Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 15 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i15.61761

PAPER

A Mobile System for Formative Assessment and Teacher-Reviewed Feedback

Abylay Yerniyazov[1], Bakyt Bakayeva[1], Zhalgasbek Iztayev[2], Pernekul Kozhabekova[2], Symbat Nurgaliyeva[1] (✉)

[1]Astana IT University, Astana, Kazakhstan

[2]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan

symbat.nurgaliyeva@astanait.edu.kz

ABSTRACT

Open-ended short responses can reveal more about student understanding than selected-response items yet scoring them during live instruction is rarely feasible. This paper presents a bring-your-own-device (BYOD) mobile classroom system that combines rubric-guided large language model (LLM) scoring of short open-ended answers, real-time teacher analytics through a classroom dashboard, and evidence-based post-quiz feedback generation. The system includes a teacher-reviewed workflow in which instructors can inspect model outputs and, when needed, adjust scores or feedback during formative classroom use. In the reported evaluation, however, teacher override was not applied or analyzed; the agreement metrics compare the system's initial rubric-guided outputs with expert reference scores. The system was evaluated in three classroom groups (N = 60 students; 480 question-level scoring cases). Three expert raters participated, each assigned to one classroom group, so each response received one expert reference score. In this test, system scores showed strong but not complete agreement with expert reference scores (QWK = 0.887; MAE = 0.089; RMSE = 0.121). Mastery-label accuracy reached 78.3%. Expert raters gave positive scores to the generated feedback reports for groundedness (M = 4.78), specificity (M = 4.37), and actionability (M = 4.57) on a 1–5 scale. Within the limits of this small-scale classroom evaluation, the findings suggest that rubric-guided LLM scoring may support formative assessment in mobile classrooms, provided teacher oversight is maintained for borderline cases.

KEYWORDS

mobile classroom, formative assessment, open-ended responses, rubric-guided scoring, teacher-reviewed assessment, evidence-grounded feedback, learning analytics dashboard, large language models (LLMs)

1 INTRODUCTION

Mobile learning can support flexible classroom access when its use rests on pedagogically informed design [1]. Recent bibliometric evidence also shows that mobile

Yerniyazov, A., Bakayeva, B., Iztayev, Z., Kozhabekova, P., Nurgaliyeva, S. (2026). A Mobile System for Formative Assessment and Teacher-Reviewed Feedback. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(15), pp. 37–50. https://doi.org/10.3991/ijim.v20i15.61761

Article submitted 2026-03-30. Revision uploaded 2026-05-26. Final acceptance 2026-05-27.

learning in higher education is increasingly associated with flexible access, interactivity, student engagement, and continuing challenges around infrastructure, digital literacy, content quality, and privacy [2]. In these environments, formative assessments can draw on mobile applications, interactive tasks, learning analytics, and timely feedback to support student achievement [3], though successful implementation depends on teachers' ability to align digital tools with pedagogical practice and interpret learning evidence [4]. AI-supported models extend this further, linking feedback and adaptive support within platform-based workflows [5]. Open-ended responses offer richer evidence of understanding than selected-response items but evaluating them at scale during live instruction remains a persistent practical challenge-especially in bring-your-own-device (BYOD) classrooms, where submissions arrive from students' own devices and rubric-based scoring must keep pace with instruction.

Recent work suggests that large language models (LLMs) hold genuine promise for rubric-based automatic scoring of written responses [6]. LLM-based short-answer grading can reduce teacher workload, but human oversight remains necessary, as model judgments may diverge from human assessments in uncertain cases [7]. Recent university-level evidence similarly cautions that GenAI scoring should not be treated as interchangeable with human judgement in high-stakes assessment, particularly where validity, reliability, fairness, and score interpretation are central concerns [8]. Adaptive approaches emphasize scoring frameworks that connect model decisions explicitly to rubric criteria or key points [9]. Comparative evidence indicates that GPT-based scoring is competitive but does not consistently outperform traditional automatic short-answer grading (ASAG) models across tasks and languages [10], and fairness across student groups remains an active concern [11]. These issues argue against fully autonomous grading, particularly when scores carry instructional consequences.

Human–AI grading research suggests that machine-generated signals can assist human graders, though teachers still need question context before relying on model outputs [12]. A teacher-in-the-loop design addresses this by keeping the educator in a position of oversight and final authority [13]. Human-centred learning analytics research similarly calls for human control, transparency, and stakeholder involvement in educational AI systems [14]. Rather than treating model outputs as final autonomous grades, such systems surface results through a teacher-facing interface so that educators can inspect scores, evidence, and feedback during formative use. Teacher-facing dashboards can help instructors translate classroom data into actionable instructional decisions [15], and recent dashboard research highlights the importance of explainability, privacy, and real-classroom deployment [16].

