A RAG-Enhanced Human-in-the-Loop Framework for Automated Assessment of Engineering Laboratory Reports

Amina Abbi[1]*, Mohammed Skouri[2], Mustapha Raoufi[3]
Laboratory of Physics High Energy and Astrophysics, Faculty of Science Semlalia, Cadi Ayyad University, Marrakech, Morocco[1, 2]
Laboratory of Materials Energy and Environment, Faculty of Science Semlalia, Cadi Ayyad University, Marrakech, Morocco[3]

***Abstract*—Recent developments in Large Language Models (LLMs) have created new opportunities to automate educational assessment and reduce workload for instructors. However, concerns regarding grading consistency, transparency, and pedagogical reliability continue to limit the adoption of fully automated assessment systems. In this study, we propose a Human-in-the-Loop framework for the automated evaluation of engineering laboratory reports based on Retrieval-Augmented Generation (RAG). The proposed framework integrates text extraction, structured content extraction, contextual retrieval from grading rubrics and laboratory resources, rubric-based evaluation, automatic feedback generation, and instructor validation into a unified grading workflow. The RAG module retrieves contextual information so that the language model can generate assessments that align with the course goals and are based on educational information about the specific assignments, thus ensuring consistent evaluation based on the rubric. The system was tested using a set of 56 laboratory reports collected from undergraduate courses in Electrical and Electronics Engineering. The experimental results indicate a strong agreement between the grades provided by the AI and the instructor, with a Pearson Correlation Coefficient of 0.988, a Mean Absolute Error (MAE) of 3.55, and a Root Mean Squared Error (RMSE) of 3.77. Besides, 89.29% of the reports were scored within ±5 points of the instructor scores. The grading time was reduced from 392 minutes to 84 minutes, a workload reduction of 78.57%. The results demonstrate that the integration of Retrieval-Augmented Generation, rubric-based evaluation, and Human-in-the-Loop validation constitutes an effective approach for AI-supported assessment of engineering laboratory reports, maintaining instructor oversight and educational integrity.**

Keywords*—Large Language Models; Retrieval-Augmented Generation; Human-in-the-Loop; automated assessment; engineering education; laboratory report evaluation; rubric-based grading; educational artificial intelligence

I. Introduction

Artificial intelligence (AI) is rapidly transforming educational technologies by enabling intelligent tutoring, personalized learning, automated assessment, feedback generation, and AI-assisted learning environments [1]– [4], [11]. Recent studies have further highlighted the growing role of AI in educational assessment, personalized feedback, and technology-enhanced learning [10], [18], [23].

Large Language Models (LLMs), one of the most significant recent advances in AI, have demonstrated remarkable capabilities in natural language understanding and generation, making them promising tools for educational assessment [1], [7], [9], [21]. Recent studies have shown that LLMs can effectively assist instructors in grading essays, short-answer questions, programming assignments, engineering assignments, and open-ended examinations while improving grading consistency and reducing instructor workload [8], [13], [16], [17], [19].

One of the hardest types of assessment is the assessment of educational assessment in engineering laboratory reports. They assess theoretical understanding, experimental procedures, technical calculations, data interpretation, engineering reasoning, and scientific communication, as opposed to the typical written assignment. The manual grading process is thus time-consuming and may suffer from inconsistency due to instructor workload and subjective judgment.

Prior research has shown the promise of LLM-based grading systems, but most methods evaluate student submissions individually without taking into account assignment-specific educational resources such as lab instructions, grading rubrics, reference solutions, and course materials. Thus, the quality of grading may be undermined by an incomplete understanding of the context and hallucinated answers. Furthermore, fully automated assessment systems are typically opaque and lack human oversight, limiting their adoption in higher education [9], [21].

Recently, a promising approach to enhance the reliability of LLM-based systems is to augment prompts with relevant external knowledge retrieved at inference time, which is known as Retrieval-Augmented Generation (RAG) [20]. In educational assessment, RAG allows language models to assess student work with resources specific to the assignment instead of relying solely on pre-trained knowledge. Recent studies have also shown the usage of context-aware assessment and context engineering to improve the reliability of grading and quality of feedback [14], [20].

However, the automated assessment of engineering laboratory reports has been relatively less explored, with an integrated framework that combines contextual retrieval, rubric-based evaluation, automated feedback generation, and Human-

*Corresponding author

in-the-Loop validation. Previous work has largely focused on essays, math assessments, programming assignments, or short-answer questions, but engineering lab reports pose additional challenges due to the need for experimental analysis and technical reasoning.

To address this gap, this study proposes an RAG-enhanced Human-in-the-Loop framework for automated assessment of engineering laboratory reports using Large Language Models. The proposed framework combines document preprocessing, structured content extraction, Retrieval-Augmented Generation, rubric-based grading, automated feedback generation, and instructor validation into a single assessment workflow. The framework was evaluated on 56 laboratory reports of the Electrical and Electronics Engineering domain. The experimental results demonstrate a high level of agreement in the grades assigned by the AI and the instructor, with a Pearson Correlation Coefficient of 0.988, a Mean Absolute Error (MAE) of 3.55 points, and an Agreement Rate of 89.29%, reducing grading time by 78.57%.

The major contributions of this work are summarized in Table I.

