# Heads we win, tails you lose: AI detectors in education.

Mark A. Bassett

*Office of Academic Quality, Standards and Integrity*
*Charles Sturt University, Australia*
*ORCID 0009-0007-0785-5844*

Wayne Bradshaw

*Library Services*
*James Cook University, Australia*
*ORCID 0000-0002-6379-0623*

Hannah Bornsztejn

*Education Portfolio*
*RMIT University, Australia*
*ORCID 0009-0007-6311-3239*

Alyce Hogg

*Division of Learning and Teaching*
*Charles Sturt University, Australia*
*ORCID 0009-0005-1126-7813*

Kane Murdoch

*Office of the Pro-Vice Chancellor and Dean of Students*
*Macquarie University, Australia*
*ORCID 0000-0002-1921-4944*

Bridget Pearce

*Academic Services*
*Brisbane Grammar School, Australia*
*ORCID 0009-0006-9950-1877*

Colin Webber

*Office of the Dean*
*SAE University College, Australia*
*ORCID 0009-0008-0201-0962*

Corresponding author:

Associate Professor Mark A. Bassett

Charles Sturt University

E: mbassett@csu.edu.au

## Abstract

The increasing use of generative artificial intelligence in student assessment has led to reliance on generative artificial intelligence detection tools by educational institutions. Unlike plagiarism detection, which identifies direct matches, AI detectors rely on unverifiable probabilistic assessments. False positives are indistinguishable from genuine cases. In this paper, we argue that generative artificial intelligence detection should not be used in education due to its methodological imperfections, violation of procedural fairness, and unverifiable outputs. Generative artificial intelligence detectors cannot be tested in real-world conditions where the true origin of a text is unknown. Attempts to validate results through linguistic markers, multiple tools, or comparisons with past work introduce confirmation bias rather than independent verification. Moreover, categorising text as human- or AI-generated imposes a false dichotomy that ignores work created with, not by, AI. Generative artificial intelligence detection also raises security concerns, as many tools lack transparency regarding data security. Academic integrity investigations must rely on evidence meeting the balance of probabilities standard, which generative artificial intelligence detection scores do not satisfy. Educational institutions and sectors should move away from punitive detection policies and focus on assessment design that integrates AI's role in learning, ensuring fairness, transparency, and pedagogical integrity.

Keywords: Artificial intelligence detection, generative artificial intelligence, academic integrity, higher education

## Introduction

Generative artificial intelligence (AI) text detection software and tools, including paraphrasing detection (hereafter, AI detectors), have become increasingly prevalent in education as institutions respond to the challenges posed by AI-generated content. Developers claim that AI detectors estimate the likelihood that a piece of writing was produced by generative artificial intelligence rather than a human, and their apparent effectiveness frequently justifies their use as a deterrent against student misconduct. However, their reliance on probabilistic inferences raises serious concerns about reliability, fairness, and due process. Instead of depending on unreliable generative artificial intelligence detection methods, educational institutions must rethink their approach to assessment design, academic integrity, and policy enforcement in an era increasingly shaped by the continued development of AI technologies.

## Problematic elements of artificial intelligence detectors

### *Unverifiable probabilistic estimates*

Generative artificial intelligence detectors estimate the likelihood that text was generated by AI by using linguistic markers that often differ between human and AI-generated text (Berber Sardinha, 2024). These markers, such as *perplexity* and *burstiness*, are combined with predictive modelling to generate probabilistic assessments of the presence and amount of AI-generated text in each document (Gehrmann et al., 2019). Probabilistic detection models underpin systems such as spam filters and medical diagnostics, achieving practical reliability despite their inherent uncertainty (Sharabov et al., 2024). These systems successfully operate on probability-based thresholds that produce ‘positive’ or ‘negative’ results without requiring absolute certainty (Sharabov et al., 2024). Like most diagnostic tests, AI detectors carry some risk of false positives and false negatives, but, unlike other probabilistic detection methods, their results cannot be independently verified in practice. In real-world conditions, no external evidence can conclusively confirm whether a flagged text was or was not AI-generated. Without a known ground truth, validation efforts rely on subjective interpretation or circular reasoning, rather than objective, independent verification.

Signal detection theory provides a useful framework for evaluating AI detectors. In controlled testing environments, developers and researchers use corpora with known proportions of AI- and human-generated content to measure detector performance. Each detection event can be classified as a hit (true positive), miss (false negative), false alarm (false positive), or correct rejection (true negative), corresponding to standard signal detection outcomes (Macmillan & Creelman, 2005; Wickens, 2002). Key performance metrics such as false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), and true negative rate (TNR) are derived from these outcomes, providing an assessment of the detector's sensitivity (its ability to correctly identify AI-generated text) and specificity (its ability to correctly reject human-written text) under *controlled conditions.* However, unlike controlled testing scenarios, real-world applications of AI detection lack ground truth data about the origin of a text, meaning there is no reliable way to verify the model's accuracy against actual conditions. Unverifiable results raise concerns about the usefulness of AI detectors outside of controlled conditions, particularly given the importance of due process in matters of academic integrity.

