Artificial Intelligence Applications And Innovations Vol. 2 No. 2 (2025)

Article history:
Received 8 February 2026
Revised 5 May 2026
Accepted 7 September 2026
Published online 1 April 2025

PeerReli-AI: AI-Enhanced Guidance for Boosting Peer Feedback Reliability

Seyede Fatemeh. Noorani[1*], Hossein. Morovvati[1], Hassan. Morovvati[1], Amir Houshang. TajFar[1]

[1] Computer Engineering and Information Technology, Payame Noor University, Tehran, Iran

*** Corresponding author email address**: sf.noorani@pnu.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Noorani, S. F., Morovvati, H., Morovvati, H., & TajFar, A. H. (2025). PeerReli-AI: AI-Enhanced Guidance for Boosting Peer Feedback Reliability.*Artificial Intelligence Applications and Innovations, 2*(2), 12-26.
https://doi.org/10.61838/jaiai.2.2.2

A B S T R A C T

Advances in artificial intelligence (AI) offer new opportunities to deliver scalable and personalized formative feedback in higher education, yet robust empirical evidence of their effectiveness remains limited. Although peer feedback can promote active learning, reflective thinking, and evaluative judgment, inconsistencies and inaccuracies in student assessments often undermine its impact. This study proposes PeerReli-AI, an AI-supported peer feedback framework that integrates automated quiz administration, peer evaluation, and AI-generated formative guidance. A messaging-bot infrastructure was used to deliver quizzes and manage anonymous peer feedback, while the Gemini large language model generated context-aware feedback on students' peer evaluations. The framework was implemented in an undergraduate Data Structures and Algorithms course across multiple iterative assessment cycles. Results from repeated-measures ANOVA and linear mixed-effects models demonstrate a significant reduction in discrepancies between peer and instructor evaluations over successive sessions, with a significant linear trend indicating progressive improvement in scoring accuracy. Analyses using both median and mean imputation for missing data yielded consistent findings, highlighting the robustness of the effect. Survey responses further indicate that students perceived the AI-generated feedback as engaging, useful, and supportive of their learning. Overall, the findings suggest that PeerReli-AI enhances the reliability and quality of peer feedback and provides a scalable approach for improving formative assessment practices in large higher education classrooms.

Keywords: *AI-Enhanced Guidance, Peer Feedback, Reliability, Formative feedback*

1. Introduction

The rapid advancement of artificial intelligence (AI) has profoundly transformed educational practices, reshaping how learning, assessment, and feedback are designed and delivered. In higher education, AI-driven systems are increasingly employed to support personalized learning, automate assessment processes, and provide timely formative feedback at scale. These technologies offer new opportunities to enhance instructional quality while addressing long-standing challenges related to scalability, consistency, and individual learner support [1, 2]. As a result, AI has emerged as a key enabler of data-informed and learner-centered educational environments.

One critical area in which AI holds considerable promise is formative assessment and feedback. Feedback plays a central role in learning by helping students identify

gaps in understanding, regulate their learning strategies, and develop higher-order thinking skills. Among various feedback approaches, peer feedback has gained increasing attention as a pedagogically powerful strategy that actively engages students in evaluating and commenting on their peers' work. By positioning students as assessors rather than passive recipients, peer feedback can foster reflective thinking, metacognitive awareness, and critical judgment skills [3, 4].

Peer feedback is commonly situated within the broader concept of peer assessment, which encompasses practices such as peer evaluation and peer grading. It is typically defined as a structured process in which students assess the quality of their peers' work using predefined criteria or rubrics [5, 6]. A substantial body of empirical research and meta-analyses has demonstrated that, when appropriately designed, peer assessment can positively influence learning outcomes, conceptual understanding, and academic performance [5]. Factors such as clarity of criteria, alignment with learning objectives, and opportunities for explanatory feedback are critical to its effectiveness [6].

Despite its pedagogical benefits, peer feedback faces persistent challenges that limit its reliability and wider adoption in formal assessment contexts. Prior studies have documented considerable variability in the accuracy and quality of peer evaluations, often arising from differences in students' domain knowledge, confidence, motivation, and evaluative skills [3]. Additionally, systematic biases such as leniency, severity, and social influence can distort peer judgments, reducing their alignment with expert or instructor assessments. Consequently, improving the consistency and reliability of peer feedback remains a central concern in educational research and practice.

Recent research has increasingly emphasized the role of technological and AI-based support in addressing these challenges. Advances in AI and natural language processing have enabled the development of systems capable of delivering scalable, timely, and context-aware feedback to learners. Within peer assessment settings, AI-generated feedback can function as a formative scaffold, guiding students toward more accurate evaluations and more constructive feedback explanations [7, 8]. Frameworks such as AI-supported peer assessment environments suggest that AI guidance can enhance evaluative judgment by modeling expert reasoning and reinforcing assessment criteria.

From a theoretical perspective, the potential of AI to improve the reliability of peer assessment can be explained through mechanisms related to evaluative judgement, cognitive scaffolding, and metacognitive regulation. Evaluative judgement refers to students' capability to make informed decisions about the quality of work based on explicit criteria and standards [9]. When students receive structured, criterion-aligned AI-generated feedback after completing peer evaluations, they are repeatedly exposed to expert-like reasoning and calibrated standards. This iterative exposure can function as a form of cognitive scaffolding, gradually aligning students' internal judgment processes with disciplinary expectations.

Furthermore, engagement with structured feedback has been shown to strengthen assessment literacy and reduce common biases such as leniency or severity effects [10]. By providing consistent, criteria-based guidance across multiple sessions, AI-generated feedback may support metacognitive monitoring, enabling students to reflect on discrepancies between their evaluations and benchmark standards. Over time, this process can reduce variability in peer ratings and enhance alignment with instructor assessments, thereby improving reliability.

However, despite growing interest in AI-assisted peer feedback, existing studies remain largely exploratory, short-term, or perception-based. Systematic reviews highlight a lack of robust empirical evidence examining how AI-generated guidance influences the development and reliability of peer assessment judgments over time, particularly within iterative instructional cycles and authentic course settings [1, 2]. Moreover, few studies explicitly account for task difficulty and individual differences when evaluating the impact of AI-supported peer feedback.

