POLICY WHITE PAPER

South African Higher Education Sector

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.

Status	Discussion draft for institutional leadership and sector bodies
Prepared for	Vice-Chancellors, DVCs, Senates, Teaching & Learning offices, the CHE, DHET, SAQA and Universities South Africa
Evidence base	30 verified peer-reviewed references; Community of Practice on AI in Higher Education, Session 2 (19 May 2026)
Anchored in	The Integrated AI Assessment Framework (Prof. Susheila Moodley, PhD, University of Johannesburg) and the FutureBanker programme
Version	1.0 · May 2026

About this white paper

This white paper sets out a policy position on the use of artificial intelligence in the grading of assessment in South African higher education. It is written from the standpoint of sector policy: its purpose is not to advocate for a particular product, but to give institutional leaders, quality-assurance bodies and faculty a defensible, evidence-led basis on which to decide whether, where and how AI should be permitted to participate in grading.

It draws on three bodies of evidence. The first is the international peer-reviewed research literature on automated and AI-assisted grading, read honestly rather than selectively. The second is the practitioner evidence surfaced by the Community of Practice on AI in Higher Education — a joint initiative of the University of Johannesburg and the FutureBanker programme — whose second session of 2026 convened sixty-seven participants to examine AI grading from South African, international and institutional vantage points. The third is the Integrated AI Assessment Framework developed at the University of Johannesburg, which is used here as a worked example of how the policy principles can be operationalised.

The document is deliberately candid about the limits of the current evidence. Where the research shows that AI grading works, it says so. Where the research shows that it does not yet work — particularly for the higher-order, judgement-laden assessment that defines qualifications at NQF Level 7 and above — it says so plainly, because a policy built on an overstated evidence base will not survive contact with an external examiner, an accreditation panel, or a court.

> **A note on scope.** This paper addresses AI in *grading and feedback* — the use of AI to mark student work and generate feedback. It does not address the separate questions of students' own use of AI in producing work, or AI in teaching delivery, except where these bear directly on how assessment should be designed and graded.

Contents

About this white paper 2
Contents 3
1. Executive summary 4
2. Purpose, audience and scope 6
3. The policy problem: why a sector response is needed now 7
4. The evidence base, honestly read 8
4.1 AI grading of higher-order assessment is not yet solved 8
4.2 The Global South is structurally absent from the evidence 8
4.3 Human-in-the-loop is the international consensus 8
4.4 Redesign, not detection 9
4.5 Local rubric architecture and tuning is the credible path 9
5. The South African dimension 10
5.1 Validity and the integrity of the NQF 10
5.2 Equity and the multilingual classroom 10
5.3 Separating reasoning from expression 10
5.4 Connectivity, data cost and the mobile-first reality 11
5.5 Data protection and POPIA 11
6. Evidence from practice: what South African institutions are already doing 12
6.1 Three vantage points on the same principle 12
6.2 What is working 12
6.3 What is not yet working 13
7. Risks, limitations and failure modes 14
8. Policy principles 15
9. Recommendations 16
9.1 For institutions — Senates and Teaching & Learning offices 16
9.2 For institutional leadership — VCs and DVCs 16
9.3 For the sector — CHE, SAQA, DHET and USAf 16
10. A phased adoption and governance framework 17
10.1 The maturity pathway 17
10.2 Non-negotiable controls at every phase from Phase 2 onward 17
10.3 Procurement and build criteria 17
11. A research and evidence agenda for the sector 19
12. Conclusion 20
13. References 21
14. Glossary of key terms 23

1. Executive summary

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

The central finding of this paper is straightforward. 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 rather than a preliminary result. When a large language model was asked to grade sixty master's-level essays already marked by academics, agreement with the human graders sat at a Cohen's kappa of 0.18: the model reliably hit the average mark but clustered everything 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 precisely the feedback that matters most for higher-order writing.

This has direct implications for South African qualifications. NQF Level 7 and above certify 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. The validity of the assessment is therefore a governance question, not merely a technical one.

At the same time, AI grading offers something South African higher education has never been able to afford at scale: specific, written feedback delivered to every student within seconds, while they still care about the topic, rather than weeks later when the misconception has hardened. The strongest version of the 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 their hand up, who in the current system leaves the tutorial with no feedback at all.