The human-generated texts used to train and test AI detectors were largely written before the advent of generative AI. For example, Turnitin tested its AI detector on '700,000 papers submitted before 2019 and therefore pre-dating GPT-3' (Turnitin, 2024). The likelihood that any of these texts were AI-generated is effectively zero. However, this testing methodology rests on the unverified assumption that student essays written before the widespread use of generative AI are directly comparable to those written today. It is far from certain that human-written text from a pre-generative AI era reflects the linguistic patterns and stylistic tendencies of contemporary student writing, which may be influenced by the AI-assisted texts they encounter.

### *Mutually exclusive linguistic markers*

A persistent flaw in the use of AI detectors is the assumption that linguistic markers of AI- and human-generated text are mutually exclusive. There is no principled reason to believe that a human cannot produce writing that contains linguistic features commonly found in AI-generated text. As exposure to AI-generated material becomes increasingly widespread, it is reasonable to expect that the linguistic patterns of human writing will shift, reflecting the influence of AI-assisted texts encountered across education, media, and everyday communication. It is therefore inevitable that some students will produce work that, despite being entirely their own, matches the statistical patterns detectors associate with AI generation. The critical issue is not whether false accusations will occur, but how frequently they will happen. Existing detector evaluations, based on student writing from a pre-generative AI era, provide no basis for estimating this risk. The presence of linguistic markers common in AI-generated text does not indicate that the text was written by AI any more than a student paper containing linguistic patterns similar to those of Shakespeare indicates that the student is Shakespeare. Regardless, many institutions continue to rely on AI detector outputs in isolation, leading to a statistical near-certainty that innocent students have been—and continue to be—wrongly found guilty of misconduct.

## Common methods used to validate artificial intelligence detector results

### *Linguistic markers*

A common method used to validate AI detector results and build a case against students is the identification of specific words, phrases, punctuation marks, or structural elements in their writing. Some argue that these so-called “AI hallmarks”, when considered alongside the detector’s result, strengthen the case for accusations of AI use. This reasoning, however, rests on the same flawed assumption as AI detectors themselves: that linguistic markers of human and AI-generated text are mutually exclusive. When an AI detector flags a text as AI-generated, staff may be tempted to search for features of the work commonly associated with generative AI as supporting evidence, including “unusually long and elaborate in formatting in answering the questions [or] near perfect punctuation and grammar” (Lewis, 2024). Upon finding such features, features that staff already suspect indicate AI use, but which are not exclusive to AI-generated content, they may present them as proof that the AI detector’s result was correct. In this scenario, there is no effort to provide *independent* verification at any stage of the process. Instead, staff reinforce their assumptions using reasoning that is entirely dependent on the AI detector’s potentially flawed output.

The presence of formulaic prose; lists; colons; em dashes; semicolons; paragraphs beginning with “Firstly”, “Secondly”, “Thirdly”; or the use of specific words like “delve” or “tapestry”, are not valid indicators that the text was AI-generated. These elements occur in AI-generated writing because they occur in the human writing that models are trained on. Stylistic elements do not, either in isolation or in conjunction with an AI detector’s result, serve as meaningful evidence that can be used to build a case against a student. Treating elements of style as evidence of AI use is not only methodologically unsound but also a clear example of confirmation bias, which can lead to wrongful accusations and serious penalties for students.

Selective attention is also at play when looking for text corroborating an AI detector’s result. Cognitive bias occurs when staff focus narrowly on information that aligns with their pre-existing suspicion (the belief that a student has used AI), while disregarding contextual or contradictory evidence that might challenge their assumption. For example, if an AI detector flags a student’s essay as AI-generated, staff treat this result as a preliminary *red flag* and begin scrutinising the text for linguistic features they associate with AI-generated writing. They might refer to lists such as GPTZero’s compilation of supposedly frequently used AI-generated phrases such as “it is essential to recognise,” “therefore,” and “in conclusion” (GPTZero, 2025). While these phrases are common in human and AI-generated writing, selective attention leads staff to interpret them as confirming AI use rather than considering their prevalence in human-authored academic work. Hunting for supporting evidence reinforces a confirmation bias loop where staff prioritise evidence that supports the AI detector’s result while overlooking counterexamples, such as personalised elements or assignment-specific context. Consequently, staff risk constructing a narrative of misconduct without fully assessing whether the available evidence substantiates the claim.

### *Multiple AI detectors*

Some educators attempt to validate an AI detector’s result by submitting the student’s text to multiple detectors, reasoning that if several tools identify the text as AI-generated, this constitutes strong evidence against the student. This approach does not provide independent verification. Instead, it amplifies the shared flaws of these tools, creating a misleading appearance of consensus. Even if all AI detectors agreed that a text was AI-generated, it would be no more validating than asking a group of phrenologists for a diagnosis—such consensus merely reflects shared flaws, not factual accuracy.

