A Critical Review of Artificial Intelligence for Assessment: From Promise to Practice

Wing Cheung Tang
BEng, MEd, PCEd, MSc, MBA, MA, PhD, RIOP, AHKPS, CMgr, FCMI, FIMA, FIMC
Head of Mathematics, Lock Tao Secondary School, Hong Kong

DOI: https://doi.org/10.51584/IJRIAS.2026.11050050

Received: 01 May 2026; Accepted: 06 May 2026; Published: 27 May 2026

ABSTRACT

The incorporation of Artificial Intelligence (AI) into evaluation processes in education, recruitment, and research has significantly accelerated, propelled by advancements in large language models (LLMs) and machine learning. This critical review amalgamates evidence from systematic reviews, empirical studies, and ethical frameworks published from 2018 to 2026 to assess AI's role in evaluation. The review looks at the technical state of AI-based assessment, talks about the problems of algorithmic bias and fairness, points out the epistemological limitations of machine scoring, talks about the gaps in regulatory and ethical accountability, and thinks about the paradox of using AI to find AI-generated content. Our analysis finds that while AI can grade consistently and quickly, the evidence for its accuracy is shaky. Most performance tests for AI (benchmarks) are not statistically sound, efforts to reduce bias are fragmented, and current quality standards fail to address AI's unique failures. Bias mitigation efforts are disjointed, and current assessment quality criteria do not adequately address AI's distinct failures.

Keywords: artificial intelligence, assessment, algorithmic bias, automated grading, large language models, ethical frameworks, human oversight

INTRODUCTION

The world of assessments is changing in a big way. Artificial Intelligence (AI) is being used to judge human performance on a scale never seen before, from automated essay scoring systems in universities to algorithmic resume screeners in hiring. The promise is appealing: machine speed, personalised feedback, and efficiency, consistency, and objectivity. But behind this promise is a complicated set of problems that have not been solved yet, such as algorithmic bias, lack of transparency, validity crises, and basic questions about what it means to judge human ability.

This critical review looks at AI's role in three areas of assessment: grading and giving feedback to students (Gnanaprakasam & Lourdusamy, 2024), screening candidates for jobs (Hawrysz, 2025), and peer review and evidence synthesis in research (Flemyng et al., 2025). The emphasis is on AI as an evaluator—systems that render or facilitate evaluative assessments of human labour.

There are five main parts to the review. Section 2 sets the standard for the technical state of the art by looking at how AI systems evaluate and the proof that they work. Section 3 looks at algorithmic bias and fairness by using systematic reviews of bias sources and ways to reduce it. Section 4 critically analyses profound epistemological issues: the obscurity of AI decisions, the constraints of proxy metrics, and the systemic weaknesses that compromise validity assertions. Section 5 talks about regulatory and ethical accountability and looks at new ways to evaluate AI in a responsible way. Section 6 examines the contradictory challenge of employing AI to identify AI-generated content within the framework of academic integrity. The conclusion brings together the results and sets priorities for future research and practice.

The main point is that AI can help with testing in real ways, but it has been adopted faster than scientists can prove it works. The field has a legitimacy gap because claims that AI is better than other technologies are often based on benchmarks that don't really measure what they say they do, and evaluations of AI's own performance

(Burnell et al., 2023) are still methodologically weak. We need a major change right away, one that puts human judgement, openness, and context over automation for automation's sake.

The Technical State of AI in Assessment: Capabilities and Limitations

A Landscape in Rapid Expansion

Testing with AI has become much more common. A comprehensive review of 49 peer-reviewed studies on Large Language Model-Powered Automated Assessment (LLMPAA) published from 2018 to 2024 revealed significant applications in reading comprehension, language education, and computer science, primarily utilising essay and short-answer formats. A narrative review of 77 fundamental studies conducted from 2018 to 2025 in higher education revealed that AI technologies possess significant potential for automating assessment processes, ensuring consistent grading, and delivering personalised feedback (Kuzminykh et al., 2024). A bibliometric and systematic literature review of 86 Scopus-indexed publications revealed a rapid growth rate of 32.0% per year, primarily centred on themes including ChatGPT, machine learning, and learning analytics.

