AI Report

We predict this text is

Human Generated

AI Probability

40%

This number is the probability that the document is AI generated, not a percentage of AI text in the document.

Plagiarism

?

The plagiarism scan was not run for this document. Go to gptzero.me to check for plagiarism.

The Pitfalls of AI Detection in Academic Writing_ Bias, False Positives, and the Need for Inclusive Assessment.docx - 8/30/2026

The Pitfalls of AI-detection in academic writing: A critical review of bias, false-positives, and inclusive assessment

Victor Angelier

March 6th, 2026

This paper examines the limitations of AI detection tools in academic settings, highlighting bias against non-native English speakers, students with disabilities, and collaborative forms of writing.
This paper adopts a review-based approach, critically synthesising existing literature and institutional practices to develop a conceptual perspective on AI detection in academic assessment.
Empirical studies show that AI text detectors exhibit an accuracy–fairness trade-off: attempts to improve detection rates can increase false positives, particularly for authors who deviate from expected language norms (Liang et al., 2023; Elkhatat, Elsaid and Almeer, 2023).
Building on institutional decisions by universities such as Vanderbilt, Curtin, and several New Zealand institutions to disable or restrict AI detection in high-stakes assessment (Vanderbilt University, 2023; Curtin University, 2024; RNZ, 2025), we argue that detector-centric policies risk undermining fairness and trust.
Drawing on anecdotal user-reported tests—including a cyber dependency preprint, team-written projects, and AI-translated professional texts—we illustrate how detectors often penalise stylistic features and translation artefacts rather than substantive originality.
In dialogue with Edward Snowden's defence of privacy as a precondition for free expression, we contend that restricting AI assistance can disproportionately burden non-native and neurodivergent writers who benefit most from language-support tools (Snowden, 2015; Zhao, Cox and Cai, 2024).
We conclude by advocating for process-oriented assessment, oral defences, and AI literacy as more inclusive alternatives to unreliable AI detection, situating detector-centric policies within wider human-rights debates on reasonable accommodation and non-discrimination in education.

The integration of artificial intelligence (AI) into academic writing has sparked intense debate, particularly around tools designed to identify AI-generated content.
Commercial detectors such as Turnitin's AI-writing module, GPTZero, Copyleaks, and Winston AI promise to safeguard academic integrity but can generate non-trivial false-positive rates that disproportionately affect marginalised groups.
Empirical work shows that GPT-detectors can be biased against non-native English writers, misclassifying a substantial share of genuine essays as AI-generated while performing far better on similar texts from native speakers (Liang et al., 2023).
A widely circulated Stanford HAI commentary summarises these concerns for a broader audience, but the underlying empirical basis remains the peer-reviewed Patterns study (Liang et al., 2023; Zou et al., 2023).

Against this backdrop, this paper examines real-world cases—such as a cyber dependency preprint, collaboratively written assignments, and AI-translated professional texts—in which detectors flagged original work as "100% AI-generated" in user-reported tests, despite the authors' and testers' documented assertion of human authorship (Peter, 2026; author's anecdotal data).
We argue that the core issue lies not in a lack of originality but in factors such as translation, stylistic simplicity, and collaborative convergence of writing styles—features that current detectors can misinterpret as hallmarks of machine production.

Furthermore, we draw parallels to broader ethical concerns, building on Edward Snowden's analogy that dismissing privacy because one has "nothing to hide" is akin to dismissing free speech because one has "nothing to say" (Snowden, 2015).
In educational contexts, access to supportive tools can function as a practical precondition for expression: when language or disability-related barriers prevent effective communication, the formal right to speak or write does not guarantee an equal ability to be heard.
Restricting AI assistance may therefore limit expression for those with barriers such as dyslexia or language challenges, stifling innovation and reinforcing inequities (Zhao, Cox and Cai, 2024; Fung et al., 2025).
Ultimately, academia should prioritise comprehension and content over authorship mechanics, using process evidence and oral explanations as verification rather than relying on opaque and error-prone algorithms.

