

DOI: 10.5281/zenodo.19511546

# AI-AUGMENTED ASSIGNMENT EVALUATION: A FRAMEWORK FOR FAIR AND EFFECTIVE ASSESSMENT IN THE ERA OF LARGE LANGUAGE MODELS

Lubna Yasmin Pinky<sup>1\*</sup>, Mohammad Ashraful Islam<sup>2</sup>

<sup>1</sup>*Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology  
University, Bangladesh*

<sup>2</sup>*Department of Computer Science and Engineering, Jahangirnagar University, Bangladesh*

Received: 03/12/2025

Accepted: 17/02/2026

Corresponding author: Lubna Yasmin Pinky  
(lubnacse@mbstu.ac.bd)

## ABSTRACT

ChatGPT and other Large Language Models such as Gemini have been quickly adopted, and have been affecting academic practice, particularly in assignment writing and grading. Even though these technologies offer a lot of efficiency to the grading process, feedback generating, and documentation, they pose serious pedagogical risks, including the absence of student engagement, excessive reliance on AI-generated content, and the inability to quantify actual learning. I propose AI-Augmented Assessment Evaluation Framework (AAEF) in this paper where automated evaluation will be integrated with human judgment to ensure fairness, transparency, and pedagogical integrity. The framework suggests two additional factors: the Pedagogical Assessment Score (PAS), assessing conceptual knowledge and the level of reasoning, and Pedagogical Risk Score (PRS) assessing AI dependency and authenticity concerns. An integrated assessment model is developed to achieve efficiency and learning results. The proposed framework is validated using a comparative study that would involve 80 undergraduate students that will be divided into three groups: traditional manual evaluation (G1), AI-only evaluation (G2), and AAEF hybrid evaluation (G3). According to the outcomes of the experiments, in spite of the fact that AI saves a lot of time (44.7 min to 8.1 min), the hybrid AAEF is better in grade consistency (89.96), retention of learning (86.42), and student satisfaction (87.81) with a safe AI dependency index (27.3). The study provides effective suggestions to educators and policymakers that can reorganize the assessment plans in the evolving AI-centered education environment.

---

**KEYWORDS:** AI-augmented evaluation, Large Language Models, automated grading, pedagogical integrity, hybrid assessment, academic fairness, Pedagogical Assessment Score, Pedagogical Risk Score

---

## 1. INTRODUCTION

The emergence of Large Language Models (LLMs) and the creation of such tools as ChatGPT and Gemini have led to a new paradigm in terms of how students write academic papers and how they are evaluated by teachers. Such models are able to generate coherent essays, program solutions, as well as generate structured documentation with minimal human intervention. Consequently, conventional forms of assessment based on assignments are progressively confronted by the challenge of separating between student-created and AI-assisted assignments.

On the one hand, LLMs can offer considerable possibilities to improve the efficiency of education. Teachers can use AI applications to perform automated grading, create feedback, conduct scoring using rubrics, and detect plagiarism. Those capabilities decrease the workload and allow to turn around faster, particularly in large classes. Conversely, the general adoption of LLMs poses significant issues of academic honesty, originality, and the decline of critical thinking abilities. Students can be too dependent on AI-generated answers, which can result in shallow knowledge and less cognitive processing.

The current evaluation systems are not made to fit this new reality. Conventional methods of evaluation mainly emphasize the ultimate result of a task, which presupposes that the work is a product of personal effort and knowledge of the student. But with the help of AI, this assumption is no longer true. It is urgent to restructure assessment systems that would consider the advantages and risks of using AI.

This gap is filled in this paper, with the proposed AI-Augmented Assignment Evaluation Framework (AAAEF) that combines automated AI-based assessment with human judgment, taking into consideration pedagogical considerations explicitly. The key findings of the research are as follows: (i) a hybrid AI-human assessment system; (ii) Pedagogical Assessment Score (PAS) as the evaluation of the depth of learning; (iii) Pedagogical Risk Score (PRS) as the evaluation of AI dependency; (iv) a mathematical model of a balanced evaluation; (v) experimental validation of traditional, AI-only, and hybrid methods; and (vi) a unified architectural framework integrating AI evaluation, pedagogical scoring, and risk-aware assessment

## 2. RELATED WORK

Automated grading, intelligent tutoring systems, and adaptive learning platforms are among the most common areas of recent research on the integration of artificial intelligence in education. The initial research

in automated assessment was conducted by Schneider et al. (2023), Tobler et al. (2023), Jauhainen et al. (2024), Wangwiwattana et al. (2023), Yeung et al. (2025) was on rule-based systems and machine learning models to grade short answers and programming assignments.

