ACADEMIA International Journal for Social Sciences
Volume 4, Issue 4, 2025 ISSN-L (Online): 3006-6638

Enhancing Teacher Feedback Using Ai-Powered Automated Grading Systems

Tazeen Huma
tazeenhuma@iqra.edu.pk
Senior Lecturer, Media Sciences Department, Faculty of Media Sciences, Iqra University Karachi.

Dr. Shakeel Ahmed
dr.shakeel.ahmed2024@gmail.com
Lecturer, Department of Sociology, University of Balochistan, Quetta, Pakistan.

Waqas Mahmood
waqas.mahmood@iub.edu.pk
Assistant Professor, Department of Educational Training, The Islamia University Bahawalpur, Pakistan.
Research Fellow, INTI International University Nilai, Malaysia. . oricid:0000-0003-4295-4454

Dawar Awan
dawar.awan@gmail.com
Department of Electrical Technology, University of Technology Nowshera.
Corresponding Author: Tazeen Huma tazeenhuma@iqra.edu.pk

Received: 20-07-2025 **Revised:** 29-08-2025 **Accepted:**25-09-2025 **Published:**05-10-2025

Abstract

This research looks into the effects of AI-assisted grading systems on feedback in higher education in Pakistan. Using a mixed methods approach, data were collected over an academic semester from 80 faculty members and 800 undergraduate students across 12 public universities in four provinces. The sample included six universities who implemented AI grading systems and six that continued with conventional grading without the AI grading systems. Data were collected from surveys of faculty and students, classroom observations, and interviews regarding feedback quality, time efficiencies, faculty satisfaction, student performance, and engagement, and faculty assessment marking and grading. The primary findings showed that the AI grading systems improved the timeliness, consistency, and personalization of student feedback significantly at the universities using them, per the SPSS data. Faculty reported that the automated grading systems improved their teaching efficiency, reduced grading time, and improved student engagement, but they faced some technological issues and advised caution regarding purely automated grading to grade, for example, a 'subjective assignment'. The qualitative data also noted challenges in using AI grading comparison to rubric criteria. In sum, the findings indicate that there is potential for AI grading to help improve feedback and assessment practices in higher education in Pakistan, although challenges remain.

Keywords: *Impact, automated, grading systems, AI technology, teacher feedback, Pakistani higher education institutions.*

INTRODUCTION

In the last couple of decades, the higher education sector has been affected by new technologies, like many other provosts and sectors of society, specifically for teaching and assessing student learning outcomes (Akintayo, Eden et al. 2024). Critiques of grading systems in educational settings remains rampant, and it has been described as slow, inequitable, inconsistent, and measured in an unnecessarily subjective manner. Contextually, this situation is worse in Pakistani universities. Grading examination scripts and assignments in manual processes is complicated by larger class sizes, excessive volumes of teaching, and less-than-helpful comments meant to enhance student learning. AI technologies now present a substantial opportunity to address the educational and academic assessment concerns just described (Crogman et al., 2023).

Automated systems under AI-assisted grading technology will reorganize how instructors evaluate student learning and give feedback. Advanced technologies, like machine learning algorithms, in addition to AI systems with natural language processing through AI systems and adaptive recognition patterns, can more accurately, effectively, and consistently evaluate student writing. In some cases, the systems may even outperform human capacity (Deepshikha, 2025). For instance, automated grading can be an effective source of feedback and can yield more detailed comments based on real-time observations while students are engaged in learning. Additionally, one of the major benefits of automated grading is the time it creates for educators to engage in more meaningful activities such as personal instruction, curricular design, and providing mentorship to students. However, there are significant downsides to consider when using automated grading systems to evaluate student learning. Important issues to consider are system reliability, context validity, and whether or not it is a need for a human evaluator to preserve the integrity of the assessment as a whole (Gnanaprakasam and Lourdusamy 2024).

Pakistan's higher education environment has specific circumstances that make the study of AI-based grading systems compelling. The country's universities are forced to address unresolved quality issues in assessment and feedback within the context of an outstandingly growing student population and severely reduced staffing levels (Tariq 2024). Academics often teach courses with well over one hundred students, which makes efficiently providing timely feedback for assignments and assessments impossible. Additional layering is added by the multiple regional languages that Pakistani universities negotiate, especially in the coproduction of teaching and assessment in Urdu and English, as they bring additional variability of standards in grading. These systemic challenges, along with the urgency to update education, suggest that using AI technologies is especially valuable in this context (Saleem, Saleem et al., 2025).