The study describes a two-mode mobile classroom system. Mode A supports in-class rubric-guided scoring and streams analytics to a teacher dashboard; Mode B generates evidence-grounded post-quiz feedback from structured rubric traces. Three research questions guide the evaluation:

- RQ1: How similar are the system's question-level scores to expert human ratings?
- RQ2: How accurately does the system assign mastery labels compared with human rubric-based labels?
- RQ3: How do expert raters evaluate the generated feedback for groundedness, specificity, and actionability?

This paper makes two related contributions. First, it proposes a transferable design model for mobile formative assessment that combines BYOD short-answer submission, rubric-guided LLM scoring, teacher-facing analytics, evidence-grounded feedback, and teacher review as a safeguard. This model is intended for classroom settings in which AI-generated scores and feedback are made visible to teachers for interpretation rather than treated as final autonomous judgments. Second, the paper reports a small-scale classroom evaluation of the system's initial AI-generated outputs, including rubric-guided scoring agreement, mastery-label assignment, and expert-rated feedback quality. A methodological contribution is the explicit separation between system capability and evaluated evidence: teacher review and score adjustment is supported by the design, but teacher override was not applied or analyzed in the reported evaluation. The results should therefore be interpreted as agreement between initial system outputs and expert reference scores, not as evidence of teacher-corrected scoring performance. This framing aligns with recent work on LLM-generated feedback and explainable AI in education [17], [18].

2 SYSTEM DESIGN

2.1 System overview

The system operates in two modes that share a rubric-based scoring backbone. Figure 1 presents the overall two-mode classroom assessment workflow. A quiz begins when the teacher creates open-ended questions, reference answers, and weighted rubric criteria. In Mode A, students submit short answers through a BYOD mobile client. Each response is evaluated against the teacher-authored rubric, and the resulting criterion-level outcomes, score percentages, and evidence excerpts are displayed on a teacher-facing dashboard [15]. This interface allows teachers to inspect the basis of the model output and, when needed in formative classroom use, adjust scores or feedback. In the present study, this adjustment function is described as part of the system design; it was not used in the reported agreement analysis. In Mode B, saved rubric traces are used to generate individualized post-quiz feedback reports that remain linked to the student's submitted evidence and rubric outcomes, following evidence-based feedback research emphasizing connection between task evidence and learning criteria [17].

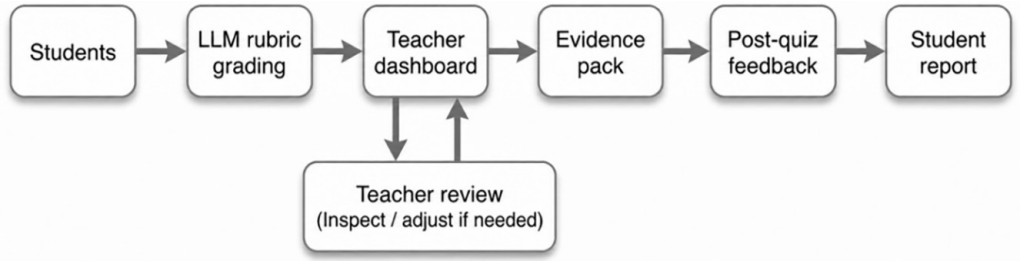

Fig. 1. Overview of the two-mode classroom assessment workflow

2.2 Mobile student client and teacher dashboard

Students use the web app on their own devices. Teachers monitor the session on a dashboard. Figure 2 shows the dashboard with student progress, completion

status, response time, and final scores. The dashboard also supports real-time score distributions, criterion-level breakdowns, participant progress, and frequent error tags per question. These views are intended to help teachers identify common misconceptions, notice borderline cases, and decide which responses need closer review during or after the quiz—consistent with work on teacher-facing learning analytics [15], [16].