These two truths are not in tension once the role of AI is correctly bounded. The policy posture this paper recommends is captured in a single principle that ran through every contribution to the Community of Practice and through the international consensus alike: the human stays in the loop. AI marks are recommendations a human signs off, never autonomous decisions. From that principle, a workable governance framework follows.

What this paper recommends

- **Permit AI-assisted grading, prohibit autonomous AI grading.** No mark at NQF Level 7 or above should be released to a student or recorded on a transcript without meaningful human sign-off.
- **Begin with formative assessment.** Build institutional trust and moderation skill in low-stakes settings before any summative use.
- **Treat detection as a dead end; redesign assessment instead.** AI-content detectors are technically unreliable and signal the wrong relationship with students. Make higher-order thinking visible in the task itself.
- **Separate reasoning from expression in rubrics.** Marking schemes that conflate conceptual understanding with polished academic English systematically disadvantage second- and third-language writers.

- **Make data governance non-negotiable.** POPIA compliance, a prohibition on third-party model training using student data, a full audit trail, and a student right to challenge any automated decision are minimum conditions of adoption.
- **Invest in local capability, not off-the-shelf deployment.** The credible path to higher-order grading runs through local rubric architecture and discipline- and language-specific tuning, not the procurement of a generic tool.
- **Build the evidence the sector lacks.** There is no peer-reviewed body of work on AI grading of higher-order, professional assessment from any African university. South Africa should produce it rather than import conclusions drawn elsewhere.

2. Purpose, audience and scope

This white paper exists to give South African higher education a shared, defensible position on a question that institutions are currently answering one ad hoc decision at a time. Individual academics are already experimenting with AI grading. Vendors are already selling into the sector. Students already assume AI is involved in how their work is judged. In the absence of policy, practice is being set by whoever moves first, and the resulting inconsistency is itself a risk to the fairness and the standing of our qualifications.

Who this paper is for

It is written for the people who hold the levers. Vice-Chancellors and Deputy Vice-Chancellors set institutional risk appetite and release the resources. Senates and Teaching and Learning committees own academic standards and assessment policy. Quality-assurance and accreditation bodies — principally the Council on Higher Education (CHE) and the South African Qualifications Authority (SAQA) as custodian of the NQF — own the integrity of the qualification. The Department of Higher Education and Training (DHET) and Universities South Africa (USAf) shape the sector environment in which all of this sits. Faculty are the people who will, or will not, make any of it work.

What is in scope

The paper addresses the use of AI to grade student assessment and to generate feedback on it, across formative and summative contexts, with particular attention to higher-order assessment at NQF Level 7 and above. It addresses the conditions — pedagogical, ethical, legal and technical — under which such use can be trusted.

What is out of scope

It does not set out to regulate students' own use of AI in producing their work, nor AI in teaching delivery, nor the wider economic and labour-market questions raised by generative AI. These matter, and they intersect with assessment, but they are separate policy conversations. Where they touch grading directly — as when a declared but AI-"polished" thesis loses the human voice that a rubric ought to reward — the paper addresses them only to that extent.

3. The policy problem: why a sector response is needed now

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

The first is demand. Class sizes in the South African system routinely run into the hundreds and, for some modules, the thousands. The honest reality of grading at that scale is that much student work is marked late, by markers, against schemes whose unwritten rules even the lecturer cannot fully articulate. Feedback that arrives two weeks after submission lands after the student has either internalised the misconception or stopped caring. 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. South African and international 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 technology is no longer hypothetical, which means 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 of these is a serious institutional exposure. Managed well, AI grading can strengthen all four. Managed by default — which is what happens in the absence of policy — it threatens all four.

> **The framing questions.** The Community of Practice opened its second session of 2026 with four questions that this paper adopts as its own. 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?

4. The evidence base, honestly read

Policy should rest on the evidence as it is, not as we would like it to be. The peer-reviewed literature on AI grading is now substantial — automated short-answer grading has been an active research field since at least 2011 — and it supports five findings that bear directly on policy. Each is stated below with the studies that ground it. The full reference list appears at the end of this paper.

4.1 AI grading of higher-order assessment is not yet solved

This is the finding that should most discipline our expectations. The methods that grade surface features well — grammar, length, lexical sophistication, the presence of expected content — are the least effective on the constructs that postgraduate rubrics actually care about: argumentation, persuasiveness, coherence of analysis, and judgement. Ke and Ng's 2019 survey of the field named higher-order, trait-specific scoring as its most under-developed area. By 2024 the picture had barely moved.