A scoping review of 131 articles in medical education indicated that 74.0% originated from nations with a notably high Human Development Index, with ChatGPT identified as the most scrutinised model (n=119). In undergraduate medical education, applications now encompass self-assessment, autonomous tutoring, and the creation of assessments, grading, clinical evaluations, assessments of procedural skills, and predictive analytics. There are a lot of different applications which show real interest, but it also makes us wonder if systematic evaluation has kept up with deployment.

Automated Essay Scoring and Short-Answer Grading

Automated essay scoring (AES) (Ramesh & Sanampudi, 2022) and automated short-answer grading (ASAG) (Bonthu, Rama Sree & Krishna Prasad, 2021) are the most advanced uses of AI in testing students. AES systems use natural language processing to compare written work to standards. They can handle thousands of papers at once. Older AES models used basic language features like word frequency, sentence length, and syntactic patterns. But new transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT) (Alaparthi & Mishra, 2020) have made it easier to understand context.

The reported concordance between AI and human evaluators fluctuates significantly. Some studies show almost perfect agreement (quadratically weighted kappa = 0.99, correlation = 0.95), while others show much lower agreement (intraclass correlation = 0.45, correlation = 0.38). This wide range of results is useful because it shows that AI's performance depends a lot on the context, the design of the prompt, the quality of the rubric, the training data, and the model's specific abilities. It does not make sense to make broad statements about "AI's grading accuracy" without giving these contextual factors.

Although NLP-based assessments (Nikhil, Annamalai & Jayapal, 2025) offer swift evaluations of content, grammar, and structure, they consistently struggle with interpreting nuanced language and creative expression. Systems trained on standardised writing samples systematically undervalue non-standard dialects, creative risk-taking, or culturally specific rhetorical strategies. Models that are optimised for statistical regularity often miss the very things that make human writing unique, like idiosyncrasy, innovation, and voice.

The Benchmark Crisis

A recent large-scale systematic review of AI benchmarks themselves may be the most damning criticism of AI assessment. The Oxford Internet Institute led a study that looked at 445 AI benchmarks, which are standardised tests used to compare and rank AI systems. The study involved 42 researchers from top global institutions. The results are very worrying. Only 16% of the studies that were looked at used statistical methods to compare the performance of models. This means that the differences between systems that were reported might not be real improvements, but just random chance. About half of the benchmarks tried to measure abstract concepts like reasoning or harmlessness without clearly saying what those words mean.

Andrew Bean, the lead author, says, "It's hard to tell if models are really getting better or just looking better without shared definitions and good measurement." The study says that if benchmarks are not based on sound science, they could give developers and regulators a false idea of how safe or capable AI systems really are. This criticism directly pertains to assertions regarding AI's efficacy in assessment: if the benchmarks employed to validate AI grading systems lack construct validity, then assertions of AI's superiority over human grading are founded on precarious grounds.

The review also found certain diseases. Some benchmarks mix up task performance with formatting compliance. For example, a model might solve a logic puzzle correctly but not meet complicated formatting requirement, which makes it look worse than it really is. Other tests show that the model's performance is fragile. For example, it might do well on short, simple math problems for kids, but if you change the numbers or words a little, it fails. This suggests that it is memorising patterns instead of really understanding them. These failure modes are not just interesting technical problems; they have real-world effects on educational assessment, where students may be punished not for getting the wrong answer but for not following the model's expected format or wording patterns.

The tests we use to measure how "smart" an AI is are often flawed. Many don't even use basic statistics to compare AI to humans. Others claim to measure "reasoning" but never define what that means. This means a company claiming their AI grader is "more accurate than a teacher" is likely using a broken test.

Benefits Acknowledged, Not Overstated

Just to be clear, AI does have real benefits. AI-assisted assessment tools make grading more consistent (Najafi et al., 2025), cut down on human bias, support automated feedback, and speed up the assessment process. Some of the benefits are less work for teachers, shorter waiting times, and better learning experiences thanks to instant, personalised feedback. These benefits are real and useful, especially in classes with a lot of students where it is not possible to grade subjective responses by hand. The issue is not whether AI is useful (clearly it is) but rather the conditions, safeguards, and legitimacy under which its judgements ought to be accepted.