Empirical evaluations show that AI detection tools exhibit a fundamental tension between accuracy and fairness.
Large-scale tests of popular detectors reveal markedly better performance on synthetic benchmark datasets than on real academic texts, where both human and AI-assisted pieces are frequently misclassified (Elkhatat, Elsaid and Almeer, 2023; Giray, 2024).
Studies of detector bias indicate that increasing sensitivity to AI text can increase false positives, particularly for authors deviating from an "expected" language profile (Liang et al., 2023; Wu et al., 2025).
Liang et al.
(2023) show that GPT detectors labelled more than half of authentic TOEFL essays by non-native writers as AI-generated, while comparable essays by native speakers were recognised nearly flawlessly (Liang et al., 2023).
These outcomes suggest that detectors penalise stylistic and perplexity patterns rather than content originality, leaving non-standard or simplified language disproportionately at risk (Liang et al., 2023; Elkhatat, Elsaid and Almeer, 2023).
This directly relates to the user-reported cases in this paper, where non-native English influences and highly structured formal prose likely contributed to false flags despite novel conceptual contributions (Peter, 2026; author's anecdotal data).

These biases have direct implications for writers with learning and language-related support needs, including dyslexia and other forms of neurodivergence.
Research on generative AI in higher education shows that students with disabilities can experience significant benefits from language models in planning, structuring, and clarifying texts; AI tools may act as cognitive and linguistic scaffolding that reduces the cognitive load of surface-level editing (Zhao, Cox and Cai, 2024).
Studies on dyslexia also report that ChatGPT-based support can contribute to improved text structure, more appropriate word choice, and greater writing confidence in both first and second languages (Fung et al., 2025).
Systematic reviews of AI-based interventions for learning disabilities conclude that such technologies can reduce barriers in written communication when embedded within careful pedagogical and ethical frameworks (Paglialunga, 2025).
Against this backdrop, documented detector bias implies that the groups who may benefit most from AI support—non-native speakers and students with disabilities—can face a heightened risk of unwarranted suspicion of "AI misuse" (Liang et al., 2023; Elkhatat, Elsaid and Almeer, 2023; Zhao, Cox and Cai, 2024).

Recent higher-education focused evaluations reinforce these concerns by systematically probing detector performance under realistic usage conditions.
In a large battery of tests across multiple commercial GenAI detection tools, Perkins et al.
show that already moderate baseline accuracy rates collapse when confronted with paraphrased, adversarially edited or partially AI-assisted student texts, and that false positives and false negatives occur at levels incompatible with high-stakes integrity decisions (Perkins et al., 2024).
The authors therefore argue that current GenAI detectors are "flawed by design" for use as primary evidence in academic-misconduct procedures, given their structural sensitivity to surface form rather than genuine authorship or understanding (Perkins et al., 2024).

In response to these limitations and fairness concerns, some universities are shifting from detection-centred models to process- and comprehension-oriented strategies.
Vanderbilt University disabled Turnitin's AI-writing detector in 2023 after internal testing and consultation, citing reliability concerns, false-positive risk at scale, and limited transparency (Vanderbilt University, 2023).
Curtin University announced that Turnitin's AI writing detection feature would be disabled from 1 January 2026, while retaining traditional similarity checking for plagiarism (Curtin University, 2024).
In 2025, multiple New Zealand universities—including Massey, Auckland, and Victoria—reported they would no longer use AI detection software for student work, judging it insufficiently conclusive for misconduct allegations and relatively easy to circumvent; instead, they emphasise document histories, oral examinations, and assessment design alongside AI literacy (RNZ, 2025; Library Learning Space, 2025).
This institutional trend reflects a growing recognition that human judgement, process documentation, and oral defences can provide more inclusive and robust integrity safeguards than automated AI detection alone.

Information-Theoretic and Practical Limits of AI-Generated Text Detection

Detection as a Statistical Decision Problem

AI-generated text detection is fundamentally a statistical classification task rather than a definitive identification process.
Detectors attempt to distinguish between two hypotheses: that a text was produced by a human writer or generated by a language model.