The clearest demonstration comes from Lundgren's 2024 study, in which GPT-4 was asked to grade sixty anonymised master's-level political-science essays previously marked by teachers. Agreement reached a Cohen's kappa of only 0.18, with percentage agreement around 35 per cent. The model reliably approximated the mean but clustered its marks in the middle of the distribution and could not engage the discipline-specific analytical criteria that distinguish levels of attainment. Steiss and colleagues found the same pattern for feedback rather than scoring earlier in 2024: human raters significantly outperformed ChatGPT on identifying weaknesses in argument structure and on prioritising higher-order revisions over surface corrections.

> **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.

4.2 The Global South is structurally absent from the evidence

Bond and colleagues' 2024 meta-systematic review, covering sixty-six prior reviews of AI in higher education, reached a direct conclusion: the literature is dominated by studies from high-income countries, ethical considerations are inconsistently reported, and the voices of students and faculty in the Global South are systematically under-represented. There is real and growing AI-grading research from developing economies — from Egypt, India, Sri Lanka, Nigeria, Mauritius, Türkiye and South Africa — but almost all of it addresses lower-order, short-answer, content-recall scoring. Even the recent South African prototype work from UNISA is positioned for assessment workload reduction in distance-learning contexts, not for higher-order grading research.

The consequence is that South African institutions cannot simply import settled conclusions. The evidence that exists was generated in different linguistic, pedagogical and infrastructural conditions, and at the lower-order end of the assessment spectrum. The asymmetry is also an opportunity, which Section 11 takes up.

4.3 Human-in-the-loop is the international consensus

The strongest voices in the field are converging on the same answer, and it is neither detection nor surveillance nor prohibition. It is human oversight built into the design. Kortemeyer's 2024 work proposes threshold-based hybrid grading, in which AI grades where its confidence is high

and humans grade where it is not. Yan and colleagues' systematic scoping review concludes that human oversight, transparency and pedagogical grounding are prerequisites for the ethical use of large language models in education. Lodge, Yang, Furze and Dawson make the conceptual case with unusual clarity: generative AI is not a calculator that automates one bounded task and leaves the cognitive landscape intact; it reshapes the entire division of labour between learner and machine, and the appropriate response is process-visible, scaffolded assessment with human judgement at the key decision points.

This consensus is convenient for policy, because it aligns the international research grain with the legal requirements of POPIA — which already entitles a person to challenge a decision based solely on automated processing — and with the professional instincts of the academics who will operate the system.

4.4 Redesign, not detection

A surveillance posture toward AI is both ineffective and corrosive. AI-content detection tools are technically unreliable: Chaka's 2023 evaluation of five such tools found inconsistent performance with substantial false-positive and false-negative rates. A false positive accuses an honest student; a false negative gives false assurance. Beyond the technical failure, a detection-led posture signals distrust and frames the institution's relationship with its students as adversarial.

The defensible alternative is to make higher-order thinking visible in the assessment itself — to set tasks where the process is part of the artefact, where AI use is named and reflected upon, and where students must demonstrate the kinds of judgement that models still cannot reliably perform. This is the position the AI Assessment Scale work of Perkins and Furze codifies, and that institutional frameworks such as UJ's operationalise.

4.5 Local rubric architecture and tuning is the credible path

The most sustained empirical programme on AI grading of higher-order science assessment — associated with Zhai and colleagues, working across the United States and China — points to a clear conclusion. 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 can be extended to local-language responses, as demonstrated for Chinese-language scientific explanations. The implication for South African institutions is that the value, and the defensible contribution, lies in the local rubric architecture and the local tuning pathway — not in deploying a generic tool and hoping it generalises to an NQF Level 7 case study written in a student's third language.

> **The five findings in one line.** AI grades surface features well and depth poorly; the Global South has not yet studied the depth question; human-in-the-loop is the consensus answer; detection does not work and redesign does; and the credible route to higher-order grading is local architecture, not off-the-shelf procurement.

5. The South African dimension

The international evidence sets the baseline. But several features of the South African system change what a responsible policy must require. These are not minor adjustments to an imported model; they are the considerations that should shape the model from the outset.