The modern LLMs have greatly improved automated evaluation with the introduction of transformer-based architectures. Xiao et al. (2024) showed that LLMs have the potential to transform the grading of essays into a grading that is more of an augmentation rather than a mere automation. Xie et al. (2024) suggested grading like a human by reconsidering automated assessment with LLMs, where rubric alignment is better. Chen et al. (2025) presented CoGrader, a team-based system of integration of LLM in assessing project reports. In the context of programming assignments, Tseng et al. (2024) designed CodEv to provide consistent and constructive feedback, Akyash et al. (2025) designed StepGrade to be able to provide grading based on the context, and Cisneros-Gonzalez et al. (2025) designed JorGPT, a grading system with an instructor. Suryakumar et al. (2025) introduced AIvaluate, a multi-agent system to evaluate answer scripts automatically with the help of semantic vectors indexing.

In the case of hybrid evaluation, Selvam and Gonzalez Vallejo (2025) specifically discussed human-in-the-loop models of ethical AI grading. Anghel et al. (2025) introduced CourseEvalAI, a rubric-based model of clear-cut evaluation of LLM. Hondarangala and Wickramaarachchi conceptualized authenticity using a knowledge based validation system. In spite of these developments, the majority of the current frameworks do not have a formal way of measuring the quality of learning and AI dependency at the same time, which is what this study addresses by introducing PAS and PRS.

## 3. PROBLEM STATEMENT

The introduction of LLMs into academic processes has radically changed the essence of assignment preparation and assessment. The conventional evaluation models presuppose that the work that is submitted is based on the personal knowledge, rationale, and individual effort of the student. This assumption is becoming untrue in the face of AI tools. It is possible to identify three challenges:

### 3.1. Lack of Authenticity Verification

The teachers cannot tell the difference between human-created and AI-aided work. The current plagiarism detection software cannot detect AI-generated text, which is always original but not

written by the student.

### 3.2. Overemphasis on Final Output

Conventional grading emphasizes the end product over the learning process and does not reflect conceptual knowledge, reasoning and cognitive involvement. Without real learning, students are able to score high marks.

### 3.3. Absence of Pedagogical Risk Control

Though AI makes things more efficient, it is not without threats, as its over-reliance, inability to think critically and superficial learning are also threats. Selvam et al. addressed that the current systems do not quantify or penalize such risks. This then requires a proper concept of balanced assessment that achieves AI-based effectiveness, keeps human judgment, quantitative learning quality, and penalizes AI abuse.

## 4. PROPOSED FRAMEWORK: AAAEF

The current section introduces the suggested AAAEF, a combination of an AI-based assessment and pedagogical metrics and human judgment that would guarantee efficiency, fairness, and learning integrity.

### 4.1. Overview and Architecture

The AAAEF is a hybrid framework that includes pedagogical metrics by integrating AI-driven assessment with human evaluation. It has three goals, which are: (i) to enhance the efficiency of evaluation; (ii) to provide fairness and consistency; and (iii) to maintain the quality of learning and academic integrity.

The general structure of AAAEF is shown in Figure 1. The framework is configured as a multi-step evaluation pipeline that handles student assignments in both automated and human-centered subsystems. The AAAEF comprises six significant elements: (1) Submission Layer - obtains assignment with optional metadata (AI usage declaration, prompt history, draft

versions); (2) AI Evaluation Module; (3) Pedagogical Assessment Module; (4) Risk Adjustment Module; (5) Human Evaluation Module and (6) Final Decision Engine.

The system starts with the submission of assignments with optional metadata like AI usage declarations and draft history. The content provided is then subject to AI-based evaluation, pedagogical assessment, and risk analysis and finally evaluated by a person. Lastly, the entire components are combined to produce the final score and feedback.

### 4.2. AI Evaluation Module

The AI module will be involved in automated assessment with the help of the tools with the support of the LLM, rubric scoring, grammar and coherence, structural assessment, and identification of AI-generated text. It generates AI Score ( $S_a$ ) and AI Feedback ( $F_b$ ), which takes much time to evaluate it and has structural consistency.

The primary goal of the module is to increase effectiveness of assessment and consistency of extensive evaluation. Nonetheless, as AI models are mostly comparing superficial characteristics, their results are followed by additional pedagogical and human examination processes.

### 4.3. Pedagogical Assessment Score (PAS)

To achieve the richness of learning, PAS is presented. PAS assesses the level of conceptual knowledge, reasoning skill, and originality actually, meaning that the evaluation will not be based on the superficiality of correct responses. The PAS measures the quality of learning without considering the superficiality of correctness and it is computed as:

$$PAS = \lambda_1 \times CU + \lambda_2 \times RD + \lambda_3 \times O \quad \dots (1)$$

where CU = Conceptual Understanding, RD = Reasoning Depth, O = Originality, and  $\lambda_1=0.40$ ,  $\lambda_2=0.35$ ,  $\lambda_3=0.25$  are weighting parameters. This aspect will make sure that the assessment system focuses on deep learning results and prevents superficial, AI-enhanced answers.

