# Heads We Win, Tails You Lose: Why AI Detection Cannot Be the Answer to Academic Integrity

White Paper

July 2026

UNIwise

# Heads We Win, Tails You Lose: Why AI Detection Cannot Be the Answer to Academic Integrity

Every semester, a growing number of higher education institutions face the same tempting proposition: run student writing through an AI detector, receive a percentage, and treat that percentage as evidence. It promises to restore certainty to a problem that feels newly chaotic. The trouble is that the promise does not hold up. Both the peer-reviewed literature and the public claims made by the leading commercial detection vendors converge on the same uncomfortable conclusion: generative AI detection is not simply imperfect, it is structurally unable to deliver what it is marketed to do.

This paper draws on a peer-reviewed analysis published in the Journal of Higher Education Policy and Management, Heads We Win, Tails You Lose: AI Detectors in Education (Bassett et al., 2026), together with independent research on detector accuracy and the marketing claims of the leading commercial AI detection tools. It sets out why AI detection fails on its own terms, why the informal methods institutions use to corroborate a detector's result compound rather than solve the problem, and why UNIwise has deliberately chosen not to build AI-generation detection into WISEflow Originality. Our position is that similarity detection, grounded in semantic comparison, exact-match and paraphrase analysis against identifiable sources, is the approach that actually meets the evidentiary standard higher education is required to work to.

## The appeal of a simple answer

Since generative AI became freely available in late 2022, institutions have faced a genuine and legitimate problem: assessments designed to certify a student's own, unaided capability can now be produced, in whole or in part, by a machine in seconds. The instinct to reach for a technical fix is understandable. A detector that returns a percentage feels objective, fast and scalable in a way that redesigning assessment does not.

But an AI detector does not identify AI use. It estimates, on the basis of statistical patterns such as perplexity (how predictable the word choices are) and burstiness (how much sentence length and structure varies), the probability that a given piece of text resembles machine-generated writing more than human writing (Bassett et al., 2026, citing Kehkashan et al., 2025; Gehrmann et al., 2019). That distinction, between detecting AI and estimating a resemblance to AI, is where almost everything that follows in this paper begins.

## Why the maths defeats the promise

Every diagnostic test, whether a spam filter, a medical screening tool or an AI detector, produces four possible outcomes: it can correctly flag a positive case, correctly clear a negative case, wrongly flag a negative case (a false positive), or wrongly clear a positive case (a false negative). Spam filters and medical screening tools work because their results can, in principle, be checked against a known reality: the email was or was not spam; the patient does or does not have the condition. AI detectors have no equivalent ground truth. Once a

piece of writing has been submitted, there is no reliable, independent way to establish what actually produced it (Bassett et al., 2026).

This absence of ground truth interacts badly with a second, more subtle problem: the base rate fallacy. A detector's advertised false positive rate does not tell an assessor the probability that a specific flagged script is actually AI-generated; that also depends on how common AI use actually is in the cohort being screened, a figure no institution can know with any precision. Bassett et al. (2026) illustrate this with a hypothetical detector that has a 1% false positive rate and a 90% true positive rate. Screened against 1,000 papers of which 100 are AI-generated, a flagged paper is genuinely AI-generated 90.9% of the time. Screened with the same detector where only 10 of 1,000 papers are AI-generated, the same flag is correct only 47.6% of the time – worse than a coin toss. The detector has not changed. Its accuracy figures have not changed. What changes is a number no institution can ever actually measure: how much AI use is really happening in that cohort. That is the base rate fallacy, and it is not a flaw that better engineering fixes. It is inherent to using a probabilistic tool in a setting where the true prevalence of the thing being detected is permanently unknown.

## The gap between marketing and evidence

The leading commercial AI detectors market themselves on very high accuracy and very low false positive rates. Those figures are worth setting directly against the independent research record.

