

Can They Spot the Fake? AI Literacy and Synthetic Media Detection among Zimbabwean University Students.

Octavious Chido Masunda

Senior Lecturer, National University of Science and Technology, Zimbabwe.

DOI: <https://doi.org/10.47772/IJRISS.2026.1026EDU0603>

Received: 16 September 2026; Accepted: 21 September 2026; Published: 01 October 2026

ABSTRACT

This study examined the extent to which AI literacy predicts Zimbabwean university students' ability to detect synthetic media. A cross-sectional survey of 518 students at the National University of Science and Technology combined a self-report AI literacy scale adapted from Ng, Leung, Chu and Qiao (2021) with an objective six-item detection task. The central finding is a structural mismatch between confidence and competence. Students rated their literacy at moderate to high levels (overall $M = 3.58$ out of 5.00), yet correctly classified only 60.3% of stimuli, with accuracy on AI-generated items barely above chance. Multiple regression showed that critical literacy was the strongest predictor of detection accuracy ($\beta = 0.33$, $p < 0.001$), while the frequency of AI tool use predicted nothing once evaluative skill was controlled for. The most consequential structural finding concerns learning channels. Social media taught detection skills to 62.0% of students, whereas formal education reached only 15.1%, meaning an unregulated informal ecology is doing the work universities have not taken up. The findings support critiques that self-report AI literacy frameworks overstate practical preparedness and indicate that detection is a trainable evaluative competency that should be embedded in curricula across disciplines rather than treated as a by-product of digital exposure.

Keywords: AI literacy, synthetic media, deepfakes, detection accuracy, Zimbabwe, digital media literacy, misinformation

INTRODUCTION

Across Sub-Saharan Africa the diffusion of smartphones and affordable mobile data has compressed a decades-long media transition into fewer than fifteen years. Zimbabwe had an estimated 6.8 million active internet users in 2024, the overwhelming majority accessing content through mobile devices (DataReportal, 2024). That acceleration brought a parallel problem. The same networks that circulate health information, electoral news and economic data also carry sophisticated synthetic content that is increasingly difficult to distinguish from authentic material. Deepfake imagery, AI-generated audio and algorithmically composed text now reach audiences without the institutional filters that once, however imperfectly, slowed the spread of fabricated material.

Zimbabwe's own information history sharpens the stakes. Digitally mediated rumour and propaganda have accompanied every major political moment since the disputed 2008 elections, when a parallel market of information flourished on mobile phones precisely because official channels were distrusted (Moyo, 2009). The state has responded to online contestation with internet shutdowns that themselves accelerated the circulation of misinformation (Mare, 2020). In such an environment, the ability to detect synthetic media is no longer a specialist skill. It is, or ought to be, a core component of what researchers variously call media literacy, digital literacy or AI literacy. Yet empirical evidence on how well ordinary users perform detection tasks is thin and geographically narrow, concentrated largely in North American and Western European samples (Groh, Epstein, Firestone and Pierson, 2022), with parallel findings on synthetic political video reported by Vaccari and Chadwick (2020). African scholarship has mapped the production and circulation of fake news and cyber-propaganda on the continent (Mare, Mabweazara and Moyo, 2019) but has said far less about whether audiences can actually recognise fabricated content when they meet it.

Zimbabwean university students represent a particularly instructive test case. They are digitally active,

politically aware and embedded in media environments where misinformation has repeatedly shaped public understanding of contested events. They are also the population most likely to be told, and to believe, that they are AI literate. Whether self-perceived literacy translates into reliable detection performance is a question that carries both scholarly and practical weight, because a population that is confident but inaccurate is arguably more vulnerable than one that knows it cannot tell.

This study addresses that question directly. The central research question is stated plainly. To what extent does AI literacy predict Zimbabwean university students' ability to detect synthetic media? Drawing on a cross-sectional survey of 518 students at the National University of Science and Technology (NUST), the study addresses this question through four subsidiary inquiries anchored in the four-aspect literacy framework developed by Ng, Leung, Chu, and Qiao (2021), which is fully introduced in the next section. First, what are the levels of AI literacy among NUST students across the technical, using, critical and ethical dimensions? Second, how accurately can students distinguish AI-generated from authentic content across a set of varied stimuli? Third, which dimensions of AI literacy are associated with detection accuracy after controlling for demographic and behavioural variables? Fourth, what AI-mediated online harms have students encountered and through what channels have they developed their detection knowledge?

The study contributes to a small but growing literature on AI literacy in African higher education and, more broadly, to debates about what being AI literate actually means when tested against objective performance rather than self-description.

LITERATURE REVIEW

Defining AI literacy

AI literacy has been conceptualised in several overlapping ways. Long and Magerko (2020) offer one of the most-cited frameworks, defining it as a set of competencies that enable individuals to evaluate AI technologies and use them effectively and ethically. Ng, Leung, Chu and Qiao (2021) propose a four-aspect model organised around knowing and understanding AI, using and applying AI, evaluating and creating with AI and engaging with AI ethically. Their framework treats ethics not as an afterthought but as a constitutive dimension of literacy itself. The present study adopts the Ng et al. (2021) model over alternatives for two reasons. It has been operationalised in survey instruments across multiple higher education contexts, which permits comparison, and its explicit separation of critical evaluation from technical knowledge and tool use maps directly onto the study's core question of which competencies actually support detection.

Concerns have been raised, however, about the gap between declarative knowledge of AI and the practical ability to act on that knowledge. Biagini (2024), reviewing the instruments available for measuring AI literacy, shows that the field relies overwhelmingly on self-perception scales and argues that what such tools capture is how literate people believe themselves to be rather than what they can demonstrably do. Self-report data of this kind can flatter respondents and overstate a population's preparedness to engage critically with synthetic content. That critique finds support in the social psychology literature. Kruger and Dunning's (1999) demonstration that individuals with limited competence systematically overestimate their ability offers one lens for interpreting mismatches between high literacy self-scores and modest detection performance, though as the discussion will show, the pattern in the present data is subtler than a simple confidence illusion.