TABLE I. SUMMARY OF THE MAIN CONTRIBUTIONS

Contribution	Description
C1	A RAG-enhanced Human-in-the-Loop framework for automated assessment of engineering laboratory reports.
C2	Integration of contextual retrieval, rubric-based grading, automated feedback generation, and instructor validation into a unified assessment workflow.
C3	Experimental validation with 56 real lab reports from Electrical & Electronics Engineering.
C4	Comprehensive evaluation with Pearson Correlation, MAE, RMSE, Agreement Rate, Maximum Error, and grading time reduction.

The rest of this study is organized as follows: Related work on AI-assisted educational assessment is presented in Section II. The proposed methodology is presented in Section III. Section IV describes the experiment setup. Section V presents and discusses the experimental results, including the limitations of the study. Finally, Section VI concludes the study and offers future research directions.

II. RELATED WORK

A. Education Assessment Using Large Language Models

Recently, Large Language Models (LLMs) have become a major research topic in educational technology, as they are capable of supporting intelligent tutoring, personalized learning, automated assessment, and feedback generation [1], [4], [11]. Recent surveys [1], [9], [11] show that automated assessment is one of the most promising applications of LLMs in education. The effectiveness of LLMs for grading short-answer questions [16], essays [8], programming assignments [17], engineering mechanics assignments [19], undergraduate calculus [5], and open-ended examinations [13] has been demonstrated in several studies. These studies consistently show high agreement between grades assigned by AI and instructors, and significantly reduce instructor workload.

But these encouraging results also highlight problems of grading consistency, hallucinations, transparency, and pedagogical reliability in prior research. Thus, current evidence points to LLMs supporting instructors rather than replacing human judgment [9], [21].

B. Human-in-the-Loop Evaluation and AI-Generated Feedback

Recent studies have increasingly advocated for Human-in-the-Loop (HITL) grading frameworks [7], [8] to improve the reliability and transparency of AI-assisted assessment. These methods combine automated assessment and instructor review, enabling instructors to verify AI-generated grades before they are posted as final results.

Moreover, several studies have shown that LLM-generated feedback can enhance student learning by offering personalized suggestions and constructive explanations [6], [15], [22], [24]. Moreover, AI-enabled pedagogical architectures have shown promising results in supporting self-regulated learning and enhancing student engagement [3]. The results suggest that a good assessment system should involve reliable grading, useful feedback, and instructor oversight.

C. Retrieval-Enhanced Generation for Education Assessment

Retrieval-Augmented Generation (RAG) has emerged as a powerful approach for improving the factuality and context-awareness of LLMs by retrieving relevant external knowledge before response generation [20]. In educational assessment, RAG allows language models to leverage lab instructions, grading rubrics, reference solutions, and course materials, which helps mitigate hallucinations and improves consistency in grading.

Context engineering has recently shown that the integration of contextual educational resources enhances the quality of AI-based assessment [14]. At the same time, research on automated assessment in the domain of STEM education emphasizes the increasing importance of domain-specific contextual information for reliable assessment [12].

D. Research Gap and Contributions

Despite many advances in AI-based assessment on educational resources, existing studies mainly focus on essays, programming assignments, mathematics or short-answer questions [5], [8], [13], [16], [17]. Engineering laboratory reports, which require evaluation of experimental procedures, technical calculations, engineering reasoning, and scientific communication, have been relatively neglected.

Additionally, most existing approaches rely solely on LLM reasoning without integrating contextual retrieval, rubric-based evaluation, automated feedback generation, and Human-in-the-Loop validation in a unified framework.

To overcome these limitations, this study proposes a RAG-enhanced Human-in-the-Loop framework for automated assessment of engineering laboratory reports using Large Language Models. The proposed framework combines Retrieval-Augmented Generation, structured content extraction, rubric-based grading, automated feedback generation, and instructor validation into one assessment workflow.

The framework is experimentally validated using 56 real Electrical and Electronics Engineering laboratory reports and evaluated using various quantitative metrics such as Pearson Correlation, MAE, RMSE, Agreement Rate, Maximum Error, and grading time reduction.

III. Methodology

A. Overview of the Proposed Framework

The proposed framework presents a Human-in-the-Loop framework for automated assessment of engineering laboratory reports enhanced by Retrieval-Augmented Generation (RAG). Unlike traditional LLM-based automated grading systems that only use student submissions, it enhances the grading process by accessing educational resources specific to assignments, such as laboratory instructions, grading rubrics, reference solutions, and course materials. This contextual information allows for a more consistent, transparent, and reliable assessment based on the intended learning outcomes.

This framework combines an end-to-end assessment workflow that includes document pre-processing, extraction of structured content, context retrieval, rubric-based assessment, automated feedback generation, support for plagiarism and validation by instructors. AI-generated grades and feedback are preliminary recommendations. Instructors maintain complete discretion to review, revise, approve, or reject the proposed assessment before the final grade is assigned, thereby ensuring pedagogical accountability and reducing automation bias.

The assessment workflow consists of eight sequential stages: 1) laboratory report submission, 2) document preprocessing and text extraction, 3) structured content extraction, 4) Retrieval-Augmented Generation, 5) LLM-based rubric evaluation, 6) automated formative feedback generation, 7) Human-in-the-Loop instructor validation, and 8) final grade publication. Fig. 1 illustrates the overall architecture of the proposed framework.