5.1 Validity and the integrity of the NQF

South African qualifications are defined by the National Qualifications Framework, and the NQF Level Descriptors specify the cognitive demand a qualification certifies. A Level 7 qualification certifies the capacity to analyse, evaluate, synthesise and exercise professional judgement in complex and unpredictable contexts. This is precisely the higher-order capability that the international evidence shows AI grades least reliably. The risk is therefore not abstract: if an AI grader clusters marks in the middle and cannot distinguish a competent analysis from an excellent one, it does not merely produce a slightly inaccurate mark — it fails to certify the very thing the qualification exists to certify. Assessment validity at NQF Level 7 and above is consequently a matter of NQF integrity, and falls squarely within the mandate of SAQA and the CHE.

5.2 Equity and the multilingual classroom

South Africa has eleven official languages, and a large 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. The strongest case made in the Community of Practice was that the students who lose the most under the current system are not the confident, well-resourced students at the front of the queue, but the first-generation students writing under time pressure in a language not their own, who have been taught that asking for help is a sign of weakness, and who leave the tutorial with no feedback at all. AI feedback, delivered to every student at once, is the first mechanism many practitioners have seen that gives those students the same feedback as the student paying for a private tutor.

The risk is the mirror image. International evidence shows that AI graders systematically misjudge non-native-English writing, and that model performance is markedly stronger in well-resourced languages. In the South African context this means a model will be more capable in English and Afrikaans than in Sesotho, isiZulu or the other official languages, simply because of training-data imbalances. A policy that ignores this would build the existing linguistic inequity into the grading layer itself. Multilingual capability is therefore not a nice-to-have feature; it is an equity requirement, and it does not mean English and Afrikaans alone.

5.3 Separating reasoning from expression

Closely related, and arguably the most consequential design insight to emerge from local practice, is the recognition that conventional rubrics quietly conflate two different things: 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, a mark goes down not because the understanding was absent but because the expression was effortful. AI grading can, if designed to, assess these two dimensions independently — grading the reasoning against the reasoning criteria and the expression against the expression criteria — and present the educator with two distinct signals to weigh. Done well, this is a genuine equity advance. Policy should encourage it and should require that institutions are explicit about how the two dimensions are weighted.

5.4 Connectivity, data cost and the mobile-first reality

Most South African students study on a phone, using standard data plans, often with intermittent connectivity. Any AI grading system adopted in this context must be mobile-first by design, must minimise data consumption, and should degrade gracefully when the connection drops — for example by saving work locally on the device. A system that assumes a laptop and uncapped broadband will widen the digital divide it claims to bridge. This is a procurement and design requirement, not an afterthought.

5.5 Data protection and POPIA

Student work, and the personal information it contains, is protected under the Protection of Personal Information Act. Two consequences follow directly. First, 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. Second, POPIA, like the GDPR, gives a person subject to a decision based solely on automated processing the right to be informed and to challenge that decision. An AI grading system must therefore be designed so that every AI-generated mark is sampleable, overridable by a human, accompanied by reasons, and formally challengeable by the student. These are not optional governance niceties; they are legal conditions of operation.

6. Evidence from practice: what South African institutions are already doing

Policy should be informed not only by the published literature but by what practitioners are actually achieving and where they are getting stuck. 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 of working platforms — a South African EdTech tool, an international university platform, and a tool built in-house at the University of Johannesburg. The session's unifying theme, "The Human in the Loop," was reinforced by every speaker. The detail below is drawn from that session and is offered as grounded, sector-relevant evidence.

6.1 Three vantage points on the same principle

Vantage point	What was demonstrated	Core claim
South African EdTech (Mindjoy, presented by Kat Morgan)	Text, voice and photo submission; native multilingual feedback including Sesotho; an "insights" layer surfacing misconception patterns across a cohort; an "opinionated" marker designed to resist the agreeableness of general models.	The point of AI marking is not speed but closing the feedback loop while the student still cares and the educator is still in charge.
International university (EduMark AI, Queen Mary University of London, presented by Dr Deepshikha Deepshikha)	Rubric-guided grading with calibratable strictness, side-by-side review of AI feedback, and an educator verification workflow before release.	"Guided grading" on an 80–20 split: AI does the repetitive 80 per cent; the educator owns the 20 per cent of review, refinement and personalisation.
Institutional build (UJ AI grading tool, presented by Dr Leon Janse van Rensburg)	Full assessment lifecycle from question generation to marking, challenge and moderation; separate AI and human marks that do not see each other, designed for research; prompt-injection protection; ethical clearance obtained.	The consistency problem is the hard one: it took two years to bring marking variation down to roughly 1–3 per cent on non-deterministic, social-science answers.