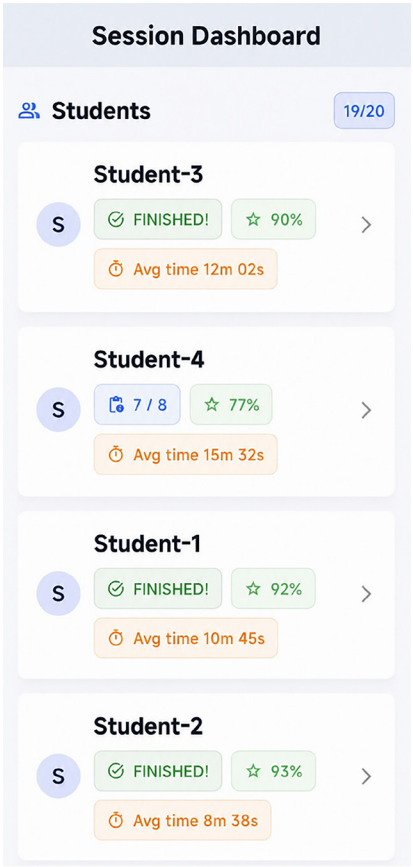

Fig. 2. Teacher dashboard interface

2.3 Rubric-guided prompting and structured outputs

Each scoring prompt combines four elements: the question text, the student's exact submission, a teacher-authored reference answer (used for factual grounding, not semantic matching), and the explicit rubric with criteria and point weights. Rather than measuring similarity to an expected response, the LLM evaluates the student's answer directly against the rubric criteria. The model returns a structured JSON object containing criterion outcomes, awarded percentages, and brief evidence excerpts from the student's response [9]. Figure 3 shows the teacher quiz creation interface used to define the question, reference answer, and weighted evaluation criteria. Figure 4 illustrates the structured JSON output produced by the scoring stage.

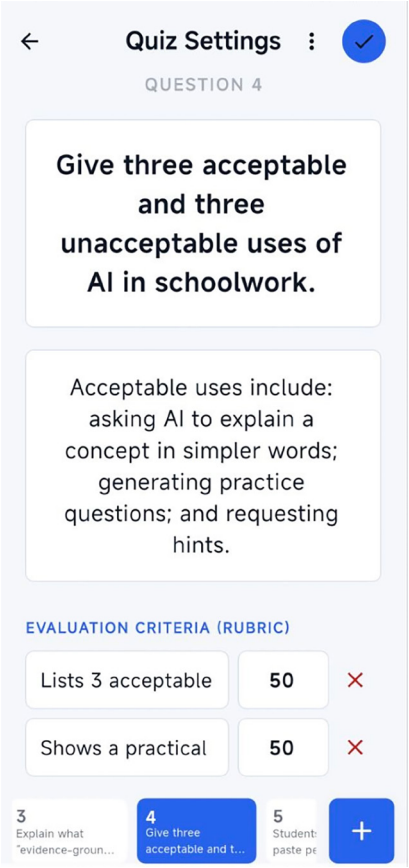

Fig. 3. Teacher quiz creation interface

```
{
  "question": "<QUESTION HERE>",
  "reference_answer": "<REFERENCE ANSWER HERE>",
  "criteria": [
    {
      "definition": "Student states the correct key idea (main claim).",
      "weight_percent": 30,
      "awarded_percent": 30,
      "student_evidence_text": "...excerpt from the student's answer..."
    },
    {
      "definition": "Student provides justification (explains why the claim is true).",
      "weight_percent": 50,
      "awarded_percent": 0,
      "student_evidence_text": null
    },
    {
      "definition": "Student includes a relevant example/detail supporting the claim.",
      "weight_percent": 20,
      "awarded_percent": 20,
      "student_evidence_text": "...excerpt from the student's answer..."
    }
  ],
  "question_awarded_percent": 50
}
```

Fig. 4. Example of structured output (JSON)

2.4 Evidence-grounded feedback generation

The feedback generation stage converts rubric scoring traces into a diagnostic report without drawing on information outside the evidence pack, limiting generated statements to what the assessment record supports. Each report provides a topic-level summary of strengths and gaps, a mastery label, and per-question guidance aligned with the rubric. Figure 5 illustrates the system interfaces used around feedback delivery, including the session lobby, student answer submission, progress summary, and AI feedback panel. This constrained design reflects the need for evidence-based feedback and explainable educational AI outputs [17], [18]. Teachers can review the rubric trace and draft reports, use them to guide direct conversations with students, and revise outputs when needed in formative practice. This review function is intended as a safeguard against unsupported or insufficiently specific feedback, although the present evaluation assessed the generated reports before any teacher revision.