Emerging AI Technologies

(a) Multimodal AI -- Recent advancements in artificial intelligence models, such as GPT-4 with vision and Gemini, demonstrate the capacity to process and interpret diverse modalities of input beyond conventional text. This capability extends the potential applications to include the automated evaluation of handwritten mathematical expressions, visual diagrams, or video-recorded demonstrations of clinical competencies. Conversely, this development introduces novel sources of bias; for instance, misinterpretation of handwriting characteristics prevalent in particular cultural contexts or associated with dyslexic individuals.

(b) Agentic Workflows -- Rather than relying on a monolithic model for score generation, subsequent assessment architectures might integrate 'AI agents.' These agents are specialized sub-modules, focusing on discrete tasks such as grammatical analysis, argumentative coherence, or creative expression, designed to interact collaboratively and critically with one another. Such a modular approach has the potential to enhance system explainability by enabling an argumentation-focused agent, for example, to pinpoint specific deficiencies. Nevertheless, this architectural shift inherently escalates both system complexity and associated computational overhead.

(c) Small Language Models (SLMs) -- In contrast to the prevailing trend of large language models (LLMs), there is an emergence of more constrained, task-dedicated models capable of execution with enhanced transparency on localized computational platforms. These Small Language Models (SLMs) potentially present an avenue for developing more interpretable and auditable assessment frameworks, given that their internal operational states are inherently more amenable to analysis compared to the opaque nature of models comprising billions of parameters.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

Systematic Problems and Partial Solutions

The Reproduction of Inequality

The promise of AI-driven assessment is based on the idea that machines are not biased, which means they don't make decisions based on personal feelings. This assumption is not true. AI systems can unintentionally encode and exacerbate biases inherent in educational data, resulting in unjust or discriminatory outcomes. A systematic literature review on fairness in educational machine learning observes that recent studies demonstrate that automated assessment tools can perpetuate or exacerbate inherent biases (Oberhauser, 2025).

A systematic literature review of machine learning algorithms in educational contexts revealed prevalent sources of bias: biased datasets, non-transparent decision-making processes, and exclusionary employment criteria if systems are inadequately designed and supervised. The applications examined encompass student dropout prediction, performance forecasting, forum post categorisation, and recommendation systems. The review found that some common ways to reduce bias are changing sample weights, using bias attenuation methods, making things fair by being aware or not aware, and learning through competition.

Fairness Metrics and Their Limits

Area Between Receiver Operating Characteristic curves (ABROCA), group difference in performance, and different disparity metrics are some of the most common ways to measure fairness. But the rise of metrics has not led to agreement. The FairAIED survey, a thorough systematic review that connects technical fairness research with educational uses, says that current surveys either look at algorithmic fairness without an educational setting or focus on educational methods while ignoring fairness. Their integrated framework looks at the sources of bias, what fairness means, how to reduce bias, how to evaluate resources, and ethical issues.

There are a lot of real-world problems to deal with. The FairAIED review (Chinta et al., 2024) looks at problems like censored or partially observed learning outcomes and the ongoing problem of figuring out how to measure and deal with the trade-off between fairness and predictive utility. Most studies found that there is no strict trade-off between fairness and accuracy. This means that algorithms that are well-designed can do both. This finding is hopeful, but it comes with a caveat: to get both fairness and accuracy, you need to plan, not just use what you have.

Domain-Specific Evidence: Language Assessment and Hiring

There is more evidence that algorithms are biased in certain testing situations. A qualitative study investigating ethical dilemmas in AI-driven language assessments revealed that automated speech recognition systems and text evaluation algorithms yielded erroneous scoring, especially for learners with non-dominant accents or varied linguistic backgrounds. This discovery has direct ramifications for educational equity: students who speak English as a second language or whose dialects deviate from standardised norms may be systematically disadvantaged by AI writing assessments.

The evidence is just as bad when it comes to hiring. A benchmarking study (Zhao et al., 2023) comparing state-of-the-art LLMs with a domain-specific hiring model on a dataset of about 10,000 real-world candidate-job pairs discovered that general-purpose LLMs had much lower predictive accuracy (ROC AUC 0.77 vs. 0.85) and much worse fairness outcomes. The best LLM had a minimum race-wise impact ratio of 0.809, while the best domain-specific model had a ratio of 0.957 (where 1.0 means perfect parity). The authors of the study warn against using off-the-shelf LLMs for important hiring tasks without strong fairness protection. They say that the bias comes from pretraining biases that keep social inequalities going.