This distinction relies on statistical patterns in word choice, syntax, and probability structure.
Theoretical work in information theory suggests that reliable discrimination becomes increasingly difficult as language models approach the statistical characteristics of human writing (Varshney, Keskar and Socher, 2020).
As the probability distributions of human and machine text overlap more closely, even an optimal passive detector operating solely on final text faces unavoidable error rates (Varshney, Keskar and Socher, 2020).

Recent work formalises this limitation via sample complexity: the quantity of text required to make reliable inferences.
Chakraborty et al.
(2024) argue that as language models improve and distributional differences shrink, the amount of text required for confident detection increases sharply.
Typical student assignments—often hundreds to a few thousand words—may not contain sufficient statistical evidence for reliable classification while maintaining low false-positive rates (Chakraborty et al., 2024).
In practice, detectors therefore operate within a difficult trade-off: raising sensitivity increases false accusations, while lowering sensitivity increases false negatives (Elkhatat, Elsaid and Almeer, 2023).

Practical deployment introduces additional failure modes due to distribution shift.
Detection systems are trained on particular datasets that rarely reflect the full diversity of real academic writing, including multilingual authorship, translation tools, collaborative editing, and discipline-specific conventions.
When detectors encounter writing conditions not represented in training data, performance can deteriorate (Wu et al., 2025).
Moreover, meaning-preserving transformations—such as paraphrasing, editing, and translation—can disrupt the statistical signatures that detectors use, reducing reliability without changing substantive authorship or originality (Wu et al., 2025; Chakraborty et al., 2024).

One proposed solution is watermarking, where language models embed detectable signals during generation.
Under controlled conditions, certain watermarking techniques can identify machine-generated text while maintaining output quality (Dathathri et al., 2024).
However, watermarking does not solve detection in open environments: edited, paraphrased, or translated text can reduce watermark robustness, and outputs from models that do not implement watermarking remain outside its scope (Zhang et al., 2024).
Watermarking therefore shifts the problem from universal detection to platform-specific verification.

Taken together, theoretical and empirical research suggests that passive AI-text detection faces both mathematical and practical limits.
As language models become more capable, statistical differences diminish, raising the baseline difficulty of detection (Varshney, Keskar and Socher, 2020; Chakraborty et al., 2024).
Meanwhile, multilingual writing, collaboration, and translation introduce variability that detectors struggle to interpret (Wu et al., 2025).
These findings imply that AI detector scores should be treated as imperfect indicators rather than dispositive evidence of misconduct.
In line with this, systematic stress-tests of widely used GenAI detection tools in higher education conclude that such systems cannot responsibly be deployed as stand-alone arbiters of misconduct, as their error patterns are tightly coupled to paraphrasing and editing practices common in legitimate student work (Perkins et al., 2024).

Policies that rely heavily on automated detection risk false positives that undermine trust and fairness, supporting a shift toward process-based assessment and oral explanations as adopted by institutions such as Vanderbilt, Curtin, and several New Zealand universities (Vanderbilt University, 2023; Curtin University, 2024; RNZ, 2025).

A central concern in institutional debates is the possibility of unsupervised, fully AI-generated papers that bypass genuine learning and distort the scholarly record.
While these risks are real, the theoretical limits of detection imply that purely detector-based responses cannot offer robust protection: as models become more human-like, it becomes harder—even in principle—to distinguish output from authentic text, and routine editing can further erode detection signals (Varshney, Keskar and Socher, 2020; Zhang et al., 2024).

These limits suggest that detector-centric enforcement is conceptually misaligned with the underlying problem. Instead of inferring authorship from final-text statistics alone, process-based mitigations focus on how work is produced.
Strategies include requiring iterative drafts with time-stamped version histories, structured writing logs and research notebooks, and targeted oral defences that probe understanding beyond the submitted text (Vanderbilt University, 2023; RNZ, 2025).
Assessment design can also hinge on locally verifiable data, personal reflection, or class-specific activities, reducing the effectiveness of one-shot external generation.
These approaches treat authorship as a verifiable process, addressing unsupervised automation risks while preserving legitimate assistive uses of AI for planning, language support, and translation.