Automated grading systems offer distinct advantages for different stakeholders in the education ecosystem. For example, students receive timely and standardized grades that affect their learning, positively affecting their grade (Vetrivel, Arun et al. 2025). Feedback that takes weeks or months to give has little value from an educational perspective because, by the time a student receives feedback on their work, they will probably have moved on to other skills and lost most of the relevant context that makes the feedback educationally valuable. With AI, students receive timely feedback that reduces learning gaps in a more dynamic way. They can adjust their learning and skill execution during their work, and even within a session. Additionally, the standardized approach of the automated systems can mitigate some of the bias and favoritism that students may perceive in their teachers, building confidence in the evaluative process (Messer, Brown et al. 2024).

Artificial Intelligence (AI) may assist professors with grading in order to free up time to fulfill the professor's primary responsibilities of creating syllabi, performing research, and mentoring students. As colleges and universities implement using AI as their assessment system, educators will need to use semi-automated tools that conform to the organization's policy manual, and assessment systems need to assist professors in maintaining a level of performance that balances organized administrative tasks with contact with students. This assistant educator approach is referred to as desk control (Jiang and Li, 2021), which facilitates the use of AI while remaining central to the educator's role.

Significant obstacles persist regarding the use of AI grading systems in Pakistani universities. These hurdles involve the fundamental technical infrastructure, such as insufficient computing systems, as well as specific problems resulting from unreliable internet access within the country. Other issues involve concerns over AI reliability in judging more nuanced subjective aspects of students' work, such as essays, research papers, and other creative work (Jiang and Li 2021). The scholarly community is still highly uncertain about the capacity of AI to 'understand' elements such as context, creativity, and critical analysis. The deployment of automated grading systems may be viewed by administrative faculty as encroaching on their professional autonomy while effectively 'de-humanizing' the grading process, which may be explained by cultural issues There are often concerns that automated grading systems will ineptly and unintelligently judge student work, which lowers the need for professional competence among faculty (Khurshid, Khurshid et al., 2024).

It is necessary to establish AI grading algorithms that are responsive to cultural and linguistic practices in Pakistan. Most automated grading systems are developed in the West, intended to work in English contexts, and therefore are not useful for Pakistan's educational context (Naseer, Saeed et al. 2025). Although some academics may adopt a Western format, their frame of reference, cultural references, and rhetorical structures are likely to be very different. Therefore, it is important to recognize that if these factors are not identified, Western-centric AI algorithms will provide inappropriate and biased responses through over- or under-valuing the text produced (Dahri, Yahaya et al. 2024).

This research focused on the integration of artificial intelligence technology to support automated grading systems for teacher feedback in universities in Pakistan. Several projects using mixed-methods research in various provinces and universities showed evidence that the AI-driven tools improved efficiency and consistency of teacher feedback, and they also highlighted challenges associated with adaptation and training. These findings contribute to the field of educational technology, in which they present new knowledge, as well as implications for use by policymakers, administrators, and educators around the world, for understanding how to adopt AI-based assessment tools. Understanding these tools is vital to their more effective integration into teaching and learning processes and structures.

Research Objectives

1. To examine the impact of artificial intelligence (AI)-powered automated grading systems on the timeliness, quality, and consistency of teacher feedback in institutions of higher education in Pakistan.
2. To examine faculty perceptions, experiences, and satisfaction with the implementation of AI grading tools in the teaching environment.
3. To investigate students' academic performance and their satisfaction with feedback received through AI-powered automated grading systems in comparison to traditional grading approaches.

Research Questions

1. How does AI-assisted automated grading systems affect the timeliness, quality and consistency of faculty feedback in Pakistani higher education.

2. What do faculty members perceive, experience, and challenge when using the AI grading system in assessment?

3. In what ways does an AI grading system affect students' academic performance and satisfaction with assessment feedback?

Importance and Purpose of the Study

This research has implications for multiple stakeholders in the higher education landscape of Pakistan. The study's results extend the global scholarship about educational technologies in higher education. The results will help policymakers and university executives to make informed decisions regarding adoption of educational technologies, including future spending and facility renovations related to modernizing assessment practices. The multi-provincial nature of the research reinforces the different realities depending on the Pakistani university foundations. This aspect of the research represents the regional differences in terms of technological readiness, faculty preparedness, and institutional culture that exist in Pakistan. This research will provide faculty members with actionable evidence of AI technologies for educational purposes. Finally, the findings of this research may assist in mitigating the workloads associated with assessment approaches. evaluation quality. Understanding the challenges and barriers that early adopters have faced will help future implementers handle the complications their colleagues faced proactively. It is necessary to understand the quality of feedback that an automated grading system may provide and the implications for students' learning outcomes. Considering these factors contributes to treating technology as an advancement to educational quality, not a detractor. This study contributes to the limited literature available regarding the implications of AI on education in developing countries. These countries are often disadvantaged by under-resourced and inequitable educational systems. The variances within these areas produce uniquely challenging implementation processes. The culturally responsive bilingual AI system discussed emphasizes improving the quality of educational resources that are aligned with local cultures. It is meant to contribute to education equity in the Global South, particularly in non-Western countries. It is critical for educational institutions and technology developers to engage with this consideration.