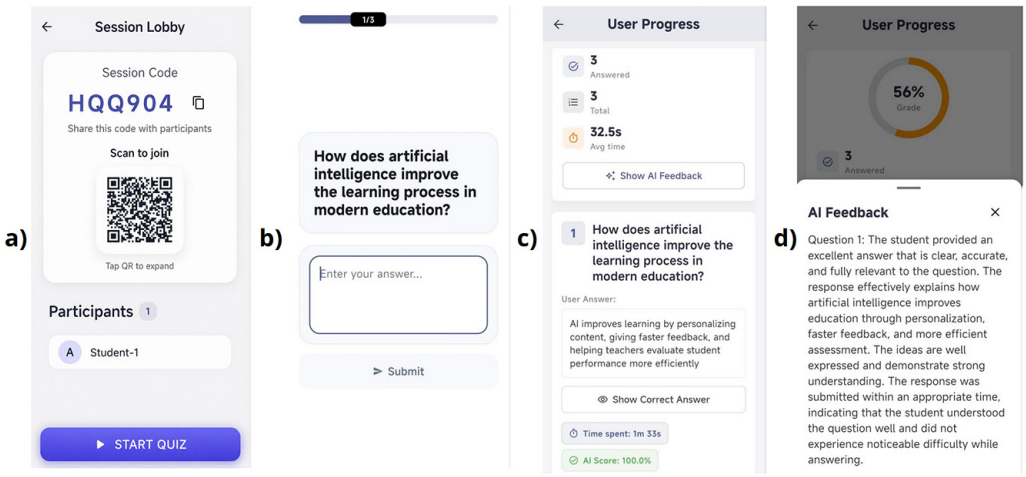

Fig. 5. System interfaces used around assessment and feedback delivery: (a) session lobby with join QR code and participant list, (b) student answer submission screen, (c) student progress and score summary, and (d) AI feedback panel

3 METHODOLOGY

3.1 Study context

The study took place during an ICT class at school. The quiz topic was "AI in Education: Uses, Limits, Ethics, and Safety." The classroom followed a BYOD model, and students used their own mobile devices to access the system and submit short open-ended responses.

3.2 Participants and dataset

Three classroom groups took part in the evaluation (N = 60). All groups completed the same quiz with eight open-ended questions. Each question had a teacher-written reference answer and a weighted rubric with several criteria. The dataset included 480 question-level scoring cases (60 students × 8 questions).

3.3 Rubric-guided scoring process

For each submission, the LLM received the question, the student's answer, the teacher's reference answer, and the weighted rubric criteria. The model returned to JSON with results for each criterion, percentages, and evidence excerpts. Each question received a final score in percentage form. Then the system assigned a mastery label. If a score was 85–100%, it was marked as Mastered. Scores of 60–84% were marked as Partial, and 0–59% as Not Yet.

3.4 Human scoring

Three expert raters participated in the human evaluation. Each was assigned to one classroom group and scored the responses of students in that group using the same teacher-authored rubric and weighting scheme as the system. Consequently, each response received one expert reference score rather than independent ratings from multiple raters. Across all three groups, expert raters provided reference scores for all 480 question-level cases. Mastery labels were derived from expert scores using the same fixed thresholds applied to system scores. Because the raters did not score a shared set of responses, the study did not estimate inter-rater reliability across common responses.

3.5 Teacher review and override in the evaluation

The implemented system includes a teacher-review function that allows instructors to inspect AI-generated scores, criterion-level evidence, and feedback, and to adjust outputs when needed during formative classroom use. In the reported classroom evaluation, this function was available as a design feature, but teacher overrides were not applied to the dataset and were not analyzed as a separate experimental condition. Therefore, all reported agreement metrics compare the system's initial rubric-guided outputs with expert reference scores.

3.6 Metrics

Agreement between system and expert scores at the question level was assessed using five metrics: Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Bias, and the share of cases within ±0.10 of the expert score. QWK measures agreement while accounting for the ordered nature of rubric-based scores; MAE and RMSE show average score differences; Mean Bias indicates directional tendencies. Mastery-label agreement was assessed using overall accuracy and a confusion matrix.

3.7 Feedback quality evaluation

Feedback reports for all 60 students were submitted for expert review, with each rater evaluating the reports for their assigned classroom group. Reports were rated on three dimensions using a 1–5 scale: groundedness (whether feedback was supported by the student's submission and rubric trace), specificity (whether feedback identified concrete strengths, gaps, or criteria rather than general comments), and actionability (whether feedback provided guidance a student could act on).

3.8 AI tool disclosure

The system uses the open-weights Llama-3.3-70b-versatile model for both rubric-guided scoring and evidence-based feedback generation. It uses a zero-shot prompting strategy that produces structured JSON outputs and grounds the evaluation in teacher-authored rubric criteria without fine-tuning.