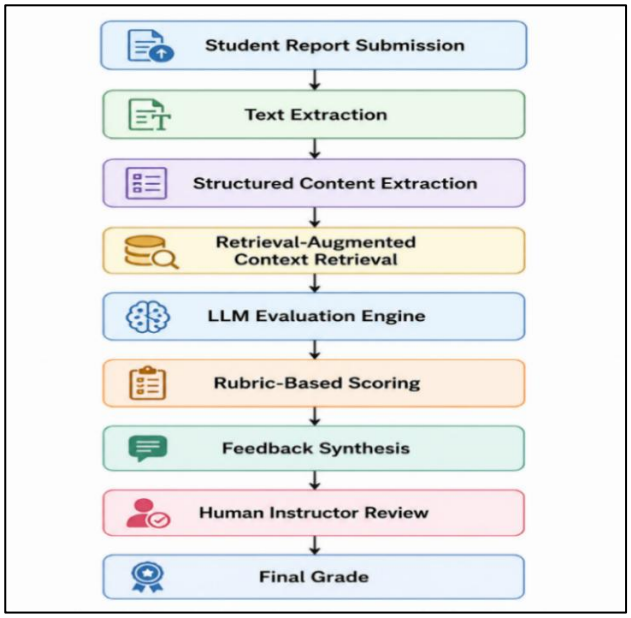

Fig. 1. Overall architecture of the proposed RAG-enhanced Human-in-the-Loop assessment framework.

B. Report Processing and Context Retrieval

Laboratory reports submitted in PDF or DOCX format are first processed to extract machine-readable content. When scanned documents are encountered, Optical Character Recognition (OCR) is applied to recover textual information. The extracted report content is subsequently prepared for automated assessment while preserving the logical organization of the laboratory report. This process provides a consistent textual representation that supports contextual retrieval and rubric-based evaluation.

The proposed framework primarily evaluates the textual content extracted from laboratory reports. Tables and textual mathematical expressions are processed through the document extraction stage and incorporated into the grading workflow whenever they are represented in machine-readable form. For graphical laboratory artifacts such as circuit diagrams, experimental plots, screenshots, and handwritten annotations, the current framework extracts only the embedded textual information when available and does not directly interpret the visual content. Consequently, grading decisions are primarily based on the extracted textual information together with the retrieved contextual resources.

To improve grading reliability and contextual understanding, the proposed framework incorporates a Retrieval-Augmented Generation (RAG) module prior to the evaluation stage. Instead of relying solely on the submitted report and the intrinsic knowledge of the Large Language Model (LLM), the framework retrieves assignment-specific educational resources that enrich the grading prompt with relevant contextual information. The knowledge base contains laboratory instructions, grading rubrics, reference solutions, course materials, and assignment metadata, enabling assessments that remain aligned with course objectives, intended learning outcomes, and predefined grading criteria while helping reduce hallucinations and improving grading consistency.

The RAG module is implemented using OpenAI's text-embedding-3-small model to generate 1536-dimensional semantic embeddings, which are indexed in a Pinecone vector database using cosine similarity. Educational resources are embedded and indexed together with their associated metadata to preserve instructional context during retrieval. For each submitted laboratory report, the retrieval query is automatically constructed from the assignment title, assignment description, and corresponding course chapter. Metadata filtering is subsequently applied to retrieve only assignment-relevant educational resources, after which the five most relevant documents are incorporated into the evaluation prompt to provide assignment-specific context for the language model.

Unlike conventional LLM-based grading approaches that evaluate reports using only the submitted content, the proposed framework combines the retrieved educational resources with grading rubrics, expected solutions, laboratory instructions, and course materials to support context-aware assessment. Consequently, laboratory reports are graded based on the desired learning outcomes and rubric criteria, not on exact textual similarity. Thus, alternative engineering approaches, experimental variations that are within an acceptable range and

alternative approaches to solutions can be fairly evaluated as long as they meet the desired learning objectives.

The assessment process combines the extracted report content with the retrieved contextual resources and evaluates all the criteria defined in the grading rubric. This allows for thorough, consistent, and pedagogically sound assessment that is aligned with the learning outcomes and instructor-defined grading standards. Fig. 2 illustrates the Retrieval-Augmented Generation process used in the proposed framework.

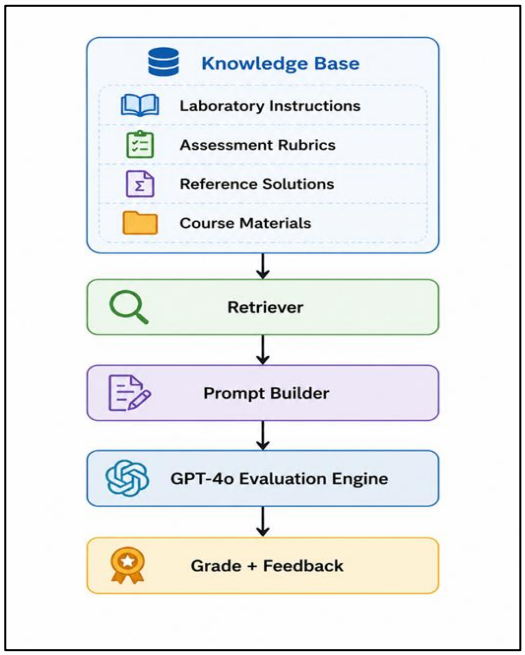

Fig. 2. Retrieval-Augmented Generation workflow for context-aware laboratory report assessment.