A different study on bias in AI-driven HRM systems (Bandara et al., 2025) found that biased data sets, unclear decision-making processes, and hiring standards that leave some people out can have a bigger impact on groups that are already at a disadvantage, such as non-binary people, women, racial minorities, and people with

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

disabilities. The study points out holes in the rules that let bias stay even though there are legal and moral rules for AI in HRM.

Intersectional and Understudied Dimensions

The literature on fairness has mostly looked at gender and race, leaving other demographic traits unexplored. A systematic review on debiasing education algorithms suggests broadening the focus beyond gender and race to encompass additional demographic characteristics and evaluating the effects of equitable algorithms on end users, as human perceptions may not correspond with algorithmic fairness standards. This is an important point: even if an algorithm is mathematically fair according to technical standards, students or candidates whose work is being judged may not think it is fair. People may lose faith in assessment systems if they think they are unfair, even if they follow the rules for fairness.

It is very important to check the fairness of data and features before checking the fairness of algorithms. Bias mitigation at the algorithmic level cannot make up for inputs that are biased. The algorithm will reproduce historical inequalities if the training data reflects them, even if corrections are made after the fact. This means that making AI assessments fair requires more than just fixing technical problems at the model level. It also means paying attention to how data is collected, how it is labelled, and the social and technical context in which assessments are used.

Epistemological and Methodological Critique

The Opacity Problem

There are more problems than algorithmic bias. A more in-depth criticism has to do with the basic epistemology of machine assessment: what does it mean for a system to "know" if a piece of work is good or bad? Most AI assessment systems are black boxes. Even when explainable AI methods are used, the explanations are often justifications that come after the fact and may not accurately show how the model made its decisions.

A common theme in systematic reviews is the lack of transparency. A thorough examination of AI-driven grading in universities revealed that a lack of transparency (Ahmed et al., 2025) is a major problem, along with algorithmic bias and worries about data privacy. The narrative review concludes that AI-driven assessment tools have the potential to change higher education for the better, but they need to be carefully combined with human expertise, strong ethical guidelines, and ongoing validation to work. Transparency is not just a technical requirement; it is also necessary for trust and responsibility.

The issue is made worse by the fact that the best models are owned by companies. When educators and students use commercial AI systems for testing, they cannot see how the system makes decisions, figure out why certain scores were given, or make sure the system is working as it should. This lack of clarity is fundamentally at odds with the openness needed for a fair assessment, especially in high-stakes situations where scores affect academic progress or job prospects.

Validity and Proxy Measures

Validity is a basic idea in educational and psychological measurement. It is the degree to which a test measures what it says it does. AI assessment systems bring up new questions about validity. Standard AES systems use simple lexical data, like how often words are used and how long sentences are, to grade essays. These characteristics might be associated with overall writing quality, but they serve as proxies rather than direct indicators of the constructs of interest (e.g., critical thinking, argumentation, creativity). A hybrid model that combines shallow linguistic features, discourse patterns, and neural context embeddings may make things better, but it doesn't solve the main problem: the model is still learning statistical patterns instead of understanding the content.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

This difference is not just for school. In a narrow technical sense, a model that gives high scores to formulaic, predictable writing and low scores to innovative, unconventional writing is not "wrong". It is correctly optimising the patterns in its training data. But if the training data follows the rules of normal writing, the model will always undervalue creativity and new ideas. The validity problem is not a technical issue but rather an inherent characteristic of statistical learning when applied to human expression.

The Absence of Critical Thinking Assessment

One of the most surprising things in literature is what is not being looked at. A recent scoping review (Simoni et al., 2025) of AI in undergraduate medical education, which examined 310 publications from a pool of 3,238 identified works, revealed that none evaluated AI's influence on critical thinking or clinical reasoning among medical students.

The review shows that AI is being used a lot more in undergraduate medical education (UGME), which has both pros and cons. However, it is still not clear what the best evidence is for its use. Some of the most important things that need to be done are to define AI skills, teaching methods, and moral rules.

