Written by a human, flagged by a machine

False positives in GenAI detection and bias against multilingual writers

www.kelvinlaw.me

Kelvin K.F. Law Associate Professor of Accounting, NTU
NTU Annual Learning and Teaching Conference • 13 May 2026

The argument in three parts

1. Detection is inference

AI detectors infer authorship from statistical regularities. They do **not** verify provenance.

2. False positives fall unevenly

Multilingual writers can face disproportionate risk. The evidence on this is documented and replicable.

3. Assessment design is the answer

Make learning visible through process evidence, oral defence, and AI-literate tasks.

The question is not “can we catch AI?”

It is “what evidence of learning should count?”

The practical problem

You open a report and see:

“AI writing detected: 47%”

What does this flag mean? Can you trust the score? What should you do next?

The vendor’s own guidance

Turnitin says an AI report may misidentify human-written, AI-generated, and AI-paraphrased text. It should not be the sole basis for adverse action.

Version-sensitive evidence

Detector outputs are **moving-target estimates**, not stable proof of provenance. Model updates can change future results. Old reports may not be retroactively refreshed.

Widely reported 2025 cases in Singapore show why the workflow matters.

How detection tools work

Perplexity

How predictable word choices are to a language model. Lower perplexity = more likely to be flagged as AI.

Burstiness

Variation in sentence length and complexity. Less variation = more likely to be flagged.

What this means in practice

- Detectors infer from text features alone. They do not verify who wrote it.
- Threshold choice changes the false positive rate.
- The same text may produce different scores across tools and across model versions.

Detection is inference about likely authorship, not proof of authorship.

What independent benchmarks say

RAID, ACL 2024

6M+ generations, 11 models, 8 domains. Detectors struggled when required to operate at very low false-positive rates. ZeroGPT plateaued at **16.9%** false positive rate.

Perkins et al., 2024

A detector reached 39.5% accuracy on unmanipulated AI samples and 67% on human-written controls. Aggregate false-accusation ratio: **15%**.

Weber-Wulff et al., 2023

14 tools tested. **None** exceeded 80% accuracy. Only five exceeded 70%. Conclusion: “neither accurate nor reliable.”

Hadra et al., 2026

Turnitin accuracy: **0.61**. Originality accuracy: 0.69. Both had near-zero recall on hybrid texts. Neither sufficient for high-stakes decisions.

Detector performance degrades exactly where academic decisions become high stakes.

The bias problem

Liang et al. (2023), *Patterns*. 7 GPT detectors, 91 TOEFL essays + 88 US student essays.

TOEFL essays (non-native English)

61.3%

false positive rate

97.8% flagged by at least one detector.

US student essays (native English)

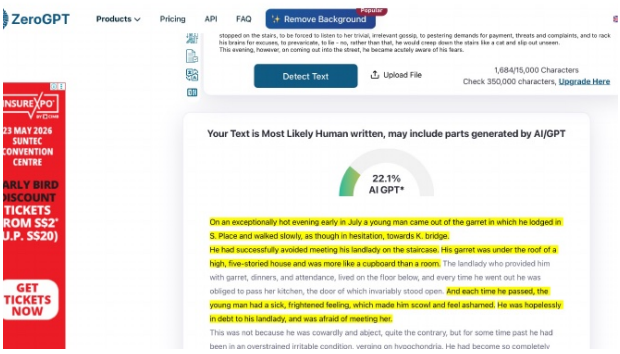
≈5%

false positive rate

Near-perfect classification.

When vocabulary was enhanced to “sound native,” the false positive rate dropped from 61.3% to 11.6%. One documented explanation, not a settled mechanism.

Or: Dostoevsky, 1866



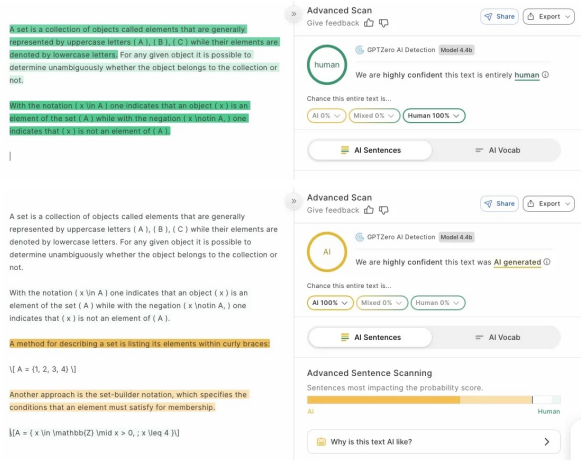
The opening of *Crime and Punishment*, pasted into ZeroGPT.

Result: **22.1%** flagged as AI-generated.

The novel was published in 1866.

If a detector flags Dostoevsky, ask what feature it is really detecting.