### *Falsified references*

Some educators use the presence of fabricated references as evidence that a student has used AI, given that some, though certainly not all, generative AI tools routinely produce fictitious citations. While it is unlikely that a student would independently fabricate references by inventing a source title, assigning it authors, and generating a non-existent DOI, it is not impossible. Therefore, the presence of falsified references is suggestive of generative AI use but is not definitive. Fortunately, in such cases, the question of AI use is irrelevant. Submitting fabricated references constitutes academic misconduct in its own right, regardless of how they were generated.

### ***Student confessions***

When confronted with an allegation of unauthorised AI use based on an AI detector's result, some students admit to using AI. This may seem to validate the detector's accuracy, but relying on such admissions as proof is a classic example of the *post hoc ergo propter hoc* fallacy. Alternatively, it might be considered an example of confession under duress. The reasoning mistakenly assumes that because the confession follows the AI detection flag, the flag must have been correct. Correlation, however, does not imply causation. Academic integrity must rise above believing in the equivalent of horoscopes, tarot cards, or Ouija boards, simply because their predictions occasionally seem to align with real events. The confession does not retrospectively validate the AI detector's result, just as a seemingly accurate horoscope does not prove astrology's legitimacy.

### ***Using generative AI***

Several techniques are being used by assessors that involve using generative AI itself to verify suspected AI use. All these techniques are fraught. Some educators attempt to determine whether a student has used AI by generating a comparison text using a Large Language Model (LLM) based on the assessment question or brief. They then compare the LLM's output to the student's work, not necessarily looking for exact matches but similarities in structure or content as indicators of AI use. This presents several methodological and logical concerns.

Confirmation bias again plays a significant role in this approach. Educators who generate an LLM response for comparison often do so under the assumption that the student has used AI. This predisposition can lead to selective attention, where similarities are highlighted while differences are overlooked. As a result, the comparison may reinforce pre-existing suspicions rather than provide objective verification.

LLMs produce text based on patterns in their training data, often adhering to predictable structures, especially when responding to standard prompts (Bender et al. 2021). A student's response may naturally follow similar patterns because it is shaped by the same assignment parameters, academic conventions, and disciplinary norms. The presence of shared features does not incontrovertibly establish AI authorship but might reflect the constraints of the task itself. Similarity in structure or argumentation also does not establish causation.

Assessments of "general similarity" and "identical structure" are subjective. Without defined benchmarks, the evaluation of whether two texts are sufficiently alike to suggest AI use will vary among assessors. What one staff member considers indicative of AI-generated text, another may see as a conventional academic response. The absence of clear standards introduces inconsistency into the evaluation process. On the other hand, setting such standards creates an arbitrary hurdle of "human-like-ness" for student writing to surmount, and what is a student to think if their writing is not viewed as sufficiently "human"?

Independent convergence of ideas is another factor to consider when comparing an AI exemplar to student writing. When responding to the same problem or topic, different authors, whether human or AI, may arrive at similar conclusions and use comparable phrasing. A broader concern is that this method may penalise students for adhering to academic conventions. If an LLM-generated response follows standard structures and argumentation patterns, and a student's work does the same, this approach risks treating conformity to disciplinary expectations as evidence of misconduct.

Another approach used by some educators is to submit the student's work to an LLM and simply ask it if it wrote the text or if the text was AI-generated. This approach is not a reliable method for detecting AI use, as it is based on the unsupported assumption that LLMs can accurately identify their outputs or distinguish between AI-generated and human-written text. LLMs cannot recognise their outputs (Raji et al., 2021). In some cases, an LLM will confidently—but wrongly—assert that it wrote a passage or that a given text is AI-generated. The model's confidence does not equate to accuracy, as LLMs lack the capacity to analyse authorship beyond pattern recognition (Raji et al., 2021). Without a mechanism to trace or verify whether a specific output was generated by AI, this method provides no meaningful evidence for evaluating academic integrity.

### *Past writing styles*

Comparing a student's past writing to their current work to detect AI use is shaped by confirmation bias. This approach assumes that differences indicate misconduct, leading staff to selectively focus on deviations while ignoring natural variation, academic growth, or contextual shifts in writing style. It is demonstrably true that writing evolves over time, due to feedback, subject familiarity, and changing assignment requirements. However, once suspicion is established, changes in clarity, structure, or vocabulary may be misinterpreted as evidence of AI use rather than legitimate progress. This bias-driven process lacks objective standards, making it unreliable and inconsistent. The method also disregards alternative explanations such as stress, illness, or "experimenting with different writing styles, genres, or linguistic patterns" (Giray et al., 2025), while reinforcing a culture of suspicion. Without transparency, students are judged against an unpublished and subjective benchmark, creating an unfair evaluative process.