To address these gaps, the present study investigates the effects of AI-generated formative feedback on the accuracy and reliability of peer feedback in an undergraduate data structures and algorithms course. By embedding AI guidance within a repeated peer feedback cycle and systematically comparing peer evaluations with instructor benchmarks across multiple sessions, this study aims to examine whether sustained exposure to AI feedback can improve students' evaluative accuracy while controlling for task difficulty and individual variability. In doing so, the study contributes empirical evidence to the emerging field of AI-enhanced peer assessment and offers practical

insights for integrating AI-supported feedback into higher education.

2. Literature Review

2.1. *Artificial Intelligence in Education*

Recent research indicates that artificial intelligence (AI) technologies are increasingly integrated into educational practices to support learning .. Reviews highlight that AI has transformed traditional teaching by enabling personalized learning, real-time feedback, and intelligent tutoring systems. These capabilities suggest that AI could similarly enhance peer feedback processes, improving reliability and engagement in student-generated evaluations [11].

Prior reviews also indicate that most AI applications have focused on higher education, particularly on assessment automation, learning analytics, and performance prediction, whereas pedagogically grounded feedback practices have received comparatively less attention. While early reviews emphasize AI's potential to support feedback, they reveal a predominant focus on automation and prediction rather than enhancing the quality and reliability of peer feedback, suggesting a critical gap in existing research.

Generative AI tools, in particular, offer scalable, personalized, and timely feedback that can enhance learner engagement and performance, especially in large or resource-constrained settings. However, concerns remain regarding the accuracy and pedagogical appropriateness of fully automated feedback, indicating that hybrid approaches combining AI and human guidance may be more effective in promoting meaningful learning experiences.

2.2. *The Role of Peer Feedback in Learning*

Peer feedback, in which learners provide comments and suggestions to each other, has long been recognized as a valuable instructional strategy in higher education. It fosters critical thinking, reflection, collaborative learning, and helps students internalize quality standards for academic work. Nevertheless, learners often struggle to provide consistent and constructive feedback due to limited feedback literacy and interpersonal dynamics, which can undermine the effectiveness of peer feedback practices [8, 12].

To improve the accuracy and quality of peer feedback, multiple strategies have been explored. For example, rubrics provide explicit evaluation criteria and performance descriptors to guide students' judgments, improving fairness and alignment with instructor scores [13-15]. Another method is using iterative review processes that allows students to revise their work based on successive evaluations, enhancing learning and internalizing assessment criteria [16]. Developing students' ability to understand, interpret, and act on feedback is another approach that improves the reliability and quality of peer evaluations [16-18].

2.3. *AI-Supported Peer Feedback*

To address challenges in peer feedback, recent studies have investigated how AI can support and enhance peer evaluation. AI can evaluate the quality, consistency, and structure of peer feedback, improving trust, consistency, and scalability of feedback systems [19, 20] [21] [22]. Explainable AI and deep learning models can transparently assess feedback quality, increasing reliability [23, 24]. Also, AI-generated scores or feedback can act as a benchmark for human peer evaluations, enhancing fairness, accuracy, and perceived validity [25] [26] [27] [7]. AI also used to motivate students, increase feedback quantity and quality, and improve learners' feedback literacy [28] [29] [30].

Table 2 provides an overview of AI-based enhancements.

Table 1. Overview of AI-Based Peer-Assessment Improvement Studies

Study	Year	AI Technique	Analyzer / Comparator	Trust / Calibration	Iterative / Multi-session	Rubric-aligned / Personalized Guidance	Telegram bot / Real-world deployment
[22]	2025	NLP framework	✓				
[31]	2024	Generative AI + NLP	✓				
[21]	2024	NLP-based feedback scaffolding	✓			✓	
[23]	2024	ML + Explainable AI	✓	✓			

[25]	2024	Generative AI		✓					
[26]	2024	AI speech scoring		✓	✓				
[24]	2024	Deep learning models	✓		✓				
[30]	2024	AI-assisted forms			✓				
[19]	2025	AI + Learning Analytics	✓		✓				
[27]	2025	Calibrated genAI	✓	✓	✓	✓			
[28]	2025	genAI + gamification			✓				
[20]	2025	AI + Learning Analytics	✓		✓				
[29]	2024	AI-supported peer review	✓		✓	✓			
[11]	2023	AI-supported peer review					✓		
[7]	2025	Generative AI vs instructor vs peer		✓					
PeerReli-AI (this study)	2026	GeminAi+Telegram bot	✓	✓	✓	✓	✓	✓	

PeerReli-AI not only employs a within-subject, multi-session experimental design and AI-guided feedback cycle, but also leverages GeminAi integrated via a Telegram bot for quiz delivery, peer feedback collection, and AI-generated formative guidance. This implementation enables scalable, real-world deployment, allowing students to interact with AI seamlessly while receiving personalized, rubric-aligned feedback that improves reliability and constructiveness across six consecutive sessions.

3. Method

3.1. Study Design

This study employed a within-subject experimental design to examine the effect of AI-generated feedback on the reliability of peer feedback. The instructional and assessment process followed the PeerReli-AI workflow, consisting of instruction, student quiz completion, peer feedback, instructor assessment, and AI-generated formative guidance, as illustrated in Figure 1. This cycle was repeated across six consecutive sessions, enabling systematic examination of changes in peer feedback reliability over time while accounting for differences in question difficulty.

