Libyan Journal of Educational Research and E-Learning (LJERE)

Volume 1, Issue 1, 2025
Page No: 42-52
Website: https://ljere.com.ly/index.php/ljere/index

Artificial Intelligence and the Future of Assessment: Opportunities for Scalable, Fair, and Real-Time Evaluation

Abdulgader Alsharif *
Faculty of Engineering, Universiti Teknologi Malaysia, UTM, Skudai 81310, Johor, Malaysia

الذكاء الاصطناعي ومستقبل التقييم: فرص التقييم القابل للتطوير والعادل والفوريّ

عبد القادر الشريف *
كلية الهندسة، جامعة التكنولوجيا الماليزية، UTM، سكوداي 81310، جوهور، ماليزيا

*Corresponding author: habdulgader@graduate.utm.my

Received: December 29, 2024	Accepted: February 06, 2025	Published: February 11, 2025

Abstract

Artificial intelligence (AI) is rapidly transforming educational assessment by enabling automated grading, adaptive testing, and personalized feedback at unprecedented scale. Research suggests that AI tools can enhance assessment scalability by automating routine grading tasks and supporting large-scale testing programs, while also enabling real-time feedback to learners (e.g., Shute, 2008; U.S. Dept. of Ed., 2023). At the same time, AI promises to improve fairness by reducing human grader bias and standardizing scoring across diverse student populations. Early deployments range from computer-scored language proficiency exams to AI-driven peer review systems in MOOCs and adaptive formative platforms. Case studies in China, the United States, and international programs demonstrate these opportunities as well as significant challenges. On the one hand, studies find high agreement between AI and human scores in some contexts (e.g., ~92% agreement in a Chinese AI grading system) and large time savings (Gradescope claims 90% faster grading). On the other hand, rigorous evaluations (e.g., UNESCO and independent studies) have identified fairness, validity, and transparency issues. This paper reviews the theoretical foundations of AI-enabled assessment and surveys global examples of AI use in evaluation. We analyze how AI can provide scalable, equitable, and real-time formative and summative assessment, and we discuss practical limitations, ethical risks (including algorithmic bias, data privacy, and digital divides), and implementation challenges. Finally, we offer recommendations for policy and practice to guide responsible, effective integration of AI into education systems.

Keywords: Artificial Intelligence in Education, AI-powered Assessment, Automated Grading, Real-time Feedback, Educational Scalability, Algorithmic Bias, Fairness in AI, AI Proctoring.

الملخص

يُحدث الذكاء الاصطناعي تحولاً سريعاً في التقييم التعليمي من خلال تمكينه للتقييم الآلي والاختبار التكيفي والتغذية الراجعة الشخصية على نطاق غير مسبوق. تشير الأبحاث إلى أن أدوات الذكاء الاصطناعي يمكن أن تعزز قابلية توسيع نطاق التقييم من خلال أتمتة مهام التقييم الروتينية ودعم برامج الاختبار واسعة النطاق، مع تمكين التغذية الراجعة الفورية للمتعلمين (على سبيل المثال، شوت، 2008؛ وزارة التعليم الأمريكية، 2023). في الوقت نفسه، يَعِد الذكاء الاصطناعي بتحسين العدالة من خلال الحد من تحيز المُصحح البشري وتوحيد التقييم عبر فئات الطلاب المتنوعة. تتراوح عمليات النشر المبكرة من اختبارات الكفاءة اللغوية المُصححة حاسوبياً إلى أنظمة مراجعة الأقران التي تعتمد على الذكاء الاصطناعي في الدورات الجماعية المفتوحة عبر الإنترنت (MOOCs) ومنصات التكوين التكيفية. تُظهر دراسات الحالة في الصين والولايات المتحدة والبرامج الدولية هذه الفرص بالإضافة إلى التحديات الكبيرة. من ناحية أخرى، وجدت الدراسات توافقاً كبيراً بين الذكاء الاصطناعي والنتائج البشرية في بعض السياقات (على سبيل المثال، توافق بنسبة 92% تقريباً في نظام التقييم الصيني بالذكاء الاصطناعي) وتوفيراً كبيراً للوقت) تزعم Gradescope أن التقييم أسرع بنسبة 90 .(%من ناحية أخرى، حددت التقييمات الدقيقة (مثل تقييمات اليونسكو والدراسات المستقلة) قضايا تتعلق بالإنصاف والصلاحية والشفافية. تستعرض هذه الورقة الأسس النظرية للتقييم المدعوم بالذكاء الاصطناعي، وتستعرض أمثلة عالمية لاستخدامه في التقييم. نحلل كيف يمكن للذكاء الاصطناعي توفير تقييمات تكوينية وختامية قابلة للتطوير ومنصفة وفورية، ونناقش القيود العملية والمخاطر الأخلاقية (بما في ذلك التحيز الخوارزمي، وخصوصية البيانات، والفجوات الرقمية)، وتحديات التنفيذ. وأخيرًا، نقدم توصيات للسياسات والممارسات لتوجيه التكامل المسؤول والفعال للذكاء الاصطناعي في أنظمة التعليم.

الكلمات المفتاحية: الذكاء الاصطناعي في التعليم، التقييم المدعوم بالذكاء الاصطناعي، التقييم الآلي، التغذية الراجعة الفورية، قابلية التوسع التعليمي، التحيز الخوارزمي، الإنصاف في الذكاء الاصطناعي، مراقبة الذكاء الاصطناعي.

Introduction

Artificial intelligence has the potential to address some of the biggest challenges in education today and accelerate progress toward equitable learning outcomes (UNESCO, n.d.). UNESCO observes that AI can innovate teaching and learning practices, helping countries achieve the Sustainable Development Goal 4 of quality education (UNESCO, n.d.). Similarly, the World Bank highlights that AI in education is "transforming education at an unprecedented pace," offering new ways to *personalize* learning, support instructors, and improve educational management (Molina et al., 2024). Indeed, global institutions recognize AI as a powerful tool for customizing instruction and assessment to individual learner needs. For example, an OECD initiative aims to build a large-scale AI assessment model that captures each learner's individual growth over time and provides personalized diagnostic feedback, rather than simply norm-referenced scoring (OECD., n.d.). These aspirations reflect a shift toward learner-centered evaluation, where AI systems can enable formative assessment loops and timely intervention.