While much of this paper focuses on AI detector limitations, the broader evidence base also highlights multi-dimensional benefits of AI assistance.
Mentally, generative AI can reduce cognitive overload for students who struggle with planning, structuring, and revision, particularly those with dyslexia or other learning differences; studies report improved confidence, reduced anxiety, and greater perceived control when AI is used as a scaffold rather than a substitute (Fung et al., 2025; Zhao, Cox and Cai, 2024).
Physically, AI-mediated tools such as speech-to-text, intelligent proofreading, and multimodal learning environments can complement assistive technologies for students with motor impairments or chronic conditions, lowering the physical strain associated with extended writing (Almgren Bäck et al., 2024; Paglialunga, 2025).
Empirical work on speech-to-text assistive technology, for example, reports improvements in written output for students with traumatic brain injury, illustrating how bypassing transcription barriers can unlock higher-level composition (Noakes, 2019).

Geopolitically, AI-assisted translation and academic-English support can reduce barriers for researchers from the Global South and non-English-dominant regions, enabling participation in international scholarship without full dependence on costly editing services.
Studies on AI-supported scientific writing report improvements in clarity, coherence, and conformity to journal expectations among non-native English-speaking researchers (da Costa, 2023; Almutairi and Reimer, 2024).
Parallel analyses describe how academics in under-resourced contexts use tools like ChatGPT to compensate for limited local supervision and editing support, enabling more equitable access to international publication and collaboration (Nobre et al., 2025; Saito et al., 2024).
These dynamics intersect with commitments to non-discrimination and equal access in higher education: restrictive or detector-heavy policies may reproduce language and disability hierarchies, whereas

process-oriented assessment and AI literacy align more closely with inclusive education norms (Liang et al., 2023; Wu et al., 2025; RNZ, 2025).

A 2026 preprint titled Cyber Dependency and Weaponization: A Framework for Assessing Critical Infrastructure Risks and State Strategies by Dr. Ada Peter introduces the Critical Infrastructure Dependencies (CID) Framework, analysing operating-system dependency patterns and their geopolitical implications, with examples including Stuxnet's exploitation of Iran's reliance on foreign industrial control systems (Peter, 2026).
In user-reported tests using tools including GPTZero, Copyleaks, iThenticate, and Winston AI, this preprint was repeatedly flagged as "100% AI-generated", despite claims of human authorship and the presence of distinct theoretical contributions (author's anecdotal data).

The paper's original elements—such as a typology categorising nations into low, moderate, high, and extreme cyber dependency—suggest detector outputs were driven less by content duplication than by surface-level features.
Likely triggers include highly regular academic structure and non-native English phrasing consistent with Global South scholarly publication pressures (Peter, 2026; author's anecdotal data).
Prior research indicates that such writing profiles can be at elevated risk of misclassification by GPT-based detectors (Liang et al., 2023).
This case therefore exemplifies how detectors can ignore conceptual originality and instead treat stylistic regularity as a proxy for AI authorship.

In another user-reported case, a collaboratively written team project—produced by four individuals without AI assistance—was flagged as "100% AI-generated" across multiple detectors (author's anecdotal data).
Academic writing often relies on shared conventions and formulaic signalling phrases, which naturally produce stylistic uniformity.
Empirical evaluations show detectors can struggle to separate conventional academic phrasing from AI-produced text, yielding misclassifications even in peer-reviewed material (Elkhatat, Elsaid and Almeer, 2023).
This case highlights how detectors can penalise collaboration, where harmonised tone is a legitimate outcome of coordination and editing.

A third set of cases concerns professional texts authored in one language and translated or refined into English using AI for cross-border correspondence and formal communication (author's anecdotal data).
Detectors classified these polished English versions as AI-generated despite the underlying ideas being wholly human-produced.
This pattern is consistent with concerns that detectors treat linguistic smoothness and error reduction as "machine-like" signals, penalising those who use tools to overcome language barriers (Giray, 2024).

Taken together, these cases illustrate how detectors can fail to distinguish originality from stylistic artefacts, disproportionately affecting non-native writers and learners who rely on assistive technologies.
While anecdotal, they align with formal evaluation studies and demonstrate the risks of treating detector outputs as decisive evidence of misconduct (Liang et al., 2023; Elkhatat, Elsaid and Almeer, 2023).