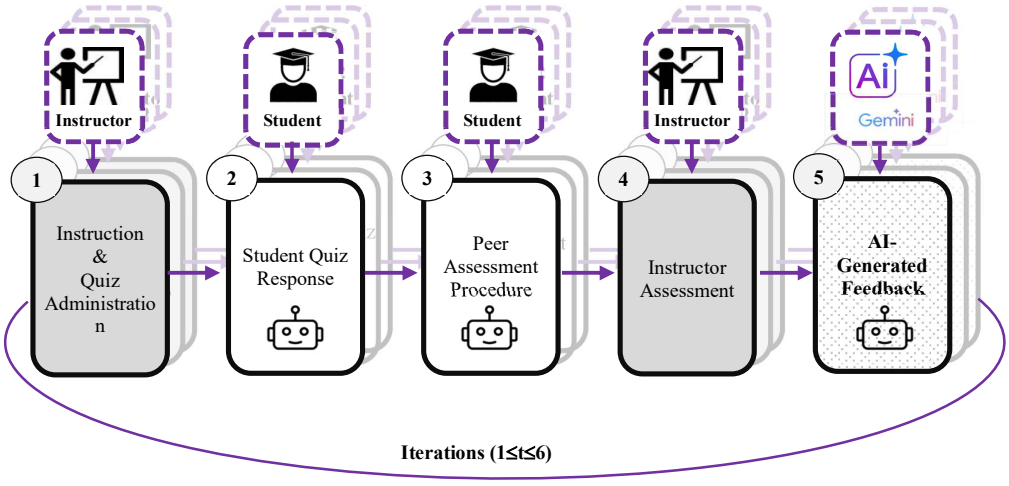

Figure 1. Workflow of the PeerReli-AI integrating AI-generated formative feedback

Step 1: Instruction and Quiz Administration

At the end of each instructional session, quizzes were scored on a 0–1 scale and administered via a Telegram bot immediately after instruction.

Each quiz was designed to cover the key concepts taught in the session. Table 2 summarizes the instructional content of the six sessions.

Step 2: Student Quiz Response

Quizzes were administered individually to students via a Telegram bot, which delivered the questions and collected responses along with timestamps and student identifiers. This setup ensured data integrity and provided the basis for peer feedback. Each student received a complete anonymized set of peer responses for evaluation. Figure 2 shows an example of the quiz interface as presented to students.

Table 2. Overview of Session Instructional Content

Session	Topic	Description
1	Time Complexity Analysis	Evaluation of the computational complexity of algorithms containing two nested loops with varying step increments.
2	Array Address Calculation	Determination of the memory address of a specific array element.
3	Stack Operations	Execution of insertion and deletion operations across multiple stacks.
4	Multiple Stack Operations	Systematic addition and removal of elements in a multi-stack environment.
5	Circular Queue	Performing enqueue and dequeue operations on a circular queue.
6	Linked List Traversal Function	Designing and implementing a function to traverse a linked list.

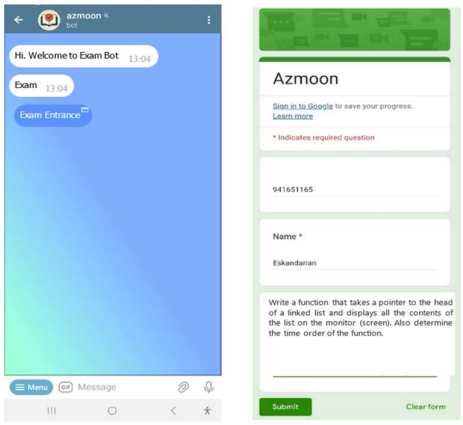

Figure 2. Example of a quiz page presented to students. The left image shows the entrance page and the right one show an example quiz page.

Step 3: Peer Feedback Procedure

Peer feedbacks were conducted using a custom script that randomly assigned each student's responses to another student for assessment. Full anonymity was maintained: evaluators did not know the identity of the quiz owner, and students did not know who evaluated their work.

Students were instructed to provide written justifications for each response rather than simply assigning scores, encouraging reflective judgment, deeper engagement, and high-quality peer feedback. Figure 3 presents an example of the peer evaluation form, including rating criteria and space for qualitative feedback. Evaluation links were automatically

distributed via email, providing secure access to the assigned peer's quiz; the code snippet responsible for this email distribution is provided in Appendix A.

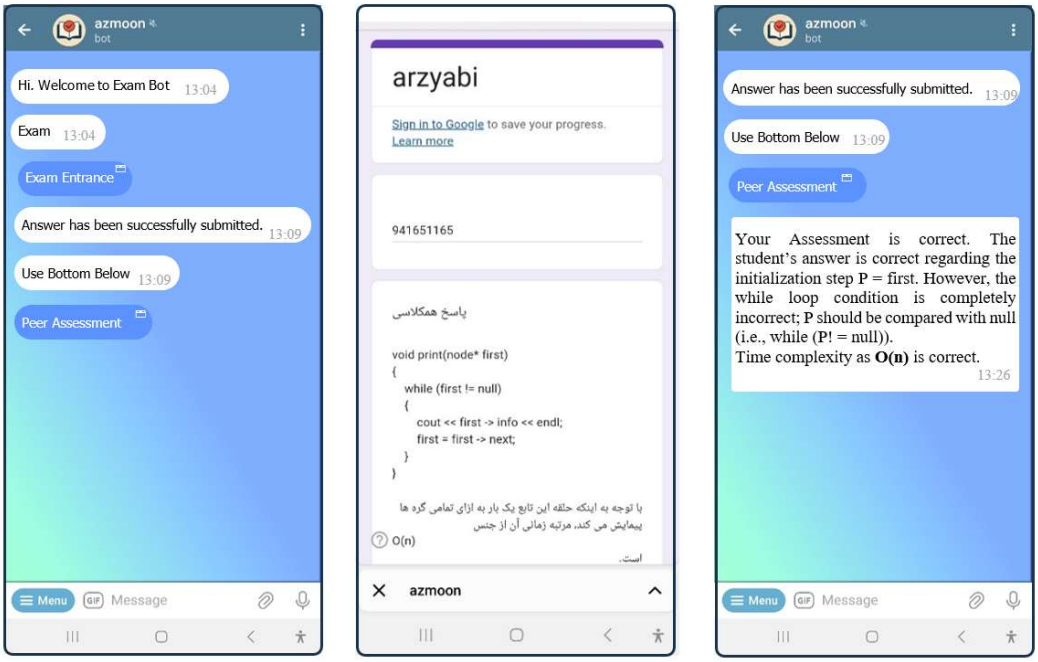

Figure 3. Example of the peer evaluation form used in the study. Example of the peer evaluation form used in the study. The left figure shows the button that directs the student to the assessment page. The middle figure shows the assessment form, and the right figure shows the AI feedback on the student's assessment.

Step 4: Instructor Assessment

The course instructor independently evaluated all quiz responses using the same scoring criteria applied to peer evaluations. These scores served as benchmark references for measuring the reliability of peer feedback at both the individual question and overall quiz levels.