LITERATURE REVIEW

In the past two decades, artificial intelligence (AI) has attracted growing interest from researchers in education. The first generation of automated grading systems focused on objective, multiple-choice, and math questions that could be scored algorithmically based on whether the answer was correct. Advances in AI for natural language processing and machine learning have opened up new possibilities for grading more complex assessments—everything from essays and short answers to creative writing. As these technologies develop and become mainstream, researchers and educators have begun to consider their implications for assessment and identify important issues of validity, reliability, and the role of automation in grading. The literature on automated grading systems presents a range of complicated possibilities to address (Chen, Zou, et al., 2022).

Studies investigating the efficiency of AI grading models have largely yielded positive outcomes, though not without some variability. In particular, when properly developed automated models have demonstrated an ability to assess specific types of assignments with marginally comparable reliability to human grading in technical writing, as well as organized essays and work that lends itself to well-defined rubric (see Kizilcec 2024). A significant advantage of automated models is the reliability (effectively fatigue-free performance) that automated systems provide, unlike human grading that can suffer from fatigue, affect, and/or unconscious biases that impact the evaluation of learning. Furthermore, the speed of evaluation with an automated model can become a significant time saver for grading related to student work because of the differential time (weeks vs. hours) to provide valuable feedback. However, some believe current Ai models are unable to provide nuanced contextual evaluations requiring across-the-board considerations of creativity, humor, irony, and deductive reasoning. The question of what types of automated evaluation of student work and which require human evaluation remains unanswered (Zhai, Chu et al. 2021).

The timing of feedback has been studied extensively in the education research field, especially in job-assessment systems. Cognitive science tells us that feedback is most effective when it is immediate or nearly immediate rather than delayed and that students need to address misunderstandings while the information is still fresh within their mind. In large university classes, feedback is regularly delayed anywhere from weeks to months. This diminishes the value of assessment because students will be unable to learn the material (Rasool, Qian et al. 2022). A valuable aspect of any job-assessment system is that the feedback is often immediate and the system can build effective cycles of learning. However, feedback may lose its value if the feedback is qualitative, generic, or vague. In those situations, it may be more beneficial to have a human sometimes respond. In short, a good feedback process is a combination of automated efficiency and human observation (Alasadi and Baiz 2023).

The management of faculty workload has been an extensive topic in the higher education literature, especially regarding assessment and grading, which is known to consume a lot of time. Research on faculty use of their time, specifically on assessment activities, finds that faculty spend as little as 20 percent (minimal) to as much as 40 percent (maximum) of their work time managing and reviewing assessments, to use their time to attend to higher-value work in research and service that benefits their students (e.g. curriculum development, mentoring, and advising). Automated grading systems are typically appealing to faculty for this reason—less work related to grading frees up time to focus on the few remaining parts of the educational experience that still require faculty engagement. However, research on the effective use of educational technology in teaching and learning has indicated that educational technology use does not arbitrarily reduce faculty workload, rather, it reshapes the work the faculty do. New responsibilities of maintaining the system, validating the results, and troubleshooting are tasks that might replace all or part of the original workload. Accordingly, it is critical to conceptualize the responsibilities of using educational technology in workload planning so that they can be accounted for to reduce potential for unrealistic planning or expectation (Miller, O'Connor et al. 2024).

The education of developing nations to Artificial Intelligence (AI) is a different process than for developed nations. Factors such as unstable electricity and unreliable internet connectivity inhibit Artificial Intelligence tools that are framed around closed, cloud-based technologies that require constant internet use. Staff and students may need support in the learning of trained technologies, which may not be the same or may take considerable time and energy committed to compliance (Hamlin 2021). Regarding acceptance and technology innovation, the culture of a society, and its attitudes towards software develops educational change, but also with regard to the ways in which people view technology, as well as their methods of teaching and learning and communications. The minimal research in the area of AI systems specifically designed for developing nations on the cusp of an innovative moment has shown that often the disconnect is attributed to not attending to contextual conditions, while most of the AI systems on the market are developed for Western nations. Developing nations must consider the opportunities to develop AI applications that are relevant to the language, culture, and economy of the developing nation (Alqahtani, et al., 2023).