### *Hidden adversarial prompts*

In a viral TikTok video, a Toronto-based English language teacher advocated for a "teacher hack" that embeds hidden "Trojan horse" prompts (also known as hidden adversarial prompts) in assessments to detect the use of generative AI (mondaysmadeeasy, 2025). Marian University has also promoted this approach that suggested staff "may be able to detect the use of AI by hiding some words in [their] assignment" and advises them to "include a word or two that would not normally be used in the essay for the assignment, then make the font white (to blend into the background) and as small as possible" (Marian University, 2023). Advocates of these "little traps" or "clever tricks" also suggest hiding an instruction that "asks for a citation from a journal that doesn't exist" (Katakam, 2024).

Laying traps for students in this way relies on deception, undermines trust between students and staff, and contradicts the principles of fair assessment and academic integrity. Furthermore, this strategy is contingent on *current* shortcomings in generative AI that can be quickly surmounted by training. This is a limited-time technique that is both ineffective and damaging to the relationships at the heart of education (Pratschke, 2024).

The issue at the heart of this method is that it assumes dishonesty by default, treating students as inherently deceptive rather than active participants in a learning environment built on mutual respect. Universities should lead by example, upholding the same ethical standards they expect from students.

## The burden of proof

The burden of proof lies with the institution to establish misconduct, not with the student to disprove the allegation. Students should not be required to prove their innocence or to respond to accusations while under investigation. Despite this fact, many innocent students feel compelled to defend themselves due to the power imbalance inherent in university disciplinary processes and the potential consequences of inaction. In practice, university procedures often diverge from the ideal application of the balance of probabilities standard. Students often lack the necessary resources, legal knowledge, or expert support to challenge allegations effectively. As a result, the misconduct process often looks as though the burden of proof has shifted onto the accused, even if this is not formally the case. The pressure to contest allegations, particularly when unrefuted claims may be treated as decisive, can create a situation that, in effect, compels students to demonstrate their innocence, even though the formal responsibility to prove misconduct rests with the university.

If a student attempts to prove they have not used generative AI, they face an inherently challenging, if not impossible task. Before considering how a student might provide such evidence, it is essential to emphasise that neither an inability nor a refusal to do so constitutes evidence of misconduct. A lack of evidence in the student's favour is not, in itself, evidence against them, nor does it strengthen the case for an allegation. Ultimately, while an absence of exculpatory evidence does not prove guilt, the way staff interpret and weigh such absences can have a significant impact on the decision-making process.

### *Evidence of process*

Drafts and revision history are often requested from students as evidence that a piece of work was developed and refined over time in response to an allegation (University of Sydney, 2022; University of Melbourne, 2023; RMIT University, 2025; Western Sydney University, 2024; University of Southern Queensland, 2025; University of Adelaide, 2025). Such requirements should be in writing and they must be provided to students in advance as part of the assessment brief. It is procedurally unfair to penalise a student for not producing documentation when they were not aware of the requirement prior to commencing work.

Software is available that claims to "bridg[e] the gap between students and educators" by surveilling students' writing process, providing staff with "video playback of draft history, along with insights into pasted text, typing patterns, and construction time" (Turnitin, 2025).

However, recent technology developments have rendered revision history an unreliable method for verifying authorship. OpenAI's Operator (OpenAI, 2025a) and Deep Research (OpenAI, 2025b) models can generate, edit, and iteratively refine a document over time, mimicking the human drafting process. The resulting revision history can be indistinguishable from that of a text written and edited by a human.

Beyond its practical limitations, mandating that students use software that surveils their every keystroke or maintain drafts and revision histories introduces risks to the educational process. Writing under the constant fear of being wrongly accused of misconduct may shape the way students approach their work. Instead of engaging freely with the material and experimenting with ideas, safe in the knowledge that their early drafts are for their eyes only, students may feel compelled to adopt formulaic, rigid structures that appear "safe" and uncontroversial. They may access a list of supposedly common AI words and phrases (GPTZero, 2025) and avoid these entirely. The defensive mindset created by attempts to hunt out generative AI use reduces assessments to exercises in compliance rather than opportunities for learning and expression. In this context, the insistence on surveillance software or keeping version histories, while procedurally advantageous, risks undermining the purpose of education.

### False positives

While developers and proponents of AI detectors often highlight low FPRs to justify their reliability, the fundamental constraint remains that no AI detection system can ever achieve a 0% false positive rate in practice, as a human could have plausibly written any text produced by generative AI. This inherent limitation renders AI detection percentage, whether high or low, statistically suggestive rather than definitive.

### *The base rate fallacy*

Consider an AI detector with a published FPR of 1%. This means that in testing, the detector incorrectly flagged 1% of human written papers as written by AI. Given this, staff might conclude that the chances of a paper being *correctly* flagged as AI-written (or containing AI-generated text) is 99%. This incorrect understanding appears to be pervasive in discussions about AI detectors in education.