Step 5: AI-Generated Feedback and Delivery

Completed peer evaluations were submitted to a large language model, Google Gemini (via the Google Gemini Chat API), to generate structured formative feedback. The model was configured within the n8n automation environment and accessed through the Google Gemini API integration.

Each AI prompt was explicitly structured to ensure replicability and consistency. The prompt defined the model's role as a university instructor in Data Structures and Algorithms and included the following elements:

The original quiz question

The instructor's reference solution

The student's submitted answer

The peer reviewer's assigned score

The peer reviewer's written justification

The model was instructed to evaluate whether the peer assessment was accurate and to provide concise, corrective formative feedback (maximum five lines), focusing on conceptual accuracy, scoring consistency, and identification of potential misunderstandings. For domain-specific computational tasks (e.g., array address calculation or Eulerian graph problems), structured computational cues were embedded within the prompt to enhance reasoning precision.

To ensure response consistency, deterministic generation settings were used (temperature set to a low value), and identical prompt templates were applied across all sessions. A complete example of the prompt template is provided in **Appendix B**, and screenshots of the workflow configuration and automation scripts are included in **Appendix A** for full technical transparency.

AI-generated feedback was automatically delivered to students via a Telegram bot immediately after the peer review phase. To prevent delivery throttling and ensure stable automation performance, a brief pause was implemented after every ten consecutive messages. This structured feedback cycle was repeated across six sessions,

enabling students to iteratively refine their peer evaluation accuracy over time.

3.2. Participants

Participants were 53 undergraduate students enrolled in the Data Structures and Algorithms course within the Computer Engineering program at Payame Noor University during the first semester of the 2025–2026 academic year. The sample included 22 male and 31 female students.

For analysis, students who were absent for more than three of the six sessions were excluded. As a result, the final dataset consisted of 46 students, some of whom were absent in one or two sessions.

3.2.1. Ethics and Data Privacy

All participants provided informed consent prior to taking part in the study. The study protocol was reviewed and approved by the Institutional Review Board (IRB) of Payame Noor University, ensuring adherence to ethical standards in research involving human subjects. Student responses, peer evaluations, and AI-generated feedback were anonymized, with all identifying information removed prior to analysis. Data were securely stored and accessed only by authorized researchers, in compliance with institutional guidelines on data privacy and protection.

3.3. Data Analysis

Figure 4 presents an example of the dataset collected during the study. For each student and quiz, the dataset recorded submission times, student identifiers, and email addresses (used to distribute peer feedback forms and deliver AI-generated guidance), students' answers, timestamps of peer evaluations, assigned evaluators, qualitative comments from peers, and the corresponding AI-generated feedback. This rich dataset enabled detailed analysis of the accuracy and reliability of peer feedback as well as the effects of AI-supported guidance on improving the quality of Evaluations.

Prior to analysis, the data underwent preprocessing. Students who were absent in more than three sessions were excluded, and the remaining data were transformed into a long format suitable for analysis in SPSS. In this format, each row corresponds to a student's evaluation in a specific session, with the following fields:

- Student: student identifier
- Session: session number (1–6)
- Diff: the difference between the student's evaluation and the instructor's evaluation
- Difficulty: the difficulty of the quiz in that session, scaled from 0 (easiest) to 1 (hardest)

Pre_AI_Feedback: a binary variable indicating whether AI feedback had been provided before that session (0 = no, 1 = yes). For Session 1, this value is 0 because the quiz and peer evaluations were completed before AI feedback was delivered; for all subsequent sessions, it is 1.

Timestamp	StudentID	StudentEmail	Answer	ReviewTime	ReviewerEmail	ReviewerComments	AIFeedBackForReviewer	InstructorScore	ReviewerScore
11/2/2025 13:10:09	7……	a……@gmail.com	P=first (While info P=null) Cout<<Info P--->next Time order=1	11/2/2025 13:17:11	azarin.m1384@gmail.com	He assigned First to p, and used a while condition; however, in the while statement, the condition should check that it is not equal to null. Time complexity: O(n) Score= 0.25	The student's answer is correct regarding the initialization step P = first. However, the while loop condition is completely incorrect; P should be compared with null (i.e., while (P != null)). Additionally, the method used to print the content and move to the next node (Cout << Info P--->next) is incorrect. The function should first print the data of the current node and then update P to P->next. The time complexity of the function is O(n), since to display all 'n' nodes, the algorithm must traverse each node exactly once. The student's evaluation in identifying the error in the while condition and stating the time complexity as O(n) is correct.	0.5	0.25

Figure 4. A screenshot of the collected data

3.4. Statistical Methods

To assess the accuracy and reliability of peer assessments, two complementary statistical approaches were employed. First, a repeated-measures ANOVA was conducted to examine whether the observed reduction in discrepancies between peer and instructor scores across the six sessions was statistically significant. This approach is particularly suitable for analyzing longitudinal within-subject data, as it accounts for the dependency of repeated observations and allows for the evaluation of polynomial trends (e.g., linear improvement) over time. Missing values due to occasional absences were imputed using both the median and the mean of each session, and analyses were conducted separately for each imputation strategy.

Second, to examine the effects of AI feedback and question difficulty on peer assessment accuracy, a linear mixed-effects model (LMM) was applied. LMMs provide a robust framework for analyzing nested or repeated measurements by including fixed effects (predictors of theoretical interest) and random effects (sources of individual variability), making them particularly suitable for educational datasets with multiple observations clustered within participants [32-34].

Additionally, the reliability of peer evaluations was quantified using Intraclass Correlation Coefficients (ICC, two-way mixed-effects, absolute agreement) and Cronbach's alpha, providing complementary measures of consistency and internal agreement across sessions. ICC values were computed for each session to evaluate agreement between peer and instructor scores at the individual question level, while Cronbach's alpha assessed the overall internal consistency of peer ratings across all quiz items. These reliability metrics offer robust evidence regarding the stability and dependability of peer assessments over time.

3.4.1. Subjective Evaluation

To complement objective measures of peer feedback reliability, a survey was conducted to capture students' perceptions and experiences with the PeerReli-AI framework, focusing on its AI-generated formative guidance. The survey included 8 statements rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) and was designed based on established technology acceptance models, including TAM (Technology Acceptance Model)[35] and UTAUT (Unified Theory of Acceptance and Use of Technology). The constructs assessed included overall user satisfaction, learner engagement, perceived learning support, perceived performance benefit, peer feedback quality, system usability, and intention to use the platform in future courses.