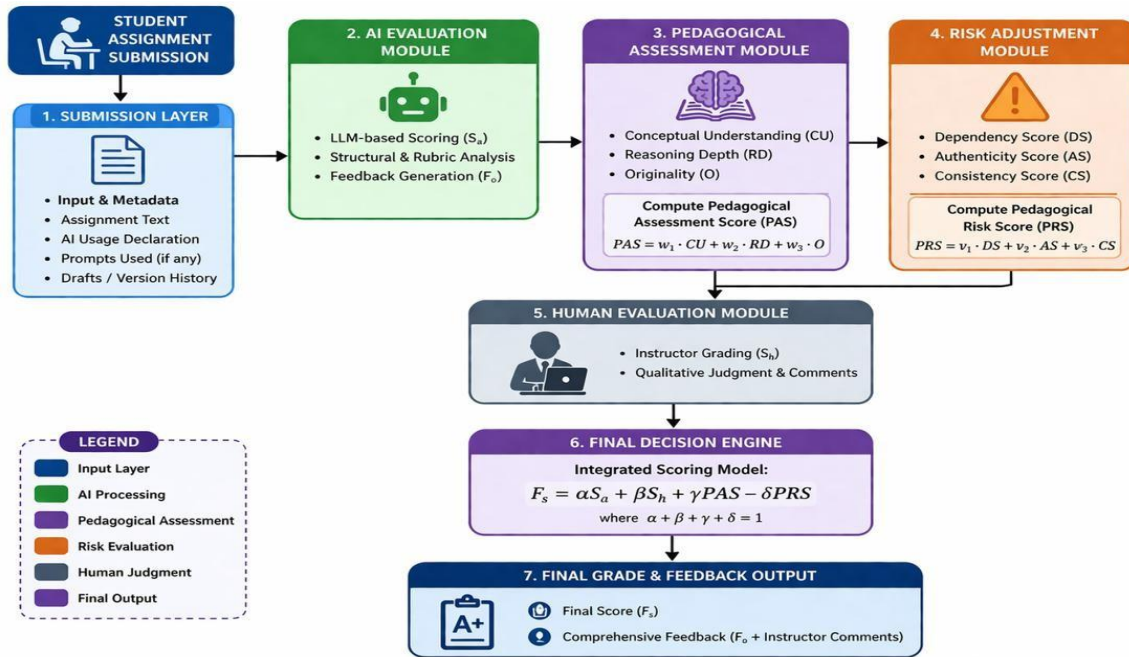


Figure 1: Architecture of the Proposed AI-Augmented Assignment Evaluation Framework (AAAEF)

#### 4.4. Pedagogical Risk Score (PRS)

Although AI makes work more efficient, it presents challenges of excessive dependence and a lack of genuineness. In response to this, the Pedagogical Risk Score (PRS) is used to measure the possible risks of AI-assisted submissions. The PRS measures the pedagogical risk posed by the use of AI:

$$PRS = \mu_1 \times DS + \mu_2 \times (1-AS) + \mu_3 \times (1-CS) \quad \dots (2)$$

where DS = Dependency Score, AS = Authenticity Score, CS = Consistency Score, and  $\mu_1=0.40$ ,  $\mu_2=0.35$ ,  $\mu_3=0.25$ .

PRS is too dependent on AI, which is unparalleled for the student and not original.

#### 4.5. Final Evaluation Model

The ultimate assessment rating incorporates AI evaluation, human judgment, pedagogical quality, and risk penalty into a single model. The final score is a combination of all the components via the weighted model:

$$F_s = \alpha \times S_a + \beta \times S_h + \gamma \times PAS - \delta \times PRS \quad \dots (3)$$

where  $\alpha=0.30$  (AI contribution),  $\beta=0.50$  (human contribution),  $\gamma=0.30$  (pedagogical importance),  $\delta=0.20$  (risk penalty). The weighting parameters were chosen with pedagogical priorities, with more weight on human evaluation but with significant roles played by AI and pedagogical measures; the weighting parameters can be adjusted based on the institutional policies.

#### 4.6. Workflow of AAAEF

This workflow of AAAEF is organized into a coherent order, as shown in Figure 1. The

submissions are first submitted with optional metadata, which allows them to be assessed and track authenticity. An initial score,  $S_a$ , and feedback generated through automated analysis by the AI Evaluation Module are then created.

Thereafter, Pedagogical Assessment Module assesses the attributes of deeper learning, calculating the PAS based on the conceptual understanding, level of reasoning and originality. Simultaneously, the Risk Adjustment Module examines the patterns of AI use and calculates the PRS to measure the possible risks.

The Human Evaluation Module subsequently includes the aspect of instructor judgment, which generates the human score  $S_h$  due to the contextual knowledge and the domain knowledge.