AI assistance can democratise written expression for individuals with dyslexia, language barriers, and other learning difficulties, in much the same way calculators expanded access to advanced mathematics.

Blocking or severely restricting such tools risks exclusion; metaphorically, it resembles demanding that a patient with chronic obstructive pulmonary disease complete a marathon unaided while others are allowed bicycles and vehicles (Zhao, Cox and Cai, 2024; Fung et al., 2025).
To avoid misinterpretation, the relevant analogy is *scaffolding rather than substitution*: legitimate assistive use supports planning, clarity, and expression, whereas outsourcing intellectual labour—delegating the core reasoning and argument—undermines learning and should remain prohibited.

Building on Snowden's claim that privacy enables free expression, access to supportive tools may similarly enable academic expression where language or disability barriers would otherwise silence or exclude (Snowden, 2015; Snowden, 2019).
Beyond accessibility, generative AI can also widen scholarly participation: non-native English researchers can refine manuscripts for international journals without full dependence on expensive editing, and students in under-resourced contexts can access feedback that would otherwise require intensive supervision (da Costa, 2023; Nobre et al., 2025).
Restrictive detector-based policies risk suppressing emergent voices not because their ideas lack merit, but because their reliance on assistive technology triggers algorithmic suspicion.

Process- and understanding-oriented assessment offers a proportionate alternative.
If a student can explain, defend, and adapt arguments under questioning, wholesale unsupervised AI generation becomes implausible regardless of detector scores.
Institutions such as Vanderbilt, Curtin, and multiple New Zealand universities have moved toward emphasising human judgement, document histories, and oral assessment as integrity safeguards (Vanderbilt University, 2023; Curtin University, 2024; RNZ, 2025).
This approach aligns with emerging recommendations that AI literacy, transparent workflows, and inclusive assessment design are more effective and just responses than reliance on unreliable binary detection outputs (Wu et al., 2025).

From a legal-normative perspective, inclusion arguments are reinforced by human-rights standards on education and disability.
Under the UN Convention on the Rights of Persons with Disabilities (CRPD), States must ensure "reasonable accommodation" in education, and UN bodies emphasise that failure to provide reasonable accommodation can amount to discrimination (UN CRPD, 2006; OHCHR, 2022).
The European Court of Human Rights has also addressed disability discrimination in education contexts, including *Çam v. Turkey*, and relevant legal anchors include Article 14 ECHR read with Article 2 of Protocol No. 1 (CGLJ, 2020; Doughty Street Chambers, 2018).

However, it is important to state the limits clearly: no appellate court has yet ruled that banning generative AI tools for disabled students violates the ECHR.
The argument advanced here is therefore best framed as an emerging and persuasive normative position: blanket bans or detector-centric policies *may* conflict with non-discrimination principles where AI tools function as effective accommodations and no equally effective alternative support is provided (Fung et al., 2025; Zhao, Cox and Cai, 2024; ENNHRI, 2023).

Many universities also publicly endorse equality, diversity, and inclusion, yet AI assessment governance has not always addressed the indirect discriminatory effects of biased detection on disabled and non-native students.

In this light, moving from detector-centric enforcement to process-based, accessible assessment can be viewed not only as a pedagogical improvement, but as a means of aligning integrity policies with established equality and inclusive education obligations (UN CRPD, 2006; OHCHR, 2022; ENNHRI, 2023).