The measurement debate matters because AI literacy is rapidly becoming a policy object. Universities, ministries and international agencies are writing AI literacy into graduate attribute statements and digital skills strategies, and what gets measured shapes what gets taught. If the instruments used to certify literacy reward self-description, institutions can declare success while leaving the underlying capabilities untouched. Long and Magerko (2020) make the related point that literacy competencies are not evenly distributed across educational tracks, which implies that generic institution-wide declarations conceal structured inequality in who can actually exercise the competencies being claimed. A framework adequate to the synthetic media era, therefore, needs at least one behavioural anchor, a task on which literacy claims can be tested rather than merely asserted. The present study provides an anchor by pairing the Ng et al. (2021) self-report dimensions with an objective detection task, making the correspondence between the two an empirical question rather than an assumption.

Synthetic media and detection research

Deepfakes and AI-generated media constitute a rapidly evolving category of synthetic content. Westerlund (2019) traces the trajectory of deepfake technology from novelty to mainstream concern, noting that the computational barrier to creating convincing synthetic faces fell sharply between 2017 and 2019. Chesney and Citron (2019) extend this analysis into the domain of political and personal harm, arguing that the ease of deepfake production fundamentally changes the threat model for disinformation, defamation and non-consensual imagery.

Groh et al. (2022) conducted one of the most rigorous empirical assessments of detection ability, testing over 15,000 participants in an online experiment. Their findings are sobering. Detection rates were only modestly above chance for AI-generated faces, and professional media experts performed no better than the general public. Two features of their results deserve emphasis because they inform the design and interpretation of the present study. Experience helped only when accompanied by feedback: participants improved across trials when told whether their judgements were correct, but not through mere exposure. Confidence was weakly coupled to accuracy, so participants who felt certain were scarcely more likely to be right. Both features suggest that detection is a trainable skill rather than an intuition that accumulates automatically with time online, which is exactly the premise on which curriculum-based interventions rest.

The difficulty of the task also has a technical basis that is worth stating. Early synthetic imagery exhibited recurrent artefacts, including malformed hands, inconsistent lighting, irregular teeth, and incoherent backgrounds, and popular detection advice still focuses on these cues. Each generation of image models weakens those cues, so heuristics learned informally against last year's fakes decay against this year's. Westerlund (2019) documented the collapse of the production barrier between 2017 and 2019, and the same trajectory has continued since then, meaning any fixed checklist of visual tells has a short shelf life. The implication for literacy education is that durable detection skills must rest on an evaluative process, meaning source checking, contextual reasoning, and calibrated suspicion, rather than on memorised artefact lists.

Vaccari and Chadwick (2020) found that exposure to deepfake political videos eroded audience confidence in video evidence more broadly, a finding with particular resonance for information environments like Zimbabwe's, where video is frequently cited as authoritative proof of contested events. Their result identifies the second failure mode of the synthetic media era. Alongside the risk of believing fabrications sits the risk of disbelieving everything, a generalised scepticism that serves manipulators almost as well because it licenses dismissal of authentic evidence. Köbis and Mossink (2021) add a useful boundary condition by showing that people cannot reliably distinguish AI-generated from human-written poetry. Although their stimuli were textual rather than visual, their study establishes that detection failure generalises across media types and is not an artefact of any single format, thereby justifying the treatment of visual detection as one instance of a broader discernment problem. Pennycook and Rand (2019) argue that analytic thinking, rather than prior beliefs or political identity, is the most reliable predictor of the accuracy of discernment across a range of media types. That claim aligns closely with the emphasis on critical evaluation in the Ng et al. (2021) framework and it generates the expectation, tested below, that the critical dimension should outpredict the others.

Information disorders in Africa

The global literature on misinformation has expanded considerably since the foundational work of Lazer et al. (2018) on the science of fake news and Wardle and Derakhshan's (2017) typology of information disorders. A distinct African research agenda has also taken shape. Mare, Mabweazara and Moyo (2019) argue that fake news and cyber-propaganda in Sub-Saharan Africa cannot be read through frameworks built for Northern media systems, pointing to the region's distinctive mix of state involvement in information manipulation, platform dependence and low institutional fact-checking capacity. Survey evidence from Kenya, Nigeria and South Africa shows that audiences believe they encounter misinformation routinely and struggle to judge the veracity of what they see (Wasserman & Madrid-Morales, 2019). Zimbabwe presents these dynamics in concentrated form. Reliance on WhatsApp as a primary news channel is high, a political history of contested information has made both official and unofficial sources objects of suspicion (Moyo, 2009) and the state has twice ordered internet shutdowns that intensified rather than suppressed the circulation of falsehoods (Mare, 2020).

What this regional literature has not yet produced is performance evidence. It documents supply, circulation and perception, but not the receiving end of the problem, namely, whether users can identify fabricated content when tested rather than asked. The distinction matters for policy. Interventions premised on supply, such as platform regulation and fact-checking, and interventions premised on reception, such as literacy education, compete for the same scarce institutional attention, and the case for the latter depends on demonstrating that reception is genuinely weak rather than merely assumed to be so.

The learning environment in which reception skills form is also distinctive. Hancock, Naaman and Levy (2020) argue that social media platforms are now the primary venues through which young people encounter and evaluate the credibility of digital content. Metzger and Flanagin (2013) add that online credibility assessment is a social process, shaped more by peer signals and platform affordances than by the systematic evaluation of content features. In media systems where formal media literacy education is thin, these informal processes are not a supplement to institutional teaching but a substitute for it. Whether that substitution produces adequate skills is an empirical question with two plausible answers. Informal learning may transmit practical, platform-current techniques faster than any syllabus could, or it may transmit a folklore of outdated visual tells and misplaced confidence. Which of these obtains in a Zimbabwean student population is precisely what a paired self-report and performance design can reveal.