C. Prompt Design and LLM-Based Evaluation

The proposed framework employs a structured prompt engineering strategy to ensure consistent, transparent, and rubric-aligned evaluation of engineering laboratory reports. Instead of using the submitted report directly, the evaluation prompt combines assignment metadata, laboratory instructions, retrieved contextual resources, grading rubrics, expected reference solutions, and student submission into a unified evaluation context. This comprehensive prompt provides the language model with assignment-specific information required to perform context-aware, rubric-aligned assessment while maintaining consistency with the intended learning outcomes.

The evaluation prompt is constructed automatically by combining the assignment title, assignment description, laboratory instructions, maximum score, grading rubric, retrieved educational resources, expected solutions, and extracted laboratory report content. The language model will be responsible for evaluating the submission based on the grading rubric defined and producing a numeric score, along with detailed formative feedback, identified strengths, weaknesses, and concrete recommendations for improvement. Such a structured prompt design facilitates grading consistency and subsequent instructor review and validation. An excerpt of the structured evaluation prompt used by the proposed framework is presented below.

Excerpt of the Evaluation Prompt Used by the Proposed Framework
System Role
You are a university professor evaluating a student's assignment submission.
Assignment Details: • Title • Description • Instructions • Maximum Score
Grading Rubric: ...
Expected Answers / Reference Solutions: ...
Course Context and Reference Materials: ...
Student Submission: ...
Carefully evaluate this submission and provide: 1. Numerical Grade 2. Overall Feedback 3. Strengths 4. Weaknesses 5. Suggestions for Improvement

The assessment stage employs GPT-4o as the primary grading model because of its strong reasoning capabilities and its ability to evaluate engineering laboratory reports containing theoretical explanations, experimental procedures, calculations, experimental analysis, and technical discussions. A low temperature setting (0.3) was chosen during grading to enhance the stability of the response and to reduce variability between repeated evaluations of similar reports.

To improve computational efficiency while maintaining high-quality formative feedback, GPT-4o-mini is employed exclusively for generating detailed feedback after the numerical grade has been determined. The generated feedback is consistent with the assigned score, as the grading decision is made before the feedback generation, and reduces the computational cost. An example of the formative feedback generated by the proposed framework is presented in Fig. 3.

Fig. 3. Example of AI-generated formative feedback.

D. Rubric-Based Assessment

The proposed framework employs a criterion-referenced rubric to ensure transparent, consistent, and pedagogically aligned assessment of engineering laboratory reports. The grading rubric consists of three complementary evaluation criteria, weighted as follows: Technical Accuracy (40%), Experimental Analysis (30%), and Completeness and Presentation (30%).

Technical Accuracy receives the highest weighting because the primary objective of engineering laboratory assessment is to evaluate students' understanding and correct application of engineering concepts, calculations, and technical reasoning. Experimental Analysis emphasizes students' ability to interpret experimental results, discuss observations, and relate practical findings to theoretical concepts. Completeness and Presentation assess the organization, clarity, and professional quality of the laboratory report, reflecting the importance of technical communication in engineering education.

During assessment, the language model considers each criterion by referring to the retrieved contextual resources, grading rubric, expected reference solutions, and the assignment-specific laboratory instructions. The weighted criterion scores are subsequently combined to produce the final assessment while maintaining consistency with the predefined learning outcomes and grading standards.

E. Human-in-the-Loop Validation

The proposed framework adopts a Human-in-the-Loop (HITL) validation strategy to ensure that the AI-assisted assessment is under the supervision of the instructor. The framework offers preliminary recommendations for the AI-generated grade and feedback to assist instructors in the grading process, rather than automatically assigning a final grade.

Following the initial evaluation, instructors review the proposed grade, generated feedback, and the corresponding laboratory report in relation to the grading rubric. They may approve the assessment without modification or revise the proposed score and feedback whenever discrepancies or pedagogical considerations are identified. Consequently, instructors retain full authority over the final grading decision while accommodating valid experimental variations and alternative solution approaches.

By requiring instructor verification before grade publication, the proposed framework mitigates automation bias and combines the efficiency of AI-assisted assessment with the professional judgment and pedagogical expertise of the instructor.

F. Academic Integrity Support

In addition to automatic assessment, the proposed framework has an academic integrity support module to assist instructors in detecting potential plagiarism prior to grading. The module does not make the determination of whether academic misconduct has occurred, but provides similarity analysis to assist instructor judgment in the assessment process.

The academic integrity module is integrated in the proposed framework but is not quantitatively evaluated in this work, as the goal is to validate the performance of the RAG-enhanced Human-in-the-Loop assessment framework.

IV. Experimental Setup

A. Dataset

The proposed framework was evaluated using 56 laboratory reports collected from undergraduate Electrical and Electronics Engineering (EEE) laboratory courses at a single higher education institution. As summarized in Table II, the dataset comprises four laboratory experiments covering analog electronics and circuit analysis. Three experiments were completed as group assignments, whereas the fourth required individual report submission.