The survey items were:

1. Satisfaction with using PeerReli-AI for peer feedback (Overall User Satisfaction – TAM).
2. Engagement and motivation from reviewing peers' submissions (Learner Engagement – UTAUT).
3. Improvement in understanding course concepts through AI-supported feedback (Perceived Learning Support – TAM/UTAUT).
4. Enhancement of assignment performance through AI guidance (Perceived Performance Benefit – TAM/UTAUT).
5. Increased overall class participation from peer feedback activities (Active Class Participation – TAM/UTAUT).
6. Encouragement to provide more accurate and thoughtful feedback (Peer Feedback Quality – Extended TAM).
7. Ease of use and intuitiveness of the PeerReli-AI platform (System Usability – TAM/UTAUT).
8. Preference to use the AI-supported feedback process in future courses (Future Use Intention – TAM/UTAUT).

Cronbach's alpha indicated good reliability of the survey ($\alpha = 0.87$). Responses were summarized using frequency counts and visualized to illustrate trends in students' subjective evaluations, with detailed distributions presented in Section 4.2.

4. Results

4.1. Descriptive Statistics

Figure 5 visually illustrates the trends in peer assessment accuracy and quiz difficulty across the six sessions.

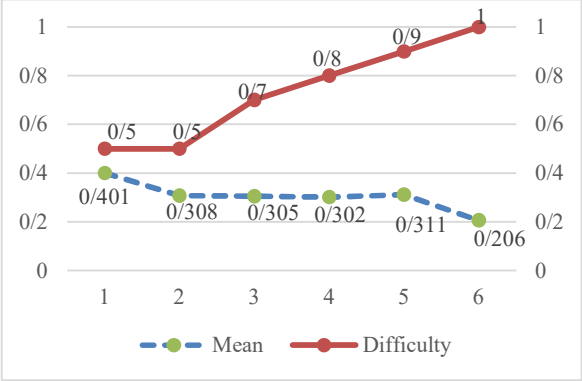

Figure 5. Dual-axis line chart illustrating mean absolute differences between peer and instructor evaluations and quiz difficulty across six sessions. The left vertical axis represents the mean absolute discrepancy, while the right vertical axis indicates quiz difficulty.

In Session 1, which served as the baseline without AI feedback, the largest mean absolute discrepancy between peer and instructor scores was observed, indicating greater variability in student evaluations. Following the introduction of PeerReli-AI in Sessions 2 through 6, the mean differences showed a consistent downward trend, despite a concurrent increase in question difficulty. Sessions 2 to 4 exhibited relatively stable discrepancies, while Session 5 showed a slight increase. Notably, Session 6 recorded the lowest discrepancy, suggesting a progressive improvement in alignment between student and instructor evaluations with sustained use of AI-supported feedback.

In Figure 5, the two vertical axes represent distinct but complementary measures. The left axis corresponds to the mean absolute difference between student and instructor evaluations, reflecting the degree of alignment in scoring across sessions. The right axis represents quiz difficulty, indicating the relative complexity of the assessment tasks administered in each session.

Overall, these descriptive results indicate that PeerReli-AI feedback is associated with enhanced alignment between peer and instructor evaluations, even under varying levels of task difficulty.

In addition to mean differences between peer and instructor scores, the reliability of peer evaluations was assessed using Intraclass Correlation Coefficients (ICC) and Cronbach's alpha. ICC values (two-way mixed-effects, absolute agreement) ranged from 0.68 to 0.82 across the six sessions, indicating moderate to excellent agreement between peer and instructor ratings. Cronbach's alpha for the overall set of peer evaluations was 0.85, reflecting high internal consistency of peer assessments across all quiz items. These metrics provide complementary evidence that the structured AI-supported feedback and repeated evaluation cycles contributed to stable and reliable peer scoring over time.

4.2. Inferential Analysis of Discrepancy Reduction Using Repeated-Measures ANOVA

To examine whether the observed reduction in mean discrepancies across sessions was statistically significant, a repeated-measures ANOVA was conducted. This approach accounts for the within-subject dependency of repeated observations and allows for evaluation of polynomial trends, such as linear improvement over time.

As described in the Methods section, students who were absent for more than three sessions were excluded from the analysis. For the remaining cases, missing values arising from occasional absences were imputed using both the median and the mean of each session, and separate analyses were conducted for each imputation strategy.

The ANOVA revealed a significant effect of session on scoring discrepancies under both imputation strategies.

In the median-imputed dataset, a repeated-measures ANOVA revealed a significant effect of session on scoring discrepancies, $F(5, 43) = 3.000$, $p = .012$ (Greenhouse-Geisser corrected: $F(4.374, 205.583) = 3.000$, $p = .016$), indicating a systematic reduction in peer assessment error across sessions (See Figure 6(a)). Polynomial contrasts further indicated a significant linear trend, $F(1, 47) = 11.005$, $p = .002$, suggesting that scoring accuracy improved progressively over time (See Figure 6(b)).

A parallel analysis using mean imputation for missing values yielded consistent results, $F(5, 43) = 2.981$, $p = .013$ (Greenhouse-Geisser corrected: $F(4.366, 205.204) = 2.981$, $p = .017$), with a significant linear trend, $F(1, 47) = 11.230$,

p = .002 (See Figure 6c,d). These results indicate that the reduction in scoring discrepancies is robust to the choice of imputation method.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Sessions	Sphericity Assumed	.922	5	.184	3.000	.012
	Greenhouse-Geisser	.922	4.374	.211	3.000	.016
	Huynh-Feldt	.922	4.879	.189	3.000	.013
	Lower-bound	.922	1.000	.922	3.000	.090
Error(Sessions)	Sphericity Assumed	14.450	235	.061		
	Greenhouse-Geisser	14.450	205.583	.070		
	Huynh-Feldt	14.450	229.297	.063		
	Lower-bound	14.450	47.000	.307		

1. (a)