| Vendor claim | Independent evidence |
|---|---|
| Turnitin states its AI writing detection carries "a less than 1% false positive rate" at document level for documents its model assesses as containing 20% or more AI writing, alongside a roughly 4% false positive rate at sentence level (Turnitin, 2023a, 2023b). | Independent testing reported by The Washington Post found Turnitin's tool "identified over half" of a 16-essay test sample "at least partly incorrectly," including flagging a wholly human-written essay as partly AI-generated (Fowler, 2023). Turnitin's own detection model was also trained and tested substantially on writing submitted before generative AI existed, which the authors argue cannot be assumed to represent how students write today (Turnitin, 2024, cited in Bassett et al., 2026). |
| GPTZero advertises "99% accuracy" and describes itself as "the only AI detector de-biased for ESL learners," claiming to hold its false positive rate for second-language writing to around 1% (GPTZero, 2026). | Liang et al. (2023), published in Patterns, tested seven commercial GPT detectors on TOEFL essays written by non-native English speakers and found a 61.3% false positive rate, against near-zero for native-speaker essays, because non-native writing tends to use more predictable vocabulary – the same trait detectors associate with AI. Weber-Wulff et al. (2023), testing 14 detection tools, found that none reached 80% accuracy and that accuracy fell further once text was paraphrased. |
| Copyleaks claims "over 99% accuracy" and an "industry-low .03% false positive rate" (Copyleaks, 2026). | The same Weber-Wulff et al. (2023) evaluation included Copyleaks among the 14 tools tested and found the field as a whole biased toward classifying text as human rather than AI, with accuracy deteriorating under paraphrasing or light |

| | |
|---|---|
| | editing – precisely the conditions real student submissions present. |

None of this means the vendors are acting in bad faith. Accuracy figures produced under controlled test conditions, with known documents and a known proportion of AI-generated text, can be entirely genuine and still tell an institution almost nothing about how the tool will perform against an unknown student submission, for the same base rate reasons set out above. The gap is not one of honesty; it is the gap between a laboratory and a classroom.

The consequences of that gap are not abstract. Weixin Liang, a co-author of the Stanford study, put it plainly: "The design of many GPT detectors inherently discriminates against non-native authors, particularly those exhibiting restricted linguistic diversity and word choice" (quoted in García Mathewson, 2023). Reporting by The Markup documented real cases of international students whose properly drafted, properly evidenced essays were flagged – students who then had to defend work they had genuinely written (García Mathewson, 2023). A tool marketed as protecting academic integrity was, in these cases, actively undermining the fairness of the process meant to protect the students it was screening.

## A false dichotomy: written by AI, or written with AI

Even a hypothetically perfect detector would run into a second, deeper problem. AI detection assumes that a piece of writing is either entirely human or entirely AI-generated. That assumption no longer describes how students actually write. Generative AI is now embedded in spelling and grammar tools, translation software, search engines and note-taking apps; brainstorming with AI, having AI restructure a paragraph, or using it to check register are all now ordinary parts of many students' writing processes, alongside entirely independent thinking (Bassett et al., 2026, citing Molenaar, 2022). A binary detector cannot represent that reality. It can only force a continuum into two boxes, and whichever box a genuinely mixed piece of work lands in will be wrong.

## The presumption of innocence, inverted

In most institutions, the standard of proof for academic misconduct is the balance of probabilities: the institution must show it is more likely than not that misconduct occurred, using evidence that is credible, relevant and probative (Bassett et al., 2026, citing Queensland University of Technology, 2023; University of York, 2023). An AI detector's output, precisely because its accuracy cannot be verified against the case in front of the assessor, does not meet that standard on its own. Nor, the authors argue, do the informal methods institutions commonly use to corroborate it: searching the text for supposed AI hallmarks such as em dashes or the word "delve" (which occur throughout human writing, because AI is trained on human writing); running the same text through several detectors, which only reproduces their shared flaws rather than independently confirming anything; generating a comparison essay with an LLM and looking for similarity of structure, which risks penalising students for following the very conventions their course has taught them; or treating a change in writing style as suspicious, when style changes for entirely ordinary

reasons such as feedback, stress or a new topic. Each of these methods looks like additional evidence. None of them is independent of the same underlying, unverifiable estimate.

The practical effect is that the burden of proof quietly shifts. A flagged student is asked to explain themselves, produce drafts, or otherwise demonstrate their innocence, even where formal policy places the burden on the institution (Bassett et al., 2026). Given that no student can conclusively prove a negative – that a piece of writing was not produced by AI – this is a burden that is, in a meaningful number of cases, impossible to discharge, however honest the student.

## Shadow detection: the legal exposure institutions are not signing up for

A separate risk sits alongside the accuracy problem, and it arises less from institutional policy than from its absence. In practice, many AI-detection episodes begin with an individual member of staff submitting a student's work to a free or consumer AI-detection website on their own initiative – behind the institution's back, in effect, rather than as part of any sanctioned process, with no data processing agreement and often without the student's knowledge (Bassett et al., 2026). UNIwise has made the same point in relation to general-purpose chatbots used for marking, and the logic applies equally to AI detectors: "Pasting student work into a general-purpose chatbot often sends personal data outside the EU, hands it to systems that may train on it, and quietly compromises both the privacy and the intellectual property of the students whose work it is. The student never agreed to that. The institution rarely knows it is happening" (UNIwise, 2026a).