Or: a math textbook, 2026



Same text, opposite verdicts.

A passage on set notation from an open math textbook. GPTZero (PRO, Model 4.4b) rates it **100% human**.

Add one sentence defining what a set is. The same tool rates the expanded passage **100% AI-generated**.

One extra sentence flips the verdict from full confidence in human authorship to full confidence in machine authorship.

Source: Antonio Lupetti (@antoniolupetti), 25 April 2026.

Consequences and cases

Adelphi: confirmed judicial reversal

A court annulled an AI-plagiarism finding and ordered the student's record cleared.

Australian Catholic University: mass referrals

Nearly 6,000 referrals in 2024, about 90% AI-related, were later described as substantially overstated. The indicator was abandoned in 2025.

UK Office of the Independent Adjudicator (July 2025)

The OIA upheld an international student's complaint after an AI-related process involving Turnitin and Grammarly. The stronger point is **procedural fairness**: fair opportunity to respond, evidence, and proportionality.

Singapore 2025

One student was reportedly cleared using process evidence. Another case involved a zero for false citations. The distinction matters: **authorship** and **fabrication** are different allegations.

Institutional movement and local alignment

Universities stepping back

- **Vanderbilt**: disabled Turnitin's AI detector for the foreseeable future.
- **University of Waterloo**: discontinued Turnitin AI detection from September 2025.

NTU's own guidance

NTU guidance already warns that AI detection tools can produce false positives and false negatives, can be bypassed, and may be biased against non-native writing patterns.

Regulators and sector bodies

TEQSA (Australia): detection is "all but impossible." JISC (UK): "never base decisions solely" on it. QAA (UK): tools are "unreliable at best."

What this talk is

Not a rejection of academic integrity. An attempt to align integrity with defensible evidence and better assessment design, consistent with NTU's own published position.

The policy paradox: “Grammarly OK, GenAI not OK”

A student’s account, OIA case

“I used Grammarly. I believed it was permitted because English was not my first language.” The student faced an AI misconduct process.

Better boundary: regulate by function, not brand

- **Mechanical**: spelling, punctuation, grammar.
- **Formative**: comments incorporated manually.
- **Substantive**: paraphrase, humanizer, rewriting.

The technical reality

Grammarly’s rewriting agents are LLM-based. Grammarly says detector scores are not an objective source of truth and that its rewrites are likely to be flagged as AI.

Rule of thumb: if the tool changes the intellectual expression, it needs disclosure and assessment-specific permission.

What employers actually want

“If we were to hire a new college graduate today,
and I have a choice between two,
one that has no clue what AI is,
and one that is expert in using AI,
I would hire the one who’s expert in using AI.”

— Jensen Huang, CEO of Nvidia, Lex Fridman Podcast #494, March 2026

We try to detect AI use. Employers already assume it.

The question is shifting from “did you use it?” to “can you use it well?”

What to do instead: make authorship visible

Sampled oral defences

Five minutes, recorded, on a rotating subset. Ask students to explain claims, sources, anomalies, and decisions. University guidance increasingly lists viva voces as an assured-assessment option. UCL caught **90%** of AI-assisted submissions this way.

Process-based assessment

Require staged artefacts: proposal, outline, annotated bibliography, drafts, revision notes, and a short reflection on tool use.

What this buys you

Less dependence on opaque classifiers. More evidence of learning over time. A fairer route for multilingual writers. Better detection of fabricated citations.

What to do instead: redesign the assessment

Secure-open split

University of Sydney's two-lane model:

- **Secure**: supervised, used for assurance.
- **Open**: AI permitted with acknowledgement, critique, and disciplinary expectations.

For accounting

Ask students to:

- trace claims to filings, standards, or data,
- identify unsupported AI-produced assertions,
- defend assumptions under questioning,
- explain when tool assistance improved or weakened judgement.

AI-integrated assessment

Students generate AI output, then critique errors, source gaps, assumptions, and hallucinations. Tests professional scepticism. Reduces incentive to hide AI use.

The goal is not to ban fluent prose. It is to reward professional scepticism.

Six recommendations

1. **Do not treat detection scores as evidence** of misconduct.
2. **Apply heightened caution** in multilingual cohorts.
3. **Prioritise process evidence** over algorithmic inference.
4. **Combine secure and open assessments.**
5. **Make policy boundaries teachable** and explicit.
6. **Reward scepticism and verification,** not fluent prose.

Fluent language is now cheap. Sound judgement is not.

Thriving together in an age of AI is not achieved by algorithmic surveillance.

It is achieved by designing assessments that make learning visible.

What is higher about higher education?

Rewarding judgement, evidence, and visible learning.

Kelvin K.F. Law
Associate Professor, Nanyang Business School, NTU
klaw@ntu.edu.sg

www.kelvinlaw.me

Open for Q&A