Detecting What Was Not Written

The Case Against the Machine-Text Detector

Debora Weber-Wulff & Johannes Krüger

Debora Weber-Wulff, Fachbereich Informatik, Hochschule für Technik und Wirtschaft Berlin, Berlin, Germany

Johannes Krüger, Fachbereich Informatik, Hochschule für Technik und Wirtschaft Berlin, Berlin, Germany

European institutions have purchased a class of instruments that claims to identify text produced by a machine, and the evidence for their accuracy does not support the use they are put to. This article examines the detector as a technology and as an administrative practice: the four properties that a detector would need, the false positive that falls on a student writing in a second language, the arms race that the vendors describe as a feature, the procedure that a score cannot support, the alternative of the process, and the governance that an institution needs before it adopts any instrument that accuses a person. It argues that the detector has no place in a procedure of misconduct, and that its only defensible use is as a reason to look at something else.

Contents **Page**

1. The Instrument and the Claim **79**
2. Four Properties a Detector Would Need **79**
3. The False Positive **80**
4. The Arms Race **80**
5. The Score in a Procedure **80**
6. The Evidence That Works **81**
7. The Cost of the Purchase **81**
8. The Vendor's Testimony **82**
9. The Student Who Is Cleared **82**
10. The Alternative Use **83**
11. The Institution That Publishes **83**
12. The Teacher Who Wants a Tool **84**
13. The Plagiarism of the Past **84**
14. The Public Position **85**
15. The Governance Before the Purchase **85**
16. The Sequence of a Policy **85**
17. The Question of Trust **86**
17. The Question of Deterrence **86**

19. Conclusions

1. The Instrument and the Claim

A detector of machine-generated text is a classifier that assigns a probability to the proposition that a text was produced by a model, and it is sold as a solution to a problem of integrity.

The classifier is trained on a corpus of texts that were and were not produced by models, and it learns features that distinguish them in that corpus. The features are statistical: the regularity of the length of the sentences, the distribution of the vocabulary, the perplexity of the sequence. A text that is unusual, uniform or exceptionally correct will be classified as machine-like, whatever produced it.

The consequence is that the instrument does not detect a machine. It detects a property of a text that correlates with a machine in the corpus on which it was trained, and the correlation is weaker than the marketing suggests.

2. Four Properties a Detector Would Need

An instrument that accuses a student must have four properties, and the detectors in use have none of them at the level that a procedure requires.

The first is sensitivity: it must identify the texts that were produced by a machine. The second is specificity: it must not identify the texts that were not, which is the property that matters most, because the cost of the error falls on an innocent person. The third is robustness: it must survive a paraphrase, a translation, a reordering and the insertion of a sentence, all of which a student can perform in a minute.

The fourth is explainability: it must be able to state what in the text produced the score, in a form that the student and an appeal body can examine. A number without a reason is not evidence, and none of the systems in commercial use supplies a reason that would satisfy a panel.

3. The False Positive

The error that matters in a procedure of misconduct is the false positive, and the studies that have examined it have found a rate that no institution would accept in another domain.

The texts most likely to be classified as machine-generated are those of a writer who is competent, consistent and writing in a language that is not her first. The regularity that a second-language writer acquires through instruction is the same regularity that the classifier has learned to associate with a model, and the association produces the error.

The consequence is that the instrument accuses disproportionately the students who have worked hardest on their language, and the disproportion falls on the international students of the institution. A rate of a few per cent applied to a cohort of ten thousand is a few hundred accusations, and the number is the measure of the harm.

4. The Arms Race

The vendors describe the contest between the detectors and the generators as an arms race, and the description is an admission that the instrument does not work.

An arms race is a relation in which each improvement of one side is answered by the other, and the equilibrium is the state in which neither has an advantage. A detector whose accuracy depends on the version of the model that produced the text is obsolete at the moment of the next release, and the institution that has purchased it has purchased a subscription to a moving target.

The instrument that would be defensible is one whose accuracy is stated for a version, tested by an independent body and withdrawn when the version changes. No commercial system is offered on those terms, and an institution that requires them will not find a supplier.

5. The Score in a Procedure

The score of a detector enters a procedure of misconduct as a piece of evidence, and the procedure treats it as though it were a document.

A panel that receives a score of ninety per cent has received a number that it cannot interpret, produced by a model that it cannot inspect, on

a text whose other features it cannot evaluate. The student is asked to demonstrate that she wrote a text, which is the demonstration of a negative, and the burden has shifted to the person accused.