Language processing issues are still significant with respect to developing AI grading systems, particularly in a multilingual setting like Pakistan (Saqlain and Shahid 2024). The most developed and sophisticated Natural Language Processing Systems have been for the English language, while other languages, including Urdu, have received very little prioritization and funding (Saqlain and Shahid 2024). Even though there are millions of Urdu speakers, it remains an under-developed, under-resourced AI language–particularly, in terms of training data and processing tools. This presents barriers in developing an AI grading system that can equally assess an assignment composed in Urdu, to the same standard as one in English. Furthermore, within the Pakistani academic context, code-switching to English is prevalent within Urdu–adding another layer of complexity to what the AI systems must be able to address. Not much empirical research is available with respect to multilingual AI in education, and this is valuable opportunity for research (Ahmed, Bhatti, et al. 2025).

The perspectives of students about learning experiences associated with automated grading and feedback are an essential area in the literature, although they may go under-valued and ignored. Some students value the fast and steady pace of feedback that is possible with automation, while other students find it valuable to uphold the argument that a machine does not truly understand students' work. Student feedback often highlights fairness issues in assessing creative and non-conventional solutions. In some cases, students' willingness to accept automated grading is related to the assignment's difficulty, the transparency of the grading outline, and visible human assessment in some capacity. Students prefer some form of combined grading, where a human reviews grading and feedback produced by a machine after the machine is used as an initial review of the work. The structure of automated grading and its use clearly have implications on acceptance and outcomes, and there is significance in how automated grading is framed, as well as the how automated grading is implemented (Ahmad, Rahmat et al. 2021).

RESEARCH METHODOLOGY

The authors used a mixed-methods research design to explore the effect of AI-powered automated grading systems on teacher feedback at universities in Pakistan. The study was conducted over one academic semester and included all four provinces of Pakistan: Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan. 80 faculty members and 800 undergraduate students from 12 public universities were involved in the study using a quasi-experimental design. AI grading tools were introduced at six universities while the other six control universities used a manual grading system. Surveys were delivered to measure the quality, turnaround time, and faculty satisfaction with the grading process before and after the introduction of AI grading tools. To assess students' perceptions on the influence of feedback and grading efficacy on their academic results a structured questionnaire was developed and administered in addition to the course assignments, tests and other relevant measures. Observations from coursework and classroom sessions assisted in assessing real-time applications of AI systems in varying disciplines. Qualitative data on faculty experience; challenges, and benefits of automated grading from higher education were obtained through semi-structured interviews with 25 faculty members in different provinces. The AI grading system was localized to meet the needs of the bilingual educational setting of Pakistani universities which employs Urdu and English language support. SPSS software was used for statistical analysis of pre- and post-intervention data. Qualitative data from interviews and observations was both analyzed thematically to identify patterns and themes.

FINDINGS AND ANALYSIS

Quantitative Data Analysis

A quantitative analysis of data collected from twelve universities in Pakistan showed a finding of significance regarding the effectiveness of AI-powered automated grading systems. The quantitative analysis considered six intervention universities which implemented AI grading tools and compared them to six control universities relying on traditional assessment methods, across several multi-variable measures, such as feedback timing, feedback quality, feedback consistency, faculty satisfaction, student performance, and performance outcomes.

Table 1: Faculty Participant Demographic Characteristics.

Characteristic	Intervention Group (n=40)	Control Group (n=40)	Total (n=80)
Gender			
Male	24 (60%)	26 (65%)	50 (62.5%)
Female	16 (40%)	14 (35%)	30 (37.5%)
Age Range			
25-35 years	12 (30%)	10 (25%)	22 (27.5%)

36-45 years	18 (45%)	19 (47.5%)	37 (46.25%)
46-55 years	8 (20%)	9 (22.5%)	17 (21.25%)
Above 55 years	2 (5%)	2 (5%)	4 (5%)

Table 1 indicates that both groups had similar gender preponderance, age; and years of teaching experience, and rank, with most participants aged 36-45 years, having 6-15 years of teaching experience, representing mid-career faculty; furthermore, the gender proportions were reflective of current gender trends in Pakistan academia with a slightly larger male in number. The demographic similarities would provide a greater confidence regarding the validity of comparisons by assuring that the observed effects could be attributed to the AI intervention rather than a result of the differences of the participants in the respective treatment groups.

