Accused by Perplexity: AI Detection as an Unvalidated Automated Decision System in Graduate Education

Heather Grizzle · May 22, 2025

AI detectors measure how unusual writing is, not who wrote it. That falls hardest on second-language writers and students using approved assistive technology.

Contents

The advice everyone is giving is wrong about the mechanism

The standard counsel to graduate students on artificial intelligence runs roughly as follows: use it as a tool rather than a substitute, do your own thinking, and understand that faculty are perceptive and will eventually notice work that lacks depth. The first two propositions are sound. The third is false, and its falseness matters, because it describes a process of human judgment that has largely been replaced.

Most institutions are not detecting artificial intelligence use through the discernment of experienced readers. They are detecting it, or believing they are, through automated classifiers embedded in submission platforms, which assign a probability score to a document and pass it into an academic integrity process. The question facing a graduate student is therefore not whether a professor will notice shallow thinking. It is whether a statistical classifier will flag her writing, and what happens to her if it does.

Novara Consulting Group's position is that AI detection in higher education is an automated decision system producing consequential decisions about individuals, that it is deployed with less validation than institutions require of almost any other instrument affecting student standing, and that its error distribution falls hardest on multilingual and disabled writers. The governance problem is institutional. The consequences are individual.

What the detector measures

An AI writing detector does not identify authorship. It estimates how statistically unusual a text is, principally through perplexity, which measures how predictable each word is given the preceding text, and burstiness, which measures variation in sentence structure and length across a document. Text that is highly predictable and structurally uniform is scored as more likely machine-generated.

That is the entire mechanism, and stating it plainly makes the problem visible. The detector is not measuring whether a person wrote something. It is measuring whether the writing is unusual, and treating unusualness as evidence of humanity. Writing that is careful, plain, structurally consistent, and lexically conservative is scored as machine-like, because that is what the model's proxy for machine-likeness is.

This produces a specific and predictable population of false accusations, and it is not a random one.

The error rates, and who carries them

Research published in *Patterns* in July 2023 by Liang and colleagues at Stanford found that detectors flagged approximately sixty-one percent of TOEFL essays written by non-native English speakers as AI-generated, against near-zero false positive rates for comparable writing by native English speakers, across the seven detectors tested. These were essays written by human beings who had taken a standardized examination to demonstrate English proficiency. The authors identified the mechanism directly: the detectors were keying on text perplexity, and writing in a second language produces lower perplexity.

Independent testing has repeatedly failed to support vendor accuracy claims. Weber-Wulff and colleagues, publishing in the *International Journal for Educational Integrity* in December 2023, evaluated a range of detection tools across multiple document types and concluded that they were neither accurate nor reliable. Vendors, for their part, generally decline to publish a false positive rate for their products, which means an institution adopting one cannot state the error rate of an instrument it is using in disciplinary proceedings.

Institutions have begun to act on this. Vanderbilt University disabled Turnitin's AI detection feature in August 2023, publishing the arithmetic that made the decision straightforward: a one percent false positive rate applied to seventy-five thousand submitted papers produces roughly seven hundred and fifty wrongful accusations a year. Other institutions have followed, and the direction of published university guidance since has been toward treating a detector score as a prompt for inquiry rather than as a finding.

One further market fact belongs in any institutional evaluation. Several providers marketing detection tools also market humanizer products designed to defeat detection, including their own. In any other assurance context, a vendor selling both the test and the means of passing it would be disqualified from supplying either.

The population this reaches, and why it is the wrong one

In any other assurance context, a vendor selling both the test and the means of passing it would be disqualified from supplying either.

The documented disparity concerns non-native English writers, and the explanation is mechanical rather than mysterious. Perplexity-based detection penalizes simpler vocabulary and more regular grammatical construction. A writer working in a second language produces exactly that profile, not because the writing is machine-generated but because constructing unusual phrasing in a second language is harder than constructing correct phrasing.

The same mechanism reaches a population that has received almost no attention, and it is the one this firm exists to consider.

A Deaf writer whose first language is American Sign Language is writing English as a second language, and the syntactic conservatism that follows is precisely what the detector scores as machine-like. A student with dyslexia using text prediction, grammar assistance, or structured writing support, all of which are ordinary and often formally approved accommodations, produces text that has been regularized by the assistive tool. A student with a motor impairment composing by speech-to-text produces text shaped by dictation software's own language model. A student using translation assistance produces output carrying the statistical signature of the translation system.

In each case, the accommodation produces the flag. The student who uses the support her institution approved is more likely to be accused than the student who did not need it, and the accusation arrives with a number attached and no explanation of what the number measures. She is then required to prove she wrote her own work, which is a demand for a negative, made of the person least positioned to satisfy it.

