Artificial Intelligence in the Grading of Higher Education Assessment

An evidence-based policy framework for the responsible adoption of AI-assisted grading at NQF Level 7 and above, with the human firmly in the loop.

Prof Susheila Moodley, Professor of Practice, College of Business and Economics, University of Johannesburg; Expert Delegate, AI and Assessment Working Group, Digital Education Council (DEC) · May 2026

1. Executive summary

Artificial intelligence has already entered the South African lecture theatre. Nearly nine in ten students (86%) report using AI in their studies, while fewer than half of academics' report having begun their own AI-literacy journey (DEC 2025, 86% of students using AI vs 40% of faculty). The adoption gap is not, in the main, evidence of resistance. It is evidence that academics have not yet seen credible proof that AI is worth the learning curve, and that they hold a reasonable suspicion that AI in grading is heading somewhere uncomfortable. Policy should take both the opportunity and the suspicion seriously.

The central finding is plain. AI is a reliable grader of surface features — fluency, grammar, structure, the presence, or absence of expected content — and an unreliable grader of analytical depth. The international evidence is now consistent enough to treat this as a finding, not a preliminary result. When GPT-4 was asked to grade sixty master's-level political-science essays already marked by academics, agreement reached a Cohen's kappa of just 0.18. The model approximated the mean but clustered marks in the middle and could not engage the discipline-specific criteria that separate a credit from a distinction. Human raters continue to outperform AI on the feedback that matters most for higher-order writing.

This has direct implications for South African qualifications. At NQF Level 7 and above, certification is of higher-order professional capability — analysis, evaluation, synthesis, and judgement. If AI grading misjudges that capability, it is not the tool that is compromised; it is the qualification, and with it the credibility of the institution and the employability of the graduate.

At the same time, AI grading offers something the system has never been able to afford at scale: specific, written feedback delivered to every student within seconds, while they still care about the topic. The strongest equity case is that AI is a force multiplier on feedback — it reaches the first-generation student writing in a third language under time pressure, the student who never puts a hand up, who in the current system leaves the tutorial with no feedback at all (Kat Morgan, CoP 2026).

These two truths are not in tension once AI's role is correctly bounded. The principle that runs through both the international consensus and the South African practitioner evidence is the same: the human stays in the loop. AI marks are recommendations a human must sign off, never an autonomous decision made by the AI

What this paper recommends.

- Permit AI-assisted grading; prohibit autonomous AI grading at NQF Level 7+.
- Begin with formative assessment before any summative use.

- Treat detection as a dead end; redesign assessment instead.
- Separate reasoning from expression in rubrics.
- Make data governance — POPIA, no third-party training, audit trail, right to challenge — non-negotiable.
- Develop and refine rubrics within your own institutional context rather than relying on generic, ready-made models. Build the evidence the sector lacks rather than importing conclusions drawn elsewhere.

2. The policy problem

Three forces are converging, and the absence of a considered policy is allowing them to converge badly.

The first is demand. South African class sizes routinely run into the hundreds, with feedback that arrives two weeks after submission landing only after the misconception has hardened. The promise of AI to close that feedback loop in seconds speaks directly to a problem the sector has lived with for decades.

The second is availability. Capable AI grading tools now exist and are being adopted. Vendors are selling into local universities; institutions are building their own tools; and faculty are quietly using general-purpose chatbots to mark, often outside any governance framework at all. The choice is no longer whether AI will participate in grading, but under what rules.

The third is risk. AI grading touches the validity of the qualification, the fairness owed to a multilingual and unequal student body, the protection of personal information under POPIA, and the trust of a public that needs to believe a degree has been earned. Each is a serious institutional exposure. Managed by default — which is what happens in the absence of policy — all four are threatened.

The framing questions.

Can AI reliably grade higher-order assessment at NQF Level 6, 7 and above? Does the answer change for African higher-education contexts, given connectivity, and data costs? How does AI grading affect students writing in their second or third language? And what governance must be in place for AI grading to be trusted — not only by students, but by the public who receive our graduates and ask whether degrees have truly been earned?

3. The evidence base

The peer-reviewed literature is now substantial — automated short-answer grading has been an active research field since 2011 — and it supports five findings that bear directly on policy.