Synthesis and Gap

Three conclusions emerge from the literature. Detection of synthetic media is difficult even for motivated and expert judges, and it improves with feedback and analytic processing rather than with raw exposure. Self-report literacy instruments dominate the field’s measurement practice and are known to capture confidence as much as competence. African audiences operate in information environments saturated with fabricated content, under platform conditions that make informal peer learning the default channel for acquiring evaluative skills. What the literature does not yet contain is a study that joins these three strands by testing, in a Southern African population, whether self-reported AI literacy corresponds to measured detection performance and by identifying which literacy dimensions carry the relationship. That is the gap this study fills, and the four research questions set out in the introduction operationalise it directly.

METHODOLOGY

The study used a cross-sectional quantitative-dominant design with task-based elements. A structured self-administered online questionnaire, built in Google Forms, combined four components. These were a demographic profile section, a self-report AI literacy scale, an objective detection task and a section on AI-mediated online harms and learning sources. Ethical clearance was granted by the NUST Institutional Review Board under approval number NUST/IRB/2026/114. Participation was voluntary and anonymous, with an informed consent statement at the opening screen and a debrief at the close. No identifying information was collected.

Data were collected in June - July 2026. The questionnaire link was shared on the WhatsApp classroom platforms, which enrol students from all NUST faculties. Because participation was open to enrolled students who chose to respond, the sampling strategy is properly described as convenience sampling. This choice has known limitations for generalisability, which are addressed in the conclusion, but it also has a compensating strength for the present purpose. A total of 518 usable responses were obtained. Table 1 presents the demographic profile of the sample.

Table 1. Demographic profile of the sample (N = 518)

Variable	Category	n	%
Faculty	Medicine	102	19.7
	Business and Economic Sciences	94	18.1
	Built Environment	66	12.7
	Information and Communication Sciences	92	17.8

	Applied Sciences	56	10.8
	Engineering	108	20.8
Age	18 to 20	187	36.1
	21 to 24	229	44.2
	25 to 29	71	13.7
	30 to 35	23	4.4
	Prefer not to say	8	1.5
Gender	Male	295	56.9
	Female	210	40.5
	Prefer not to say	13	2.5
Year of study	Part 1	98	18.9
	Part 2	115	22.2
	Part 3	127	24.5
	Part 4	108	20.8
	Part 5	43	8.3
	Postgraduate	27	5.2
Primary device	Smartphone	378	72.9
	Laptop or desktop	112	21.6
	Tablet	15	2.9
	Campus computer	13	2.5
AI tool use	Frequently	167	32.2
	Occasionally	198	38.2
	Rarely	97	18.7
	Never	56	10.8

The AI literacy scale comprised eleven items grouped across four dimensions drawn from Ng et al. (2021). These were technical understanding (3 items), using AI tools (2 items), critical evaluation (3 items) and ethical engagement (3 items). All items used a five-point Likert response format ranging from strongly disagree (1) to strongly agree (5). Internal consistency across the full scale was strong (Cronbach's alpha = 0.90), exceeding the conventional threshold of 0.70 (Nunnally, 1978).

The detection task presented six stimuli. Three were authentic images drawn from public domain photographic archives (a portrait, a news scene and a landscape) and three were AI generated images produced with freely available image generation tools (two synthetic faces and one fabricated event scene). The stimulus set was deliberately composed of content types that circulate widely on Zimbabwean social media platforms, namely portraits, news imagery and event scenes. Students judged each item as real or AI generated. Each correct classification earned one point, yielding scores from 0 to 6. The forced-choice design follows the paradigm used by Groh et al. (2022). A private answer key was maintained by the researcher and correct answers were not revealed within the form.

The final section asked students to indicate which AI mediated online harms they had encountered, with multiple selections permitted, and to identify the sources through which they had learned to assess digital content authenticity. Three open-ended questions invited students to describe their detection strategies, additional harms not listed and recommendations for improving AI literacy education.

Descriptive statistics were computed for all scale items and demographic variables. Pearson correlation coefficients were calculated among the four literacy dimensions and between each dimension and the detection score. A multiple linear regression model was estimated with detection score as the dependent variable and the four literacy dimensions, gender, age, faculty (Engineering dummy with Business and Economic Sciences as the reference category) and AI tool use frequency as predictors. Group differences in detection performance were

examined by gender (independent samples t-test) and faculty (one-way ANOVA). Open responses were analysed inductively following the six-phase thematic analysis procedure outlined by Braun and Clarke (2006), moving from familiarisation and initial coding to theme development and review.

RESULTS

AI literacy levels (RQ1)

Table 2 presents means and standard deviations for all scale items and the four composite dimension scores. The ethical dimension returned the highest mean (dimension $M = 4.04$), indicating that students hold strong normative commitments about responsible AI use. The technical dimension followed ($M = 3.81$), then the using dimension ($M = 3.41$), while the critical dimension scored lowest ($M = 3.15$). The overall scale mean was 3.58 ($SD = 0.57$). The internal ordering is itself informative. Students rate their principles more highly than their evaluative skills, and the single lowest-scoring item on the entire scale (Q16, $M = 2.89$) is the one asking directly whether they can identify signs that an image or video may be AI generated. Even at the level of self-description, students sense that recognition is their weakest capacity.