Tests of Within-Subjects Contrasts

Source	Sessions	Type III Sum of Squares	df	Mean Square	F	Sig.
Sessions	Linear	.647	1	.647	11.005	.002
	Quadratic	.001	1	.001	.017	.896
	Cubic	.264	1	.264	3.609	.064
	Order 4	.001	1	.001	.008	.929
	Order 5	.010	1	.010	.230	.634
Error(Sessions)	Linear	2.763	47	.059		
	Quadratic	2.794	47	.059		
	Cubic	3.432	47	.073		
	Order 4	3.343	47	.071		
	Order 5	2.118	47	.045		

2. (b)

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Sessions	Sphericity Assumed	.914	5	.183	2.981	.013
	Greenhouse-Geisser	.914	4.366	.209	2.981	.017
	Huynh-Feldt	.914	4.869	.188	2.981	.013
	Lower-bound	.914	1.000	.914	2.981	.091
Error(Sessions)	Sphericity Assumed	14.406	235	.061		
	Greenhouse-Geisser	14.406	205.204	.070		
	Huynh-Feldt	14.406	228.827	.063		
	Lower-bound	14.406	47.000	.307		

3. (c)

Tests of Within-Subjects Contrasts

Source	Sessions	Type III Sum of Squares	df	Mean Square	F	Sig.
Sessions	Linear	.650	1	.650	11.230	.002
	Quadratic	4.586E-5	1	4.586E-5	.001	.978
	Cubic	.252	1	.252	3.410	.071
	Order 4	.003	1	.003	.042	.839
	Order 5	.009	1	.009	.208	.651
Error(Sessions)	Linear	2.720	47	.058		
	Quadratic	2.800	47	.060		
	Cubic	3.468	47	.074		
	Order 4	3.337	47	.071		
	Order 5	2.081	47	.044		

4. (d)

Figure 6. Session-wise peer assessment discrepancies across six sessions. Panels (a) and (b) show the median-imputed dataset: (a) mean discrepancies per session, (b) linear trend in discrepancy reduction. Panels (c) and (d) show the mean-imputed dataset: (c) mean discrepancies per session, (d) linear trend in discrepancy reduction. Across both imputation strategies, discrepancies decreased systematically, indicating progressive improvement in scoring accuracy over time.

4.3. Linear Mixed-Effects Analysis

A linear mixed-effects model was used to examine how AI-generated feedback and question difficulty influenced the accuracy of peer assessments. The outcome variable was the absolute discrepancy between peer and instructor scores. AI feedback and difficulty were specified as fixed effects, while individual differences in scoring behavior were captured through random intercepts for students. Repeated measurements across sessions were modeled to account for within-participant dependence.

Type III Tests of Fixed Effects[a]				
Source	Numerator df	Denominator df	F	Sig.
Intercept	1	44.206	164.100	.000
Difficulty	3	84.595	1.292	.282
Pre_AI_Feedback	0	.	.	.

a. Dependent Variable: Diff.

Figure 7. Fixed effects estimate from the Linear Mixed-Effects model predicting absolute differences between peer and instructor scores (Diff).

As is shown in Figure 7, the analysis revealed that question difficulty did not significantly predict peer–instructor discrepancies ($F = 1.29$, $p = 0.282$), indicating that increased task complexity did not adversely affect assessment accuracy.

Although a formal statistical test of the AI feedback effect was not feasible, descriptive evidence consistently showed reduced discrepancies following the introduction of AI-supported feedback.

4.4. Subjective Evaluation

Students' perceptions of the PeerReli-AI framework were assessed using an 8-item Likert-scale survey (1 = Strongly Disagree, 5 = Strongly Agree), as described in Section 3.4. Figure 8 presents the distribution of responses.

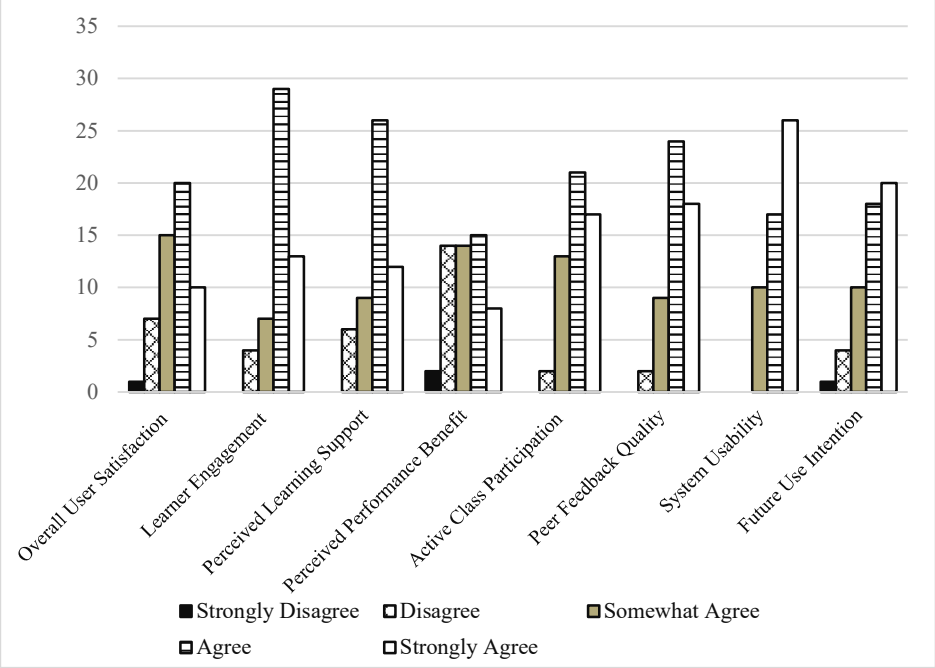

Figure 8. The Results of subjective evaluation

The majority of students reported positive perceptions of PeerReli-AI. Most participants selected Agree or Strongly Agree regarding engagement (Questions 2 and 5), perceived learning support (Question 3), performance improvement (Question 4), and platform usability (Questions 7 and 8). Responses to Question 6 further indicate that PeerReli-AI encouraged students to conduct more careful and thoughtful peer evaluations.

Only a small minority of students expressed disagreement, and very few selected Strongly Disagree, reflecting minimal negative perceptions of the system.

Overall, these results suggest that PeerReli-AI not only contributed to improved accuracy in peer assessments but also enhanced student engagement, satisfaction, and perceived learning support across the sessions.