Two distinct legal problems follow. The first is a straightforward data protection failure. Student coursework is personal data, and submitting it to a third-party processor without a data processing agreement, without a documented legal basis, and often without visibility over where the data is stored or for how long, is very difficult to reconcile with an institution's obligations as a data controller under the GDPR. Bassett et al. (2026) note that some detection services store submitted work on overseas servers (Copyleaks, 2022; GPTZero, 2024), where local data protection standards may not match those the student's home institution is bound by, and that many providers are silent on retention periods or deletion rights. When a member of staff runs a script through a browser-based detector independently of the institution's own systems and contracts, the institution typically has no visibility that an international transfer of student data has taken place at all – until, potentially, a regulator or a student asks.

The second is an intellectual property problem that is easy to overlook. A student's essay, dissertation or code is their own copyright-protected work. Submitting it to a third-party AI system, particularly one whose terms of use reserve the right to retain or use submitted content to improve its models, risks handing value in that work to a commercial third party without the student's informed consent. Institutions that would never allow a member of staff to publish a student's unpublished dissertation on a public website without permission are, in practice, exposed to something structurally similar every time a script is pasted into an uncontracted AI detector. None of this is an argument that only applies to a handful of individual staff members acting alone; it is a governance gap institutions need to close deliberately, not a workaround to quietly tolerate.

## Explainability and the EU AI Act: the black box problem is now a compliance problem

The case against AI detection is no longer only a methodological one. Under the EU Artificial Intelligence Act (Regulation (EU) 2024/1689), AI systems used to evaluate learning outcomes, or to monitor and detect prohibited behaviour of students during tests and assessments, sit within Annex III, the list of AI uses classified as high-risk (European Parliament and Council, 2024, Annex III; UNIwise, 2026b). An AI detector deployed to flag suspected AI-generated submissions is a natural fit for that description. The application date for these obligations has recently moved, under the Digital Omnibus agreement of 7 May 2026, from August 2026 to December 2027 (UNIwise, 2026b) – but as UNIwise has argued elsewhere, the substance of the obligations has not changed, only the date on which non-compliance starts to bite.

Article 13 of the Act requires that a high-risk AI system be sufficiently transparent to enable deployers to interpret its output and use it appropriately, with instructions covering the system's capabilities, its limitations and how its output should be interpreted (European Parliament and Council, 2024, Art. 13). Article 86 goes further, giving any individual subject to a decision based on such a system's output, where that decision significantly affects their rights, the right to obtain from the deployer – the institution, not the vendor – clear and meaningful explanations of the role the AI system played and the main elements of the decision taken (European Parliament and Council, 2024, Art. 86). Article 26 places the burden of human oversight, monitoring and logging on the deployer, and Article 27 requires a Fundamental Rights Impact Assessment before such a system is used at all.

This is precisely where AI detection collapses on its own terms. Bassett et al. (2026) establish that a detector's output is an unverifiable statistical estimate with no ground truth to check it against; there is, by definition, no reasoning trail connecting a percentage score to the words on the page in a way a human can independently audit. That is not a documentation gap a better vendor manual can close. It means an AI detector cannot generate the clear and meaningful explanation Article 86 requires for the student whose case depends on it, and it cannot give the marker the interpretable output Article 13 requires from the person who is supposed to exercise oversight over it. A tool that cannot explain itself to the person applying it, and cannot explain itself to the person it is applied to, sits uneasily within a regulatory framework built specifically around both of those requirements – in a sector, education, that the Act names explicitly.

## Where plagiarism actually comes from

UNIwise's own research into academic misconduct points to a further reason AI detection is the wrong tool for the job: it is aimed at the wrong target. Our white paper Ways in which Students Plagiarise, based on qualitative interviews across multiple higher education institutions, finds that the single largest source of plagiarised content is not obscure corners of the open internet, and certainly not generative AI, but other students at the same institution – current classmates, students on parallel groups or cohorts, and work drawn from institutional archives of earlier submissions. Peer-to-peer copying, whether through shared notes, assignments recycled between cohorts, or direct collusion, is consistently the

pattern our research finds institutions actually contend with, far more than external or AI-generated content.

That finding matters directly for how detection should be built. A tool tuned to guess whether a machine wrote a passage does nothing to catch the far more common case of one student's work matching another's from the same course, the same cohort, or a previous year's intake. Catching that requires a source pool that actually includes institutional archives and prior submissions, together with cross-cohort and cross-flow comparison – exactly the kind of verifiable, source-backed matching similarity detection is built to do, and that a probabilistic AI-generation score cannot do by design.