6.2 What is working

- **Closing the feedback loop.** The single most valued capability is fast, specific, written feedback delivered while the topic is still live — turning a missed tutorial into the equivalent of a private one.
- **Surfacing cohort-level misconceptions.** Analysis across hundreds of submissions can reveal patterns invisible to a marker working paper by paper — for instance, that a large share of a class could define two concepts but never learned to compare them — allowing a lecturer to reteach precisely.
- **Instant re-marking when a rubric changes.** When a marking criterion is corrected mid-moderation, an entire cohort can be re-marked against the new criterion almost immediately, which is simply not feasible by hand.
- **Native multilingual feedback.** Feedback generated directly in the student's language, rather than translated, preserves meaning — with the honest caveat that quality is uneven across languages.

- **Rubric design as a forcing function.** 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.

6.3 What is not yet working

- **Consistency across non-deterministic instances.** Because each marking run is effectively a fresh instance, the system behaves like a thousand different markers making different assumptions each time. Encoding a discipline's unwritten rules tightly enough to hold variation within a few per cent is the current frontier of engineering work, and the practitioner consensus is that the problem is not yet fully solved.
- **Weakness in lower-resource languages.** Models are functional but markedly more formal and less reliable in South African languages other than Afrikaans, owing to training-data imbalance. This is an industry-wide limitation, not a single-vendor flaw.
- **The agreeableness problem.** General-purpose models are trained to be helpful and agreeable, which is a liability in a marker: left uncorrected they award everyone a mild, middling grade with vague praise. Making a marker willing to say "this is wrong, and here is why" requires deliberate engineering.
- **Computation versus language.** Language models can be "talked into" arithmetic errors; reliable STEM grading requires **processing math algorithmically** and feeding the result to the model, rather than trusting its own calculations.
- **The loss of human voice.** Markers are encountering AI-"polished" student writing whose monotonous cadence and repetitive structure can make genuinely good work read as flat — a new challenge that rubrics will need to accommodate by valuing the student's own voice and the originality of ideas separately from surface polish.

7. Risks, limitations and failure modes

The following risks are real, are documented either in the literature or in local practice, and must be actively managed rather than assumed away. The mitigations listed are taken up in the recommendations and the governance framework that follow.

Risk	Why it matters	Primary mitigation
Invalid grading of higher-order work	AI clusters marks in the middle and cannot distinguish levels of analytical attainment, compromising the validity of NQF 7+ qualifications.	Mandatory human sign-off; prohibition on autonomous summative grading; local validation against human marks.
Linguistic inequity	Models perform worse in lower-resource South African languages, building existing inequity into the grading layer.	Multilingual capability requirement; separation of reasoning from expression; sampling of marks across language groups.
Inconsistency	Non-deterministic models make different assumptions on each run, so the same script can receive materially different marks.	Pre-deployment consistency testing against a published variance threshold; fresh-instance variation monitored over time.
Data-protection breach	Student personal information may be exposed or used to train third-party models, in breach of POPIA.	Data-protection impact assessment; contractual prohibition on third-party training; institutional data-residency terms.
Loss of student trust	Only around a third of students perceive AI grading as fair; trust is a precondition for adoption, not a by-product of it.	Transparency about where AI is used; reasons provided with every mark; a working right to challenge.
Automation of judgement	Over-reliance erodes the academic judgement the system is meant to support, and lets students disengage if they sense no one is reading their work.	Visible human involvement; AI Steward oversight; assessment redesign that keeps thinking in the loop.
Vendor lock-in and opacity	Procuring a closed off-the-shelf tool surrenders control of rubric logic and audit access.	Procurement criteria below; preference for local rubric architecture and exportable audit trails.

8. 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. Specific recommendations and the governance framework derive from them.

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. “Meaningful” excludes rubber-stamping at scale.
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 AI use 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 handled under a documented data-protection governance framework 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.

9. Recommendations

The recommendations are addressed to those who can implement them. They are intended to be adopted together; piecemeal implementation tends to retain the risks while forgoing the benefits.

9.1 For institutions — Senates and Teaching & Learning offices

1. **Adopt an AI-assisted assessment policy** that codifies the seven principles above, defines permitted and prohibited uses by NQF level and assessment stakes, and is owned by Senate rather than by individual faculties.
2. **Mandate human sign-off for all summative grading** and prohibit autonomous AI grading at NQF Level 7 and above outright.
3. **Require a data-protection impact assessment** before any AI grading tool processes student work, contract for institutional data residency and a prohibition on third-party model training.
4. **Revise rubric standards** to separate reasoning from expression and to make space for the student's own voice and the originality of ideas.
5. **Integrate AI grading into existing moderation and external-examination processes** rather than creating a parallel track; AI marks must be sampleable and auditable by moderators and external examiners.