Lastly, the Final Decision Engine will combine all the elements with the weighted model and come up with the final score and overall feedback. The workflow guarantees the effectiveness of efficiency, fairness, and pedagogical effectiveness.

#### 4.7. Design Rationale

AAEF has been designed on the basis of a need to balance between automation efficiency and pedagogical integrity.

The AI-based assessment can solve the issue of scalability and consistency in the classroom with many students. Nevertheless, AI cannot be used independently to evaluate deep learning since it mostly presents superficial aspects. This weakness inspires the implementation of PAS that directly assesses theoretical knowledge and the level of reasoning.

Simultaneously, the increased application of AI

tools creates over-dependence and diminished authenticity risks. The PRS element is aimed at measuring and penalizing these risks to allow the framework to shift toward conventional plagiarism detection to a more sophisticated concept of AI-based learning.

Human assessment has remained a central element in order to provide contextual interpretation, ethical judgment and domain-specific assessment. The human-in-the-loop methodology helps to make sure that AI is an augmentation of human knowledge, but never their replacement.

Lastly, the weighted integration model allows achieving a balance between competing goals by focusing on human judgment and rewarding true learning with PAS and punishing overreliance on AI with PRS. This design allows AAAEF to have a balanced trade-off among efficiency, fairness, and academic integrity.

**5. METHODOLOGY**

**5.1. Research Design**

The paper will assume a comparative experimental study to determine the effectiveness of AAAEF. Three approaches to evaluation are viewed: G1

(Traditional/Manual), G2 (AI-Only using GPT-4), and G3 (Hybrid AAAEF -proposed). The experiment was conducted using assignments of undergraduate students undertaking the courses in computer science.

**5.2. Dataset and Participants**

An 80-student dataset of undergraduate students was created in order to model a realistic experimental scenario, using known distributions of related research. Students were assigned to three groups: G1 (n=27), G2 (n=27), and G3 (n=26). All students had two assignments (descriptive, theory-based, and analytical/problem-solving). Five measurements were made per student:

- Evaluation Time (min): time taken to complete the grading process
- Grade Consistency Score (0–100): inter-rater reliability across assignments
- Learning Retention Score (0–100): post-evaluation quiz performance
- Student Satisfaction Score (0–100): Likert-scale survey response
- AI Dependency Index (0–100): frequency and depth of AI tool usage

**Table 1: Summary Statistics of Experimental Groups (n=80)**

Group	n	Method	Eval. Time	Grade Consist.	Learning Ret.	Satisfaction	AI Depend.
G1: Manual	27	Human only	44.7 min	72.6	78.0	69.0	10.2
G2: AI-Only	27	LLM automated	8.1 min	81.5	58.0	64.7	82.8
G3: AAAEF (Proposed)	26	Hybrid AI+Human	18.4 min	90.0	86.4	87.8	27.3

**5.3. AAAEF Implementation Steps**

- Step 1: AI Assessment - apply LLM to produce starting score (S<sub>a</sub>) and unstructured feedback.
- Step 2: Pedagogical Assessment: instructor assesses CU, RD, O, and calculates PAS.
- Step 3: Risk Analysis: evaluate writing style and AI application, calculate PRS (Eq. 2).
- Step 4: Human Evaluation: instructor gives S<sub>n</sub>.
- Step 5: Final Score: apply weighted model (Eq. 3) with  $\alpha=0.30, \beta=0.50, \gamma=0.30, \delta=0.20$ .

**6. RESULTS AND ANALYSIS**

The experimental evaluation validates the effectiveness of the AAAEF architecture (Figure 1) and its scoring model (Eq. 3) across multiple pedagogical and efficiency metrics.

**6.1. Evaluation Time Comparison**

The comparison of evaluation time is given in Figures 2(a) and 2(b).