3.9 Ethics and consent

The study was conducted as part of regular classroom instruction. Permission to conduct the classroom activity and to use anonymized student responses for research and evaluation purposes was obtained from the school administration in accordance with the procedures applied by the school and the authors' institution and, where applicable, national requirements for this classroom-based educational activity. The dataset used for analysis did not include student names, official grades, contact details, school identifiers, or other directly identifying information.

Because the activity was conducted during normal instruction and the analysis used anonymized classroom responses, separate ethics committee review was not required under the applicable school and institutional procedures, and no formal waiver document was issued. Students were informed that their anonymized responses could be used for research and evaluation purposes. Separate written consent or assent was not required under the procedure applied to this anonymized classroom evaluation.

4 RESULTS

4.1 Rubric scoring agreement (RQ1)

Across the 480 question-level scoring cases, system scores showed encouraging but incomplete agreement with expert reference scores within this dataset (refer to Table 1). A QWK of 0.887 indicates encouraging agreement for rubric-based percentage scores in this evaluation context. The MAE (0.089) and RMSE (0.121) indicate relatively small average differences between system and expert reference scores, although not all responses were closely matched. The negative mean bias (–0.069) shows that the system slightly underscored compared with the expert. Overall, 62.9% of scores were within ±0.10 of the expert reference score, while the remaining 37.1% differed by more than ±0.10. This indicates that a meaningful subset of responses still requires closer instructional review.

Table 1. Question-level scoring agreement (N = 480)

Metric	Value
Quadratic Weighted Kappa (QWK)	0.887
Mean Absolute Error (MAE)	0.089
Root Mean Square Error (RMSE)	0.121
Mean Bias	–0.069
Proportion within ±0.10	62.9%

4.2 Mastery-label classification (RQ2)

Mastery-label accuracy across the 60 students was 78.3%. Table 2 shows the full confusion matrix. Most disagreements occurred near category boundaries. Among expert-labeled Partial cases, seven were assigned to Not Yet by the system and one was assigned to Mastered. In addition, five expert-labeled Mastered cases were assigned to Partial. Agreement was strongest for the Not Yet category, where all 15 expert-labeled cases received the same system label. This pattern suggests that rubric-guided LLM scoring was more consistent with expert labels in clear cases and less consistent at category boundaries.

Table 2. Confusion matrix: system vs. expert mastery labels (N = 60)

	AI: Mastered	AI: Partial	AI: Not Yet
Human: Mastered	17	5	0
Human: Partial	1	15	7
Human: Not Yet	0	0	15

4.3 Feedback quality (RQ3)

Expert ratings of the 60 student feedback reports are presented in Table 3. Groundedness had the highest mean score (4.78). It was followed by actionability (4.57) and specificity (4.37). In this sample, expert raters judged the feedback positively, especially in relation to its grounding in student responses and rubric traces. The somewhat lower rating for specificity suggests that some reports described gaps in general terms rather than pointing to particular rubric criteria or student phrasing.

Table 3. Expert evaluation of generated feedback reports (N = 60)

Dimension	Mean Rating (1–5)
Groundedness	4.78
Specificity	4.37
Actionability	4.57

5 DISCUSSION

5.1 Interpretation of findings

A QWK of 0.887 across 480 question-level cases is encouraging for a zero-shot rubric-guided LLM approach applied in a live classroom context. The slight underscoring tendency (mean bias = –0.069) is consistent with variation reported in prior LLM scoring studies, where performance depends on prompts, rubrics, tasks, and grading contexts [6]. Evidence from another short-answer grading context similarly shows that LLM-generated scores may diverge from human judgments in some cases, reinforcing the need for careful teacher review [7]. In this study, 62.9% of system scores were within ±0.10 of the expert reference score, while 37.1% were outside that range. This pattern suggests that the system may be useful for accelerating formative scoring, but not for replacing professional judgment in cases where student answers are incomplete, ambiguous, or close to rubric boundaries.

Mastery-label accuracy of 78.3% is meaningfully higher than chance but still leaves roughly one in five students assigned a label that differed from the expert reference. As the confusion matrix shows, disagreements occurred mainly near mastery-label thresholds, especially between Partial and Not Yet, with additional Mastered-to-Partial cases and one Partial-to-Mastered case. This suggests that the system was more consistent with expert labels in clear-cut cases and less consistent at category thresholds—precisely where teacher review and human–AI grading support are most important [12], [13].