Despite these opportunities, AI's rapid emergence has raised concerns. Critics caution that algorithmic systems might *exacerbate* existing inequities if not carefully designed, for example by reinforcing biases or limiting students' control over learning (Etoile Partners., 2024). UNESCO reports that fewer than 10% of schools and universities worldwide have formal policies or guidance for the ethical use of AI, indicating a policy vacuum as AI tools spread into classrooms (UNESCO., 2023). In this context, rigorous research and policy attention are needed. This paper examines both the theoretical promise and the practical realities of AI-based assessment. We first review AI approaches to evaluation (automated scoring, adaptive testing, AI tutors, etc.), then explore opportunities for scalable, fair, and real-time assessment. We present case studies and global examples (Table 1) that illustrate how AI is being deployed. We then discuss ethical issues, limitations, and implementation challenges. We offer recommendations for research, policy, and practice to ensure that AI serves to *enhance* educational evaluation rather than undermine it.

AI in the classroom can facilitate active learning and formative assessment by monitoring student responses and offering hints or corrections in real time. For example, smart tutoring systems can analyze a student's problem-solving steps and prompt targeted questions, while automated essay-grading tools can highlight errors and suggest improvements to writing drafts. This integration of AI into everyday lessons enables *ongoing assessment for learning*, rather than relying solely on periodic tests (Litman et al., 2021). The theoretical benefits include increased engagement (students see the impact of feedback immediately) and better adaptation (the system adjusts to each learner's pace). Empirical studies of intelligent tutoring systems support these claims: meta-analyses find ITS approaches significantly more effective than traditional instruction on average (U.S. Department of Education, 2023), and many K-12 programs report higher mastery rates when AI mentors supplement human teaching. However, leveraging these benefits at scale requires careful design to ensure the AI's feedback is pedagogically sound and context-aware.

AI Assessment Technologies: Foundations and Typology

AI-driven assessment encompasses a range of tools that operate at different points in the learning process. Broadly, these include:

- **Automated scoring systems**: Algorithms that evaluate student work such as essays, short answers, or code. These use natural language processing, computer vision, or model analysis to assign scores and sometimes generate feedback. Well-known examples include ETS's e-rater for writing and Turnitin's AI tools. Hybrid models can combine rules-based features with deep learning. Research indicates that *hybrid approaches* may balance accuracy and interpretability. Nevertheless, automated scorers remain prone to errors in nuanced judgments and may need human oversight.
- **Computer-adaptive testing (CAT)**: Systems that adapt item difficulty to each test-taker's estimated ability, providing a tailored assessment path. While CAT existed before modern AI, new AI methods can improve item selection and skill inference (e.g., through Bayesian or reinforcement learning models). This yields more precise ability estimates in fewer items, which is highly *scalable* for large exams. CAT exemplifies how AI can make summative assessments more efficient.
- **Intelligent tutoring and formative analytics**: Platforms that continuously assess learners through tasks and activities, using AI to identify misconceptions and recommend content. These tools offer real-time formative feedback and guide personalized learning paths. For example, an AI tutor might analyze a student's solution to a math problem, detect a specific error, and provide a targeted tutorial on that concept. Studies show that real-time AI feedback can sustain student engagement and improve understanding (U.S. Department of Education, 2023).
- **Peer assessment and collaboration support**: AI can even enhance peer grading by selecting reliable peer reviewers, adjusting peer scores for consistency, and providing AI-generated feedback to supplement human

reviews. Recent reviews find that most studies report **AI improves the quality and fairness of peer assessment**, although many research gaps remain (Topping et al., 2025).

These technologies can operate during learning (formative) or after instruction (summative). Importantly, AI can also integrate with item banks and psychometric models to analyze large-scale assessment data, detect patterns (such as item drift or cheating), and continuously improve tests. By reframing assessment as an ongoing, interactive process, AI-based systems promise richer data on learning. Yet from a theoretical standpoint, this raises key questions about *validity* and *equity*. An AI system must demonstrate that its scores or feedback align with human judgments of competence, without introducing systematic bias. The next sections examine these opportunities and challenges in detail.