9.2 For institutional leadership — VCs and DVCs

6. **Establish an AI Steward model.** Place trained faculty champions across schools to support colleagues, oversee local practice, and act as the human accountability layer between the tool and the institution. This distributes capability rather than centralising it in a unit that faculty experience as remote.
7. **Resource faculty AI literacy as a first-order investment,** recognising that the adoption gap is driven by lack of evidence and time, not by resistance, and that the academics who must operate the system are the ones who will determine whether it succeeds.
8. **Invest in local capability —** rather than treating AI grading as a procurement problem solved by buying a generic tool.

9.3 For the sector — CHE, SAQA, DHET and USAf

9. **Issue sector guidance** affirming that autonomous AI grading of higher-order assessment is not currently consistent with the maintenance of NQF standards, and that human-in-the-loop grading is the expected norm.
10. **Clarify quality-assurance expectations** for AI-assisted grading within existing accreditation and review processes, including audit-trail and moderation requirements, so that institutions are not left to improvise.
11. **Convene shared infrastructure.** Multilingual capability and local adaptation are expensive and duplicative if every institution builds alone. The sector should explore shared, POPIA-compliant capability — particularly for South African languages — as a public good. Fund and prioritise the research agenda set out in Section 11, so that South African policy rests on South African evidence.

10. A phased adoption and governance framework

Trust in AI grading is earned in sequence, not granted at the outset. The framework below sets out a maturity pathway that begins where the stakes are lowest and the evidence strongest, and advances only as institutional capability and local validation accumulate. No institution should begin at Phase 2, and the practitioner consensus is emphatic that an institution should never "start with the final exam and work backwards."

10.1 The maturity pathway

Phase	What happens	Gate to the next phase
Phase 0 *Readiness*	Senate policy adopted; data-protection impact assessment completed; AI Stewards appointed and trained; faculty literacy programme begun. No student work graded by AI yet.	Policy, DPIA and Steward network in place; faculty literacy under way.
Phase 1 *Formative only*	AI used for formative feedback only, with no marks of record. Educators develop moderation skill and co-design rubrics with the system. Trust and evidence accumulate at zero transcript risk.	Demonstrated educator confidence; rubrics co-designed; consistency behaviour understood.
Phase 2 *Assisted summative*	AI assists summative grading at lower NQF levels first, always with human sign-off, full audit trail and integration into moderation. Variance monitored against a published threshold.	Variance within threshold; moderation and external examiners satisfied; student-appeal route working.
Phase 3 *Higher-order summative*	AI assists summative grading at NQF Level 7+ for the first time, using locally contextualized rubric architecture, with mandatory human decision on every mark and ongoing local validation against human raters. Never autonomous.	Continuous local validation; never delegated to the tool alone.

10.2 Non-negotiable controls at every phase from Phase 2 onward

- Human sign-off on every mark of record, with the reviewer accountable for the decision.
- A full, exportable audit trail: every AI mark sampleable, the AI's reasoning visible, and the human override recorded.
- A student right to be informed that AI was involved and to challenge the resulting mark, with a defined turnaround.
- Consistency testing before deployment and monitoring thereafter, against a variance threshold the institution publishes and defends.
- Sampling of marks across language groups to detect differential performance.
- Integration with moderation and external examination, not a parallel process.

10.3 Procurement and build criteria

Whether an institution buys, builds or adapts, the following criteria translate the principles into a checklist that procurement and IT governance can apply. They are drawn from the trust requirements articulated by practitioners and from POPIA.

Criterion	Requirement
POPIA compliance	Student and staff data sit under the institution's data agreement; a DPIA is completed before processing.
No third-party training	Contractual guarantee that institutional data is never used to train

Criterion	Requirement
	external models.
Audit and override	Every mark is sampleable, accompanied by reasons, and overridable by a human, with the audit trail exportable for external examiners.
Human-in-the-loop by design	The system cannot release a mark of record without a human decision; autonomous mode is disabled for summative use.
Multilingual capability	Demonstrated capability across relevant South African languages, with honest disclosure of performance differences.
Reasoning / expression separation	Ability to assess conceptual reasoning and expressive fluency as distinct, separately weighted signals.
Mobile-first and low-bandwidth	Usable on a phone on a standard data plan, with local save and graceful degradation offline.
Consistency controls	Evidence of variance within a stated threshold across repeated runs and across models where applicable.
Insight layer	Cohort-level analysis that surfaces misconception patterns to inform teaching, not merely a mark per script.