Feedback quality ratings were consistently positive. Groundedness scored highest (4.78), which may reflect the constrained generation design limiting the model to information already in the rubric trace and evidence pack, consistent with the emphasis on evidence-based feedback in classroom LLM use [17]. Specificity rated somewhat lower (4.37), indicating that feedback occasionally described gaps at a general rather than criterion-specific level. These results are promising within this sample but should not be taken as evidence that the system avoids unsupported claims in other contexts—particularly given that explainability remains an ongoing concern in educational AI [18].

5.2 Classroom use and teacher-reviewed assessment

The system should therefore be understood as a formative decision-support tool rather than an autonomous grading system. In classroom practice, the dashboard can support teachers in monitoring completion, identifying recurring misconceptions, recognizing borderline mastery cases, and deciding which responses warrant closer attention during or after the lesson. Because rubric-based evidence is presented alongside scores, the interface is intended to support informed teacher review rather than a simple accept-or-reject decision. This design is consistent with legal and human-centered arguments for meaningful teacher involvement in automatic assessment systems [13], [14].

The teacher-reviewed workflow should therefore be understood as a design safeguard rather than as an empirically evaluated intervention in this study. Teacher overrides were not applied to the analyzed scores. Therefore, the reported QWK, MAE, RMSE, bias, and mastery-label accuracy reflect the system's initial rubric-guided outputs rather than teacher-corrected results. This distinction is central to the interpretation of the findings. The present study examines the extent to which the system's initial AI-generated outputs corresponded with expert reference scores. However, the study does not assess whether teacher review leads to improvements in scoring accuracy, feedback quality, student learning, or classroom decision-making. Future work should examine how teachers use the dashboard in classroom practice, which AI-generated outputs they choose to adjust, and how these adjustments affect the feedback presented to students and the planning of subsequent instruction.

5.3 Transferability, risks, and scaling considerations

The transferable contribution of the study lies in the design model rather than in the specific performance level observed in one classroom dataset. The system brings together BYOD-based mobile submission, rubric-guided LLM scoring, teacher-facing analytics, evidence-grounded feedback, and teacher review as a safeguard. Together, these components respond to practical needs found in a range of educational settings. Prior research on mobile learning supports the use of device-based access when it is guided

by clear pedagogical design [1]. Studies of formative assessment in mobile environments also emphasize the importance of interactive tasks, timely feedback, and learning analytics [3]. In addition, research on teachers' technology use suggests that successful implementation depends on the alignment between digital tools, pedagogical goals, and data literacy [4]. In the Kazakhstani context, work on AI-driven technology integration in STEM education further points to the value of practical, technology-supported learning designs for student engagement and conceptual understanding [19]. A study on intelligent mobile systems for student performance evaluation further highlights the role of data-driven assessment, automated learner-performance analysis, and mobile interfaces in supporting academic decision-making [20]. Teacher-facing dashboard research also supports the use of analytics to make classroom data more actionable [15], [16].

At the same time, rubric-based automation introduces practical risks. One practical risk is that teachers may place too much confidence in numeric dashboard scores when these scores appear more precise than the underlying evidence allows. This is especially important for borderline responses or for answers that reveal aspects of understanding not fully captured by the rubric. Fixed mastery thresholds may also introduce threshold effects, because even small score differences can shift a student from one category to another. For this reason, rubric quality is central to the usefulness of the approach. If criteria are unclear, unevenly weighted, or fail to address common misconceptions, the system may produce scores and feedback that appear reasonable but remain incomplete. Although evidence-grounded feedback can reduce unsupported comments, teacher review is still needed to judge whether the feedback is specific, instructionally appropriate, and consistent with the student's actual response.

Scaling this approach would also require careful attention to teacher workload, teacher preparation, model consistency, privacy, infrastructure, and validation across different subjects and languages. Work on interactive mobile case technologies in informatics teacher training suggests that technology-supported learning activities can strengthen communication, creativity, and problem-solving competencies among future computer science teachers [21]. While dashboards can make it easier to inspect large sets of responses, they may also introduce new review demands, particularly when many cases are flagged for attention or when teachers need to revise vague feedback. Model behavior may vary across topics, prompt formats, languages, and student writing styles, so local validation remains necessary before broader deployment. Privacy and infrastructure are also important in BYOD classrooms, where student devices, network reliability, and institutional data-protection requirements can affect implementation. For these reasons, the present results should be interpreted as evidence from a small-scale classroom evaluation of one design model, not as proof of general performance across educational contexts.