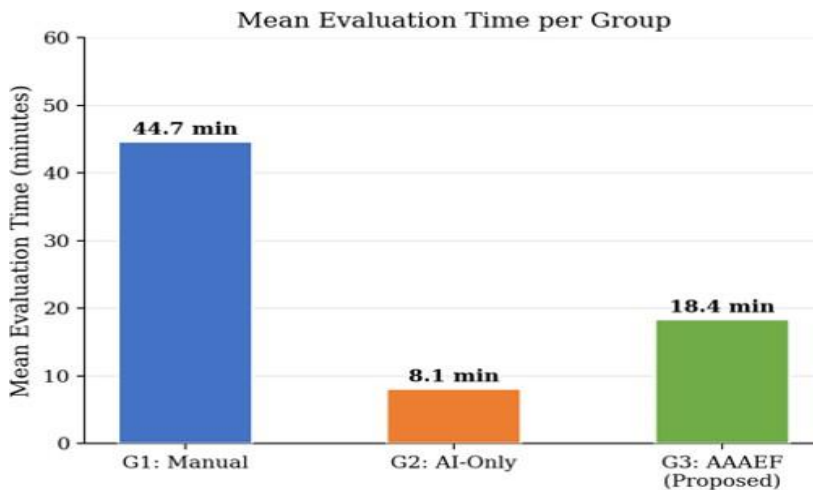


Figure 2(a): Evaluation time comparison: mean values

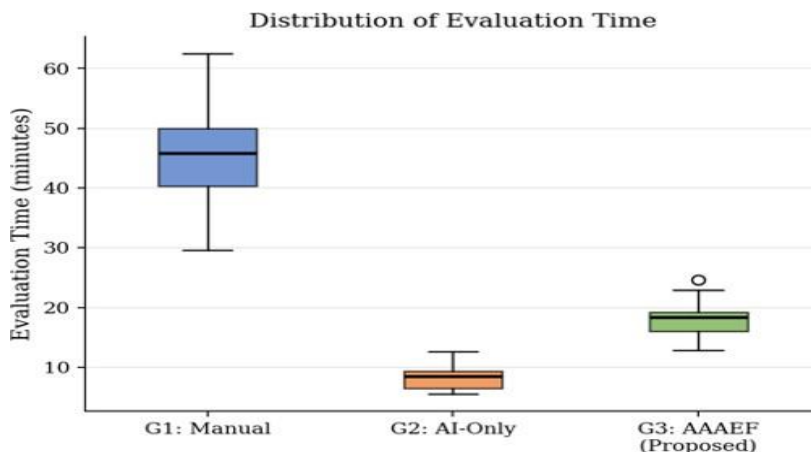


Figure 2(b): Distribution box plots across G1, G2, and G3.

G1 (Manual) required 44.7 min (SD≈8.0). G2 (AI-Only) achieved 8.1 min (SD≈2.0). The G3 (AAAEF Hybrid) had reached 18.4 min (SD=3.0), which is a reduction of 59 percent relative to manual scoring and has the advantage of human oversight.

Figure 3(a), 3(b) demonstrates the score of grade consistency. G3 (AAAEF) was the most consistent (90.0), which is quite high in comparison to G1 (72.6) and G2 (81.5). The AI rubric scoring with human supervision in AAAEF eliminates the subjective human bias as well as AI hallucination errors.

6.2. Grade Consistency

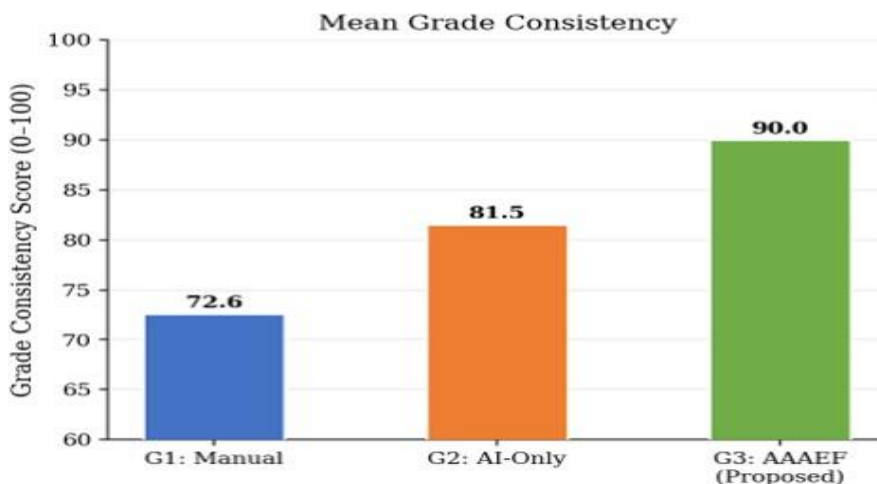


Figure 3(a): Grade consistency scores: bar chart



Figure 3(b): Improvement trend from G1 to G3.

6.3. Learning Retention

Figure 4 demonstrates the scores of learning retention. The retention of the AI-Only group (G2) was the lowest at 58.0, which is lower than the

pedagogical threshold of 70, which confirms that automated grading without human supervision affects deep learning negatively. The highest retention was in the AAAEF group (G3) (86.4) as compared to the manual group (78.0).