The lack of research on critical thinking assessment is not an oversight; it signifies a genuine limitation of existing AI systems. You can't just match patterns to think critically. It entails metacognition, contextual evaluation, the assessment of conflicting evidence, and the acknowledgement of personal fallibility. These are exactly the abilities that even the most advanced statistical models lack. Using AI to test critical thinking could mean measuring something else and calling it critical thinking.

The Accuracy–Fairness–Explainability Trilemma

The previous sections show that there is a trilemma. AI assessment systems must make trade-offs between three important traits: accuracy (how well they agree with human judgements or the truth), fairness (how well they do not show systematic bias across demographic groups), and explainability (how clear their decision-making processes are) in Figure 1. No system can perfectly meet all three needs at the same time.

Think of it like a project manager: you can have it fast, cheap, or high-quality, but you can only pick two. With AI assessment, you can prioritize accuracy, fairness, or explainability, but building a system that excels at all three is impossible given current technology.

(a) Systems that maximise accuracy (like fine-tuned BERT models with almost perfect human agreement) may not be fair because they learn and amplify biases that are already in the training data.
(b) Fairness-optimized systems (like those with adversarial debiasing or equalised odds constraints) might give up accuracy or explainability as the model gets more complicated and harder to understand.

(c) Explainable systems, like linear models or decision trees, usually lose accuracy because they can't capture the complex patterns in human writing.

This trilemma is not a short-term technical problem that needs to be fixed. It is a structural constraint resulting from the characteristics of statistical learning in the context of human evaluative judgements. Understanding this trilemma has real-world consequences: different testing situations need different trade-off setups.

Low-stakes formative feedback may emphasise explainability; high-stakes standardised testing may emphasise fairness; research applications may emphasise accuracy. It's wrong to think that one configuration works for all situations.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

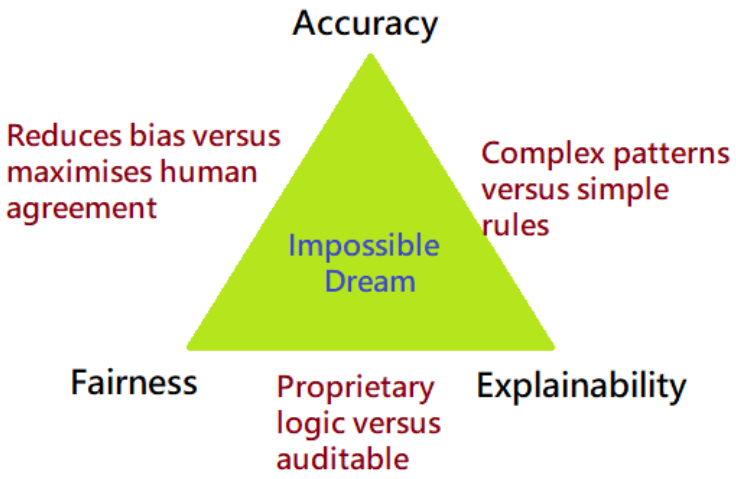

Figure 1: Accuracy-Fairness-Explainability Trilemma

Ethical Frameworks and Regulatory Accountability

Emerging Certification Standards

As worries about AI ethics grow, several frameworks for responsible AI assessment have come up. The IEEE CertifAIEd program is a globally recognised way to certify autonomous intelligent systems. It looks at twenty-six ethical values, such as fairness, trust, dignity, and avoiding discrimination. Assessors use the ethical risk profile to figure out which of the four main criteria suits (transparency, accountability, algorithmic bias, or privacy) the product fits into. The program is very similar to global rules like the EU AI Act, which says that risk assessments should be based on "appropriate technical tools and benchmarks."

At the AAAI/ACM Conference on AI, Ethics, and Society, a new AI auditing tool was introduced that looks at systems in several important areas, including fairness, explainability, robustness, transparency, bias, sustainability, and compliance with the law. This framework lets you look at multiple dimensions at the same time, unlike other tools that only look at one dimension at a time. This makes AI deployment completer and more accountable. A validation study conducted across multiple countries established a seven-dimensional AI Responsibility framework (Jaturat, Na-Nan & Hu, 2026) that includes privacy and security, transparency and accountability, employment impact, sustainability, user-centred design, social impact, and innovation and adaptation.