Our other research – Why Students Turn to Plagiarism and Understanding the Reasons Behind the Rise of Academic Misconduct – points to a related conclusion: misconduct is driven at least as much by academic pressure, unclear expectations around paraphrasing and referencing, and genuine misunderstanding of the grey zones around collaboration, as by deliberate dishonesty. A detection tool aimed at catching students after the fact, and aimed at the wrong source of copying in the first place, does nothing to resolve the confusion, workload pressure or unclear guidance our research identifies as root causes – and a false accusation arising from that tool actively damages the trust that better assessment design and clearer communication depend on.

## Where UNIwise stands: evidence over inference

UNIwise has deliberately chosen not to build generative AI detection into WISEflow Originality, and the reasoning set out in this paper is a large part of why. An unverifiable, probabilistic estimate of whether a machine produced a piece of text cannot meet the evidentiary standard that academic integrity processes are required to apply. Building product features on top of that estimate – additional AI hallmarks, comparison essays, multiple-tool consensus – would only add further unverifiable layers to an already unverifiable foundation.

Instead, WISEflow Originality is built around the kind of evidence a misconduct case can actually stand on: verifiable textual relationships between a student's submission and an identifiable source. It uses AI-driven semantic similarity to detect paraphrasing and meaning-level overlap across more than 50 languages, not to guess whether a machine wrote the words, but to show, with a traceable source and sentence-level highlighting, where a submission matches something that already exists. Matches are grouped into Exact, Strong and Potential categories precisely so that assessors are shown the strength of the evidence rather than a single opaque percentage, and every match can be traced back to its source for verification – the independent confirmation that AI detection, by its nature, cannot offer.

That distinction matters more than it might first appear. A similarity match against an identifiable source – a previous submission, a published article, a web page – is falsifiable: the assessor can open the source and compare it directly against the submission. A generative AI detection score is not falsifiable in the same way; there is no source document to compare it against, only a statistical inference about a process that has already finished

and left no trace. The first kind of evidence can, in the language Bassett et al. (2026) use throughout their analysis, meet a balance-of-probabilities standard. The second, by its own admission from the researchers who study it most closely, cannot.

This is also why WISEflow Originality is built with EU-centric, privacy-by-design data handling, granular control over what is scanned, indexed or shared, and full audit trails for every administrative decision. If a tool is going to sit inside a formal misconduct process, the institution needs to be able to show exactly what happened to the evidence, not simply trust a vendor's accuracy claims.

Because our own research shows peer-to-peer copying, not machine generation, to be the dominant real-world pattern, WISEflow Originality's source pools are built to include institutional archives and enable cross-flow and cross-institution matching, so a match can be traced to the actual earlier submission it resembles, rather than inferred from a statistical hunch about authorship.

## Fairness, proportionality, and whose responsibility this is

It is worth stepping back from the technical and legal detail to say plainly what is at stake for students. Being told that an algorithm has flagged your own, honestly written work as machine-generated is not a neutral administrative moment. Bassett et al. (2026) describe the practical effect on students under investigation: a power imbalance that leaves them feeling compelled to defend themselves against an accusation that, by the researchers' own analysis, can be neither conclusively proven nor conclusively disproven. The stress, the reputational risk, and – for international students in particular – the fear of consequences for their visa status (García Mathewson, 2023) fall on students regardless of whether they did anything wrong, simply because a probabilistic tool generated a number.

Cheating in some form has always existed in higher education, long before generative AI, and institutions have always had to exercise professional academic judgement to identify and respond to it. What blanket AI detection does is convert that proportionate, targeted task into an indiscriminate one: every student's submission is run through a tool that cannot reliably distinguish the honest majority from the small minority who may have used AI improperly, and every student is placed under a low-grade suspicion that the evidence in this paper shows the tool cannot actually justify. The effect is to make suspects of an entire student population in order to catch a few – a poor trade even before the false positive problem is taken into account.

That reasoning also settles the question of where responsibility for adapting to generative AI belongs. It is the institution's task to respond to the arrival of generative AI: through assessment design that is robust to it, through clearer guidance on where AI assistance is and is not appropriate, and through investment in the kind of academic judgement and conversation no detector can replace. It is not the student's task to prove a negative about a tool whose own creators cannot verify its accuracy under real conditions. Placing that burden on students, however informally, gets the responsibility for managing this transition backwards.