The FPR of a detector reveals nothing about the probability that a flagged paper is AI-generated. To calculate this likelihood, one needs to know the FPR, the TPR *and* the base rate—the actual proportion of AI-generated papers in the population. The problem facing the practical use of AI detectors is obvious. We do not know the proportion of papers that were AI-generated in real-world scenarios. Therefore, in practice, we can *never* know the probability that a flagged paper is actually AI-generated.

Consider the following hypothetical example as an indication of the meaninglessness of AI-detector accuracy rates in practice. A hypothetical AI detector has an FPR of 1% and a TPR of 90% (in testing, it correctly flagged 90% of papers written by AI as such). 1000 papers were submitted to the detector, with 100 being AI-generated and 900 human-generated. Given these base rates, the probability that a paper identified as AI-generated is AI-generated is 90.9%. Yet, if 300 papers were AI-generated, the probability of the detector being correct is 97.5%. If 10 were AI-generated, the probability is 47.6%. Even when the FPR and TPR are known, the likelihood of a given flagged paper being AI-generated changes depending on the base rate, which cannot be known in practice.

### *False negatives*

Let's consider another hypothetical scenario where an AI detector somehow manages to achieve a 0% FPR. Even with this tool in hand, educational institutions still have a significant problem that undermines their core mission: that of false negatives. False negatives occur when an AI detector fails to identify AI-generated text. While the detector has a 0% FPR, it can still produce false negatives, particularly since AI detectors are easy to circumvent (Perkins et al., 2024). A tool that misses AI-generated text cannot be considered reliable, especially in high-stakes academic or professional settings. This reflects a classic "miss" in signal detection theory; a failure to detect a present signal, which undermines confidence in the tool's utility (Wickens, 2002).

While proponents of AI detectors might argue that their supposed overall effectiveness justifies their use, the presence of false negatives changes the nature of what these tools are actually detecting. Students who are adept at circumventing AI detectors can exploit their weaknesses, creating a disparity in how misconduct is identified and addressed. From this perspective, AI detectors do not identify AI use, so much as students' lack of skill with using generative AI or lack of access to more powerful generative AI tools.

Generative AI's outputs can easily be altered to evade detection (Perkins et al., 2024). However, excessive manipulation defeats the purpose if the resulting text is clearly unnatural or incoherent. Submitting AI-generated content that has been distorted into gibberish does not serve as meaningful evidence that AI detectors can be effectively bypassed. This raises an important question about how assessment validity is affected when AI-generated content can pass detection without any modification.

Consider a scientific report as an example assessment. In this case, the assessment instructions state that AI use is prohibited—even though this prohibition is impossible to enforce unless the assessment is supervised. Teaching staff submit all students' reports through our hypothetical 0% FPR AI detector, which flags 23 of the 50 reports as AI-generated. As our hypothetical AI detector has a 0% FPR, these 23 students are referred for academic misconduct. The other 27 are graded as per policy. In this example, however, 11 of the 27 reports are 100% AI-generated. These 11 are False Negatives, the other 16 are True

Negatives, and the 23 referred cases are True Positives, with no False Positives. At this point, staff may claim that they would be able to tell if 11 of the assessments were 100% AI-generated. Indeed, recent studies have shown that staff who frequently use generative AI for writing tasks "are highly accurate and robust detectors of AI-generated text without any additional training" (Russell et al., 2025). However, given that any text could have plausibly been written by a student, then no matter how good someone is at detecting AI, the detection is ultimately meaningless.

## The false dichotomy of artificial intelligence detection

Up until this point, we have been ignoring a major flaw underpinning all efforts to detect AI-generated material. We have been operating under the assumption that text is either entirely human-written or entirely AI-generated. This assumption is based upon a false dichotomy that does not accurately reflect the reality of how generative AI is used in contemporary writing, where human and AI contributions to a work are intermingled. The insistence that work must be categorised as either human or AI-generated ignores the blurred boundaries between these two modes of text production and renders the notion of AI detection conceptually flawed from the outset. Students' work is frequently created *with*, not *by,* generative AI. The use of these tools involves a hybrid process where AI assistance is incorporated in a variety of ways and at various stages in the writing process. This fluidity makes the binary approach of AI detection tools not merely inadequate but meaningless. The attempt to draw a strict line between human and AI writing creates more problems than it solves, fostering a climate of suspicion while failing to address the real challenges posed by AI in writing, assessment, or education at large.

## The boundaries of assessment

Determining the limits of AI use in education depends on establishing when an assessment begins. Institutional policies restricting unauthorised generative AI use usually refer to assessment-related activities, using phrases such as “in assessment” (Charles Sturt University, 2020), “to complete an assessment task” (University of Sydney, 2022), and “in an academic exercise” (Macquarie University, 2023). These terms imply a clear boundary but fail to specify when assessment formally starts. The issue is not simply one of policy wording but of conceptual ambiguity. If students are permitted to use AI outside of assessment but not within it, enforcement depends on identifying a precise threshold that, in practice, remains undefined.