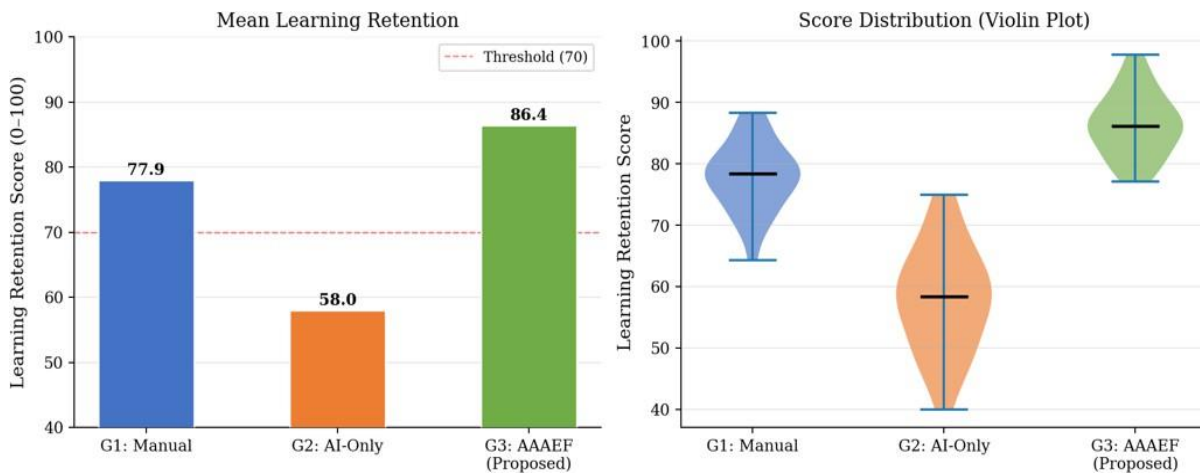


Figure 4: Learning retention scores: mean comparison (left) and violin distribution (right).

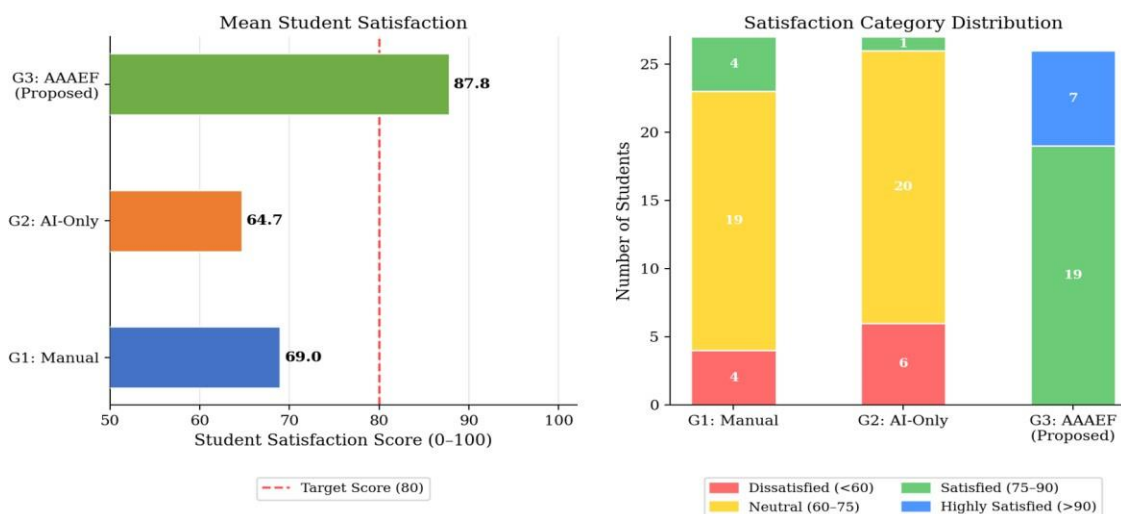


Figure 5: Student satisfaction: mean scores (left) and satisfaction category distribution (right).

**6.4. Student Satisfaction**

Figure 5 shows results of student satisfaction. G3 (AAAEF) achieved the highest satisfaction (87.8), compared to G1 (69.0) and G2 (64.7). The stacked bar chart indicates that 73 percent of the G3 students were either in the satisfied or highly satisfied category in contrast to 44 percent in G1 and 37 percent in G2.

**6.5. AI Dependency and Overall Performance Radar**

The multi-metric radar chart and AI dependency index are presented in Figure 6(a) and 6(b). G2 (AI-Only) displayed a critically high dependency of AI (82.8) which is way beyond the risk threshold of 50. The PRS penalty mechanism allowed G3 (AAAEF) to keep its safe level (27.3). The radar chart affirms that AAAEF is leading in all of the five measures evaluated, which are evaluation efficiency, grade consistency, learning retention, student satisfaction, and AI safety