Table 2. AI literacy scale item and dimension means and standard deviations (N = 518)

Dimension	Item	Statement (abbreviated)	M	SD
Technical	Q10	I can explain how AI systems process data	3.6	0.74
	Q11	I understand how machine learning models learn from examples	3.7	0.73
	Q12	I am aware of common AI applications in everyday life	4.03	0.71
		Dimension mean	3.81	0.72
Using	Q13	I regularly use AI tools for study or work tasks	3.53	0.88
	Q14	I can adapt AI tools to suit different purposes	3.31	0.9
		Dimension mean	3.41	0.86
Critical	Q15	I can evaluate whether AI generated content is credible	3.35	0.85
	Q16	I can identify signs that an image or video may be AI generated	2.89	0.86
	Q17	I verify the source of online content before sharing it	3.13	0.84
		Dimension mean	3.15	0.82
Ethical	Q18	I believe AI developers have a duty to prevent misuse	3.93	0.71
	Q19	The spread of AI generated misinformation is a serious societal problem	4.13	0.67
	Q20	I have a personal responsibility to avoid spreading AI generated falsehoods	3.8	0.71
		Dimension mean	4.04	0.73
Overall scale			3.58	0.57

Detection performance (RQ2)

The mean detection score across the sample was 3.62 out of 6 ($SD = 1.54$), equivalent to correctly classifying 60.3% of stimuli. Figure 1 shows the percentage of students who correctly identified each of the six stimulus items and Figure 2 presents the full distribution of scores.

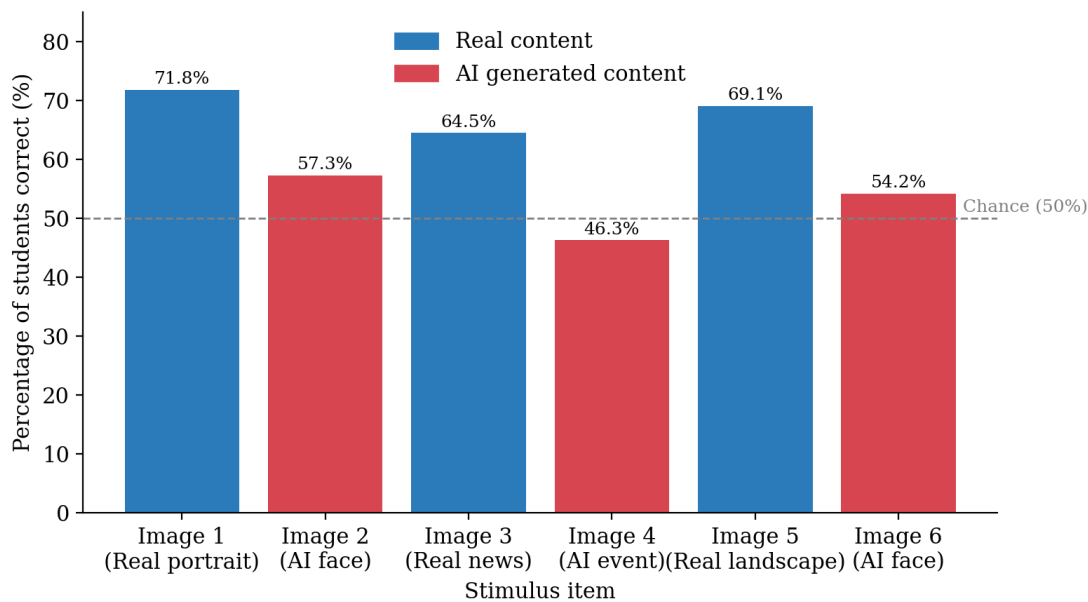


Figure 1. Percentage of students correctly identifying each stimulus item. The dashed line marks the 50% chance threshold. Real items are shown in blue and AI generated items in red.

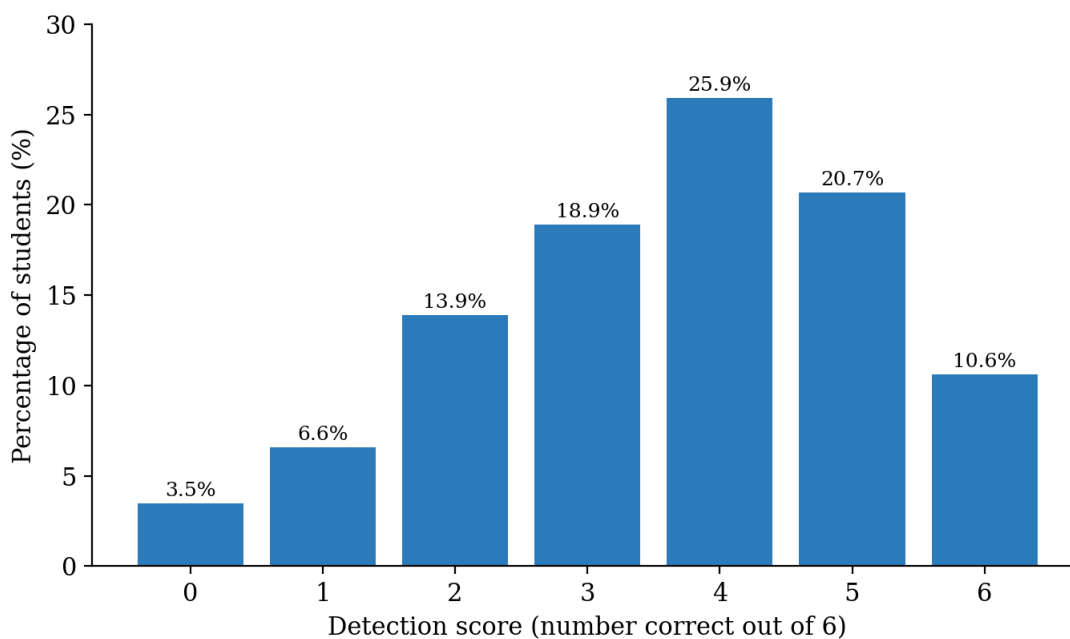


Figure 2. Distribution of detection scores (0 to 6 correct). The modal score was 4, achieved by 25.9% of students. Only 10.6% achieved a perfect score of 6.