The consequence is a procedure in which the standard of evidence is lower than the regulation states and in which the appeal body, if there is one, has no basis on which to reverse. The instrument that a regulation should state is simple: no finding of misconduct may rest on the output of a detector, and the detector may be used only to prompt an inquiry that must find other evidence.

6. The Evidence That Works

The evidence that a procedure can rely on is the evidence of the process, and it is available in any institution that designs its assessments to produce it.

The drafts with the comments, the version history, the logbook, the data set that the student collected, the notes of a supervision, the oral defence of a text: each of them connects the work to a person, and each of them is a document that a panel can read. The evidence is not certain and it is inspectable, which is the property that a procedure requires.

The instrument is to make the process the ordinary requirement of an assessment, so that the question of authorship is answered before it is asked. An institution that requires the process does not need the detector, and an institution that needs the detector has not required the process.

7. The Cost of the Purchase

The cost of a detector is a licence, and the licence is the smallest part of the expenditure that the instrument causes.

The larger costs are the training of the staff who use it, the cases that it generates, the appeals that it loses, the time of the office that processes them, and the damage to the students who are accused and cleared. The last is not in the budget and it is the largest.

The instrument is a total cost statement, prepared by the office that would administer the cases, before the purchase. A statement that includes the expected number of false positives and the cost of the ap-

peals converts the decision into an arithmetic exercise, and the arithmetic is not favourable.

8. The Vendor's Testimony

The evidence for the accuracy of the detectors comes, in most of the literature available to institutions, from the vendors themselves.

A study conducted by the company that sells the instrument, on a corpus that the company has assembled, with a metric that the company has chosen, is a marketing document with a method. It is not worthless and it is not independent, and an institution that requires an independent evaluation before a purchase is not being unreasonable.

The instrument is a rule of procurement: no instrument that can lead to a finding against a student is purchased without an evaluation by a body that has no interest in the result, published with its method. A rule of that kind is a sentence in the procurement code and it is the minimum of academic governance.

Claim	Reality	Consequence
Identifies machine text	Detects regularity	Accuses the consistent writer
Robust to paraphrase	Defeated by a translation	Obsolete with each release
Objective score	An unexplained number	The burden shifts to the student
Deters misconduct	Changes the method	The process is what deters

Table 1 What a detector claims and what it delivers

9. The Student Who Is Cleared

A student who is accused and cleared has been through a procedure that no finding can repair, and the institution rarely records the case.

The accusation is known to the teacher, frequently to the cohort, and always to the student, who has spent weeks assembling a defence against a number. A student who is cleared in the third month of a semester has lost the semester, and the loss is not a sanction and it is the effect of the instrument.

The instrument is a rule that an inquiry opened on the basis of a detector is closed within ten days if no other evidence is found, and that the record of the inquiry is deleted when it is closed. The two provi-

sions are the minimum that an institution owes to a person it has suspected without a basis.

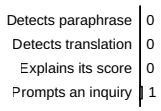


Figure 1 What a detector can and cannot do

10. The Alternative Use

There is a use of the detectors that is defensible and it is the opposite of the one for which they are sold.

A teacher who reads a submission that seems inconsistent with the previous work of a student may use a detector as one of several reasons to look closer: at the drafts, at the version history, at the previous assessments. The detector is a prompt and not a proof, and the inquiry that follows is an inquiry into the process.

The instrument is to write that use into the guidance: the detector may be consulted by a teacher, its result may not be communicated to a student as a finding, and any inquiry must rest on other evidence. A guidance of that kind gives the teacher an instrument and it denies the instrument the authority that it does not deserve.

11. The Institution That Publishes

An institution that uses a detector and does not publish the number of its accusations and their outcomes has adopted an instrument whose effect it does not know.

The four figures that a serious institution would publish are the number of inquiries opened on the basis of a detector, the number that found other evidence, the number that were closed without a finding, and the median duration. The four together describe the cost of the instrument in the lives of students.

The publication is the instrument through which the institution learns, and it is also the instrument through which a vendor's claim is tested against a record that the vendor does not control.

12. The Teacher Who Wants a Tool

The demand for a detector comes from the teachers, and the demand is a symptom of a design problem rather than a request for a technology.

A teacher who has two hundred submissions and a week to mark them cannot examine the process of each. She wants an instrument that tells her where to look, and the detector is offered as that instrument. The need is real and the detector is a poor answer to it.

The instrument that answers the need is a design: a smaller number of assessments with a process that can be examined, a sample of the submissions read closely, and a viva for the cases that matter. The three require hours, and the hours are the resource that the institution must supply if it wants the teacher to stop looking for a machine to do it.