Table 2: Pre-Intervention Timeliness of Feedback Comparison

Feedback Return Time	Intervention Group (n=40)	Control Group (n=40)	p-value
Mean Days to Return Graded Assignments	18.4 (SD=4.2)	17.8 (SD=4.5)	0.523
Assignments Returned Within 1 Week	8%	10%	0.742
Assignments Returned Within 2 Weeks	32%	35%	0.668
Assignments Returned After 3 Weeks	12%	10%	0.753

Table 2 shows the baseline feedback timeliness. No differences were observed between groups. Both groups displayed the same patterns of delayed feedback, averaging just over 2.5 weeks. A small amount of assignments was returned within a week. This indicates systemic issues with feedback regarding timeliness within Pakistani universities. Statistically, it is the similarity of baseline experience between groups that allows us to examine the impact of the intervention in response to the application of AI. If subsequent differences emerge, they can represent changes due through intervention rather than differences inherent in the groups prior to baseline.

Table 3 Post-Intervention Feedback Timeliness Comparison

Feedback Return Time	Intervention Group (n=40)	Control Group (n=40)	p-value
Mean Days to Return Graded Assignments	3.2 (SD=1.8)	18.1 (SD=4.3)	$<0.001^{***}$
Assignments Returned Within 1 Week	89%	9%	$<0.001^{***}$
Assignments Returned Within 2 Weeks	11%	33%	$<0.001^{***}$

The introduction of AI resulted in meaningful changes to the timeliness in reporting on the intervention group. From Table 3, we see that the mean return time improved substantially, from 18.4 days to 3.2 days (an 83% improvement) with a highly statistically relevant p-value > 0.001. Notably, 89% of the intervention group participants had their assignments returned within one week compared to 9% of the control group. The differences in the mean return times and the 1-week reporting period are consistent within the control group across each measurement, confirming that similar study evaluations in the intervention group were not simply a result of the absolutely unique characteristics of the 3 assignments selected by the study evaluator. Rather, participants in the intervention group were meaningfully more timely in the feedback reporting timeline, a particularly significant improvement to the feedback mechanism since providing feedback in a timely manner is critical to effective learning.

Table 4: Faculty Perceptions of the Quality of Feedback!

Quality Indicator	Intervention Pre (Mean±SD)	Intervention Post (Mean±SD)	Control Pre (Mean±SD)	Control Post (Mean±SD)
Feedback Specificity (1-5 scale)	2.8±0.6	4.1±0.5***	2.7±0.7	2.8±0.6
Feedback Consistency (1-5 scale)	2.9±0.5	4.3±0.4***	2.8±0.6	2.9±0.5
Overall Feedback Quality (1-5 scale)	2.9±0.5	4.0±0.4***	2.8±0.6	2.9±0.5

Note: ***p<0.001

Faculty self-reporting both demonstrated statistically significant improvement in the intervention group (p<0.001) across all quality areas. For example, the specificity of feedbacks increased from 2.8 to 4.1 in the intervention group and a consistency score showed the largest disparities of change from 2.9 to 4.3. Meanwhile, the control group retained consistent scores. These findings suggest that AI systems, in addition to keeping up our rapid feedback response rate, improved perceived quality on several areas of enhancement, improved action ability and provided clearer guidance to facilitate students' improvement.

Table 5: Faculty Time Use Changes

Activity	Intervention Pre (Hours/Week)	Intervention Post (Hours/Week)	Change	Control Pre	Control Post
Grading Assignments	12.4±3.2	5.8±2.1***	-53.2%	12.1±3.4	12.3±3.2
Providing Feedback	4.6±1.8	2.2±1.1***	-52.2%	4.4±1.9	4.5±1.8
Individual Student Consultation	3.8±1.5	6.4±1.9***	+68.4%	3.9±1.6	4.0±1.5

Note: ***p<0.001

Table 5 describes institutional changes to faculty workload across the intervention and control group. The total amount of time spent grading decreased by 53% in the treatment group, meaning, on average, that faculty were able to reclaim 8.8 hours of productive time per week. The vast majority of the reclaimed faculty time was spent on lesson planning, one-on-one support and consultation time (68.4% increase), and research. For faculty in the control group, there was virtually no change in the allocation of time spent on teaching, research and student success. Overall, faculty were able to recover valuable time for planning, research, and student consultations.