Performance was uneven across stimuli. The real portrait produced the highest correct identification rate at 71.8%, while the AI-generated event scene yielded 46.3%, below the chance threshold. Three of the six items were correctly identified by fewer than 58% of students, confirming that AI-generated imagery poses a genuine detection challenge rather than a trivial task.

The unevenness has a systematic structure that raw accuracy conceals. Figure 3 aggregates performance by stimulus type. Mean accuracy on the three authentic items was 68.5%, whereas mean accuracy on the three AI-generated items was 52.6%, barely above chance. This sixteen-point asymmetry indicates a response bias toward judging content as real rather than a uniform inability to discriminate. Students appear to extend a default presumption of authenticity to what they see, which means the aggregate score of 60.3% flatters their ability to do what matters most in practice: catching the fake rather than confirming the genuine. Errors of this kind are also the more consequential in real information environments, since misclassifying a fabricated image as real is

the error that misinformation exploits.

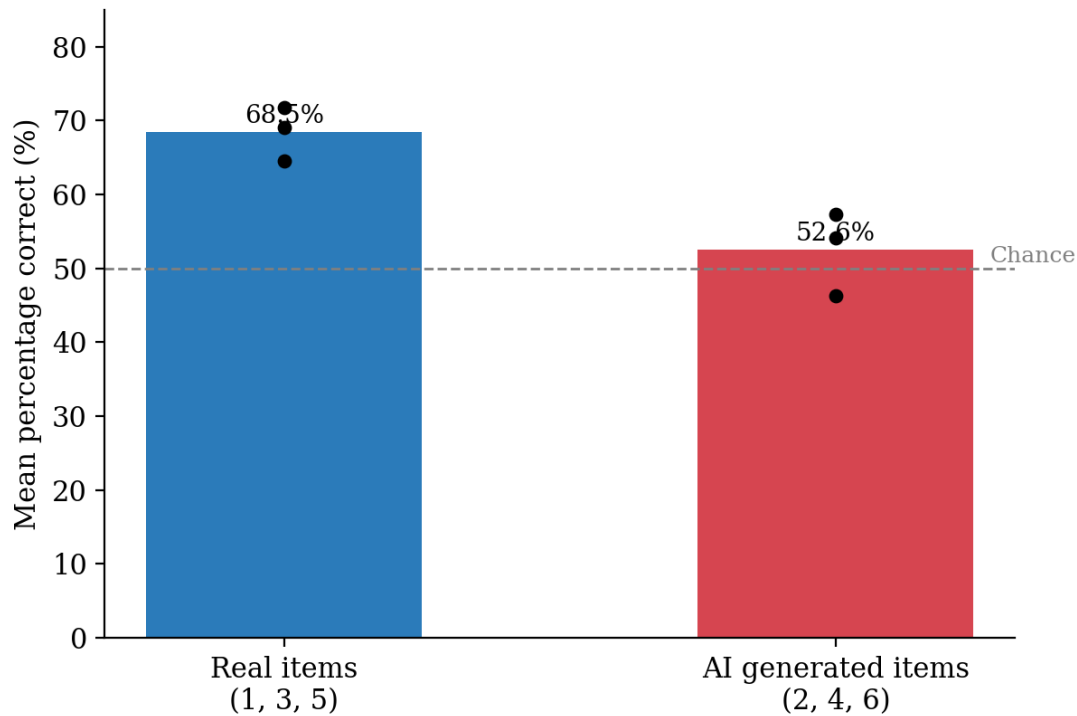


Figure 3. Mean detection accuracy for authentic items and AI-generated items. Black points show the three individual item accuracies within each category. The dashed line marks the 50% chance threshold.

Score distribution skewed toward the middle range. The modal score was 4 out of 6, achieved by 25.9% of students. Only 10.6% scored a perfect 6, while 10.0% scored 1 or below. An independent samples t-test revealed no significant gender difference in detection performance ($p = 0.404$), suggesting that ability is distributed broadly across gender lines.

Literacy dimensions and detection accuracy (RQ3)

Table 3 presents the Pearson correlation matrix among the four literacy dimensions and the detection score. Critical literacy showed the strongest bivariate association with detection performance ($r = 0.36$, $p < 0.001$), followed by technical literacy ($r = 0.26$, $p < 0.001$). The ethical and using dimensions showed weaker associations ($r = 0.17$ and $r = 0.14$ respectively). The four dimensions intercorrelate moderately ($r = 0.33$ to 0.53), consistent with a common underlying construct but not so strongly as to suggest redundancy.

Table 3. Pearson correlations among literacy dimensions and detection score (N = 518)

	Technical	Using	Critical	Ethical	Detection
Technical	1				
Using	0.53***	1			
Critical	0.46***	0.43***	1		
Ethical	0.41***	0.33***	0.44***	1	
Detection	0.26***	0.14**	0.36***	0.17***	1

Note. ** $p < 0.01$. *** $p < 0.001$.

Table 4 presents the multiple regression model. The overall model was significant ($F(8, 509) = 12.44$, $p < 0.001$) and explained 16.4% of variance in detection scores (adjusted R squared = 0.150). Critical literacy emerged as the strongest predictor (beta = 0.33), followed by technical literacy (beta = 0.14) and Engineering faculty membership (beta = 0.11), all significant at $p < 0.01$. Full coefficients appear in Table 4.

The using dimension deserves particular note. Its bivariate correlation with detection was positive yet its regression coefficient is negative and non-significant. The sign reversal indicates a suppression pattern. Whatever modest association tool use has with detection at the bivariate level is carried entirely by its overlap with critical and technical literacy, and once those are held constant, frequent tool use contributes nothing further and trends mildly negative. Familiarity with AI as a producer does not transfer to competence in recognising AI as an artefact. The same conclusion follows from the AI tool use frequency variable, which was likewise non-significant.