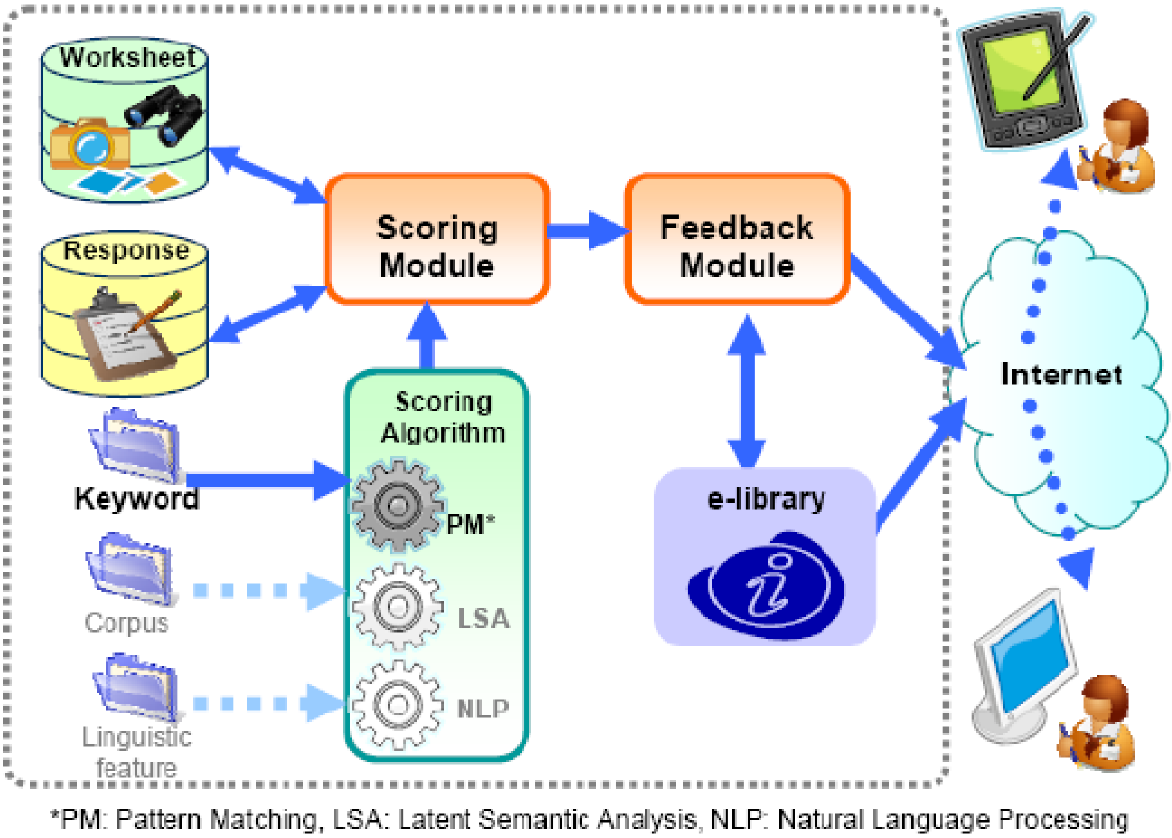

Figure 1 Architecture of an Automated Scoring System.

Opportunities: Scalability, Fairness, and Real-Time Feedback

AI's entry into assessment opens several major opportunities for educational systems:

Scalability and Efficiency

AI systems can grade or analyze student work at massive scale with minimal human labor. For example, automated essay scorers have been deployed to evaluate millions of student responses in national exams and classroom settings. In China, one AI grading project reportedly achieved *92% agreement* with human teachers in a 120 million-student trial. Similarly, automated mathematics and STEM code graders can instantly score homework for huge online courses. Tools like Gradescope (developed at UC Berkeley) claim to *reduce grading time by up to 90%* in large engineering and science classes (Wiggers, K., 2018). Such efficiency gains free instructors to focus on higher-level feedback and pedagogical tasks.

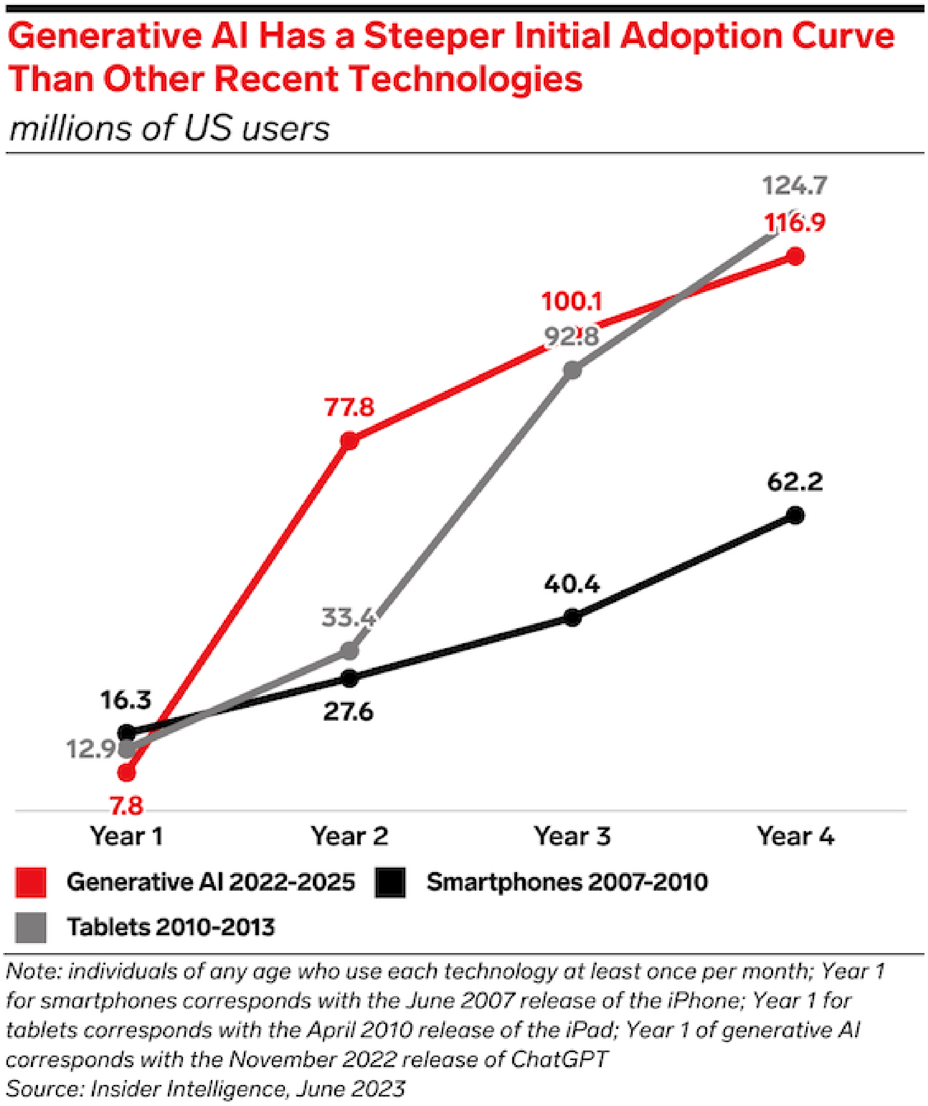

Figure 2 Illustrative Trend of AI Adoption in Educational Assessment (Source: Voicebot.ai, 2023).

Visualization of the growth of AI usage in education. Current data confirm rapid growth: for instance, UNESCO notes that chatbots like ChatGPT have reached 100 million users (fastest diffusion of any app) (UNESCO., 2023), and surveys indicate increasing interest in AI tutors. The scalability of AI is not only technical but also geographic: AI-based platforms can be deployed across many schools, languages, and countries without hiring more examiners. An OECD initiative explicitly aims to provide "in-depth analysis at both national and school levels" using AI, allowing for richer insights from diverse assessments (OECD., n.d.).

Fairness and Objectivity

A key promise of AI assessment is *reducing human bias and inconsistency*. When properly calibrated, an AI marker applies the same criteria to every student's work. In high-stakes testing (e.g., language proficiency exams), this can enhance reliability. A study of the Pearson PTE Academic English test found that candidates valued AI's consistency and reduced human bias (Leaton Gray, S., 2025). More broadly, automated scoring can ensure that all responses are judged by identical rubrics, which is hard to guarantee with multiple human graders. Adaptive testing also promotes fairness by adjusting difficulty to ability, so that each student has an appropriate challenge level.