The laboratory reports include theoretical explanations, experimental procedures, calculations, observations, experimental analysis, tables, and conclusions, providing representative examples of undergraduate engineering laboratory documentation. These reports were utilized to assess the accuracy, consistency, and efficiency of grading of the proposed RAG-enhanced Human-in-the-Loop assessment framework.

TABLE II. Dataset Composition

Laboratory	Topic	Assignment Type	Reports
TP1	Diodes	Group	8
TP2	Common Emitter Amplifier	Group	8
TP3	Operational Amplifier Applications	Group	8
TP4	Capacitor Charging and Discharging	Individual	32
Total			56

B. Experimental Approach

Instructor-assigned grades (on a 100-point scale) were used as the reference standard for evaluating the proposed framework. Each laboratory report was first graded manually by the course instructor according to the predefined assessment rubric. The same reports were subsequently evaluated using the proposed RAG-enhanced Human-in-the-Loop framework following the workflow illustrated in Fig. 1.

For each submission, the framework extracted the report content, retrieved assignment-specific contextual resources via the RAG module, and generated an initial grade and formative

feedback. Next, the AI-generated grades were compared to the instructor-assigned grades to evaluate grading accuracy, consistency, and operational efficiency. During the experimental evaluation, instructor grades served exclusively as the reference for quantitative analysis. Although the proposed framework incorporates Human-in-the-Loop validation, instructor intervention was intentionally omitted during performance evaluation so that AI-generated grades could be compared directly with the reference grades under identical conditions. Instructors have full authority to review, modify, approve, or reject AI-generated assessments before the final grade is released in operational use.

C. Evaluation Metrics

The performance of the proposed framework was evaluated using six complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Maximum Error, Agreement Rate (±5 points), Pearson Correlation Coefficient, and Grading Time Reduction. These metrics were selected to assess grading accuracy, agreement with instructor-assigned grades, and operational efficiency.

Pearson Correlation Coefficient was used to measure the linear agreement between AI-generated and instructor-assigned grades, while MAE, RMSE, and Maximum Error quantified grading deviations. The Agreement Rate represents the percentage of laboratory reports for which the AI-generated grade differed by no more than ±5 points from the instructor-assigned grade. Grading Time Reduction was used to evaluate the efficiency gained through AI-assisted assessment.

To further assess the robustness of the proposed framework, statistical significance testing, p-values, and 95% confidence intervals were computed for the primary evaluation metrics. These complementary statistical measures provide additional evidence regarding the reliability and consistency of the proposed RAG-enhanced Human-in-the-Loop assessment framework.

D. Implementation Details

The proposed framework was implemented in Python using OpenAI GPT-4o as the primary model for rubric-based grading and GPT-4o-mini for formative feedback generation. Retrieval-Augmented Generation (RAG) was implemented using the OpenAI text-embedding-3-small embedding model, which generates 1536-dimensional semantic embeddings. These embeddings were indexed in a Pinecone vector database using cosine similarity to enable efficient semantic retrieval of assignment-specific educational resources.

Educational resources, including laboratory instructions, grading rubrics, reference solutions, and course materials, were indexed as complete semantic documents without additional text chunking because these resources are relatively short and self-contained. For each laboratory report, the retrieval query was automatically constructed from the assignment title, assignment description, and corresponding course chapter. Metadata filtering was subsequently applied to retrieve the five most relevant educational resources, which were incorporated into the evaluation prompt together with the grading rubric, expected reference solutions, laboratory instructions, and the extracted laboratory report content.

Table III summarizes the implementation parameters used during the experimental evaluation.

TABLE III. IMPLEMENTATION CONFIGURATION

Parameter	Value
Programming Language	Python
Primary Grading Model	GPT-4o
Feedback Model	GPT-4o-mini
Embedding Model	OpenAI text-embedding-3-small
Embedding Dimension	1536
Vector Database	Pinecone
Similarity Metric	Cosine Similarity
Indexed Documents	Complete Semantic Documents
Retrieved Documents	Top-5
Retrieval Filtering	Assignment/Chapter Metadata
Assessment Strategy	RAG-Enhanced Rubric-Based Evaluation
Human Validation	Human-in-the-Loop

V. RESULTS AND DISCUSSION

A. Overall Grading Performance

The proposed framework was evaluated using 56 undergraduate Electrical and Electronics Engineering laboratory reports. Table IV summarizes the quantitative performance of the proposed framework by comparing AI-generated grades with instructor-assigned grades on a 100-point grading scale.

TABLE IV. PERFORMANCE OF THE PROPOSED FRAMEWORK

Metric	Value
Number of Reports	56
Grading Scale	0–100 points
Mean Instructor Grade	83.64
Mean AI Grade	80.09
Mean Absolute Error (MAE)	3.55 points (95% CI: 3.23–3.89)
Root Mean Squared Error (RMSE)	3.77 points (95% CI: 3.43–4.11)
Maximum Error	6 points
Agreement Rate (±5 points)	89.29%
Pearson Correlation	0.988 (95% CI: 0.980–0.993, $p < 0.001$)
Manual Grading Time	392 min
AI-assisted Grading Time	84 min
Time Reduction	78.57%

The proposed framework achieved a Pearson Correlation Coefficient of 0.988 (95% CI: 0.980–0.993, $p < 0.001$), indicating a statistically significant and very strong agreement between AI-generated and instructor-assigned grades. The Mean Absolute Error (MAE) was 3.55 points (95% CI: 3.23–3.89), while the Root Mean Squared Error (RMSE) was 3.77 points (95% CI: 3.43–4.11). Furthermore, 89.29% of the laboratory reports received AI-generated grades within ±5

points of the instructor-assigned grades, demonstrating high grading consistency.