Table 4. Multiple linear regression predicting detection score (N = 518)

Predictor	B	SE	beta	t	p
Constant	0.38	0.621		0.611	0.541
Technical literacy	0.318	0.115	0.142	2.751	0.006
Using literacy	-0.145	0.09	-0.08	-1.605	0.108
Critical literacy	0.626	0.093	0.329	6.694	< 0.001
Ethical literacy	-0.008	0.109	-0.003	-0.069	0.945
Gender (male = 1)	0.095	0.127	0.031	0.753	0.451
Age	0.025	0.019	0.054	1.325	0.185
Engineering faculty	0.426	0.155	0.112	2.74	0.006
AI tool use frequency	-0.033	0.065	-0.021	-0.504	0.614

Note. R squared = 0.164. Adjusted R-squared = 0.150. F(8, 509) = 12.44, p < 0.001. The reference category for faculty is Business and Economic Sciences.

A one-way ANOVA showed significant differences among faculty in detection performance (F(5, 512) = 3.91, p = 0.002). Engineering students achieved the highest mean score (M = 3.98), followed by Information and Communication Sciences (M = 3.85) and Medicine (M = 3.58). Business and Economic Sciences students returned a mean of M = 3.52, while Applied Sciences students returned the lowest mean (M = 2.98).

AI Mediated Harms and Learning Sources (RQ4)

Figure 4 summarises responses to the harm-exposure and learning-source questions. Encountering false news stories was the most commonly reported harm (72.0%), followed by exposure to fake images or videos (65.1%). Just over 48% reported receiving scam messages and 31.1% had experienced some form of account impersonation. Only 12.0% reported no prior exposure to any AI-mediated harm, which shows the extent to which the student population is already operating in a synthetic media environment.

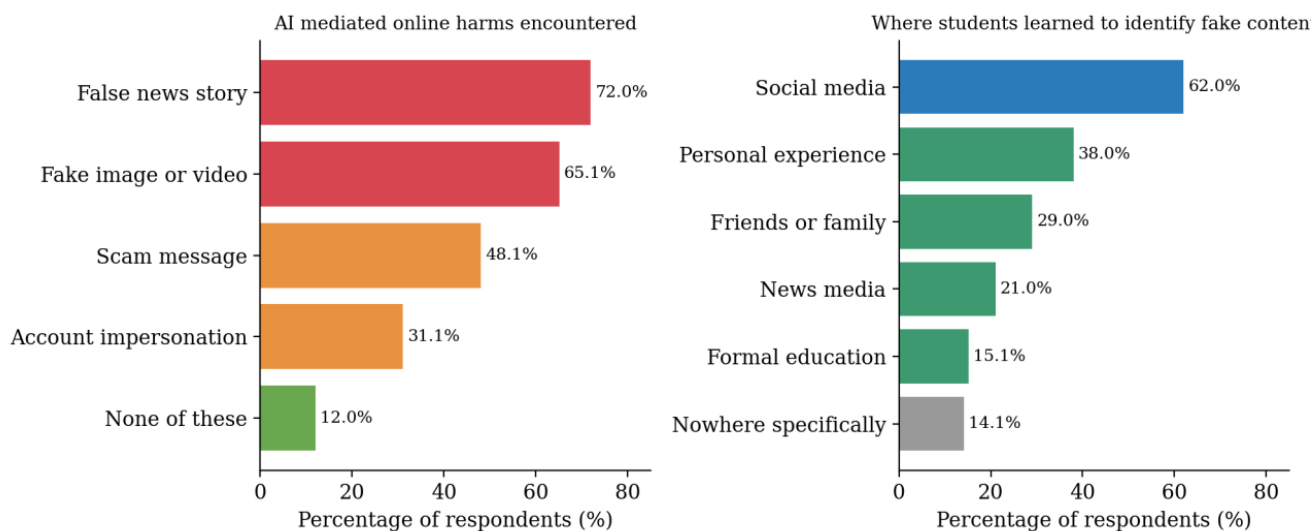


Figure 4. Prevalence of AI-mediated online harms encountered (left panel) and reported sources through which students learned to identify fake content (right panel). Multiple selections were permitted.

On the learning sources question, social media was cited by 62.0% of students, far outpacing any other channel. Personal experience ranked second at 38.0% and friends or family at 29.0%. Formal education was cited by only 15.1% and 14.1% of students reported they had not encountered any guidance on identifying fake content from any source. The dominance of informal and social channels over institutional learning is a structurally significant finding that bears directly on recommendations for curriculum design.

Qualitative themes

Analysis of the three open-ended questions produced a set of recurring themes, organised in Table 5. The themes illuminate the gap between normative awareness and systematic practice that the quantitative results suggest.

Table 5. Qualitative themes from open responses (N = 518)

Theme	Source	Illustrative student response
Physical anomaly detection	Q29	“I look at the teeth and hands. AI still gets those wrong most of the time.”
Contextual inconsistency	Q29	“Something looked off about the background, like the shadows did not match the light source.”
Emotional intuition	Q29	“I cannot always explain it but some images just feel wrong to me.”
Verification behaviour	Q32	“I search for the original source on Google Images before I believe or share anything.”
Curriculum integration	Q34	“Media literacy should be a compulsory module in all faculties, not just communication studies.”
Practical workshop training	Q34	“We need hands-on workshops where we actually practice spotting fakes, not just theory lectures.”