Corbin et al. (2025b) highlight the uncertainty that rules about generative AI use create for both students and staff. Students, lacking explicit guidance, develop their own interpretations of when assessment begins and, by extension, when generative AI use is permissible. Some view AI-assisted brainstorming and editing as legitimate preparatory activities, while others fear that any engagement with AI before submission could be classified as misconduct. The absence of clear institutional boundaries forces students to navigate this ambiguity independently, leading to inconsistent self-regulation.

For educators, a lack of clarity on when and how AI can be legitimately used results in inconsistent enforcement. If assessment begins the moment students receive a task, AI-assisted research or planning could be deemed a violation. If assessment is defined strictly as the production of the final submission, then AI use in early drafting stages may be acceptable. Few institutions, however, do an adequate job of specifying which activities fall within the scope of assessment, and it is usually left to educators to interpret and enforce poorly defined boundaries. We find ourselves in a fragmented landscape where some staff attempt to prohibit the use of AI at any stage of the process, while others regulate only its use to produce final outputs. Without a formally recognised threshold, enforcement is subjective, and students face different standards of practice depending on who is marking their work.

The question of when assessment begins is further complicated by the nature of AI itself. Unlike most traditional forms of academic misconduct, generative AI use exists along a continuum. At what point does AI-assisted research become AI-generated content? When does generative AI-enhanced editing transition from editing to writing? Without a clear boundary, these distinctions remain arbitrary, and enforcement—if it is even possible to detect—depends on subjective judgements rather than principled criteria. Molenaar (2022) supports this view, proposing a model of hybrid human–AI intelligence that describes varying degrees of control and automation. This framework helps explain how authorship and agency shift as the locus of control moves from the student to the AI system.

The attempt to enforce a rigid boundary between human and AI-assisted writing not only fails in practice but also contradicts the principles of inclusive education. Many students, particularly those with disabilities, experience difficulties with written expression despite having a clear understanding of the content. When generative AI tools became widely available, they offered a powerful means for these students to express their ideas more effectively. For some, AI is not a shortcut but an essential tool akin to spell checkers, voice-to-text software, or screen readers. Yet rather than adapting assessment policies to acknowledge these realities, institutions are focused on penalising students for seeking assistance with their written expression, conflating such uses with academic misconduct.

The AI Assessment Scale (AIAS) (Furze, 2024a) attempts to categorise AI use in education, defining five levels of integration ranging from full prohibition (No AI) to unrestricted collaboration (AI Exploration). While it offers a structured framework for discussing AI's role in assessment, it does not function as an enforceable standard as it "attempt[s] to elicit compliance through language alone, without corresponding mechanisms to enforce those boundaries" (Corbin et al., 2025a). The authors acknowledge this limitation, conceding that the AIAS cannot prevent students from using AI in ways that fall outside its prescribed boundaries (Furze, 2024b)

On a fundamental level, the AIAS relies on the false premise that AI use in assessment can be segmented into discrete stages. AI Planning (Level 2), for example, allows AI for brainstorming but prohibits its use in drafting, while AI Collaboration (Level 3) permits AI-assisted writing as long as students critically engage with the output (Furze, 2024a). As we have already demonstrated, no clear boundary exists between work conducted by a human and an LLM. A student who generates structured notes using AI may need to engage in only minimal revision to convert those notes into a submittable output, such as an essay. The question of whether their use aligns with Level 2 or crosses into Level 3 is disputable. Similarly, distinguishing between superficial edits and substantive engagement in AI-assisted drafting is a subjective exercise, dependent on human judgment rather than verifiable criteria.

The difficulty of drawing definitive lines around generative AI use exposes a core problem with current AI policies. If educational institutions intend to regulate generative AI use in assessment, they must first define when assessment begins. In practice, this is an inherently absurd line to draw.

## Security risks

Submitting assessments to AI detection websites can pose a serious security risk for students. Many platforms do not provide clear policies on where data is stored, who can access it, or whether it is shared with third parties. This lack of transparency raises concerns about privacy, intellectual property, and institutional compliance with data protection laws. If universities do not assert ownership of student writing, it is quite possible that submitting work to a generative AI tool may amount to a breach of copyright in some cases. A greater risk is the uncertainty around data storage and retention. Many AI detection tools do not specify how long they keep submitted work or whether students can request its deletion.

In the event of a data breach, student submissions could be exposed to the public, creating risks for academic integrity and personal privacy. Some AI detection tools may expose student work to unauthorised access, including the possibility of misuse or commercial exploitation. Without transparency about where and how submissions are stored, there is no way to ensure that student work is not repurposed for other uses, including inclusion in databases accessible to contract cheating services.

Some detection services store student work on overseas servers (GPTZero, 2024; Copyleaks, 2022), adding another layer of risk to their use. Data protection laws vary by country, meaning that student work stored on an international server may not be protected by the same privacy standards that apply in their home institution. Some jurisdictions allow governments to access stored data without requiring consent from the user. Others permit commercial use of stored documents in ways that students may not expect or consent to. Once a submission is held in an international database, it may be impossible to track how it is used or whether it can ever be removed.