These frameworks are real steps forward, but they are not always put into practice in the same way. Certification is optional, enforcement is weak, and many schools and companies that hire people don't have the time or money to do regular ethical audits of their AI assessment tools. There is a big difference between how many frameworks are available and how many people use them.

Human Oversight as a Non-Negotiable Requirement

One recommendation that comes up in all the systematic reviews is that people need to be in charge. The LLMPAA systematic review (Emirtekin, 2025) finds that LLMPAA technologies have potential, but their effectiveness depends on the situation. Human oversight is necessary to make sure that the results of assessments are fair and reliable. The review of AI-based grading in universities comes to the same conclusion: for it to work, it must be carefully combined with human knowledge.

There are many reasons why people need to watch over things. It protects against mistakes and hallucinations in algorithms. It lets people make judgements based on the situation which AI systems can't do. It keeps people accountable by making sure that human decision-makers are still responsible for the results of assessments,

rather than letting technical systems take over. And it keeps the educational relationship between teachers and students, which is based on relationships, not transactions.

The practical application of human oversight differs according to context. In automated grading systems, a mixed method that marks unclear answers for teacher review looks like it could work well. A study using a new NLP-based automatic grading system (Ayaan & Ng, 2025) used rule-based threshold logic to give zero, partial, or full marks based on the semantic score and word count. It also flagged answers that were unclear for the teacher to look over. This "teacher-in-the-loop" model is a responsible middle ground between fully automating and not automating at all.

Regulatory Gaps and Recommendations

Even though people are starting to realise that regulation is necessary, there are still big holes. An inquiry into bias within AI-driven HRM systems reveals regulatory deficiencies that enable bias to endure, notwithstanding the existence of legal and ethical frameworks governing AI in HRM. The study offers implementable strategies for enhancing inclusivity in AI-driven HRM practices, such as frequent audits, clear documentation, and stakeholder engagement.

The literature suggests several recommendations: long-term evaluations of AI assessment tools in classrooms; the creation of culturally responsive AI models; a more thorough examination of ethical risks, including bias and data privacy; the combination of AI-guided assessment frameworks with human-led pedagogical judgement; and ongoing validation to guarantee fair and effective educational results. A systematic review on debiasing education algorithms suggests that fairness of data and features should be evaluated prior to algorithmic fairness, that the scope of education fairness studies should be broadened beyond gender and race, and that the effects of fair algorithms on end users should be assessed.

The problem is that suggestions are not orders. Voluntary compliance will not be enough if there are no standards that can be enforced and ways to hold people accountable. The lesson from other fields, like financial auditing and medical device regulation, is that real change needs independent third-party assessment and legal responsibility for harm. It is still not clear if such mechanisms will be created for AI assessment.

AI Assessing AI-Generated Content

The Integrity Crisis

The rise of generative AI has made it harder than ever to keep academic integrity. A study from the UK's Higher Education Policy Institute in 2025 found that 88% of students had used generative AI in the past year to help them finish their homework (Freeman, 2025). There are worries about the validity of many current assessment methods in higher education because students could use ChatGPT and similar tools to finish those assessments.

The integrity crisis is more than just cheating. It concerns the essential legitimacy of evaluation. If students can write passing essays with little effort or understanding, then the test isn't measuring what it says it is. Conventional assessment formats (take-home essays, standardised problem sets, reflective writing) are becoming more susceptible.

The Detection Arms Race

The answer has been clear: the creation of AI detection tools (Chaka, 2024) that can find text that was made by AI. This is a paradox, though: AI is now being used to find AI, and both sides of the evaluation process are becoming automated. The arms race dynamics are like those in plagiarism detection, but the stakes are higher because the technology behind it is changing quickly.

There are limits to what the detection tools can do. They make mistakes by flagging human writing as AI-generated and missing AI-generated content. For people who don't speak English as their first language and for

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

shorter texts, they are less reliable. And you can get around them by using paraphrasing tools, translation, or just changing the way things are laid out. The detection approach may be the most important thing because it puts the burden of proof on students, which makes things adversarial and breaks down trust.