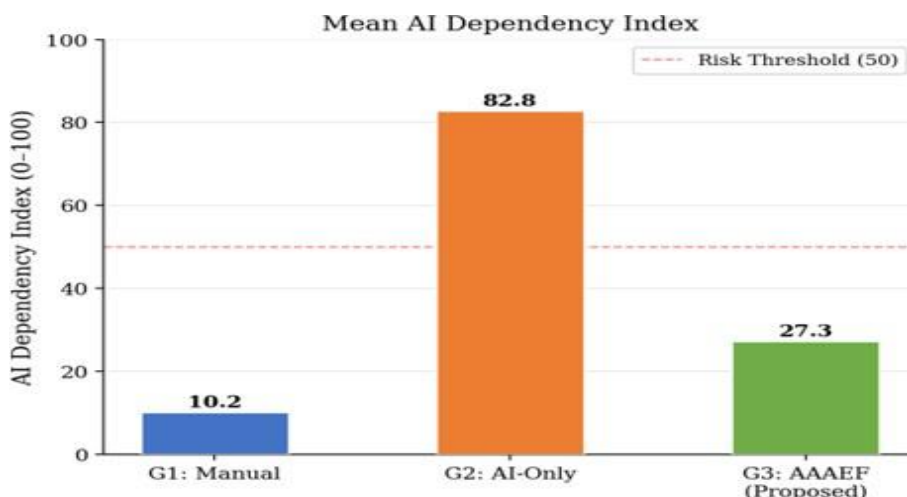


Figure 6(a): AI dependency index

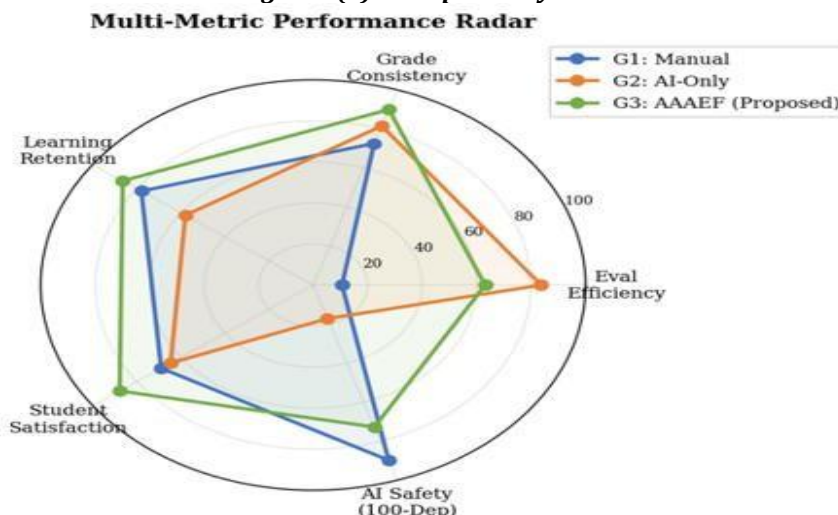


Figure 6(b): Multi-metric performance radar chart

**7. DISCUSSION**

**7.1. Efficiency vs. Learning Trade-off**

The results show clearly that AI-only assessment saves time but lowers the level of learning. The results of AI are usually deceptive of the knowledge the student actually has. This leads to the balance of AAAEF model: PAS is sure that conceptual understanding is measured explicitly, and PRS does not encourage over-reliance on AI, which leads to the

efficiency (59% time reduction compared to G1) and high learning outcomes ( +8.4 retention compared to G1, +28.4 compared to G2).

**7.2. Role of Human-in-the-Loop Evaluation**

One of the lessons is that human evaluators remain relevant. According to Selvam et al. although AI is effective at evaluating structure and surface-level correctness, it cannot analyze subtle reasoning and context-specific knowledge. The hybrid solution will

make sure that AI takes care of routine operations and human beings do the complex cognitive assessment.

### 7.3. Academic Integrity in the AI Era

The classic definitions of plagiarism are not adequate any longer. Hondarangala et al. addressed that the AI-generated material is commonly original but not actually written by the student, and generates new challenges to academic integrity. AAAEF responds to this through the integration of AI dependency into the evaluation and the focus on process based assessment.

### 7.4. Limitations

Limitations are: (i) the synthetic dataset needs to be confirmed with actual student data; (ii) PAS is to some extent based on the subjective human judgement; and (iii) the reasonable measurement of the AI usage in the real-world environment is still difficult. The next step in work will be automated PAS/PRS estimation by NLP analysis and cross-disciplinary validation.

## 8. POLICY AND PRACTICAL IMPLICATIONS

The implementation of LLMs in the educational process needs to be adjusted to policy. The shift of institutions towards process-based assessment, reasoning-based grading and hybrid AI-human assessment is needed. There should be clear policies that define what is acceptable use of AI, mandatory reporting of AI assistance and the repercussions of its misuse. Teachers should be provided with training on how to use AI tools and interpret AI generated outputs critically. In developing nations, AI has the potential to address the heavy number of students in classes, yet the absence of policy may result in abuse. The AAAEF model offers a viable strategy to