If student work is to be submitted to third-party websites or tools, the safest way to do so is to ensure the institution has a formal contract with the provider. AI detectors—or other generative AI tools—are no different in this regard. Without a binding agreement that guarantees data security, privacy, and control over stored work, universities cannot justify exposing students to the risks outlined above.

### Balance of probabilities

In many universities, the standard of proof for academic misconduct is the common law *balance of probabilities* standard (also known as the preponderance of evidence), rather than the criminal law standard of *beyond a reasonable doubt* (Western Sydney University, 2023; Queensland University of Technology, 2023; University of York, 2023; University of Alberta, n.d.; University of Kent, n.d.). The balance of probabilities requires the university to demonstrate, based on actual evidence rather than subjective impressions, that it is *more likely than not* that the student committed academic misconduct. Evidence supporting the university's claim must be credible, relevant, and probative. Types of evidence that fail to meet this standard include AI detector results, single or multiple; linguistic markers; comparing student work to an AI-generated response; generative AI's claim that the text is AI-generated; changes in writing style compared to past work; student silence in response to allegations; confessions under pressure; and absence of drafts or revision history. The above is a list of *indicators*, flags that may warrant further investigation. But even when combined, any combination of evidence from this list will never reach the level of "more likely than not".

### *Students' right to silence*

Like all accused in legitimate legal systems, students under investigation must be afforded the right to silence; they do not have to speak on their own behalf. Furthermore, a refusal to respond does not tip the scales against them. Staff can request that students speak to their work, but there is a clear line between requesting an oral response to verify that a student has met the learning outcomes and questioning a student about academic integrity-related issues. The critical distinction here is that a student's right to silence exists if they are under investigation. As part of an assessment, *Academic* staff can request a student to speak to their work and provide additional information confirming they have met the learning outcomes. A student can refuse this request, but it would result in failing the assessment. Such a case is a grading issue rather than an integrity one. Of course, refusing to respond or not being able to speak to their own work is strong evidence that might be used against a student in the event of an academic integrity investigation. Once placed under investigation, however, students have the right to silence.

## Conclusion

AI detection in education is not merely flawed; it is conceptually unsound. These tools operate on unverifiable probabilistic assessments that cannot meet the evidentiary threshold required for academic integrity investigations. The attempt to categorise text as either human- or AI-generated ignores the fluid reality of contemporary writing, where AI-assisted work exists along a continuum. Efforts to validate AI detection through linguistic markers, multiple tools, or comparisons with past work amount to confirmation bias rather than independent verification. Worse, reliance on AI detectors and surveillance software fosters a climate of suspicion, undermining student trust and eroding the integrity of assessment itself.

Institutions must accept that AI detection is an unworkable solution to a problem that cannot be solved through surveillance and punishment. The focus must move from detection and enforcement to assessment design that recognises AI's role in learning and the reality that unsupervised assessments cannot be secured. The continued use of AI detectors exposes students to procedural injustices and signals a fundamental misunderstanding of education's purpose. AI detection does not safeguard academic integrity; it undermines it.

*Disclosure of interest*: The authors report no competing interests to declare.

*Disclosure of use of generative AI tools*: Generative AI (ChatGPT-4o) was used in the preparation of this manuscript to edit, review, and critique the document.

Acknowledgement: The authors thank Professor Jason M. Lodge for his assistance in conceptualising this research.

## References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. Applied Corpus Linguistics, 4(1), 100083. https://doi.org/10.1016/j.acorp.2023.100083

Charles Sturt University. (2020). Student Misconduct Rule 2020. CSU Policy Library. https://policy.csu.edu.au/document/view-current.php?id=501

Copyleaks. (2022). Terms of use. Copyleaks. https://copyleaks.com/termsofuse

Corbin, T., Dawson, P. and Liu, D. (2025): Talk is cheap: why structural assessment changes are needed for a time of GenAI, Assessment & Evaluation in Higher Education. https://doi.org/10.1080/02602938.2025.2503964

Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025). 'Where's the line? It's an absurd line': Towards a framework for acceptable uses of AI in assessment. Assessment & Evaluation in Higher Education, 1–13. https://doi.org/10.1080/02602938.2025.2456207

Furze, L. (2024, August 28). Updating the AI assessment scale. Leon Furze. https://leonfurze.com/2024/08/28/updating-the-ai-assessment-scale/

Furze, L. (2024, August 9). Can the AI assessment scale stop students cheating with AI? Leon Furze. https://leonfurze.com/2024/08/09/can-the-ai-assessment-scale-stop-students-cheating-with-ai/

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. arXiv. https://doi.org/10.48550/arxiv.1906.04043

Giray, L., Sevnarayan, K., & Ranjbaran Madiseh, F. (2025). Beyond policing: AI writing detection tools, trust, academic integrity, and their implications for college writing. Internet Reference Services Quarterly, 1–34. https://doi.org/10.1080/10875301.2024.2437174