The mean instructor-assigned grade was 83.64 points, whereas the mean AI-generated grade was 80.09 points, indicating an average difference of 3.55 points. This result suggests a slight tendency of the AI to assign more conservative grades than the instructor. Nevertheless, the very high correlation coefficient, together with the narrow confidence intervals and relatively small grading errors, indicates that the proposed framework closely reproduces instructor grading decisions while maintaining consistent ranking of student performance across the evaluated reports.

B. Agreement with Human Evaluation

Fig. 4 illustrates the relationship between AI-generated grades and instructor-assigned grades for the 56 laboratory reports. Most data points are closely distributed around the line of perfect agreement, demonstrating strong consistency between the two grading approaches.

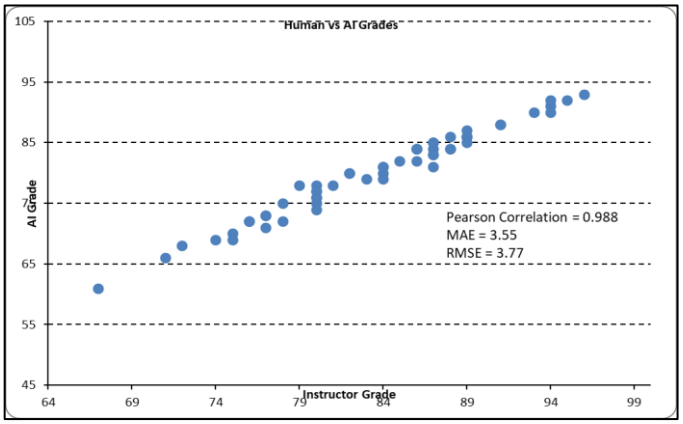

Fig. 4. Correlation between instructor-assigned grades and AI-generated grades.

Overall, 89.29% of the laboratory reports received AI-generated grades within ±5 points of the instructor-assigned grades, while the maximum observed difference was 6 points. The mean AI-generated grade (80.09) was slightly lower than the mean instructor-assigned grade (83.64), corresponding to an average difference of 3.55 points. This observation suggests a slight conservative grading bias of the proposed framework rather than inconsistent assessment. Nevertheless, the very high agreement between AI-generated and instructor-assigned grades indicates that this bias has only a limited impact on the overall grading consistency and reliability of the proposed framework.

C. Impact of Retrieval-Augmented Generation

The Retrieval-Augmented Generation (RAG) module retrieves lab instructions, grading rubrics, reference solutions, and course materials relevant to each lab report, providing assignment-specific contextual information throughout the assessment process. The contextual resources are provided to the evaluation prompt, allowing the language model to execute rubric-aligned and course-specific assessment consistent with the intended learning outcomes.

The high correlation between AI-generated and instructor-assigned grades shows the effectiveness of the proposed RAG-enhanced Human-in-the-Loop framework as an AI-assisted assessment system. Figure 5 shows an example of contextual information retrieved by the RAG module during the evaluation process.

D. Human-in-the-Loop Validation

The experimental results demonstrate that the proposed framework provides reliable preliminary assessments that can effectively support instructor decision-making. The high agreement between AI-generated and instructor-assigned grades indicates that the framework can substantially reduce instructor workload while maintaining grading consistency.

Within the proposed Human-in-the-Loop workflow, AI-generated grades and formative feedback are presented to the instructor as recommendations rather than final decisions. Instructors retain full authority to review, revise, approve, or reject the proposed assessment before the final grade is released. This validation process preserves pedagogical accountability while reducing the potential impact of automation bias in educational assessment.

E. Grading Efficiency

In addition to grading accuracy, the proposed framework substantially reduced the instructor's grading workload. The 56 laboratory reports were evaluated manually in 392 minutes, whereas the AI-assisted evaluation completed the same grading task in 84 minutes, which is a reduction in grading time of 78.57%.

The reported grading time includes only the assessment process itself, encompassing laboratory report evaluation, score generation, and feedback generation. Initial system development, RAG knowledge base construction, document indexing, and system configuration were performed offline and were therefore not included in the reported grading time. Fig. 5 compares the time required for manual and AI-assisted grading.

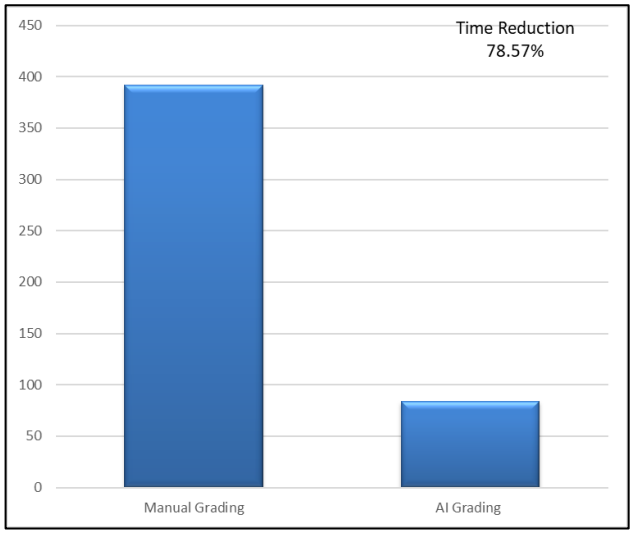

Fig. 5. Manual and AI-assisted grading time comparison.