However, fairness is a complex issue that goes beyond consistency. Recent research (Litman et al., 2021) specifically examined algorithmic fairness in automated essay scoring. They report that different AES models (feature-based, neural, hybrid) showed systematic biases: for instance, scores varied by students' race, gender, or socioeconomic background (Litman et al., 2021). In other words, an AI could inadvertently learn to favor writing styles or vocabulary more common to certain groups. In one analysis, even a well-known Chinese AES (iWrite) was found to fail validity for L2 learners, yielding scores that did not predict human judgments. These findings underscore that AI systems trained on historic data can inherit existing inequities. Thus, achieving fairness requires careful design: diverse training data, bias mitigation strategies, and rigorous fairness testing are essential.

When issues are addressed, AI may even improve equity. For example, adaptive quizzes can automatically present scaffolded practice to students who struggle, helping them catch up. AI tutors can provide extra support in under-resourced contexts (e.g., rural areas) where human expertise is scarce. If well-implemented, AI could help standardize education quality across schools. As Etoile Partners emphasizes, AI in education must meet "stringent fairness and trustworthiness standards" (as envisioned in the EU AI Act) to avoid "reinforcing biases or limiting opportunities based on opaque decision-making" (Etoile Partners., 2024). A balanced view is that AI *can* enhance fairness through consistency and personalization, but it also poses new fairness risks that require institutional oversight and policy safeguards.

Real-Time Feedback and Personalization

Traditional assessments often provide feedback only after grading is complete, limiting their formative value. AI enables immediate feedback on complex tasks. For instance, an AI system can analyze a student's drawn graph or solve a language-speaking task and instantly suggest corrections. Real-time feedback is crucial for engagement: educational theory shows that timely responses help students correct mistakes while the concept is fresh (U.S. Department of Education, 2023). U.S. ED's recent report lists "providing real-time feedback" as a core dimension of technology-enhanced assessment.

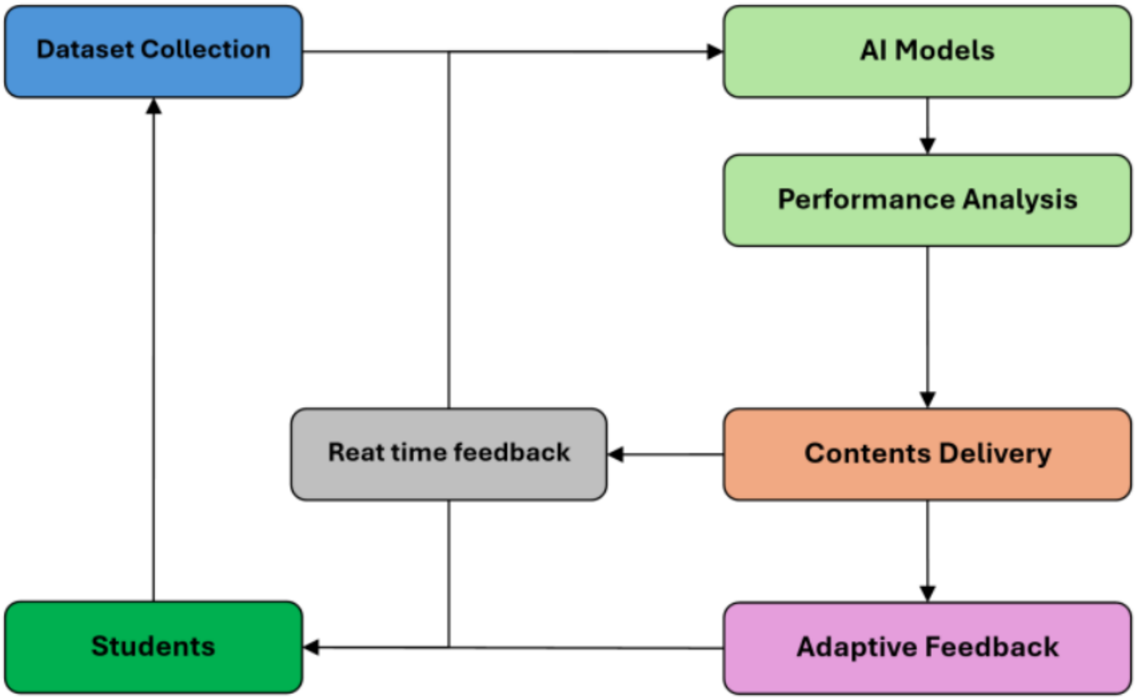

Figure 3 AI-Powered Learning Pathways and Real-Time Feedback.

Concrete examples abound. In online homework platforms, students often get instant scoring and hints generated by algorithms (e.g., auto-graded fill-in quizzes or AI tutors that monitor keystrokes). AI can guide a student through iterative drafts of an essay: systems may point out grammatical errors or logical gaps as soon as they occur, allowing immediate revision. In language learning, AI pronunciation checkers give on-the-spot feedback on phonetics and fluency. Even in large lectures, clicker questions scored by AI can yield class-level analytics that the instructor uses to adjust instruction in real time. The interactivity of AI-driven formative assessment transforms evaluation into a part of learning, rather than a separate endpoint.

Personalization goes hand in hand with real-time feedback. Because AI systems gather detailed data on each learner, they can tailor tasks dynamically. An AI-based quiz might notice a student consistently misses a concept and then serve more remedial questions on that topic. Such personalization can make assessment *scalable* in the sense that it adapts to many students simultaneously without extra teacher effort. Empirical studies of AI tutors (sometimes called "intelligent tutoring systems") show large learning gains when feedback is both immediate and customized to the learner's current understanding (e.g., Kulik & Fletcher, 2016).

Table 1 Examples of AI-based assessment applications worldwide. All examples are drawn from published studies or reports.