AI detectors, affected by documented bias and non-trivial false-positive rates, risk penalising those who deviate from narrow stylistic norms—especially non-native writers and students with disabilities (Liang et al., 2023; Zhao, Cox and Cai, 2024).
The cyber dependency preprint, alongside collaborative projects and translated professional texts, illustrates how translation and style can overshadow originality when algorithms are treated as arbiters of authorship.
Theoretical results further indicate that as language models become more human-like, reliable detection becomes increasingly constrained, making detector-centric enforcement a brittle strategy (Varshney, Keskar and Socher, 2020; Chakraborty et al., 2024; Zhang et al., 2024).
Empirical stress-testing in university settings reaches a similar conclusion: across days of benchmarking multiple commercial platforms on authentic and adversarially modified assignments, Perkins et al.
find detector outputs too unstable and manipulable to support high-stakes allegations (Perkins et al., 2024).
Embracing AI assistance within transparent, process-oriented assessment frameworks—and verifying authorship through oral defences and iterative drafts—offers a more inclusive and pedagogically sound alternative.
This approach echoes Snowden's argument about protecting the conditions of free expression, reframed here as equitable access to assistive technologies in academic writing, consistent with human-rights commitments to inclusive education and reasonable accommodation (Snowden, 2015; UN CRPD, 2006; OHCHR, 2022).
Recent institutional trends—from Vanderbilt to Curtin and New Zealand universities—suggest an ongoing shift toward limiting high-stakes AI detection while strengthening assessment design, process evidence, and AI literacy (Vanderbilt University, 2023; Curtin University, 2024; RNZ, 2025).
Future research should systematically document these policy experiments and develop guidelines ensuring AI enhances rather than restricts scholarly discourse, while keeping integrity policies aligned with non-discrimination and equality-of-access obligations.

Perkins M Roe J Vu B H Postma D Hickerson D McGaughran J and Khuat H Q.
(2024) 'GenAI detection tools, adversarial techniques and implications for inclusivity in higher education', arXiv preprint arXiv:2403.19148.
Available at: https://arxiv.org/abs/2403.19148 (Accessed: 5 March 2026).AI, Algorithmic, and Automation Incidents Repository (2024) 'Universities disable Turnitin AI detection in students' best interests'.
Available at https www aiaaic org aiaaic repository ai algorithmic and automation incidents universities disable turnitin ai detection in students best interests Accessed 5 March 2026.

Almgren Bäck, G. et al.
(2024) 'Dyslexic students' experiences in using assistive technology to support written language skills: a five-year follow-up', Disability and Rehabilitation: Assistive Technology, 19(3), pp.
123 145.
Available at https www tandfonline com doi full 10 1080 17483107 2022 2161647 Accessed 5 March 2026.

Almutairi, S. and Reimer, T. (2024) 'GenAI as a toolkit for English academic writing among international students', Journal of English for Academic Purposes (preprint).
Available at https files eric ed gov fulltext EJ1482961 pdf Accessed 5 March 2026.

Chakraborty, S. et al.
(2024) ‘On the possibilities of AI-generated text detection’, in Proceedings of the 41st International Conference on Machine Learning (ICML).
Available at https proceedings mlr press v235 chakraborty24a html Accessed 5 March 2026.

CGLJ (2020) ‘ECtHR: Lack of educational support for autistic student violated rights’.
Cambridge Globalist Law Journal.
Available at https cglj org 2020 09 17 ecthr lack of educational support for autistic student violated rights Accessed 5 March 2026.

Curtin University (2024) ‘Update on Turnitin AI detection tool’.
Available at https www curtin edu au news oasis news update on turnitin ai detection tool Accessed 5 March 2026.

da Costa, C. (2023) ‘The use of artificial intelligence to improve the scientific writing of non-native English speakers’, Revista da Associação Médica Brasileira, 69, e20230560.
Available at https pmc ncbi nlm nih gov articles PMC10508892 Accessed 5 March 2026.

Dathathri, S. et al.
(2024) ‘Scalable watermarking for identifying large language model outputs’, Nature.
Available at https www nature com articles s41586 024 08025 4 Accessed 5 March 2026.

Doughty Street Chambers (2018) ‘Strasbourg case: disabled student excluded from university education’.
Available at https insights doughtystreet co uk post 102epfy strasbourg case disabled student excluded from university education Accessed 5 March 2026.

EdTech Innovation Hub (2025) ‘Curtin University to disable Turnitin AI detection tool in 2026 amid reliability concerns’.
Available at https www edtechinnovationhub com news curtin university to disable turnitin ai detection tool in 2026 as debate over reliability continues Accessed 5 March 2026.

Elkhatat, A.M., Elsaid, K. and Almeer, S. (2023) ‘Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text’, International Journal for Educational Integrity, 19(1), pp. 1 20.
Available at https link springer com article 10 1007 s40979 023 00140 5 Accessed 5 March 2026.