Assessment Redesign as the Sustainable Response

The long-term answer to generative AI is not to improve detection but to improve the design of assessments. Researchers at Singapore Management University investigated how well ChatGPT did on spreadsheet modelling problems (Cheong, 2025). They found that tests can be made GAI-resistant by requiring application, analysis, evaluation, and creation (the higher cognitive levels in Bloom's Taxonomy) rather than just simple recall and understanding of facts. Problems that are unique to a situation and can't be solved just by matching patterns are still strong enough to handle AI help.

The main point is that AI makes us think again about what tests are for. If the goal is to see how well students can give formulaic answers to questions they know are coming, then AI will do better than students, and the test is useless. AI can be a tool instead of a threat if the goal is to measure higher-order thinking, real-world performance, and contextual judgement. However, this is only true if the assessment tasks are set up in this way. This necessitates pedagogical creativity rather than technical remedies.

The Deep Contradiction

There is a big contradiction in the talk about AI and academic integrity as shown in Figure 2. On the one hand, schools are making tools to find AI and rules to stop students from using it. But those same schools are also using AI increasingly to grade student work. It is very unfair. Students are punished for using AI to do their work, but teachers are encouraged to use AI to grade it. This asymmetry is not necessarily incoherent (different roles have different responsibilities) but it does raise questions about consistency and legitimacy. If AI is not reliable for testing, then why is it okay for teachers to use it? If AI is trustworthy, why are students punished for utilising it? The solution is to make a clear difference between generation and evaluation, but in practice, this difference is often not clear, and the public conversation rarely talks about the tension.

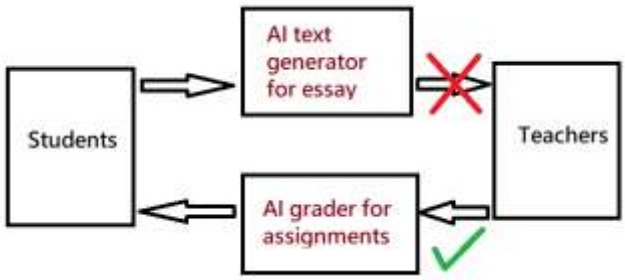

Figure 2: Integrity Paradox

CONCLUSION

This critical review has looked at AI in assessment from the points of view of technical, ethical, epistemological, and regulatory issues. The evidence supports a nuanced position: AI can be a useful assistive tool for assessment, particularly for low-stakes, high-volume tasks where consistency and efficiency are paramount. Assertions regarding AI's superiority over human judgement are exaggerated and frequently based on methodologically deficient standards. There are still big problems with bias, lack of transparency, and validity, and there isn't enough research on how to test critical thinking.

Five priorities arise for forthcoming research and practice.

(a) Make the science of AI evaluation stronger -- It is unacceptable that only 16% of AI benchmarks use statistical methods to compare how well different models work. The field requires common definitions of assessment

constructs, stringent validation protocols, and clear reporting standards. Researchers assessing AI evaluation systems ought to adhere to the same methodological criteria as those evaluating human assessment systems; indeed, elevated standards are justified due to the opacity of AI decision-making.

(b) Put fairness first in all areas of intersectionality -- The existing literature primarily emphasises gender and race, neglecting other demographic characteristics. Fairness research needs to grow to include things like socioeconomic status, disability, language background, and where you live. Furthermore, fairness metrics must be validated against stakeholder perceptions, not solely technical compliance.

(c) Make sure that people oversee high-stakes testing -- The general agreement among systematic reviews is clear: human oversight is necessary. This is not a short-term fix until AI gets better; it is a permanent need based on the limits of statistical learning when it comes to human judgement. Regulation ought to mandate substantive human oversight, especially for decisions that have considerable implications for individuals.

(d) Change the way we test for the AI age -- The long-term answer to generative AI is not to find it but to change it. Assessment tasks ought to prioritise advanced cognitive skills, genuine performance, and contextual judgment, abilities that are uniquely human. This needs new ways of teaching and training for teachers, not just technical fixes.

(e) Create ways to hold people accountable that can be enforced -- Voluntary ethical frameworks are not enough. The field needs independent audits, legal responsibility for algorithmic harms, and real penalties for not following the rules. The lesson from other regulated areas is that self-regulation is not regulation at all if there is no enforcement.