Country/Region	AI Assessment Tool	Context / Subject	Reported Outcome / Notes
China	Custom AES (English essays)	K-12 writing courses	~92% agreement with teachers in large trial; aims to reduce teacher grading time (Wiggers, K., 2018).
China	iWrite Automated Scoring	College English as L2	Provides instant holistic scores and feedback, but validation study found scores did not predict human ratings.
USA	Gradescope (AI-assisted grading)	University STEM courses	Claims to reduce grading time by ~90%, improving consistency. Used widely in large courses (Wiggers, K., 2018).
Global (Pearson)	PTE Academic (fully AI-scored test)	High-stakes English exam	AI scoring yields fast, consistent results; candidates value objectivity but some doubt nuances (intonation, etc.).
OECD project	AI Personal Diagnostic Tool	International testing (Pilot)	In development: merges AI with psychometrics to personalize OECD assessments. Supports formative use.
MOOCs / Online	Intelligent Tutoring Systems	Various subjects (math, languages)	Reported 10–30% improvement in learning outcomes compared to traditional practice (meta-analyses). Offers real-time hints.
Latin America	AI diagnostic reports (World Bank program)	Regional education systems	Emerging use: providing teachers with AI-generated analytics from student data to guide instruction.
Africa	Mobile learning AI tutor (pilot)	Early-grade math (e.g., Nigeria)	Improved engagement noted; connectivity and teacher training remain challenges. Data still limited.

Case Studies and Global Initiatives

Several notable initiatives illustrate how AI is currently used in educational assessment:

- **High-Stakes Exams (Pearson PTE)**: The Pearson Test of English (PTE) Academic is one of the world's largest AI-scored language exams. It uses automated scoring for speaking, writing, listening, and reading tasks. A recent UCL study of PTE found that candidates appreciated AI's objectivity and fast scores, yet many were unsure how speech nuances (like intonation or accent) were evaluated (Leaton Gray, S., 2025). The PTE case highlights both benefits (consistency, speed, scalability) and trust issues: examinees wanted more transparency on how spoken responses were rated. Pearson continues to refine its AI models and provide detailed feedback on test preparation, reflecting an iterative design process.
- **National Systems (China)**: China has aggressively deployed AI grading in schools. As VentureBeat (citing SCMP) reported, roughly 25% of Chinese schools (about 60,000) piloted a deep-learning essay-grading system by 2018. This system, trained on millions of scored essays, reportedly achieved ~92% agreement with human graders (Wiggers, K., 2018). The Chinese government's involvement and data resources make this a testbed for large-scale AI assessment. Complementing this, research in China (Qian et al. 2020) evaluated a commercial AES system (iWrite) and found it lacking validity for second-language writing (Litman et al., 2021), suggesting cautious use. These developments show both the scale of AI adoption and the need for independent validation.
- **K–12 and Higher Ed (USA)**: In the US, AI tools like Gradescope and various edtech platforms support teachers. For example, Gradescope uses machine learning to group similar problems and pre-grade them,

cutting human effort. Another avenue is AI-enabled formative platforms (e.g., DreamBox in math) that adapt lessons and quizzes in real time. Ongoing projects (often supported by NSF or Dept. of Ed) are also piloting AI analytics for student progress. The growing use of AI coexists with debates: some educators warn that over-reliance on algorithms may deskill teaching or lead to "teaching to the AI" (preparing students for machine graders). Pedagogical integration remains a challenge.

-

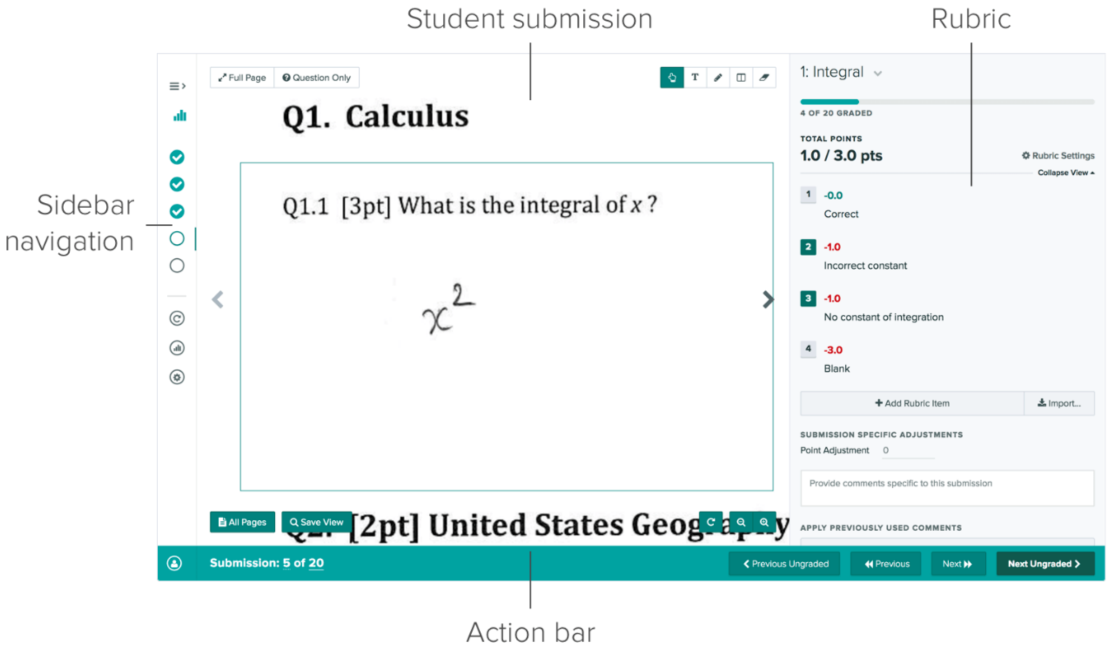

Figure 4 The image above illustrates how AI tools like Gradescope assist teachers by automating the grading of assignments. Using machine learning, Gradescope groups similar problems and pre-grades them, saving time and reducing manual effort. Such AI tools represent the growing integration of AI in education for grading and real-time feedback (Source: Gradescope).