Table 6: Student Academic Performance Outcomes

Performance Measure	Intervention Group (n=400)	Control Group (n=400)	Effect Size (Cohen's d)	p-value
Mean Course Grade (0-100)	72.6±11.4	68.9±12.3	0.31	0.001**
Assignment Scores (0-100)	74.8±10.2	70.3±11.8	0.41	<0.001***
Assignment Resubmission Rate	42.5%	18.7%	-	<0.001***

Note: **p<0.01; ***p<0.001

Table 6 presented measurable improvements in student achievement. Students who received intervention student earned significantly higher course grades and assignment scores, as well as moderate effect sizes. The assignment resubmission rate increased by a factor of two, suggesting that quick feedback resulted in meaningful changes and improvements in skill development. The findings from Table 6 suggest that AI-Augmented Feedback can support ongoing assessment and promote iterative learning strategies that are valuable and important practice but often impractical using traditional grading methods.

Table 7: Student Satisfaction with Feedback

Satisfaction Indicator	Intervention Group (Mean±SD)	Control Group (Mean±SD)	p-value
Timeliness of Feedback (1-5)	4.4±0.6	2.9±0.8	<0.001***
Clarity of Feedback (1-5)	4.2±0.7	3.1±0.7	<0.001***
Overall Satisfaction (1-5)	4.2±0.6	3.1±0.7	<0.001***

Note: ***p<0.001

Table 7 indicated that students in the experimental group scored higher on all dimensions of satisfaction than control group students. The feedback prompt (4.4 as compared to 2.9) stood out as, both students viewed AI feedback as fair and more helpful. Overall satisfaction was also indicative of positive perceptions as the experimental group reported an average score of 4.2, while control students averaged 3.1. This suggests students realized the benefits of using AI feedback even when believing that feedback was useful or accurate based on their own responses, even though there are suspicions about student resistance to automated assessments.

Table 8: Faculty Satisfaction and Usability

Measure	Mean±SD	Percentage Agreement
Time Saved (1-5)	4.5±0.6	92.5% (Agree/Strongly Agree)
Overall Satisfaction (1-5)	4.0±0.7	80.0% (Agree/Strongly Agree)

Reliability of AI Grading (1-5)	3.6±0.8	67.5% (Agree/Strongly Agree)
Technical Support Adequacy (1-5)	3.4±1.0	60.0% (Agree/Strongly Agree)

Faculty satisfaction is summed in Table 8, and time savings received the highest rating (4.5, 92.5% agreement). Overall satisfaction was strong (4.0, 80% positive), but in contrast, lower ratings for reliability and technical support showed areas for improvement. Reliability issues highlighted the faculty's uncertainty related to trusting AI for complex assignments. Additionally, technical support adequacy received the lowest rating, indicating support systems and infrastructure needed to be developed to implement AI most effectively.

Qualitative Analysis

Contextual qualitative insights were gleaned from semi-structured interviews with 25 faculty members. We conducted thematic analysis and five main themes emerged in relation to inquiry participant's understandings of the experiences of implementation in Pakistani universities.

Theme 1: Relief of Grading Burden

The most prominent theme was the relief of grading written assignments. Faculty described feeling tired of the amount of written assignments to grade, particularly those that had a large number of students. As one associate professor noted, "grading would take an entire weekend and I wouldn't feel able to do anything else." Several described their experience after the implementation and explained it in metaphors of burdens being lifted. A lecturer from Sindh shared that she was able to give thoughtful feedback to every assignment, instead of just comments from a rubric, due to the fact that she was less pressed for time. This liberation included both emotional relief and psychological release, and several acknowledged stress reducing, job satisfaction increasing, and general well-being increasing. Some participants indicated guilt, however, felt for being able to avoid their own responsibilities for grading.

Theme 2: Trust and Concerns Regarding Reliability

Concern about the reliability of AI, even when results appear to be good, remained an issue, particularly in subjective tasks. Several faculty members described a process of reviewing every AI-assigned grade by comparing collaboration with their judgment and recording it over various assignments and time. Only after the faculty member and AI assessments agreed over time for a number of assignments, and in some cases multiple times, were instructors more comfortable with the results; however, there was still concern that AI could generate a 'bad' assessment. One instructor from Balochistan conveyed discomfort with AI's ability to displace or detect creativity and critical thinking that may diverge from norms. Some of the faculty reported that in cases where assignments were so poorly assessed by AI, the only solution was to have a 'human touch'. Such concerns were strongest with assignments in the humanities and social sciences where the assessment of the quality of the argument and contextual relevance were needed. Again, faculty emphasized that they were using AI as a tool to support their decision-making process: AI would generate a provisional assessment, but the faculty member would always assess and/or modify the provisional assessment. Faculty indicated they preferred a hybrid approach where AI assessments are reviewed and adjusted by the faculty member prior to finalizing the assignment.