GPTZero. (2024). GPTZero terms of use. GPTZero.me. https://gptzero.me/terms-of-use.html#userdata

GPTZero. (2025). Discover the most common AI vocabulary words. GPTZero. https://gptzero.me/ai-vocabulary

Katakam, M. (2024, February 9). AI or not AI? That is the question: Unravelling the modern Trojan horse in our digital world. Medium. https://medium.com/@maheshkatakam/ai-or-not-ai-that-is-the-question-unraveling-the-modern-trojan-horse-in-our-digital-world-16caf89bf5cf

Lewis, C. (2024, April 24). Yale SOM student suspended over alleged AI use sues university. New Haven Register. https://www.nhregister.com/news/article/yale-som-student-suspended-alleged-ai-use-sues-20206927.php

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide (2nd ed.). Lawrence Erlbaum Associates.

Macquarie University. (2023). Academic Integrity Policy. Policy Central. https://policies.mq.edu.au/document/view.php?id=3

Marian University. (2023). LibGuides: Artificial intelligence in education: Detection. Marian.edu. https://libguides.marian.edu/c.php?g=1321167&p=9721351

Molenaar, I. (2022). Towards hybrid human–AI learning technologies. European Journal of Education, 57(4), 542–555. https://doi.org/10.1111/ejed.12527

mondaysmadeeasy. (2025). TikTok - Make your day. TikTok.com. https://www.tiktok.com/@mondaysmadeeasy/video/7304804982673476870

OpenAI. (2025). Introducing Operator. OpenAI.com. https://openai.com/index/introducing-operator/

OpenAI. (2025). Introducing deep research. OpenAI.com. http://openai.com/index/introducing-deep-research/

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). GenAI detection tools, adversarial techniques and implications for inclusivity in higher education [Preprint]. arXiv. https://arxiv.org/abs/2403.19148

Pratschke, B. M. (2024). Generative AI and education. Springer Nature. https://doi.org/10.1007/978-3-031-67991-9

Queensland University of Technology. (2023). Manual of policies and procedures: E/2.1 Academic integrity. https://mopp.qut.edu.au/document/view.php?id=15

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 207–217). ACM. https://doi.org/10.1145/3442188.3445898

RMIT University. (2025). Academic integrity. https://www.rmit.edu.au/students/my-course/assessment-results/academic-integrity

Russell, J., Karpinska, M., & Iyyer, M. (2025). People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. arXiv. https://doi.org/10.48550/arxiv.2501.15654

Sapling. (n.d.). AI detector (GPT / ChatGPT). Sapling.ai. https://sapling.ai/ai-content-detector

Sharabov, M., Tsochev, G., Gancheva, V., & Tasheva, A. (2024). Filtering and detection of real-time spam mail based on a Bayesian approach in university networks. Electronics, 13(2), 374. https://doi.org/10.3390/electronics13020374

The University of Adelaide. (2025). Suspecting academic misconduct. https://www.adelaide.edu.au/learning/resources-for-educators/academic-integrity/steps-in-the-process/suspecting-academic-misconduct

Turnitin. (2024). Turnitin's AI writing detection model architecture and testing protocol. Turnitin.com. https://www.turnitin.com/whitepapers/turnitin-ai-writing-detection-model-architecture-and-testing-protocol

Turnitin. (2025). Turnitin Clarity: Student writing transparency. Turnitin.com. https://www.turnitin.com/campaigns/clarity/

University of Alberta. (n.d.). Proving academic misconduct: Information for instructors. https://www.ualberta.ca/en/dean-of-students/media-library/documents/academic-integrity/provingmisconduct.pdf

University of Kent. (n.d.). Academic integrity: Glossary of terms. https://student.kent.ac.uk/studies/academic-integrity/list-of-terms

University of Melbourne. (2023). Student academic integrity policy (MPF1310). https://policy.unimelb.edu.au/MPF1310/

University of Southern Queensland. (2025). Responding to an allegation of academic misconduct. https://www.unisq.edu.au/current-students/academic/academic-integrity/responding-to-allegation

University of Sydney. (2022). Academic Integrity Policy 2022. https://www.sydney.edu.au/policies/showdoc.aspx?recnum=PDOC2012/254

University of York. (2023). Academic misconduct policy 2023–24. https://www.york.ac.uk/media/abouttheuniversity/supportservices/academicregistry/registryservices/sca/guidetoassessment/University-of-York-Academic-Misconduct-Policy-2023-24.pdf

Western Sydney University. (2023, March 15). Student misconduct rule procedures. https://policies.westernsydney.edu.au/document/view.current.php?id=348

Western Sydney University. (2024). Academic integrity guide. Learning Futures. https://www.westernsydney.edu.au/learning_futures/home/teaching_support/academic_integrity_guide

Wickens, T. D. (2002). Elementary signal detection theory. Oxford University Press.