- **International Collaboration (OECD and UNESCO)**: Organizations are proposing frameworks and tools. The OECD's ongoing project (Developing an International Large-Scale AI Tool) aims to produce an AI-enhanced assessment model for research and practice (OECD., n.d.). UNESCO, in its *AI and Education* guidance, stresses equity and inclusion. For example, UNESCO's recent survey (1,000+ institutions) revealed widespread use of AI by learners (e.g., ChatGPT) but a lack of formal policies(UNESCO., 2023). UNESCO's *AI Competency Frameworks* (2021–2023) also advocate teaching AI literacy to students and educators. These international efforts underscore a dual focus: leveraging AI's capabilities while establishing ethical and pedagogical guardrails.
- **Case Study (Peer Assessment Enhancement)**: An emerging trend is AI in peer review. The RIPPLE system (Canada/Europe) uses AI to calibrate peer grades and recommend high-quality resources based on student-generated content (Topping et al., 2025). By having AI assess the consistency of peer feedback and highlight discrepancies, such systems aim to maintain grading fairness in large classes. Though not yet mainstream, early studies (e.g., Khosravi et al. 2025) show AI can increase reliability of peer feedback and flag anomalous scores. This represents AI supporting the instructor rather than replacing the learner.

These examples (summarized in Table 1) show AI being applied at multiple levels: from classroom quizzes to national exams and global research programs. In many cases, pilot programs report significant time savings and engagement benefits, but they also note challenges in preparation. For instance, the PTE study found that students invest effort learning how to "optimize" responses for the AI grader, raising equity questions if some students can afford better coaching (Leaton Gray, S., 2025). Likewise, while Chinese AI graders accelerate feedback in remote areas, educators in China have raised concerns about the opacity of the algorithms (especially given some developers' military ties) (Wiggers, K., 2018). In sum, while AI assessments are technically powerful, their social impacts and contextual fit vary by setting.

Limitations and Ethical Concerns

Despite the promise of AI, several critical limitations and ethical issues must be addressed:

- **Bias and Fairness**: As noted earlier, AI models can encode biases present in training data. If a historical dataset under-represents certain student groups, the AI may score them unfairly. Also, subtle cultural or linguistic differences may be misjudged. Algorithmic transparency is thus crucial: students and educators should understand, at least at a high level, how decisions are made. Providing model "rationales" or allowing audits can build trust. Without accountability, AI could *exacerbate* inequities. UNESCO warns that large language models in education can entrench social biases if not properly governed (Etoile Partners., 2024). Regular *algorithmic audits* and bias testing (as advocated in the EU's AI Act) are necessary to ensure fairness.

- **Validity and Reliability**: AI grading must accurately measure what it intends to measure. Some AI approaches (especially those trained on proxy tasks) might rate writing for style rather than content mastery. In the iWrite case, the system's inability to generalize beyond its training data made it an unreliable predictor of human scores (U.S. Department of Education, 2023). Similarly, voice-analysis AI for speaking tests might over-emphasize accent or tone rather than language proficiency. Validity evidence (alignment with learning objectives, consistency across formats) must be gathered for AI tools just as for traditional assessments. AI's opaque nature makes this challenging: ensuring reliability may require hybrid models where difficult cases are flagged for human review.
- **Data Privacy and Security**: AI systems rely on student data (responses, behavior logs, biometrics). Safeguarding this data is an ethical imperative. Many current systems collect personal information for personalization; insufficient safeguards could lead to breaches or misuse. The EDDS partners highlight that AI in education is still "technologically immature… with serious privacy, security, and fairness risks" (Etoile Partners., 2024). In practice, many countries (unlike, say, the EU) lack clear regulations on educational data use. Without strong governance (encryption, anonymization, consent), students' privacy could be violated. Indeed, UNESCO reports that *only a minority of institutions have any policy* on AI use (UNESCO., 2023). As a result, there is a duty to build privacy protections into every AI assessment platform by design (e.g., using federated learning or opt-in data sharing).
- **Equity and Access (Digital Divide)**: Not all learners have equal access to the technology required for AI-based assessment. Students in rural or low-income schools may lack devices or high-speed internet. Educause notes that *paid AI tools could exacerbate the digital divide*, giving advantaged students further boost unless mitigated (Davidoff, M., 2024). For example, if an assignment is graded by a cloud-based AI, students without reliable internet or devices may struggle to submit or retrieve feedback. Efforts to broaden access (e.g., providing free AI platforms or offline-capable tools) are needed to prevent new inequalities.
- **Academic Integrity and Misuse**: Paradoxically, AI can both prevent and enable cheating. AI proctoring systems aim to detect dishonest behavior during online exams (e.g., unusual eye movement or device use), raising privacy and bias concerns (face recognition can be error-prone for darker skin tones). Meanwhile, generative AI like ChatGPT can produce essays or answers on demand, challenging traditional assessment validity. As students gain access to powerful AI writing assistants, educators must rethink what assessments can measure. Some institutions are already banning AI or redesigning tasks (oral exams, projects) to mitigate misuse. These issues are closely tied to fairness: not all students have equal familiarity with or access to AI tools. Policy must address AI literacy and responsible use, so that AI becomes a learning aide rather than a shortcut.
- **Teacher and Student Attitudes**: Finally, trust and acceptance can limit AI deployment. In the UCL-Pearson study, some test-takers found it "unsettling" that no human was grading them (Leaton Gray, S., 2025). Emotional and psychological aspects matter. If students or teachers distrust AI, they may reject useful tools. Building confidence may require human-AI collaboration models (e.g., AI suggests a grade that a teacher then confirms). Training teachers to interpret AI reports and integrating AI into pedagogy are nontrivial tasks that need support.

Implementation Challenges

Deploying AI-driven assessment at scale faces practical hurdles beyond the technology itself:

- **Technical Infrastructure**: Many educational institutions, especially in low-resource settings, lack the requisite hardware, bandwidth, or technical support. High-quality AI systems (e.g., speech graders) may require powerful servers and stable internet. Where infrastructure is weak, a "digital divide" will limit the reach of AI. Even simple AI tools often need recent devices; students using old smartphones or shared computers may not benefit. Solving this requires investment in connectivity and devices (a policy issue) as well as designing AI apps that are lightweight and offline-capable where possible.
- **Cost and Sustainability**: Developing robust AI models is expensive (data collection, computing, maintenance). While open-source tools exist, many offerings are proprietary. Schools in developing countries may struggle to afford licensing fees. Moreover, costs recur as models need updating (e.g., to include new languages or curricular changes). Projects like World Bank's AI education initiative highlight the need for sustainable financing models (open platforms, public-private partnerships) to ensure that AI tools do not become a luxury of wealthy regions.
- **Data Challenges**: AI depends on quality data (examples of student work, labeled by experts). Many languages and disciplines have scarce labeled datasets, limiting AI's applicability. In practice, AI systems

developed in one context (say, English essays) may not transfer to others (other languages or cultural contexts) without new data. This slows international scalability. Collaborative data-sharing initiatives could help, but they raise privacy and ownership questions.