5.4 Limitations

This study has several limitations. It included three classroom groups from one institution and covered one quiz topic, so the findings should be interpreted cautiously. Three expert raters participated, but each rater scored only one group; consequently, each response had one expert reference score rather than multiple independent ratings. Future studies should include a shared set of responses scored by multiple experts to estimate inter-rater consistency. The study also used one model setup, Llama-3.3-70b-versatile with zero-shot prompting, so results may differ with other models, prompts, rubrics, subjects, or languages. Feedback quality was evaluated for one quiz topic, and the constrained generation approach reduces but does not eliminate the risk of unsupported or insufficiently specific statements.

Future work should include larger and more diverse samples, multi-subject and multi-institution validation, analysis of teacher review behavior, and assessment of student learning outcomes following feedback delivery.

6 CONCLUSION

This paper presented a BYOD mobile classroom system and design model that integrates rubric-guided LLM scoring, teacher-facing learning analytics, evidence-grounded feedback, and teacher review as a safeguard for formative assessment. The system was evaluated in three classroom groups (N = 60 students; 480 question-level scoring cases). Its initial rubric-guided outputs showed encouraging but incomplete agreement with expert reference scores within this dataset (QWK = 0.887; MAE = 0.089), while mastery-label accuracy reached 78.3%. Generated feedback reports also received positive expert ratings for groundedness, specificity, and actionability.

The contribution of the study is therefore twofold: it presents a transferable mobile formative assessment workflow linking BYOD submissions, rubric traces, dashboard analytics, and evidence-based feedback; and it reports a small-scale classroom evaluation of the system's initial AI-generated outputs. Although the system design supports teacher override, this function was not applied or analyzed in the reported evaluation. The findings suggest that rubric-guided LLM scoring may support formative classroom assessment when used under teacher oversight; however, they do not provide a basis for fully autonomous grading. Broader claims would require further validation across subjects, languages, institutions, student groups, and teacher-review practices.

7 REFERENCES

[1] S. Moya and M. Camacho, "Leveraging AI-powered mobile learning: A pedagogically informed framework," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100276, 2024. https://doi.org/10.1016/j.caeai.2024.100276

[2] A. D. Samala, S. A. Papadakis, and S. Rawas, "Global insights into mobile learning in higher education: A PRISMA-guided bibliometric analysis from 2007 to 2023," *International Journal of Educational Reform*, advance online publication, 2025. https://doi.org/10.1177/10567879251341869

[3] V. Efrianova, M. Yaakob, A. A. Salameh, K. C. Hussin, and N. A. M. Zaki, "Formative assessment of student's academic achievements in mobile learning environments," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 11, pp. 52–63, 2024. https://doi.org/10.3991/ijim.v18i11.49045

[4] K. Børte, S. Lillejord, J. Chan, B. Wasson, and S. Greiff, "Prerequisites for teachers' technology use in formative assessment practices: A systematic review," *Educational Research Review*, vol. 41, p. 100568, 2023. https://doi.org/10.1016/j.edurev.2023.100568

[5] P. T. D. Thuy, N. T. Van, N. H. Trang, N. M. Thuong, V. T. T. Hien, and V. Q. Chung, "Designing an AI-supported formative assessment model for pre-service mathematics teacher self-study in Vietnam," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 22, pp. 50–68, 2025. https://doi.org/10.3991/ijim.v19i22.57723

[6] G.-G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai, "Applying large language models and chain-of-thought for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100213, 2024. https://doi.org/10.1016/j.caeai.2024.100213

[7] C. Grévisse, "LLM-based automatic short answer grading in undergraduate medical education," *BMC Medical Education*, vol. 24, p. 1060, 2024. https://doi.org/10.1186/s12909-024-06026-5

[8] G. Zacharis and S. Papadakis, "Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment," *Educational Process: International Journal*, vol. 19, 2025. https://doi.org/10.22521/edupij.2025.19.591

[9] H. Wang *et al.*, "LLMarking: Adaptive automatic short-answer grading using large language models," in *Proceedings of the 12th ACM Conference on Learning @ Scale (L@S 2025)*, 2025, pp. 105–115. https://doi.org/10.1145/3698205.3729551

[10] R. Ferreira Mello *et al.*, "Automatic short answer grading in the LLM era: Does GPT-4 with prompt engineering beat traditional models?" in *Proceedings of the 15th International Conference on Learning Analytics & Knowledge (LAK '25)*, 2025, pp. 93–103. https://doi.org/10.1145/3706468.3706481

[11] L. Rodrigues, C. Xavier, N. Costa, D. Gašević, and R. Ferreira Mello, "Is GPT-4 fair? An empirical analysis in automatic short answer grading," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100428, 2025. https://doi.org/10.1016/j.caeai.2025.100428