5. Conclusion

This study introduced PeerReli-AI, a hybrid framework that integrates AI-generated formative feedback with incentive-compatible peer assessment mechanisms to support the accuracy and reliability of peer evaluations. Empirical results suggest that PeerReli-AI was associated with reduced discrepancies between peer and instructor scores across multiple sessions, improved the consistency of peer assessments, and maintained these trends even as question difficulty increased. Furthermore, students reported high engagement, perceived learning support, and satisfaction, indicating positive reception and usability of the platform.

These findings suggest that combining structured AI guidance and strategic incentives may enhance the quality and reliability of peer assessment in higher education. However, given the within-subject quasi-experimental design, causal inferences cannot be definitively established. Future research should examine the framework's generalizability in larger and more diverse classes, explore the integration of advanced AI models for personalized and adaptive feedback, and develop dynamic, performance-sensitive incentive mechanisms to maintain high-effort evaluations over time. Additionally, improvements in platform usability, including real-time dashboards, gamification, and analytics tools, may further enhance student engagement and the visibility of feedback for both students and instructors.

Authors' Contributions

All authors equally contributed to this study.

Declaration

None.

Transparency Statement

None.

Acknowledgments

None.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] R. Luckin, W. Holmes, M. Griffiths, and L. Forcier, "Artificial intelligence in education: Promises and implications for teaching and learning," *This book discusses various aspects of AI in education, including its role in assessment and exam proctoring,* 2022.

[2] N. Winstone and D. Carless, *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge, 2019.

[3] M. M. Ashenafi, "Peer-assessment in higher education–twenty-first century practices, challenges and the way forward," *Assessment & Evaluation in Higher Education,* vol. 42, no. 2, pp. 226-251, 2017, doi: 10.1080/02602938.2015.1100711.

[4] F. Berkmans, M. Bigerelle, J. Lemesle, L. Nys, M. Wieczorowski, and C. Brown, "Peer Assessment in Interdisciplinary Learning: Measuring Reliability and Engaging Critical Thinking," *Thinking Skills and Creativity,* p. 101950, 2025.

[5] K. S. Double, J. A. McGrane, and T. N. Hopfenbeck, "The impact of peer assessment on academic performance: A meta-analysis of control group studies," *Educational Psychology Review,* vol. 32, no. 2, pp. 481-509, 2020, doi: 10.1007/s10648-019-09510-3.

[6] K. J. Topping, "Peer assessment," *Theory into Practice,* vol. 48, no. 1, pp. 20-27, 2009, doi: 10.1080/00405840802577569.

[7] M. Usher, "Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education," *Assessment & Evaluation in Higher Education,* pp. 1-16, 2025.

[8] S. M. Evangelou and M. Xenos, "Peer Assessment in Education: A Seven-Year Implementation and Iterative Development of a Customizable Digital Tool," *Next Research,* vol. 2, no. 4, p. 101000, 2025, doi: 10.1016/j.nexres.2025.101000.

[9] D. Carless and D. Boud, "The development of student feedback literacy: Enabling uptake of feedback," *Assessment & evaluation in higher education,* vol. 43, no. 8, pp. 1315-1325, 2018.

[10] A. Lizzio and K. Wilson, "Feedback on assessment: Students' perceptions of quality and effectiveness," *Assessment & evaluation in higher education,* vol. 33, no. 3, pp. 263-275, 2008.

[11] A. V. Y. Lee, "Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation," *Studies in Educational Evaluation,* vol. 77, p. 101250, 2023.

[12] Y. D. K. Sinaga, E. Arliani, J. C. Ngala, and N. L. I. T. Agustina, "Accuracy of Self-Assessment and Peer Assessment in Learning: A Systematic Literature Review," *Jurnal Paedagogy,* vol. 11, no. 2, pp. 312-322, 2024, doi: http://doi.org/10.33394/jp.v11i2.9417.

[13] E. Panadero, A. Jonsson, L. Pinedo, and B. Fernández-Castilla, "Effects of Rubrics on Academic Performance, Self-Regulated Learning, and Self-Efficacy: a Meta-analytic Review," *Educational Psychology Review,* vol. 35, p. 113, 2023, doi: http://doi.org/10.1007/s10648-023-09823-4.

[14] M. Charalambides and R. Panaoura, "Rubric-Guided Peer Review in Higher Education," *[Journal Name],* 2025. [Online]. Available: https://example.org/rubric-guided-peer-review-2025.

[15] S. Pericleous, E. Tsolaki, M. Charalambides, and R. Panaoura, "The Impact of Rubric-Guided Peer Review and Self-Assessment in an Engineering Math University Course," *Education Sciences,* vol. 15, no. 11, p. 1433, 2025, doi: http://doi.org/10.3390/educsci15111433.

[16] W. Wu, J. Huang, C. Han, and J. Zhang, "Evaluating peer feedback as a reliable and valid complementary aid to teacher feedback in EFL writing classrooms: A feedback giver perspective," *Studies in Educational Evaluation,* vol. 73, p. 101140, 2022, doi: http://doi.org/10.1016/j.stueduc.2022.101140.

[17] N. T. Kerman, S. K. Banihashem, and M. Karami, *Online peer feedback in higher education: A systematic review of practices*. 2024.

[18] F. Y. Yu, Y. H. Liu, and K. Liu, "Online peer-assessment quality control: A proactive measure, validation study, and sensitivity analysis," *Studies in Educational Evaluation,* vol. 81, p. 101341, 2023, doi: http://doi.org/10.1016/j.stueduc.2023.101279.

[19] K. J. Topping, E. Gehringer, H. Khosravi, S. Gudipati, K. Jadhav, and S. Susarla, "Enhancing Peer Assessment with Artificial Intelligence," *International Journal of Educational Technology in Higher Education,* vol. 22, no. 1, p. 3, 2025, doi: https://doi.org/10.1186/s41239-024-00501-1.

[20] Y. Zhang, X. Chen, and M. Li, "Incorporating AI and learning analytics to build trustworthy peer assessment systems," *arXiv preprint,* 2025. [Online]. Available: http://arxiv.org/abs/2501.01234.

[21] K. Guo, M. Li, and X. Chen, "EvaluMate: Using AI to support students' feedback provision in peer assessment for writing," *Assessing Writing,* vol. 55, p. 100841, 2024, doi: https://doi.org/10.1016/j.asw.2024.100841.