- **Regulation and Standards**: Currently, there is no global standard for AI assessment tools. A school or test board may adopt one vendor's system without clear benchmarks of performance. The EU AI Act (coming into force 2025) will classify certain educational AI as "high-risk," requiring risk assessments and user information. Other regions have yet to follow suit. In the absence of regulations, institutions must self-regulate (e.g., establish review boards, pilot studies). This can be onerous and uneven. Harmonizing standards (perhaps led by UNESCO or OECD) would help guarantee minimum quality and fairness measures worldwide.
- **Human Capacity and Training**: Many educators lack familiarity with AI tools. Effective use of AI assessment requires training teachers and administrators to interpret AI outputs and integrate them pedagogically. Professional development programs are needed. Simultaneously, curriculum must adapt: students themselves should learn about AI (UNESCO's AI competency frameworks emphasize this (UNESCO, n.d). Without this, even the best AI tools may go unused or misused.
- **Cultural and Language Issues**: AI models often reflect the culture and idioms of their training data. For instance, an AI grader trained on Western essay samples might misinterpret an African or Asian student's rhetorical style. Ensuring cultural fairness requires local adaptation of AI systems. Similarly, testing for fairness must include multilingual and multicultural validations.

These challenges highlight that technology alone cannot fix educational problems. Implementation requires systemic change. Yet the breadth of interest (from local schools to international agencies) suggests momentum is building. Overcoming these hurdles will likely involve phases of pilot-testing, stakeholder feedback, and policy support (analogous to how national digital education initiatives have evolved).

Ethical Considerations

The ethical landscape of AI in assessment is multifaceted. Key principles (often derived from AI ethics guidelines) include transparency, equity, accountability, and student rights:

- **Transparency and Explainability**: Stakeholders (students, parents, educators) should know *how* decisions are made. "Black-box" scoring breeds mistrust. For example, if a student's essay receives an unexpectedly low grade, they and their teacher need some explanation (e.g., "Your analysis lacked textual evidence" or "Structural issues noted"). Some automated systems now include rationale components or allow users to see feature scores. Explainability is partly a technical challenge (interpretable AI) and partly a policy one (reporting standards).
- **Student Autonomy and Consent**: Students should have a say in how AI is used in their education. For instance, they might be able to opt out of AI exam proctoring or request human review of AI assessments. Automatic data collection for AI analytics raises the question of informed consent, especially for minors. Policies like the EU's GDPR and national privacy laws (e.g., COPPA in the U.S.) impose rules, but enforcement in schools is often weak. Ethical practice would involve clear consent forms and data rights.
- **Bias Mitigation**: Institutions must actively look for and correct biases. This includes not only testing data fairness (gender, race, disability) but also avoiding "construct misalignment." For example, over-emphasizing speed of response may disadvantage some learners (e.g., English learners or those with disabilities). AI designers should incorporate fairness definitions suitable for education contexts (such as equal success rates across groups) and involve diverse stakeholders in design.
- **Privacy and Security**: Ethical guidelines demand robust privacy protection. This means secure storage of assessment data, minimal data retention, and strong anonymization. The EDDS partners' scorecard highlighted that many AI systems currently score poorly on privacy standards (Etoile Partners., 2024). Ethical use requires compliance with laws, but also beyond-compliance practices (e.g., Privacy Impact Assessments for new AI tools).
- **Avoiding Unintended Consequences**: The history of educational reform warns that good intentions can backfire. For example, if an AI tool inadvertently pressures students (24/7 monitoring) or if reliance on AI reduces human teaching quality, the net effect may be negative. Continuous monitoring of outcomes is essential. AI tools should augment, not replace, the human judgment of educators.

The ethical framework for AI assessment is still maturing. UNESCO's Recommendation on the Ethics of AI (2021) and other guidelines stress fairness and inclusion in the deployment of educational AI (UNESCO, n.d.). However, the recent UNESCO survey indicates most institutions are "in the wilderness" regarding generative AI governance. Proactively establishing ethics committees, audit procedures, and accountability mechanisms is crucial as AI tools become integrated.

Policy and Practice Recommendations

To harness the benefits of AI assessment while mitigating risks, we offer the following recommendations:

1. **Establish Clear Guidelines and Standards**: National education authorities and international bodies (OECD, UNESCO) should define standards for AI assessment tools. These include technical benchmarks (accuracy, bias metrics) and process norms (pilot testing, data privacy protocols). Certification or quality labels for approved AI assessment platforms could guide adoption. For example, AI models might be required to pass independent validation against diverse test sets before deployment.
2. **Invest in Infrastructure and Access**: Governments must invest in digital infrastructure (devices, connectivity) and ensure equitable access to AI tools. This may involve subsidizing laptops/tablets and AI-enabled software in underserved schools. Public-private partnerships (e.g., with edtech companies) could provide scaled licensing. As Educause argues, making advanced AI tools freely available to low-income students can reduce the divide (Davidoff, M., 2024). Targeted funding is essential so that AI does not widen existing gaps.
3. **Prioritize Human–AI Collaboration**: Emphasize AI as a support for teachers, not a replacement. Professional development programs should train educators to interpret AI feedback and integrate it into instruction. Pilot programs should involve teachers from the outset, collecting their input on tool design. Encouraging co-design (teacher+developer partnerships) will ensure AI assessments align with pedagogical goals.
4. **Ensure Ethical Training and AI Literacy**: Curricula should include AI literacy for students and teachers. Understanding AI's capabilities and limits will help users engage critically. Ethics training (e.g., recognizing bias, data rights) is also important. UNESCO's AI competency frameworks advocate this dual focus (UNESCO, n.d.). In practice, schools might hold workshops on "Responsible AI use" just as they do for Internet safety.
5. **Support Ongoing Research and Evaluation**: Policymakers should fund academic research on AI in assessment. Long-term studies (beyond short pilots) are needed to measure impact on learning outcomes, equity, and student wellbeing. Data from deployed AI tools should be continuously analyzed. Collaboration between universities, industry, and schools (like the UCL–Pearson partnership) can generate evidence. Transparent publication of findings (including failures) will accelerate collective learning.
6. **Develop Inclusive Data Practices**: Encourage the creation of diverse, representative datasets for AI development. International consortia can share anonymized assessment data to improve AI models across languages and cultures. Such collaboration must respect privacy (e.g., federated learning models). Policy can mandate data sharing under safe frameworks to prevent siloed, biased AI.
7. **Regulate High-Stakes Uses Carefully**: AI used in critical exams should be subject to stricter oversight. Until ethical and validity issues are fully addressed, fully automated scoring for high-stakes certification should include human-in-the-loop checks. Hybrid scoring models (AI plus human rater on a subset of items) can balance efficiency and reliability. Institutions like ETS have long used such hybrid methods; new AI policies might require them for accountability.
8. **Promote Open and Transparent AI Solutions**: Whenever possible, use open-source or transparent AI tools in education. Closed proprietary systems hinder scrutiny. Governments and NGOs can fund open AI assessment platforms that allow auditing by researchers. This will help build trust and prevent "black box" scenarios.