F. Discussion

Experimental results indicate that the combination of Retrieval-Augmented Generation (RAG), rubric-based assessment, and Human-in-the-Loop validation provides an effective solution for the automated assessment of engineering laboratory reports. The proposed framework achieved strong agreement with instructor-assigned grades while substantially reducing grading time, highlighting its potential as an AI-assisted decision-support tool that improves instructor efficiency while preserving instructor oversight.

The reported results reflect the performance of the proposed framework as an integrated assessment system. Although the RAG module provides assignment-specific contextual grounding during assessment, the present study did not include a quantitative comparison between the proposed RAG-enhanced framework and an equivalent LLM-based grading approach without Retrieval-Augmented Generation. Consequently, the independent contribution of the RAG module cannot be isolated from the overall framework performance.

Furthermore, the experimental evaluation was conducted on 56 laboratory reports collected from undergraduate Electrical and Electronics Engineering courses at a single higher education institution. Therefore, the reported results should be considered as an initial validation of the proposed framework. Future work will investigate the proposed framework using larger, multi-institutional, and multidisciplinary datasets, perform controlled comparisons between RAG-enhanced and non-RAG grading approaches, evaluate the effectiveness of the academic integrity support module, and explore the integration of multimodal vision-language models to enhance the interpretation of graphical laboratory artifacts, including circuit diagrams, experimental plots, screenshots, and handwritten annotations.

VI. Conclusion and Future Work

This study presented a Retrieval-Augmented Generation (RAG)-enhanced Human-in-the-Loop framework for the automated assessment of engineering laboratory reports using Large Language Models. By integrating structured content extraction, contextual retrieval, rubric-based grading, automated formative feedback generation, academic integrity support, and instructor validation into a unified workflow, the proposed framework provides a comprehensive AI-assisted assessment solution for engineering education.

Experimental evaluation on 56 undergraduate Electrical and Electronics Engineering laboratory reports demonstrated that the proposed framework can provide accurate, consistent, and efficient AI-assisted assessment while preserving instructor oversight. The framework achieved a Pearson Correlation Coefficient of 0.988 ($p < 0.001$), a Mean Absolute Error (MAE) of 3.55 points, an Agreement Rate of 89.29%, and a grading time reduction of 78.57%, demonstrating its potential to improve assessment efficiency while maintaining grading quality and supporting reliable formative assessment in engineering education.

Future work will focus on validating the proposed framework using larger multi-institutional and multidisciplinary datasets, conducting controlled comparisons with equivalent LLM-based grading approaches without Retrieval-Augmented Generation, evaluating the effectiveness of the academic integrity support module, and investigating the integration of multimodal vision-language models to further enhance the assessment of graphical laboratory artifacts.

References

[1] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, "Large language models for education: A survey and outlook," IEEE Signal Processing Magazine, vol. 42, no. 6, pp. 51–63, Nov. 2025, doi: 10.1109/MSP.2025.3594309.

[2] N. Harsha Vardhan Reddy, N. Thirumal Reddy, M. Bharath, N. Hemanth, D. P. Dharnasi, B. Nirmala, and A. Jitendra, "AI based learning assistant using machine learning," International Journal of Engineering & Extended Technologies Research, vol. 8, no. 2, Mar.–Apr. 2026, doi: 10.15662/IJEETR.2026.0802003.

[3] G. D. O. Lima, J. A. R. Costa, F. A. Dorça, and R. D. Araújo, "An AI-supported pedagogical architecture to foster self-regulated learning in virtual environments," Computer Applications in Engineering Education, vol. 34, no. 1, Nov. 2025, doi: 10.1002/cae.70118.

[4] H. Li, L. Xiao, K. Lu, D. Li, Z. Zhang, and Q. Xia, "Exploring the application of large language models (LLMs) in data structure instruction: An empirical analysis of student learning outcomes in computer science," Information, vol. 17, no. 4, p. 353, 2026, doi: 10.3390/info17040353.

[5] S. Dakshit and S. S. Roy, "Framework for evaluating LLM performance in undergraduate calculus," Informatics, vol. 13, no. 6, p. 82, Jun. 2026, doi: 10.3390/informatics13060082.

[6] F. A. Zampirolli, F. Teubl, P. H. Pisani, and T. A. P. e Silva, "Intelligent feedback for individualized introductory programming exercises," Computer Applications in Engineering Education, vol. 34, no. 1, Dec. 2025, doi: 10.1002/cae.70132.

[7] A. Vanhoyweghen, V. Holst, M. Mobini, L. Van de Voorde, T. Vanleke, B. Verbruggen, B. Verbeken, A. Algaba, S. Verboven, M.-A. Guerry, F. Van Droogenbroeck, and V. Ginis, "Human-in-the-loop LLM grading for handwritten mathematics assessments," arXiv preprint arXiv:2603.13083, Mar. 2026.