Theme 3: Technical and Infrastructure Challenges

Interviews with faculty members highlighted multiple infrastructure and technology challenges. Internet connectivity was an ongoing issue. The server dropped downloads when more than one instructor uploaded assignment at the same time. One professor from Balochistan reported she visits her colleagues' house for improved internet when available. Instructors reported discomfort and issues with the interface but improved as they gained practice. However, there were issues such as applications crashing, which led to instructors worrying about whether their submissions were being processed, and lost data and slow, or no technical support, left instructors grading assignments manually, as a temporary solution. Certain bilingual feature functionalities posed higher challenges. As an example, the assessment of output varied widely when students switched back and forth between Urdu and English with regards to AI usage and application.

Theme 4: Broadened Pedagogical Opportunities

The broadened pedagogical opportunities that teachers experienced from introducing AI were common themes during interviews with teachers. The time saved exploring avenues afforded the option of larger pedagogical risk taking. Several teachers modified their courses to include more formative assessments because they were confident AI would assist with the grading. One professor from Khyber Pakhtunkhwa added weekly writing assignments that were previously not possible because students would now have more practice. Moreover, many teachers were able to spend more time with individual students meeting their needs, and providing targeted assistance for the students that the AI identified as needing additional support. Similarly, a few teachers shared that they brought in pedagogy that they considered before, including flipped classrooms or project-based learning, simply because assessments were complicated. There were a few cautions regarding use of AI in assessment and appropriate evaluative decisions of humans; however, the focus for a lot of pedagogy had moved away from relying solely on high-stakes assessment. While considering these pedagogical moments, another notable section from the interviews was around the cultural and language frameworks of integrating AI.

Theme 5: Cultural and Linguistic Considerations

The integration of AI into coaching additionally uncovered demanding situations related to bilingualism in Pakistani better education. Faculty cited numerous pupil options in instructional writing: a few desired simplest Urdu, whilst others used each languages. Faculty valued AI's bilingual functions however stated that English exams seemed extra polished. They criticized AI's coping with of field-precise Urdu words, pointing out that human intervention turned into necessary. One Sindh professor defined AI remarks that supplied context however covered examples that clashed with nearby culture. This confirmed AI have to be higher tuned to help multilingual classrooms.

DISCUSSION

This study's findings imply that AI-automatic grading structures enhance turnaround time, comments quality, college workload, and workload control at Pakistani universities. Despite those benefits, a few obstacles remain. The remarks turnaround time dropped sharply from over 18 days to approximately three days—a transformative extrade with widespread pedagogical implications. Prompt remarks enables college students cope with misconceptions whilst the fabric is fresh, main to higher studying. Faculty mentioned better pride because of time stored in grading, as AI alleviates workload because of large instructions and heavy coaching loads. However, waiting for era on my own to remedy workload troubles is unrealistic. Success relies upon on placing a stability among pedagogical integration, technical infrastructure, expert development, and cultural alignment. Concerns approximately AI's reliability in grading complicated checks display the want for human judgment and hybrid models. AI's choppy overall performance with English and Urdu displays that maximum herbal language processing equipment are constructed for English, posing dangers of inequitable AI use.

Student overall performance progressed fairly and significantly, specially in non-stop tests in comparison to very last exams. This indicates that AI-powered remarks more often than not helps college students' ongoing mastering instead of their overall performance on high-stakes tests.

CONCLUSION

In Pakistani universities, using AI-powered automatic grading structures has proven each effectiveness and challenges. Among the challenges, computerized grading structures, which make use of AI technologies, provide promising answers to longstanding evaluation troubles. These troubles consist of receiving well timed comments, handling workload, and keeping consistency. Receiving comments in days in preference to weeks became a dramatic change. This ensured that grading and coaching practices align with the significance of well timed comments for scholar learning. Faculty workload became notably reduced, as computerized structures took over grading tasks. This allowed educators to cognizance on sports requiring their expertise, inclusive of operating with college students individually, innovating on pedagogy, and carrying out research. However, college participants document that AI grading structures aren't but dependable sufficient to absolutely automate grading processes. These structures lack the very last integration of grading and remarks, that is the maximum subjective aspect, as very last judgments nonetheless want to be made through educators. The handiest method moves a stability among the performance of AI grading and human involvement. Success relies upon on hybrid structures, expert support, a tradition that embraces innovation, and powerful coaching.