Two of the dominant detection strategies, noticing physical anomalies and relying on contextual incongruity, are consistent with the analytic mode of processing that Pennycook and Rand (2019) associate with accurate discernment. The prevalence of emotional intuition as a stated strategy is more ambiguous. Where intuition tracks genuine visual familiarity with photographic conventions, it may be productive, but in a social media environment saturated with AI-generated content, habituated exposure can attenuate rather than sharpen anomaly detection (Groh et al., 2022). Intuition is also exactly the strategy that the observed response bias would predict, since a default feeling that content is authentic produces correct answers on real items and errors on fabricated ones.

Student recommendations for curriculum integration and practical workshops were strikingly consistent across faculties. Respondents were broadly aware that formal education had not adequately equipped them and were specific about what they wanted: applied, task-oriented training rather than conceptual introductions to AI as a phenomenon.

DISCUSSION

The central finding of this research is a structural mismatch between self-reported AI literacy and actual detection performance. NUST students expressed moderate to high confidence across all four literacy dimensions yet classified only 60.3% of stimuli correctly, a margin that leaves substantial room for misclassification in real-world conditions, and their accuracy on the fabricated items specifically was barely distinguishable from guessing. This pattern is precisely the dynamic that Biagini (2024) identifies in reviewing the field’s measurement tools. Scales built on self-perception capture how literate people believe themselves to be and those beliefs can float free of the verifiable skills the scales are intended to proxy. The Dunning-Kruger framing is tempting here but imprecise. The data do not show that the least knowledgeable students are the most overconfident, and indeed, students gave their lowest self-rating on the one item asking directly about recognising AI-generated imagery. What the data show instead is a ceiling on how much self-reported awareness of any kind can buy in a task that demands perceptual discrimination. Students know roughly where their

weakness lies. They simply have nowhere institutional to take that knowledge.

Two alternative readings of the mismatch deserve consideration before the interpretation is accepted. One is that the detection task understates true ability because six static images are a narrow and difficult test. This cannot be ruled out entirely, but it runs counter to the evidence, since static images are generally easier to scrutinise than the compressed, fast-scrolling video content that dominates actual feeds, suggesting the task is more likely to overstate everyday performance than to understate it. The other is that the sample was unmotivated and answered carelessly. The score distribution argues otherwise. Performance clustered in a middle band with clear item-level structure, real items were identified at rates well above chance and the response pattern tracked stimulus type systematically, none of which is the signature of random responding.

The regression model clarifies which literacy dimensions are doing functional work. Critical literacy, the capacity to evaluate and interrogate AI generated content, was the strongest predictor of detection performance after controlling for demographics and behaviour. That result aligns with Pennycook and Rand's (2019) argument that analytic processing, not domain knowledge per se, drives accurate discernment. Technical literacy contributed additionally, though at a weaker effect size. The suppression pattern around the using dimension is the theoretically interesting null. Frequent engagement with AI tools showed a small positive bivariate association with detection yet contributed nothing once evaluative skill was held constant, trending mildly negative. Comfort with AI production does not generate AI detection competence and may even coexist with a habituated fluency that makes synthetic output feel unremarkable. For universities this null carries a warning. Strategies that equate AI literacy with hands-on tool exposure, currently the path of least resistance in curriculum reform, are unlikely to produce citizens who can recognise fabrication.

The faculty differences warrant close attention and more than one explanation. Engineering students outperformed other students by approximately 1.0 points on a 6-point scale, a gap large enough to be practically meaningful. The disciplinary interpretation is that Engineering programmes include components of signal processing, image analysis and computational reasoning that scaffold the systematic visual scrutiny detection requires. A rival interpretation is selection rather than training. Engineering admission is among the most competitive at the university, so the faculty effect may partly reflect prior general ability rather than anything the curriculum does. The present cross-sectional design cannot adjudicate between these accounts, and honesty requires saying so. What can be said is that the effect survives controls for age, gender and tool use, that the disciplinary account generates a testable prediction the selection account does not, namely that detection ability should improve across year levels within Engineering, and that either account supports the same practical conclusion. Detection ability is unevenly distributed across the university and its distribution follows structural lines rather than individual effort. Long and Magerko (2020) anticipated exactly this when they argued that AI literacy competencies track educational pathways and that universities have treated AI literacy as a generic graduate attribute rather than a skill requiring targeted pedagogical investment.

The harm exposure data locate these findings in their environment. Fewer than one in eight students reported encountering no AI mediated online harm, so the great majority are already navigating a space seeded with synthetic content. False news stories and fake images were the dominant harms, matching the typology Wardle and Derakhshan (2017) established and the structural dynamics of misinformation spread described by Lazer et al. (2018). The pattern also extends the African evidence base. Wasserman and Madrid-Morales (2019) found high perceived exposure to misinformation among audiences in Kenya, Nigeria and South Africa. The present data add the performance dimension that perception studies cannot supply and show that high exposure coexists with weak detection, which is the most dangerous combination the literature describes.

Perhaps the most actionable finding is the learning channel distribution. Social media is the dominant teacher of detection skills in this sample, cited by 62% of students, while formal education ranked fifth at 15.1%. Hancock et al. (2020) observed that young people treat platform-mediated peer signals as credibility proxies, and Metzger and Flanagin (2013) showed that online credibility judgement leans on social heuristics rather than systematic content evaluation. The present data suggest that in the absence of formal media literacy education those informal heuristics become the primary framework for navigating synthetic content. A counterargument deserves engagement here. Informal peer learning is not inherently inferior, since it is fast, free, adaptive and often carries exactly the platform-specific knowledge formal curricula lack, and some of the concrete strategies

students reported, such as reverse image searching, plainly circulate through these channels. The problem is not informality as such. The problem is unevenness and the absence of any quality assurance, so that whether a given student acquires reverse image search or acquires a vague trust in gut feeling is left to algorithmic chance. An institution that leaves detection training to platforms has not avoided teaching. It has delegated teaching to an environment optimised for engagement rather than accuracy.