[22] G. C. Zapata, B. Cope, and M. Kalantzis, "Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda," *arXiv preprint,* 2025. [Online]. Available: http://arxiv.org/abs/2509.15035.

[23] R. Wlodarski, L. da Silva Sousa, and A. Connell Pensky, "Automatic Analysis of Peer Feedback using Machine Learning and Explainable Artificial Intelligence," *arXiv preprint,* 2025. [Online]. Available: http://arxiv.org/abs/2504.02962.

[24] K. Guo, M. Li, and X. Chen, "Sustainable education on improving the quality of peer assessment: design and implementation of an online deep learning-based peer assessment system," *Assessing Writing,* 2024. [Online]. Available: http://doi.org/10.1016/j.asw.2024.100841.

[25] X. Li, "Feedback sources in essay writing: peer-generated or AI-generated feedback?," *arXiv preprint,* 2024. [Online]. Available: http://arxiv.org/abs/2408.05678.

[26] X. Li, "Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement," *2nd International Conference on Design Science,* 2024. [Online]. Available: http://example.org/ai-speaking-assessment-2024.

[27] G. C. Zapata, B. Cope, M. Kalantzis, and D. Searsmith, "Calibrated Generative AI as Meta-Reviewer: A Systemic Functional Linguistics Discourse Analysis of Reviews of Peer Reviews," *arXiv Preprint,* vol. abs/2509.15035, 2025. [Online]. Available: https://arxiv.org/abs/2509.15035.

[28] R. Wlodarski, L. da Silva Sousa, and A. Connell Pensky, "Level Up Peer Review in Education: Investigating genAI-driven Gamification System and Its Influence on Peer Feedback Effectiveness," *arXiv,* vol. 2504.02962, 2025. [Online]. Available: https://arxiv.org/abs/2504.02962.

[29] W. Wu, J. Huang, and C. Han, "Using AI-supported peer review to enhance feedback literacy: An investigation of students' revision of feedback on peers' essays," *Computers & Education,* 2024. [Online]. Available: http://doi.org/10.1016/j.compedu.2024.104676.

[30] X. Li, "The Influence of AI-assisted Peer Evaluation Forms on the Evaluation of Design Project-Based Students Cooperative Learning," *2nd International Conference on Design Science,* 2024. [Online]. Available: http://example.org/ai-evaluation-forms-2024.

[31] X. Chen, Y. Zhang, and M. Li, "Effects of an AI-supported approach to peer feedback on feedback quality and writing ability," *Computers & Education,* vol. 189, p. 104676, 2024, doi: https://doi.org/10.1016/j.compedu.2024.104676.

[32] T. F. Jaeger, "Random Effects: A Tutorial and Elementary Model," *Wiley Interdisciplinary Reviews: Cognitive Science,* 2018.

[33] H. Singmann and D. Kellen, "An introduction to mixed models for experimental psychology," in *New methods in cognitive psychology*: Routledge, 2019, pp. 4-31.

[34] V. A. Brown, "An introduction to linear mixed-effects modeling in R," *Advances in Methods and Practices in Psychological Science,* vol. 4, no. 1, p. 2515245920960351, 2021.

[35] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly,* pp. 319-340, 1989. [Online]. Available: https://www.jstor.org/stable/249008.

Appendix

Appendix A: Script for Randomly Emailing Quiz Responses to Students for Peer Evaluation

```
{
  "parameters": {
    "jsCode": "const students = items.map(item => ({\n  email: item.json['Column
    1'],\n  answer: item.json.Comments\n}));\n\n// رندوم‌سازی\nconst shuffled =
    students.sort(() => Math.random() - 0.5);\n\n// ساخت جفت‌ها بدون اینکه کسی
    پاسخ خودش را بگیرد\nconst assignments = [];\n\nfor (let i = 0; i < shuffled.
    length; i++) {\n  let reviewer = shuffled[i];\n  // پیدا کردن اولین reviewee
    که خودش نباشد\n  let revieweeIndex = (i + 1) % shuffled.length;\n
    assignments.push({\n    json: {\n      reviewer: reviewer.email,\n
    reviewee_answer: shuffled[revieweeIndex].answer\n    }\n  });\n}\n\nreturn
    assignments;\n"
  },
  "type": "n8n-nodes-base.code",
  "typeVersion": 2,
  "position": [
    384,
    -176
  ],
  "id": "83b31bac-1b81-4b87-bab7-1e99cf492a7e",
  "name": "Code in JavaScript",
  "executeOnce": false
},
```


Appendix B: Relevant JavaScript Code Segments for AI-Based Peer Assessment Feedback


```
{
  "parameters": {
    "options": {}
  },
  "type": "@n8n/n8n-nodes-langchain.lmChatGoogleGemini",
  "typeVersion": 1,
  "position": [
    720,
    240
  ],
  "id": "8ec957d8-3aa8-4519-9a74-62d8ed235cde",
  "name": "Google Gemini Chat Model",
  "credentials": {
    "googlePalmApi": {
      "id": "jOsCBquyNUAMJD6b",
      "name": "Google Gemini(PaLM) Api account"
    }
  }
},
```

```
"Google Gemini Chat Model": {
  "ai_languageModel": [
    [
      {
        "node": "AI Agent",
        "type": "ai_languageModel",
        "index": 0
      }
    ]
  ]
},
```

```
{
  "parameters": {
    "promptType": "define",
    "text": "=این پاسخ خود دانشجو است:{{ $json['پاسخ همکلاسی'] }}\nاین پاسخ همکلاسی خود دانشجو نسبت به ارزیابی است: {{ $json['ارزیابی شما'] }}",
    "options": {
      "systemMessage": "=درست ارزیابی کرده یا نه و تو براش در حد 5 خط به دانشجو توضیحات دقیق اون جواب رو بدی\nسلام شما یک استاد ساختمان داده هستی که"
    }
  },
  "type": "@n8n/n8n-nodes-langchain.agent",
  "typeVersion": 2.2,
  "position": [
    816,
    32
  ],
  "id": "463a4e05-2f63-4dbf-9fe2-b7e3bf18335f",
  "name": "AI Agent"
},
```