ENNHRI (2023) ‘Artificial intelligence and its impact on the rights of persons with disabilities’.
ENNHRI–CoLab Policy Paper.
Available at https ennhri org wp content uploads 2023 12 PP AI and disability rights ENNHRI CoLab Jerneja Turin pdf Accessed 5 March 2026.

et al.
(2025) ‘A study on using ChatGPT to help students with dyslexia learn Chinese and English writing’, SAGE Open, 15(2), 21582440251358269.
Available at https journals sagepub com doi 10 1177 20965311251358269 Accessed 5 March 2026.

Giray, L. (2024) ‘The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism’, The Serials Librarian, 85(3), pp.
181 189.
Available at https www tandfonline com doi abs 10 1080 0361526X 2024 2433256 Accessed 5 March 2026.

Library Learning Space (2025) ‘New Zealand universities give up using software to detect AI in students’ work’.
Available at https librarylearningspace com new zealand universities give up using software to detect ai in students work Accessed 5 March 2026.

Liang, W. et al.
(2023) ‘GPT detectors are biased against non-native English writers’, Patterns, 4(7), 100779.
Available at https pmc ncbi nlm nih gov articles PMC10382961 Accessed 5 March 2026.

Noakes M A.
(2019) ‘Speech-to-text assistive technology for the written expression of students with traumatic brain injury’, Psychology in the Schools, 56(10), pp.
1721 1738.
Available at https pubmed ncbi nlm nih gov 31697151 Accessed 5 March 2026.

Nobre, F. et al.
(2025) ‘How ChatGPT empowers academics in the Global South’, Scientometrics (ahead of print).
Available at https pubmed ncbi nlm nih gov 40836854 Accessed 5 March 2026.

OHCHR (2022) ‘Humanity should get the best from AI, not the worst – UN disability rights expert’.
Available at https www ohchr org en stories 2022 05 humanity should get best ai not worst un disability rights expert Accessed 5 March 2026.

Paglialunga, A.
(2025) ‘The effectiveness of artificial-intelligence-based interventions for students with learning disabilities: A systematic review’, Frontiers in Education, 10, 12385150.
Available at https pmc ncbi nlm nih gov articles PMC12385150 Accessed 5 March 2026.

Peter, A.
(2026) ‘Cyber dependency and weaponization: A framework for assessing critical infrastructure risks and state strategies’, SSRN preprint.
Available at: https://www.ssrn.com/abstract=5076760 (Accessed: 5 March 2026).

RNZ (2025) ‘Universities give up using software to detect AI in students’ work’.
Available at https www rnz co nz news national 574517 universities give up using software to detect ai in students work Accessed 5 March 2026.

Snowden, E. (2015) ‘Edward Snowden: a right to privacy is the same as freedom of speech – video interview’, The Guardian, 22 May.

Available at https www theguardian com us news video 2015 may 22 edward snowden rights to privacy video Accessed 5 March 2026.

Snowden, E. (2019) 'Edward Snowden on surveillance and free speech', Columbia News.
Available at https news columbia edu news snowden mass surveillance knight institute jameel jaffer amy davidson sorkin Accessed 5 March 2026.

UN CRPD (2006) Convention on the Rights of Persons with Disabilities.
Available at https www ohchr org en instruments mechanisms instruments convention rights persons disabilities Accessed 5 March 2026.

and Socher, R. (2020) 'Limits of detecting text generated by large-scale language models', in Information Theory and Applications Workshop (ITA).
Available at: https://arxiv.org/abs/2002.03438 (Accessed: 5 March 2026).

Vanderbilt University (2023) 'Guidance on AI detection and why we're disabling Turnitin's AI detector'.
Available at https www vanderbilt edu brightspace 2023 08 16 guidance on ai detection and why were disabling turnitins ai detector Accessed 5 March 2026.

Wu, J. et al.
(2025) 'A survey on LLM-generated text detection: necessity, methods, and future directions', Computational Linguistics, 51(1), pp.
275 327.
Available at https direct mit edu coli article 51 1 275 127462 A Survey on LLM Generated Text Detection Necessity Accessed 5 March 2026.