[8] C. Zhang, K. P. Ju, X. Lu, Y.-C. G. Yen, and J. M. Rzeszotarski, "EvaluAId: Human-AI collaborative evaluation of open-ended student essays," in Proc. CHI Conf. Human Factors in Computing Systems (CHI '26), 2026, Art. no. 867, pp. 1–20, doi: 10.1145/3772318.3790814.

[9] M. Jukiewicz and M. Wyrwa, "Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing feedback," Applied Sciences, vol. 16, no. 2, p. 680, 2026, doi: 10.3390/app16020680.

[10] O. Niyibizi, "Impact of large language models on personalized learning, assessment automation, and student outcomes in higher learning institution," Journal of Technology-Assisted Learning, vol. 2, no. 1, pp. 47–61, Mar. 2026, doi: 10.70232/jtal.v2i1.22.

[11] S. Filippi and B. Motyl, "Large language models (LLMs) in engineering education: A systematic review and suggestions for practical adoption," Information, vol. 15, no. 6, p. 345, 2024, doi: 10.3390/info15060345.

[12] M. O. Omopekunola, "Large language model-based automated item generation in STEM assessments: Historical mapping and a scoping review of empirical studies," Journal of Educational Technology Development and Exchange, vol. 19, no. 2, pp. 141–165, Jun. 2026, doi: 10.18785/jetde.1902.07.

[13] J. S. Jauhiainen and A. Garagorry Guerra, "Evaluating open-ended high-stakes examinations with LLMs: Alignment between ChatGPT-4o and human grading in high and low-resource languages," Frontiers in Digital Education, vol. 3, no. 2, p. 17, 2026, doi: 10.1007/s44366-026-0091-1.

[14] Z. Tian, A. Liu, L. Esbenshade, M. Xiao, Z. Zhang, Y. Lápicus, T. Han, K. He, and M. Sun, "Creating and evaluating K-12 GenAI assessment graders through context engineering," arXiv preprint arXiv:2606.12422, 2026, doi: 10.48550/arXiv.2606.12422.

[15] N. Abbas and E. Atwell, "Cognitive computing with large language models for student assessment feedback," Big Data and Cognitive Computing, vol. 9, no. 5, p. 112, 2025, doi: 10.3390/bdcc9050112.

[16] O. Henkel, A. Boxer, L. Hills, B. Roberts, and Z. Levonian, "Can large language models make the grade? An empirical study evaluating LLMs'

ability to mark short answer questions in K-12 education," in Proc. 11th ACM Conf. Learning @ Scale (L@S '24), 2024, pp. 300–304, doi: 10.1145/3657604.3664693.

[17] Q. Zhou, Z. Wang, Y. Li, and W. K. Chan, "Impacts of histories and models on LLM grading: A study in advanced software engineering courses," arXiv preprint arXiv:2606.08400, Jun. 2026, doi: 10.48550/arXiv.2606.08400.

[18] K. Parmar, D. Palaniappan, T. Premavathi, and R. Jain, "Revolutionizing educational assessments with AI technology," in Rethinking the Pedagogy of Sustainable Development in the AI Era, pp. 227–252, 2025, doi: 10.4018/979-8-3693-9062-7.ch011.

[19] A. Mowafy, M. Talebi-Kalaleh, M. Sabek, H. Peng, M. Aqib, S. M. Adeeb, M. M. Elgammal, and C. Pettit, "Automated grading of engineering mechanics assignments using large language models and computer vision: A work in progress," in Proc. 2025 ASEE Annual Conference & Exposition, Montreal, Canada, Aug. 2025, doi: 10.18260/1-2--55492.

[20] Y. Huang and J. X. Huang, "A survey on retrieval-augmented text generation for large language models," ACM Computing Surveys, vol. 58, no. 12, Art. no. 300, pp. 1–38, 2026, doi: 10.1145/3805774.

[21] D. Deepshikha, "A systematic review on the future of educational assessment: AI-driven grading and personalised feedback in higher education," Artificial Intelligence in Education, vol. 2, no. 2, pp. 75–115, 2026, doi: 10.1108/AIIE-03-2025-0036.

[22] X. Li, D. Chen, D. Tuffley, G. Scott, G. Tuxworth, T. Yao, and G. Torrisi-Steele, "Evaluating AI-generated feedback for formative assessment in higher education," International Journal of AI in Pedagogy Innovation and Learning Futures, vol. 2026, no. 1, Mar. 2026, doi: 10.46787/ijaipil.v2026i1.6962.

[23] N. T. Mai, W. Cao, and Q. Fang, "A study on how LLMs (e.g. GPT-4, chatbots) are being integrated to support tutoring, essay feedback and content generation," Journal of Computing and Electronic Information Management, vol. 18, no. 3, 2025, doi: 10.54097/6r6yhn67.

[24] J. A. Khan, M. Yaqoob, M. Tasadduq, H. S. Dar, and A. Ahsan, "Personalized and constructive feedback for computer science students using the large language model (LLM)," arXiv preprint arXiv:2510.11556, Oct. 2025, doi: 10.48550/arXiv.2510.11556.