RECOMMENDATIONS

RECOMMENDATIONS

In mild of those findings, establishments must enforce a phased pilot method. This will permit gadget refinements and allow customers to alter to new practices earlier than complete implementation. Early pilot packages need to contain a various blend of disciplines and college with various stages of technological comfort. This technique will assist monitor contextual challenges. Institutions need to put money into reliable, high-pace net for sustainable AI operations. Any interruptions would require contingency planning. AI expert improvement should be complete and ongoing, now no longer only a one-time event. Programs ought to integrate era schooling with the instructional motives for evaluation layout and AI tools. Institutions want to offer responsive technical aid and clean problem-fixing frameworks. Continuous development of AI Urdu structures is important to make sure truthful evaluation in Urdu and different languages. Institutions have to set up clean rules to outline which exams are automatic and which require human oversight, particularly for high-stakes decisions. Monitoring of AI device overall performance and ongoing threat exams are vital. This will assist become aware of bias or unequal effect on precise scholar businesses and help in assembly dreams for fairness and quality.

REFERENCES

Ahmad, S. F., et al. (2021). "Artificial intelligence and its role in education." Sustainability **13**(22): 12902.

Ahmed, W., et al. (2025). "The Role of AI in Documenting Pragmatic Markers and Semantic Structures in Undocumented Pakistani Languages." Journal of Arts and Linguistics Studies **3**(2): 1891-1906.

Akintayo, O. T., et al. (2024). "Evaluating the impact of educational technology on learning outcomes in the higher education sector: A systematic review." International Journal of Management & Entrepreneurship Research **6**(5): 1395-1422.

Alasadi, E. A. and C. R. Baiz (2023). "Generative AI in Education and Research: Opportunities, concerns, and solutions." Journal of Chemical Education **100**(8): 2965-2971.

Alqahtani, T., et al. (2023). "The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research." Research in social and administrative pharmacy **19**(8): 1236-1242.

Chen, X., et al. (2022). "Two decades of artificial intelligence in education." Educational Technology & Society **25**(1): 28-47.

Crogman, H. T., et al. (2023). "Ungrading: The case for abandoning institutionalized assessment protocols and improving pedagogical strategies." Education Sciences **13**(11): 1091.

Dahri, N. A., et al. (2024). "Investigating AI-based academic support acceptance and its impact on students' performance in Malaysian and Pakistani higher education institutions." Education and Information Technologies **29**(14): 18695-18744.

Deepshikha, D. (2025). "A comprehensive review of AI-powered grading and tailored feedback in universities." Discover Artificial Intelligence **5**(1): 251.

Gnanaprakasam, J. and R. Lourdusamy (2024). The role of AI in automating grading: Enhancing feedback and efficiency. Artificial Intelligence and Education-Shaping the Future of Learning, IntechOpen.

Hamlin, L. (2021). "A model for graduate nursing faculty teaching workload." Journal of Professional Nursing **37**(2): 244-248.

Jiang, Y. and B. Li (2021). "Exploration on the teaching reform measure for machine learning course system of artificial intelligence specialty." Scientific Programming **2021**(1): 8971588.

Khurshid, S., et al. (2024). "Learning for an uncertain future: artificial intelligence a challenge for Pakistani education system in the era of digital transformation." Qualitative Research Journal.

Kizilcec, R. F. (2024). "To advance AI use in education, focus on understanding educators." International Journal of Artificial Intelligence in Education **34**(1): 12-19.

Messer, M., et al. (2024). "Automated grading and feedback tools for programming education: A systematic review." ACM Transactions on Computing Education **24**(1): 1-43.

Miller, E. E., et al. (2024). "Assessment of non-time-intensive changes to a management course to address course evaluation scores." Currents in Pharmacy Teaching and Learning **16**(12): 102170.

Naseer, M. A., et al. (2025). "Navigating the integration of artificial intelligence in the medical education curriculum: a mixed-methods study exploring the perspectives of medical students and faculty in Pakistan." BMC Medical Education **25**(1): 273.

Rasool, U., et al. (2022). "Written corrective feedback strategies: A systematic review." Voyage Journal of Educational Studies **2**(2): 67-83.

Saleem, T., et al. (2025). "Integrating AI in Pakistani ESL classrooms: Teachers' practices, perspectives, and impact on student performance." PloS one **20**(9): e0333352.

Saqlain, M. and A. Shahid (2024). "Assessing content and language integrated learning (CLIL) effectiveness in Tesol contexts." Journal of Applied Linguistics and TESOL (JALT) **7**(4): 113-125.

Tariq, M. U. (2024). AI-Driven Learning and Engagement in Higher Education, IGI Global.

Vetrivel, S., et al. (2025). Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments. Adopting Artificial Intelligence Tools in Higher Education, CRC Press**:** 41-61.

Zhai, X., et al. (2021). "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020." Complexity **2021**(1): 8812542.