Zhang, H. et al.
(2024) 'Impossibility of strong watermarking for language models', in Proceedings of the 41st International Conference on Machine Learning (ICML).
Available at https proceedings mlr press v235 zhang24o html Accessed 5 March 2026.

Zhao Y Cox A and Cai J.
(2024) 'Exploring the affordances of generative AI in academic writing for students with disabilities', Studies in Higher Education (ahead of print).
Available at https eprints whiterose ac uk id eprint 224322 3 use 20of 20generative 20ai 20by 20disabled 20students 20accepted 20version pdf Accessed 5 March 2026.

Zou, J. et al.
(2023) 'AI-detectors biased against non-native English writers', Stanford HAI.
Available at https hai stanford edu news ai detectors biased against non native english writers Accessed 5 March 2026.

(2025) 'Bridging the language gap: enhancing academic performance of non-native students with AI-powered translation', Studies in Computational Technologies, 5(1).
Available at https sct ageditor ar index php sct article view 1426 Accessed 5 March 2026.

Saito, T. et al.
(2024) ‘Advancing Global South university education with large language models’, International Journal of Educational Technology in Higher Education (preprint).
Available at: https://arxiv.org/abs/2410.07139 (Accessed: 5 March 2026).

FAQs

What is GPTZero?

GPTZero is the leading AI detector for checking whether a document was written by a large language model such as ChatGPT. GPTZero detects AI on sentence, paragraph, and document level. Our model was trained on a large, diverse corpus of human-written and AI-generated text with support for English, Spanish, French, German, and other languages. To date, GPTZero has served over 17 million users around the world, and works with over 100 organizations in education, hiring, publishing, legal, and more.

When should I use GPTZero?

Our users have seen the use of AI-generated text proliferate into education, certification, hiring and recruitment, social writing platforms, disinformation, and beyond. We've created GPTZero as a tool to highlight the possible use of AI in writing text. In particular, we focus on classifying AI use in prose. Overall, our classifier is intended to be used to flag situations in which a conversation can be started (for example, between educators and students) to drive further inquiry and spread awareness of the risks of using AI in written work.

Does GPTZero only detect ChatGPT outputs?

No, GPTZero works robustly across a range of AI language models, including but not limited to ChatGPT, GPT-5, GPT-4, GPT-3, Gemini, Claude, and AI services based on those models.

What are the limitations of the classifier?

The nature of AI-generated content is changing constantly. As such, these results should not be used to punish students. We recommend educators to use our behind-the-scene Writing Reports as part of a holistic assessment of student work. There always exist edge cases with both instances where AI is classified as human, and human is classified as AI. Instead, we recommend educators take approaches that give students the opportunity to demonstrate their understanding in a controlled environment and craft assignments that cannot be solved with AI. Our classifier is not trained to identify AI-generated text after it has been heavily modified after generation (although we estimate this is a minority of the uses for AI-generation at the moment). Currently, our classifier can sometimes flag other machine-generated or highly procedural text as AI-generated, and as such, should be used on more descriptive portions of text.

I'm an educator who has found AI-generated text by my students. What do I do?

Firstly, at GPTZero, we don't believe that any AI detector is perfect. There always exist edge cases with both instances where AI is classified as human, and human is classified as AI. Nonetheless, we recommend that educators can do the following when they get a positive detection: Ask students to demonstrate their understanding in a controlled environment, whether that is through an in-person assessment, or through an editor that can track their edit history (for instance, using our Writing Reports through Google Docs). Check out our list of several recommendations on types of assignments that are difficult to solve with AI.

Ask the student if they can produce artifacts of their writing process, whether it is drafts, revision histories, or brainstorming notes. For example, if the editor they used to write the text has an edit history (such as Google Docs), and it was typed out with several edits over a reasonable period of time, it is likely the student work is authentic. You can use GPTZero's Writing Reports to replay the student's writing process, and view signals that indicate the authenticity of the work.

See if there is a history of AI-generated text in the student's work. We recommend looking for a long-term pattern of AI use, as opposed to a single instance, in order to determine whether the student is using AI.