## What this means in practice for institutions

Institutions considering how to respond to generative AI in assessment can draw a straightforward line from the evidence above. An AI detection score should not be used at all – not even informally, as something to casually raise with a student. An indicator with no verifiable evidence behind it cannot be resolved one way or the other: raising it opens a discussion nobody can actually settle, and does nothing but place a student under suspicion with no way to clear their name. That is not a cautious use of the tool; it is a pointless one dressed up as due diligence. Reserve formal misconduct processes for evidence that can actually be checked against an identifiable source, such as text-matching and paraphrase detection against known documents.

Put the effort that would otherwise go into chasing detector scores into what actually works: talking to students directly and early, rather than confronting them with a percentage; guiding them on how to use AI well rather than policing them for using it at all; building AI use into assessment design from the outset instead of trying to detect it after the fact; and setting out, assignment by assignment, in plain and explicit terms, exactly where AI assistance is permitted and where it is not. And treat the underlying pressures our own research has identified – unclear expectations, workload and grey zones around collaboration – as at least as important a target for intervention as any detection technology.

## Conclusion

The central finding of Heads We Win, Tails You Lose is that AI detection is not a technology that needs refining; it is a category of tool built on a question – did a machine write this – that cannot be answered reliably once a text has left the hands of its author. Vendors will continue to publish strong accuracy figures, and those figures may even be true under the conditions in which they were measured. But a true laboratory result does not transfer to a real classroom, where the true origin of a text is, and will remain, unknown. UNIwise's position follows directly from that evidence: build assessment integrity tools around comparisons that can be verified, not estimates that cannot, and put the effort saved into the assessment design and student support that address why students turn to shortcuts in the first place.

It also means recognising who that effort belongs to. Responding to generative AI is properly the institution's responsibility to carry – through assessment design, clear guidance and sound academic judgement – not a burden to pass down to the students it exists to educate.

## References

Bassett, M. A., Bradshaw, W., Bornsztejn, H., Hogg, A., Murdoch, K., Pearce, B., & Webber, C. (2026). Heads we win, tails you lose: AI detectors in education. Journal of Higher Education Policy and Management. https://doi.org/10.1080/1360080X.2026.2622146

Copyleaks. (2026). AI Detector – Free AI checker for ChatGPT, GPT-5, Gemini & more. https://copyleaks.com/ai-content-detector

European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Articles 13, 26, 27 and 86, Annex III. Official Journal of the European Union. https://artificialintelligenceact.eu/

Fowler, G. A. (2023, June 2). Turnitin says its AI cheating detector isn't always reliable. The Washington Post. https://www.washingtonpost.com/technology/2023/06/02/turnitin-ai-cheating-detector-accuracy/

García Mathewson, T. (2023, August 14). AI detection tools falsely accuse international students of cheating. The Markup. https://themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating

GPTZero. (2026). GPTZero – AI detector. https://gptzero.me/

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Turnitin. (2023a, March 16). Understanding false positives within our AI writing detection capabilities. https://www.turnitin.com/blog/understanding-false-positives-within-our-ai-writing-detection-capabilities

Turnitin. (2023b, June 14). Understanding the false positive rate for sentences of our AI writing detection capability. https://www.turnitin.com/blog/understanding-the-false-positive-rate-for-sentences-of-our-ai-writing-detection-capability

UNIwise. (2026a). Assist, don't dictate: where AI belongs in assessment – and where it doesn't [Blog post]. https://uniwise.eu/resources/blog/assist-dont-dictate-where-ai-belongs-in-assessment-and-where-it-doesnt

UNIwise. (2026b). The EU AI Act and assessment: December 2027 is not a snooze button [Blog post]. https://uniwise.eu/resources/blog/the-eu-ai-act-and-assessment-december-2027-is-not-a-snooze-button

UNIwise. (n.d.). Understanding the reasons behind the rise of academic misconduct [White paper]. https://uniwise.eu/white-papers-and-reports/understanding-the-reasons-behind-the-rise-of-academic-misconduct

UNIwise. (n.d.). Ways in which students plagiarise [White paper]. https://uniwise.eu/white-papers-and-reports/ways-in-which-students-plagiarise

UNIwise. (n.d.). Why students turn to plagiarism [White paper]. https://uniwise.eu/white-papers-and-reports/why-students-turn-to-plagiarism

UNIwise. (n.d.). WISEflow Originality. https://uniwise.eu/solutions/originality

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19, 26. https://doi.org/10.1007/s40979-023-00146-z