3.1 AI grading of higher-order assessment is not yet solved.

This is the finding that should most discipline our expectations. Ke and Ng's 2019 survey named higher-order, trait-specific scoring as the field's most under-developed area, and by 2024 the picture had barely moved. Lundgren's 2024 study put GPT-4 to grading sixty master's-level political-science essays already marked by teachers; agreement reached a Cohen's kappa of only 0.18, with the model unable to engage the discipline-specific analytical criteria that distinguish levels of attainment. Steiss and colleagues found the same pattern for feedback: human raters significantly outperformed ChatGPT on identifying weaknesses in argument structure.

Policy implication. Autonomous AI grading of higher-order, NQF Level 7+ assessment is not supported by the current evidence and should not be permitted. AI may assist; it may not decide.

3.2 The Global South is structurally absent from the evidence.

Bond and colleagues' 2024 meta-systematic review, covering sixty-six prior reviews, concluded that the literature is dominated by studies from high-income countries and that voices from the Global South are systematically under-represented. The AI-grading research that exists from Egypt, India, Sri Lanka, Nigeria, Mauritius, Türkiye, and South Africa is entirely focused on lower-order, short-answer scoring. Because South African institutions cannot simply import ready-made answers, they are free to build something genuinely their own.

3.3 Human-in-the-loop is the international consensus.

The strongest voices in the field converge on the same answer, and it is neither detection nor surveillance nor prohibition. Kortemeyer proposes threshold-based hybrid grading; Yan and colleagues' scoping review concludes that human oversight, transparency and pedagogical grounding are prerequisites for the ethical use of large language models in education; and Lodge, Yang, Furze and Dawson argue that generative AI reshapes the cognitive division of labour and that process-visible, scaffolded assessment with human judgement at the decision points is the appropriate response. Assessment records — marks, examiner comments, moderator reports, portfolio's of evidence — constitute personal information under POPIA 4 of 2013, and learners have the right to access and correct records held about them. Where an institution restricts access to assessment data needed to mount a meaningful appeal, it risks acting inconsistently with both POPIA's lawful processing conditions and the constitutional right to just administrative action under PAJA. Assessment data governance and appeals procedures must therefore be aligned, or the right to appeal becomes procedurally hollow.

3.4 Redesign, not detection

A surveillance posture is both ineffective and corrosive. Chaka's 2023 evaluation of five AI-content detection tools found inconsistent performance with substantial false-positive and false-negative rates — false positives accuse honest students; false negatives give false assurance. The defensible alternative is to make higher-order thinking visible in the assessment itself, where AI use is named and reflected upon, and where students must demonstrate the kinds of judgement that models still cannot reliably perform.

3.5 Local rubric architecture and tuning is the credible path.

The most sustained empirical programme on AI grading of higher-order science assessment — Zhai and colleagues, working across the United States and China — shows that off-the-shelf models are unreliable on higher-order rubric dimensions; models fine-tuned on discipline-specific data substantially close the gap with human raters; and the approach extends to local-language responses. The value is not in the AI tool itself — it is in the locally designed rubrics that tell the tool what good looks like in a South African context, and in the deliberate process of refining its outputs against NQF standards, QCTO requirements, and local assessment practice until its judgements can actually be defended.

4. The South African dimension

The international evidence sets the baseline. Five features of the South African system change what a responsible policy must require.

4.1 NQF validity

The NQF Level Descriptors define South African qualifications. A Level 7 qualification certifies the capacity to analyse, evaluate, synthesise, and exercise professional judgement in complex contexts — precisely the higher-order capability that AI grades least reliably. If an AI grader clusters marks in the middle and cannot distinguish a competent analysis from an excellent one, it does not produce a slightly inaccurate mark; it fails to certify what the qualification exists to certify. Validity at NQF Level 7+ is therefore a matter of NQF integrity, within the mandate of SAQA and the CHE.

4.2 Multilingual equity

South Africa has eleven official languages, and a substantial proportion of students are assessed in their second, third or even fourth language. This is where the equity argument for AI grading is strongest and the equity risk is sharpest. AI feedback delivered to every student at once is the first mechanism many practitioners have seen that gives the multilingual, first-generation, time-pressured student the same feedback as the student paying for a private tutor. The mirror risk is that AI graders systematically misjudge non-native-English writing, and that model performance is markedly stronger in well-resourced languages. Multilingual capability is not a feature; it is an equity requirement — and it does not mean English and Afrikaans alone.