The response bias finding sharpens what needs to be taught. Students' default presumption ran toward authenticity, with fabricated items accepted as real far more often than real items were rejected as fake. Vaccari and Chadwick (2020) warn of the opposite failure: a generalised scepticism in which exposure to deepfakes erodes trust in authentic evidence. Both failures are live in Zimbabwe, where video has been central to contested political narratives, and they define the pedagogical target between them. The goal of detection training is calibrated judgement rather than maximal suspicion, and calibration is a trainable evaluative skill (Masunda, 2024), which is exactly what the regression results say distinguishes successful detectors.

CONCLUSION

This study examined the extent to which AI literacy predicts Zimbabwean university students' ability to detect synthetic media. The answer is partially, unevenly and through one dimension above the others. The overall association is real and statistically strong, but the effect operates primarily through critical evaluation skills rather than through technical knowledge or tool familiarity. Students who systematically interrogate content outperform those who rely on technical awareness or intuition alone, and the frequency of AI tool use adds nothing once evaluative skill is accounted for.

Four contributions stand out. First, the study extends the AI literacy and detection literature to Southern Africa, moving the regional research agenda from documenting the supply of fabricated content to measuring the audience's capacity to recognise it. Second, it quantifies a capability gap between self-reported literacy and objective performance, including a marked bias toward presuming authenticity, that has policy-relevant implications for how AI literacy is taught and assessed. Third, it identifies critical literacy as the key functional dimension, indicating where investment will pay off for curriculum designers. Fourth, it documents the informal learning ecology through which most students currently develop detection skills, providing a baseline against which the impact of formal interventions can be measured.

Several limitations should be acknowledged. The sample is a convenience sample drawn from service course enrolments at a single university, which limits generalisability within Zimbabwe and across the region, even though it spans all six faculties. The detection task used six static images rather than the dynamic video deepfakes increasingly prominent in circulation, and a short task measures detection with limited precision. The literacy measure is self-report, so part of the literacy-detection relationship may reflect shared variance in reflective self-awareness, and one critical dimension item (Q16) asks directly about the ability the detection task measures, which risks inflating that dimension's association with the outcome. Re-estimation excluding this item, alongside signal detection modelling that separates sensitivity from response bias, is a priority for subsequent analysis of these data. The model explained 16.4% of variance, so most of what drives detection remains unmeasured, with cognitive reflection, prior deepfake exposure and visual media experience the obvious candidates for future models. Future research should also incorporate video stimuli, expand to multiple institutions and adopt longitudinal designs capable of capturing change following interventions. In Zimbabwe specifically, fielding detection tasks around electoral cycles, when synthetic content surges, would test whether the presumption of authenticity documented here holds at its highest stakes.

The practical recommendation is straightforward, even if implementation is not. Universities in Zimbabwe and comparable contexts should move AI literacy education beyond declarative content and beyond tool exposure. Detection is a trainable, assessable evaluative competency, and responsibility for training it sits with faculty curriculum committees and university teaching and learning offices rather than with any single department. The students themselves asked for exactly this in their open responses, applied and task-based training embedded across disciplines rather than confined to communication programmes. The curriculum case for delivering it is strong and the cost of continuing to outsource it to social media is already visible in these results.

REFERENCES

1. Biagini, G. (2024). Assessing the assessments: Toward a multidimensional approach to AI literacy. *Media Education*, 15(1), 91-101. <https://doi.org/10.36253/me-15831>
2. Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
3. Chesney, B. and Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753-1820. <https://doi.org/10.15779/Z38RV0D15J>
4. DataReportal. (2024). Digital 2024: Zimbabwe. <https://datareportal.com/reports/digital-2024-zimbabwe>
5. Groh, M., Epstein, Z., Firestone, C. and Pierson, E. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. <https://doi.org/10.1073/pnas.2110013119>
6. Hancock, J. T., Naaman, M. and Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1), 89-100. <https://doi.org/10.1093/jcmc/jzz022>
7. Köbis, N. and Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553. <https://doi.org/10.1016/j.chb.2020.106553>
8. Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
9. Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J. and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096. <https://doi.org/10.1126/science.aao2998>
10. Long, D. and Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-16. <https://doi.org/10.1145/3313831.3376727>
11. Mare, A. (2020). State-ordered Internet shutdowns and digital authoritarianism in Zimbabwe. *International Journal of Communication*, 14, 4244-4263.
12. Mare, A., Mabweazara, H. M. and Moyo, D. (2019). "Fake news" and cyber-propaganda in Sub-Saharan Africa: Recentering the research agenda. *African Journalism Studies*, 40(4), 1-12. <https://doi.org/10.1080/23743670.2020.1788295>
13. Masunda, O. (2024). Peace Education 4.0: A Curriculum Framework for Africa. *Academia Lasalliana Journal of Education and Humanities*, 6(1). <https://doi.org/10.55902/SJCE7127>
14. Metzger, M. J. and Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59, 210-220. <https://doi.org/10.1016/j.pragma.2013.07.012>
15. Moyo, D. (2009). Citizen journalism and the parallel market of information in Zimbabwe's 2008 election. *Journalism Studies*, 10(4), 551-567. <https://doi.org/10.1080/14616700902797291>
16. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W. and Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
17. Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
18. Pennycook, G. and Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50. <https://doi.org/10.1016/j.cognition.2018.06.011>
19. Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media and Society*, 6(1). <https://doi.org/10.1177/2056305120903408>
20. Wardle, C. and Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.
21. Wasserman, H. and Madrid-Morales, D. (2019). An exploratory study of "fake news" and media trust in Kenya, Nigeria and South Africa. *African Journalism Studies*, 40(1), 107-123.

<https://doi.org/10.1080/23743670.2019.1627230>

22. Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39-52. <https://doi.org/10.22215/timreview/1282>