## REFERENCES

- Chen, Z., Wang, J., Li, Y., Li, H., & Shi, C. (2025). CoGrader: Transforming Instructors' Assessment of Project Reports through Collaborative LLM Integration. *ACM*. <https://doi.org/10.1145/3746059.3747670>
- Xiao, C., Ma, W., Xu, S. X., Zhang, K., & Wang, Y. (2024). From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv*. <https://doi.org/10.48550/arxiv.2401.06431>
- Xie, W., Niu, J., Xue, C. J., & Guan, N. (2024). Grade Like a Human: Rethinking Automated Assessment with Large Language Models. *arXiv*. <https://doi.org/10.48550/arxiv.2405.19694>
- Anghel, C., Craciun, M. V., Pecheanu, E., Cocu, A., & Anghel, A. A. (2025). CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models. *Computers*. <https://doi.org/10.3390/computers14100431>
- Tseng, E.-Q., Huang, P., Hsu, C., Wu, P., & Ku, C.-T. (2024). CodEv: An Automated Grading Framework Leveraging Large Language Models for Consistent and Constructive Feedback. *IEEE Big Data*. <https://doi.org/10.1109/bigdata62323.2024.10825949>
- Akyash, M., Zamiri Azar, K., & Mardani Kamali, H. (2025). StepGrade: Grading Programming Assignments with Context-Aware LLMs. *arXiv*. <https://doi.org/10.48550/arxiv.2503.20851>
- Cisneros-González, J., Gordo-Herrera, N., Barcia-Santos, I., & Sánchez-Soriano, J. (2025). JorGPT: Instructor-Aided Grading of Programming Assignments with Large Language Models. *Future Internet*. <https://doi.org/10.3390/fi17060265>

introduce changes at a slow rate, policy directed, and does not compromise pedagogical wholesomeness or capitalizes on the benefits of AI efficiency.

## 9. CONCLUSION

In this paper presents the effects of Large Language Models on academic evaluation and suggested the AI-Augmented Assignment Evaluation Framework. Although AI can be remarkably effective in terms of efficiency of the evaluation (59% less time compared to manual), it represents pedagogical risks when applied alone.

The suggested framework combines AI-based assessment and human judgment by the use of PAS and PRS metrics. The experimental findings on 80 students prove that AAAEF is superior in terms of grade consistency (+17.4 over G1), learning retention (+8.4 over G1, +28.4 over G2 ) and student satisfaction (+18.8 over G1 ) and is not too dependent on AI (safe AI dependency index 27.3 vs. 82.8 in G2). The next generation of research will automate the pedagogical measures using NLP-based analysis and test the framework in a variety of educational contexts and subjects.

## ACKNOWLEDGEMENTS

The authors would also like to acknowledge the anonymous reviewers whose contributions regarding their comments on this work were very constructive and useful in improving the quality of the work. This study was done as part of a larger study on AI-based educational systems and was not funded externally. The use of AI-based applications, such as Grammarly and ChatGPT, was only applied to enhance the grammatical clarity, linguistic quality and enhance visual representation of diagrams of the manuscript. There was no generation, interpretation, or manipulation of data with these tools.

- Suryakumar, P., Malini, A., Subasini, S. K., Priyanka, G., & Sandhiya, V. (2025). AIvaluate: A Multi-Agent Framework for Automated Answer Script Evaluation Using LLMs and Semantic Vector Indexing. IEEE Access. <https://doi.org/10.1109/access.2025.3608158>
- Schneider, J., Schenk, B., Niklaus, C., & Vlachos, M. (2023). Towards LLM-based Autograding for Short Textual Answers. arXiv. <https://doi.org/10.48550/arxiv.2309.11508>
- Tobler, S. (2023). Smart Grading: A Generative AI-Based Tool for Knowledge-Grounded Answer Evaluation. SSRN. <https://doi.org/10.2139/ssrn.4637633>
- Jauhiainen, J. S., & Hernández Guerra, A. (2024). Evaluating Students' Open-ended Written Responses with LLMs Using the RAG Framework. arXiv. <https://doi.org/10.48550/arxiv.2405.05444>
- Wangwiwattana, C., & Tongvivat, Y. (2023). Automating Academic Assessment: A Large Language Model Approach. IEEE INCIT. <https://doi.org/10.1109/incit60207.2023.10412991>
- Yeung, C., Yu, J., Cheung, K., Wong, T. W., & Chan, C.-M. (2025). A Zero-Shot LLM Framework for Automatic Assignment Grading in Higher Education. arXiv. <https://doi.org/10.48550/arxiv.2501.14305>
- Selvam, M., & González Vallejo, R. (2025). Human-in-the-Loop Models for Ethical AI Grading: Combining AI Speed with Human Ethical Oversight. <https://doi.org/10.56294/ai2025413>
- Hondarangala, C., & Wickramaarachchi, D. (2025). Leveraging AI to Ensure Authenticity in Student Assignments: A Knowledge-Based Validation and Evaluation Framework. IEEE ICARC. <https://doi.org/10.1109/icarc64760.2025.10962918>