An institution that has approved an assistive technology accommodation and simultaneously runs a detector penalizing its output has built a contradiction into its own disciplinary process. It is unlikely to have noticed, because the disability services office and the academic integrity office do not typically review each other's instruments. Postsecondary institutions carry obligations under the Americans with Disabilities Act and Section 504 of the Rehabilitation Act to provide academic adjustments and to avoid practices that screen out students with disabilities. A detection practice that systematically flags the products of approved accommodations sits uncomfortably against both.

The evidentiary standard is lower than for anything comparable

Consider what an institution would require before acting on any other automated instrument affecting a student's standing.

An admissions test would need published validity evidence, established reliability, documented performance across demographic groups, and a defensible cut score. A proctoring tool triggering an integrity referral would at minimum require that the recording be reviewable. A disability determination would require documentation and a named decision-maker. An institution making a consequential decision on an unexplained numeric score, produced by an instrument whose error rate the vendor will not disclose, would ordinarily be understood to have acted arbitrarily.

AI detection scores are treated differently, and the reason appears to be genre rather than principle. They arrive inside a plagiarism tool institutions already trust, they look like the similarity index they sit beside, and they were adopted through procurement rather than through academic governance. The similarity index at least points at a source document a student can inspect and contest. The AI score points at nothing. There is no source, no passage, no comparison, and no mechanism by which the student can examine the basis of

the allegation.

Better institutional practice treats a detector output as triage rather than evidence: a prompt to open a conversation in which the student explains her argument, walks through her drafting, and discusses her sources. That is the correct posture, and it is worth noting what it concedes. If the detector's output is only a reason to look more closely, and the looking is what actually determines the outcome, then the detector is a mechanism for deciding which students get investigated. Given the error distribution above, it is a mechanism for deciding that multilingual and disabled students get investigated more often.

What a graduate student should actually do

The useful advice is not to work harder at sounding human, which is both demeaning and unachievable. It is to hold the record, for the same reason any institution facing an unexplained automated determination should hold the record: the accusation is unfalsifiable without process evidence, and process evidence has to be created before it is needed.

Draft in an environment that retains version history, and do not disable it. A document with a two-month revision trail is dispositive in a way no argument about writing style will ever be. Keep research notes, annotated sources, and outlines, and keep them dated. Where an assistive or language support tool is used, disclose it in advance under the course's stated policy rather than after a flag, because prior disclosure converts a defense into a record. Where an accommodation involves writing support, ensure the disability services office has documented it, and understand that this documentation is now also integrity documentation. And be able to explain the argument aloud, not because a professor's perception is the safeguard, but because the structured conversation is where these matters are actually resolved.

None of this is about deception. It is what any person should do when a system that cannot be inspected may make a claim about them that they will be asked to rebut.

What an institution should establish

Four questions, which an academic integrity office should be able to answer before a detector output reaches a student.

1. What is the instrument's false positive rate, on what population, established by whom. If the vendor will not state it, that is the finding, and it should be recorded as one rather than treated as an absence of information.
2. What is the disaggregated performance across the populations the institution serves, including multilingual students, international students, and students using approved assistive technology. An institution serving these populations and unable to answer has not evaluated the tool for its own context.
3. What evidentiary weight is assigned to the output, stated in writing, and does any adverse finding rest on it. If a score alone can initiate a proceeding, the institution has adopted an unvalidated instrument as a trigger for discipline.
4. Who reviews the interaction between approved accommodations and detection outputs. If nobody does, the institution should assume the interaction is adverse, because the mechanism guarantees it.

Conclusion

The advice that graduate students should think for themselves, defend their own ideas, and develop expertise rather than shortcuts is correct and would be correct in any era. It is also not responsive to the risk students actually face, which is not that they will be found shallow but that they will be flagged by a classifier measuring statistical unusualness and asked to prove a negative.

That risk is distributed unevenly, and its distribution is the opposite of the one integrity policy is meant to produce. The students most likely to be wrongly accused are those writing in a second language, those signing in a first one, and those using the accommodations their institutions formally approved. They are being investigated more often not because they are more likely to cheat, but because a proxy for machine writing turns out to be a proxy for writing under constraint.

Institutions did not choose this outcome. They acquired it, in a procurement decision made without validation evidence, and they will keep producing it until someone asks the questions that should have been asked before deployment.

References

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. "GPT Detectors Are Biased Against Non-Native English Writers." *Patterns* 4, no. 7 (July 2023).

Weber-Wulff, Debora, et al. "Testing of Detection Tools for AI-Generated Text." *International Journal for Educational Integrity* 19 (December 2023).

Vanderbilt University Center for Teaching, statement on disabling Turnitin's AI detection feature, August 2023.

Americans with Disabilities Act of 1990 and Section 504 of the Rehabilitation Act of 1973, as applied to postsecondary institutions and academic adjustments.