[12] Y. Li *et al.*, "Can AI support human grading? Examining machine attention and confidence in short answer scoring," *Computers & Education*, vol. 228, p. 105244, 2025. https://doi.org/10.1016/j.compedu.2025.105244

[13] L. Colonna, "Teachers in the loop? An analysis of automatic assessment systems under Article 22 GDPR," *International Data Privacy Law*, vol. 14, no. 1, pp. 3–18, 2024. https://doi.org/10.1093/idpl/ipad024

[14] R. Alfredo *et al.*, "Human-centred learning analytics and AI in education: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100215, 2024. https://doi.org/10.1016/j.caeai.2024.100215

[15] Z. Wang, W. Lin, and X. Hu, "Self-service teacher-facing learning analytics dashboard with large language models," in *Proceedings of the 15th International Conference on Learning Analytics & Knowledge (LAK '25)*, 2025, pp. 824–830. https://doi.org/10.1145/3706468.3706491

[16] L. Cabral, R. Pinto, and G. Gonçalves, "AI-powered learning analytics dashboards: A systematic review of applications, techniques, and research gaps," *Discover Education*, vol. 4, 2025. https://doi.org/10.1007/s44217-025-00964-y

[17] J. Meyer *et al.*, "Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100199, 2024. https://doi.org/10.1016/j.caeai.2023.100199

[18] G. Türkmen, "The review of studies on explainable artificial intelligence in educational research," *Journal of Educational Computing Research*, vol. 63, no. 2, pp. 277–310, 2025. https://doi.org/10.1177/07356331241310915

[19] M. Serik, K. Tleuzhanova, and S. Nurgaliyeva, "Enhancing master's-level STEM education through AI-driven IoT projects: A Kazakhstan experiment," *International Journal of Information and Education Technology*, vol. 15, no. 10, pp. 2086–2094, 2025. https://doi.org/10.18178/ijiet.2025.15.10.2407

[20] A. Herayono, M. Anwar, E. Tasrif, and Q. N. M. Batubara, "Intelligent mobile system for student performance evaluation: Model testing using structural equation modeling," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 20, no. 4, pp. 23–36, 2026. https://doi.org/10.3991/ijim.v20i04.60101

[21] R. Akhitova, A. Alzhanov, and V. Borisov, "Enhancing informatics teacher training with interactive mobile case technologies," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 13, pp. 111–132, 2025. https://doi.org/10.3991/ijim.v19i13.50985

8 AUTHORS

Abylay Yerniyazov is an undergraduate student in Software Engineering at Astana IT University, Astana, Kazakhstan. His interests include mobile application development, artificial intelligence, machine learning, and educational software development (E-mail: 230500@astanait.edu.kz).

Bakyt Bakayeva, MS in Computer Science, is a Senior Lecturer at the School of Software Engineering, Astana IT University, Astana, Kazakhstan. Her scientific and professional interests include mobile app development, data management systems, machine learning, data science, educational technologies, and the development of interactive learning materials for future IT specialists. She also explores how artificial intelligence, data-driven approaches, and well-structured digital applications can improve the quality, accessibility, and effectiveness of computer science (E-mail: bakyt.bakayeva@astanait.edu.kz).

Zhalgasbek Iztayev, a candidate of pedagogical science, is an Associate Professor of Computer science in the Higher school of Computer Engineering and Energetic at M. Auezov South Kazakhstan University in Shymkent, Kazakhstan. His research interests include education in Computer science, Programming languages, optimization methods, and computer and mathematical modeling (E-mail: zhalgasbek.Iztaev@auezov.edu.kz).

Pernekul Kozhabekova, a candidate of technical science in computer and mathematical modeling, is docent in the Higher school of Computer Engineering and Energetic at M. Auezov South Kazakhstan University in Shymkent, Kazakhstan. Her research interests include information systems analysis and modeling, the development of modern data-driven technologies, optimization methods, and computer and mathematical modeling (E-mail: pernekul.kozhabekova@auezov.edu.kz).

Symbat Nurgaliyeva, PhD in Computer Science, is an Assistant Professor at the School of Software Engineering, Astana IT University, Astana, Kazakhstan. Her research focuses on improving the quality of education for future computer science teachers through the theoretical and practical integration of robotics technologies, with particular emphasis on the use of mobile robots in the educational process. Her scientific interests include educational robotics, mobile robot development, machine learning, and neural networks (E-mail: symbat.nurgaliyeva@astanait.edu.kz).