In fact, assessment is not just a technical issue of measurement. It is a human activity that is built on trust, authority, and care. Outsourcing assessments to AI is not just a way to save time; it changes the way assessment works. This change needs to be looked at carefully, not just accepted without question. The issue is not if AI can evaluate, but if it ought to, and under which circumstances, with what protection, and for what purposes. These are questions for more than just engineers; they are also questions for teachers, government officials, and regular people.

REFERENCES

1. Ahmed, S., Awan, D., Adil, M., & Ahmad, A. (2025). Ethics and Transparency in AI, Transparency, and Accountability in Higher Education. Social Science Review Archives, 3(4), 2729-2741.
2. Alaparthi, S., & Mishra, M. (2020). Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv preprint arXiv:2007.01127.
3. Ayaan, A., & Ng, K. W. (2025). Automated grading using natural language processing and semantic analysis. MethodsX, 14, 103395.
4. Bandara, R. J., Biswas, K., Akter, S., Shafique, S., & Rahman, M. (2025). Addressing algorithmic bias in AI-driven HRM systems: Implications for strategic HRM effectiveness. Human Resource Management Journal, 35(4), 1047-1063.
5. Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2021). Automated short answer grading using deep learning: A survey. In International cross-domain conference for machine learning and knowledge extraction (pp. 61-78). Cham: Springer International Publishing.
6. Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., ... & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. Science, 380(6641), 136-138.
7. Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. Journal of Applied Learning & Teaching, 7(1), 115-126.
8. Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. arXiv preprint arXiv:2407.18745.
9. Cheong, M. L. (2025). ChatGPT's performance evaluation in spreadsheets modeling to inform assessments redesign. Journal of Computer Assisted Learning, 41(3), 1-17.

10. Emirtekin, E. (2025). Large language model-powered automated assessment: a systematic review. Applied Sciences, 15(10), 5683.
11. Flemyng, E., Noel-Storr, A., Macura, B., Gartlehner, G., Thomas, J., Meerpohl, J. J., ... & Grainger, M. (2025). Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. Campbell Systematic Reviews, 21(4), cl2-70074.
12. Freeman, J. (2025). Student Generative AI Survey 2025. Oxford: Higher Education Policy Institute.
13. Gnanaprakasam, J., & Lourdusamy, R. (2024). The role of AI in automating grading: Enhancing feedback and efficiency. In Artificial intelligence and education-shaping the future of learning. IntechOpen.
14. Hawrysz, L. (2025). Artificial Intelligence in Candidate Screening: Opportunities and Challenges. Organization & Management, Series No. 228, 203-217.
15. Jaturat, N., Na-Nan, K., & Hu, B. (2026). Measuring AI responsibility: A cross-country validation of a multidimensional framework. International Journal of Information Management Data Insights, 6(1), 100388.
16. Kuzminykh, I., Nawaz, T., Shenzhang, S., Ghita, B., Raphael, J., & Xiao, H. (2024). Personalised feedback framework for online education programmes using generative AI. arXiv preprint arXiv:2410.11904.
17. Najafi, F. T., Pabba, V. R., Subramanian, R., & Vidalis, S. M. (2025). AI-Assisted Grading–A Study on Efficiency and Fairness. In 2025 ASEE Southeast Conference.
18. Nikhil, V., Annamalai, R., & Jayapal, S. (2025). NLP-driven approaches to automated essay grading and feedback. In Adopting Artificial Intelligence Tools in Higher Education (pp. 99-117). CRC Press.
19. Oberhauser, H. J. (2025). Bias in Artificial Intelligence: Exploring its role in institutional discrimination and strategies for mitigation (Master's thesis, Universidade NOVA de Lisboa, Portugal).
20. Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.
21. Simoni, J., Urtubia-Fernandez, J., Mengual, E., Simoni, D. A., Royo, M., Egaña-Yin, D., ... & Pereira, J. L. (2025). Artificial intelligence in undergraduate medical education: an updated scoping review. BMC Medical Education, 25(1), 1609.
22. Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., ... & Gu, Q. (2023). Beyond one-model-fits-all: A survey of domain specialization for large language models. arXiv preprint arXiv, 2305.