:** Fair and consistent assessments can increase students' learning motivation, as they feel rewarded based on their efforts and abilities.
4. **Improved Learning Outcomes:** With more immediate and personalized

feedback, students can more quickly correct mistakes and improve their learning outcomes.

C. For Researchers

1. **Science Development:** This research can make significant contributions to the development of automated assessment technologies, particularly in the context of education.
2. **Basis for Further Research:** The results of this study can be the basis for further research that can further explore the use of AI technology in various aspects of education.
3. **Improved Research Quality:** By developing applications that have a direct impact on the educational process, researchers can improve the quality and relevance of their research.
4. **Interdisciplinary Collaboration:** This research opens up opportunities for collaboration between disciplines, such as information technology, education, and psychology, to create comprehensive and innovative solutions.