These recommendations aim to guide responsible practice. The goal is not to slow innovation, but to align it with educational values. By embedding fairness and human agency at the core, AI can indeed broaden access to quality assessment.

Conclusion

Artificial intelligence offers transformative opportunities for educational assessment: it can make evaluation faster, more consistent, and more personalized. As this review has shown, AI-driven tools are already enabling large-scale automated grading, adaptive testing, and instant feedback in diverse contexts worldwide. When properly harnessed, these capabilities can enrich the learning process, support teachers, and free up resources for deeper pedagogical work.

However, reaping these benefits requires vigilance. AI assessment is still an emerging field with many open questions about validity, bias, and ethics. Evidence from studies (e.g., Litman et al., 2021; Qian et al., 2020) warns against blind trust in AI scores (U.S. Department of Education, 2023). Ethical principles demand transparency, fairness, and respect for student rights. Practical hurdles from digital access to teacher training must be overcome through thoughtful planning and investment.

Ultimately, the future of AI in assessment depends on how educational systems respond today. Policymakers, educators, technologists, and communities must collaborate to shape AI tools that empower all learners. With robust research, clear guidelines, and a focus on inclusion, AI can complement human judgment and creativity in the classroom, rather than replace them. If implemented wisely, AI-based assessment can contribute to *scalable,* *fair, and real-time evaluation* that benefits students globally. The stakes are high, but so is the potential: AI should be a catalyst for better education, ensuring that no learner is left behind in the digital age.

References

1. Etoile Partners. (2024, March 8). *Ensuring fairness in AI for education: A framework for trustworthy systems*. EDDS – Education Data & Digital Services. https://www.edds-education.org/post/ensuring-fairness-in-ai-for-education-a-framework-for-trustworthy-systems
2. Leaton Gray, S. (2025, April 3). *The future of AI in high-stakes testing: The fairness question*. IOE (UCL) Blog. https://blogs.ucl.ac.uk/ioe/2025/04/03/the-future-of-ai-in-high-stakes-testing-the-fairness-question
3. Litman, D., Zhang, H., Correnti, R., Matsumura, L., & Wang, E. (2021). A fairness evaluation of automated methods for scoring text evidence usage in writing. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021* (Vol. 12748, pp. 255–267). Springer. https://doi.org/10.1007/978-3-030-78292-4_21
4. Molina, E., Cobo, C., Pineda, J., & Rovner, H. (2024). *AI revolution in education: What you need to know*. World Bank. https://documents1.worldbank.org/curated/en/099734306182493324/pdf/IDU152823b13109c514ebd19c241a289470b6902.pdf
5. Hung, P. H., Hwang, G. J., Lin, Y. F., Wu, T. H., & Su, I. H. (2013). Seamless connection between learning and assessment-applying progressive learning tasks in mobile ecology inquiry. *Journal of Educational Technology & Society*, *16*(1), 194-205.
6. OECD. (n.d.). *Developing an international large-scale AI tool for educational assessment and personalised learning*. https://www.oecd.org/en/about/projects/developing-an-international-large-scale-ai-tool-for-educational-assessment-and-personalised-learning.html
7. Pearson Education. (2021). *Computer-based tests and machine marking: Candidates' perceptions and beliefs about the test taking experience* (M. Richardson, S. Leaton Gray, J. Popov, B. Maddox). [Unpublished report].
8. Voicebot.ai. (2023). *Generative AI adoption accelerates past smartphones and tablets*. Retrieved from https://voicebot.ai
9. Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's automated essay scoring system iWrite. *Journal of Educational Computing Research, 58*(4), 771–790. https://doi.org/10.1177/0735633119881472
10. U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. https://www.ed.gov/ai-report
11. UNESCO. (n.d.). *Artificial intelligence in education*. UNESCO. Retrieved July 1, 2025, from https://www.unesco.org/en/digital-education/artificial-intelligence
12. UNESCO. (2023, June 1). *UNESCO survey: Less than 10% of schools and universities have formal guidance on AI*. UNESCO News & Events. Retrieved from https://www.unesco.org/en/articles/unesco-survey-less-10-schools-and-universities-have-formal-guidance-ai
13. Wiggers, K. (2018, May 28). *Chinese schools are testing AI that grades papers almost as well as teachers*. VentureBeat. https://venturebeat.com/ai/chinese-schools-are-testing-ai-that-grades-papers-almost-as-well-as-teachers/
14. Davidoff, M. (2024, November 11). *How broadening AI access can help bridge the digital divide*. EDUCAUSE Review. https://er.educause.edu/articles/2024/11/how-broadening-ai-access-can-help-bridge-the-digital-divide
15. Topping, K. J., Gehringer, E., Khosravi, H., Gudipati, S., Jadhav, K., & Susarla, S. (2025). Enhancing peer assessment with artificial intelligence. *International Journal of Educational Technology in Higher Education*, *22*(1), 3.
16. Gradescope. (n.d.). *Grading submissions*. Gradescope. Retrieved May 13, 2025, from https://guides.gradescope.com/hc/en-us/articles/22249389005709-Grading-Submissions