4.3 Separating reasoning from expression

Conventional rubrics quietly collapse two different things into one score: whether the student understood the concept, and whether the student could express it in polished academic English. For a second- or third-language writer thinking and translating against the clock, the mark goes down not because the understanding was absent but because the expression was strained. AI grading can — if designed to — assess these two dimensions independently and present the educator with two distinct signals to weigh. Policy should encourage this and require institutions to be explicit about how the two are weighted.

4.4 Connectivity and the mobile-first reality

Most South African students' study on a phone, on standard data plans, often with intermittent connectivity. Any AI grading system must be mobile-first by design, must minimise data consumption, and should degrade gracefully when the connection drops. A system that assumes a laptop and uncapped broadband will widen the digital divide it claims to bridge.

4.5 Data protection and POPIA

Student work is protected under the Protection of Personal Information Act. Student work and educator content must sit under the institution's own data agreement and must not be used to train third-party models — a student's unpublished work should never resurface in a commercial model months later. POPIA also gives a person subject to a decision based solely on automated processing the right to challenge that decision. Every AI-generated mark must therefore be sampleable, overridable by a human, accompanied by reasons, and formally challengeable.

5. Evidence from practice

The second session of the Community of Practice on AI in Higher Education, held on 19 May 2026, examined AI grading through three live demonstrations — a South African EdTech tool (Mindjoy), an international university platform (EduMark AI, Queen Mary University of London), and a tool built in-house at the University of Johannesburg, College of Business and Economics. The COP's 2026 theme, the human in the loop, was reinforced by every speaker.

What is working. Closing the feedback loop — fast, specific, written feedback delivered while the topic is still live — is the single most valued capability. Cohort-level analysis surfaces misconception patterns invisible to a marker working paper by paper. When a rubric is corrected mid-moderation,

an entire cohort can be re-marked against the new criterion almost immediately. Native multilingual feedback preserves meaning. And having to explain a marking scheme to a system that has never sat in a marking meeting forces the unwritten rules of a discipline into the open — a benefit even where the AI never marks a single script.

What is not yet working. Achieving consistency remains the central challenge. Every time the AI grades, it behaves like a different marker with a different set of assumptions. Models are markedly weaker in South African languages other than Afrikaans, owing to training-data imbalance. General-purpose models are trained to be agreeable, which is a liability in a marker and requires deliberate engineering to overcome. Because language models follow linguistic patterns rather than calculating, a plausible sounding but wrong answer is always a risk. Reliable STEM grading therefore separates the two tasks — computation is managed by a dedicated tool, and the verified result is then passed to the model to interpret and contextualise. And markers are increasingly encountering AI-polished writing whose monotonous cadence can flatten genuinely good work — a challenge the marking rubric will need to accommodate by valuing the student's own voice distinctly.

6. Policy principles

The following seven principles are intended to be stable even as the technology changes. They should anchor institutional assessment policy and inform the work of quality-assurance bodies.

1. **The human stays in the loop.** AI grading produces recommendations that a qualified human reviews and signs off. No mark at NQF Level 7 or above is released or recorded without meaningful human decision.
2. **Validity before efficiency.** Speed and workload relief are welcome side effects, never the justification. The first test of any AI grading practice is whether it preserves the validity of the assessment for the level of qualification concerned.
3. **Equity by design.** Systems must be assessed for differential performance across languages and student groups, and rubrics must separate conceptual reasoning from expressive fluency so that multilingual students are not penalised for the act of translating under pressure.
4. **Transparency and the right to challenge.** Students must know when AI has been involved in grading their work, must receive the reasons for a mark, and must be able to challenge any automated decision — consistent with POPIA.
5. **Redesign over surveillance.** Academic integrity is protected by designing assessment so that higher-order thinking is visible and the AI used is named and reflected upon, not by deploying unreliable detection tools that treat students as suspects.
6. **Data sovereignty and protection.** Student and staff data remain under institutional control, are never used to train third-party models, and are managed under a documented data-protection regime with a full audit trail.
7. **Evidence-led adoption.** Adoption proceeds in phases, beginning with formative use, and is accompanied by local validation and a willingness to publish honestly what works and what does not.

These principles are the spine. *The recommendations, procurement criteria, phased governance pathway, research agenda, and full reference base are set out in the extended version of this white paper.*