11. A research and evidence agenda for the sector

Policy made today rests on evidence generated elsewhere, at the lower-order end of the assessment spectrum, in other languages and other systems. That is a fragile foundation. The same review that documents the absence of Global South voices — Bond and colleagues, 2024 — also tells the sector what to do about it: increase ethics, collaboration and rigour, and generate the missing evidence.

The intersection of three established research conversations is genuinely unoccupied. The international literature is asking the right questions about validity, fairness and higher-order assessment. The developing-economy literature has produced real work but has not engaged the higher-order end. And no peer-reviewed body of empirical work addresses AI grading of NQF Level 7+ professional, case-based assessment in a South African or broader African context. That intersection is where the sector's contribution lies.

The Integrated AI Assessment Framework developed at the University of Johannesburg — which synthesises the Perkins and Furze AI Assessment Scale, Bloom's revised taxonomy, the South African NQF Level Descriptors and Fink's significant-learning dimensions — offers a worked model, grounded in the real assessments of the FutureBanker programme at NQF Level 7. It is presented here not as the only approach but as a credible template: a framework that does not pretend AI can grade higher-order work autonomously, that keeps the human in the loop, and that is designed from the outset to be studied. If pursued honestly over the next two to three years, with publication of both successes and failures, work of this kind would produce the first sustained body of empirical evidence on AI grading of higher-order professional assessment in an African higher-education context.

The agenda in brief. Validate AI grading against human marks at NQF Level 7+ on real professional assessments; test differential performance across South African languages; study the student-trust and right-to-challenge experience; and publish the consistency, validity and equity findings openly so that the sector — not a single institution — builds the evidence base.

12. Conclusion

AI is already grading student work in South African higher education. The only open question is whether it does so under policy or by default. This paper argues for policy — and for a particular kind of policy, one that neither bans the technology out of fear nor adopts it out of enthusiasm, but bounds its role precisely to what the evidence supports.

That boundary is clear. AI is a capable assistant and an unreliable judge. It can close the feedback loop, reach the students the current system fails, surface what a cohort has misunderstood, and force a discipline's tacit rules into the open. It cannot, on the current evidence, be trusted to certify higher-order professional capability on its own — which is exactly what an NQF Level 7 qualification exists to do. The human stays in the loop not as a transitional courtesy but as a permanent feature of a valid system.

Handled this way, AI grading is an opportunity for South African higher education to do something the international literature has not yet done: to produce local evidence about a local problem, in a multilingual and unequal context where the validity of assessment matters enormously for graduate employability and institutional credibility. The intellectual gap is real. The work to close it has to be ours.

— end of policy white paper —

13. References

The thirty references below were individually verified by direct lookup of the publisher page, ACL Anthology, arXiv, Springer, ScienceDirect, PLOS, Frontiers, Wiley, AAAI proceedings or PubMed, with author names, titles, journals, volumes, page numbers and DOIs checked against the source (verification date 18 May 2026).

1. Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: HLT, 752–762. Portland: ACL.
2. Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. Computer Speech & Language, 28(4), 833–857.
3. Abdul Salam, M., Abd El-Fatah, M., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. PLOS ONE, 17(8), e0272269.
4. Kadupitiya, J. C. S., Ranathunga, S., & Dias, G. (2016). Sinhala short sentence similarity calculation using corpus-based and knowledge-based similarity measures. Proceedings of WSSANLP 2016, COLING 2016, 44–53. Osaka.
5. Saha, S., Dhamecha, T. I., Marvaniya, S., Foltz, P., Sindhgatta, R., & Sengupta, B. (2019). Joint multi-domain learning for automatic short answer grading. arXiv:1902.09183.
6. Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2021). Automated short answer grading using deep learning: A survey. Machine Learning and Knowledge Extraction, CD-MAKE 2021, LNCS 12844, 61–78. Springer.
7. Ndukwe, I. G., Daniel, B. K., & Amadi, C. E. (2019). A machine learning grading system using chatbots. Artificial Intelligence in Education, AIED 2019, LNCS 11626, 365–368. Springer.
8. Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. Education and Information Technologies, 27, 5573–5604.
9. Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. Procedia Computer Science, 169, 726–743.
10. Twabu, K., & Nakene-Mginqi, M. (2024). Developing a design thinking artificial intelligence driven auto-marking/grading system for assessments to reduce the workload of lecturers at a higher learning institution in South Africa. Frontiers in Education, 9, 1512569.
11. Mayfield, E., & Black, A. W. (2020). Should AI systems be used to automatically score essays? Considering the value of educational measurement and the potential for bias. Frontiers in Education, 5, 97.
12. Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), ACL 2019, 1–10. Florence.
13. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32, 1052–1092.
14. Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55, 2495–2527.
15. Madnani, N., & Cahill, A. (2018). Automated scoring: Beyond natural language processing. Proceedings of the 27th International Conference on Computational Linguistics (COLING), 1099–1109. Santa Fe: ACL.

16. Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22466–22474.
17. Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education, 21, 4.
18. Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. FinTech and Artificial Intelligence for Sustainable Development, 387–409. Palgrave Macmillan.
19. Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching, 6(2), 94–104.
20. Thaldar, D., Botes, M., Badru, T., Chenia, H., Duma, S., Dlamini, S., Amin, K., Hugo, J., Govender, S., Vosloo, J., Koorbanally, N., & Chuturgoon, A. (2025). Generative AI governance in higher education: A case study from Africa. Frontiers in Political Science, 7, 1666661.
21. Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. Journal of Educational Computing Research, 58(1), 87–125.
22. Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19), 6300–6308.
23. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050.
24. Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894.
25. Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. Discover Artificial Intelligence, 4, 47.
26. Stowell, J. R., & Zhu, J. (2025). Evaluating ChatGPT for automated creation and grading of essay questions in higher education. Teaching of Psychology (advance online publication).
27. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55(1), 90–112.
28. Lodge, J. M., Yang, S., Furze, L., & Dawson, P. (2023). It's not like a calculator, so what is the relationship between learners and generative artificial intelligence? Learning: Research and Practice, 9(2), 117–124.
29. Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. Computers and Education: Artificial Intelligence, 6, 100210. See also Yang, J., Latif, E., He, Y., & Zhai, X. (2025), Fine-tuning ChatGPT for automatic scoring of written scientific explanations in Chinese, arXiv:2501.06704; and Zhai, X., He, P., & Krajcik, J. (2022), Applying machine learning to automatically assess scientific models, Journal of Research in Science Teaching, 59(10), 1765–1794.
30. Lundgren, M. (2024). Large language models in student assessment: Comparing ChatGPT and human graders. arXiv:2406.16510. Source of the 0.18 Cohen's kappa finding on master's-level essays.

14. Glossary of key terms

AES (Automated Essay Scoring). The use of computational methods to assign scores to extended written responses.

AIAS (AI Assessment Scale). A scale, associated with Perkins and Furze, describing levels of permitted AI involvement in an assessment, from no AI to full AI collaboration.

AI Steward. A trained faculty champion placed within a school to support colleagues in AI integration and to act as the human accountability layer between an AI tool and the institution.

ASAG (Automated Short-Answer Grading). The use of computational methods to grade short, constructed responses; an active research field since at least 2011.

Bloom's revised taxonomy. A hierarchy of cognitive processes from remembering and understanding through to analysing, evaluating and creating; higher-order processes are those at the upper end.

Cohen's kappa. A statistic measuring agreement between two raters beyond what chance would predict; values near 0 indicate near-chance agreement, values near 1 indicate near-perfect agreement.

DPIA (Data-Protection Impact Assessment). A documented assessment of the data-protection risks of a processing activity, completed before processing begins.

Formative / summative assessment. Formative assessment supports learning and does not count toward a final mark; summative assessment counts toward a mark of record.

Higher-order assessment. Assessment of analysis, evaluation, synthesis and professional judgement, as distinct from recall and comprehension; characteristic of NQF Level 7 and above.

Human in the loop. The principle that AI produces recommendations a qualified human reviews and signs off, rather than autonomous decisions.

NQF (National Qualifications Framework). The South African framework that classifies qualifications by level; the Level Descriptors specify the cognitive demand each level certifies. SAQA is its custodian.

POPIA. The Protection of Personal Information Act, South Africa's data-protection law, which among other things entitles a person to challenge a decision based solely on automated processing.