13. The Plagiarism of the Past

The detectors of machine text have been preceded by the detectors of copied text, and the lesson of the earlier generation is directly applicable.

The similarity services are useful and they are not evidence of misconduct: a high similarity score can be produced by a correct citation, a standard method section or a shared template. The institutions that learned to read the score as a reason to look, and not as a finding, have a practice that works. The institutions that treated the score as a finding have produced cases that they lost.

The instrument is the same rule for both generations of the technology: a score is a prompt, the finding rests on the comparison of the texts, and the student is shown the comparison and invited to explain it.

14. The Public Position

A university's position on the detectors is a public statement about how it treats its students, and the statement is read by applicants.

An institution that announces that it uses a detector to protect the integrity of its awards has announced that it will accuse a proportion of its students on the basis of an instrument whose error rate it has not measured. An institution that announces that no finding rests on a score alone has announced a standard that a student can rely on.

The instrument is a paragraph in the regulation of assessment, published on the page that a student reads before she enrols, with the rule and the reason. A paragraph of that kind costs nothing and it distinguishes an institution from its competitors.

15. The Governance Before the Purchase

An institution needs four decisions before it adopts any instrument that can accuse a person, and the four belong to different bodies.

The first is the academic decision: what the instrument is for, taken by the council that governs assessment. The second is the legal: whether the instrument is compatible with the rights of the student and with the law of data protection, taken by the office that advises the institution. The third is the ethical: whether the error falls on a group that the institution has a duty to protect, taken by the body that governs equality.

The fourth is the financial: what the instrument costs over five years, including the cases, taken by the office that holds the budget. A purchase that has not passed the four has been made by an office that had the authority to buy and not the authority to decide.

16. The Sequence of a Policy

An institution that has purchased a detector and now doubts it can withdraw in a sequence, and the sequence begins with the cases.

The first step is the audit of the cases in which the detector was used, with the outcome of each. The second is the rule that no finding rests on a score alone, adopted by the academic council. The third is the requirement of the process in the assessments of the programmes with

the largest cohorts. The fourth is the non-renewal of the licence at the end of its term.

A sequence of that kind takes a year and it costs less than the licence. The audit is the part that produces the argument, because a table of cases is more persuasive than a principle.

17. The Question of Trust

The use of the detectors is a symptom of a relation between the institution and its students that has become a relation of suspicion.

An institution that trusts its students designs assessments that make the trust reasonable. An institution that suspects them installs instruments that make the suspicion visible, and the students respond to the visibility rather than to the rule. The consequence is a climate in which the honest student is treated as a suspect and the dishonest one learns to defeat the instrument.

The instrument is to state the relation in the policy: the institution assumes the integrity of its students, it designs its assessments to make that assumption verifiable, and it accuses a student only on evidence that it can show her. A statement of that kind is the foundation on which any procedure rests.

17. The Question of Deterrence

The argument that survives when the accuracy of a detector is challenged is deterrence: that the instrument prevents the conduct whether or not it proves it.

The argument is plausible and it is testable, and the evidence that would test it is the trend of the conduct in the institutions that have adopted the instrument. The vendors do not publish it, the institutions do not measure it, and the claim therefore rests on the impression of the staff who report that the submissions have changed.

The change that the staff report is real and it is not deterrence: it is the substitution of one method for another, as the earlier generation of similarity software produced a change in the way that text was copied rather than a reduction in the copying. An instrument that changes the method has not prevented the conduct and it has moved it to a place where the institution can see it less.

The instrument is to measure what the adoption changed: the number of the inquiries, the number of the findings, the kinds of conduct that the staff report, in the years before and after. A measurement of that kind is the only evidence about deterrence that an institution can obtain, and it is available to any institution that keeps a record of its cases.

19. Conclusions

The detector of machine-generated text has no place in a procedure of misconduct, and its only defensible use is as one of several reasons for a teacher to look more closely at a process.

Three conclusions follow. The rule must be written: no finding may rest on the output of a detector, the instrument may prompt an inquiry, and the inquiry must close within ten days if no other evidence is found. The alternative must be built: a process of drafts and defences in the assessments that matter, funded in hours, so that the question of authorship is answered by evidence that a panel can read. And the purchase must be governed: an academic, a legal, an ethical and a financial decision, taken before the licence and reviewed at its end, with the cases published.

An institution that has done the three has protected its awards and its students. An institution that has installed a detector has installed a suspicion and given it a number.

References

[1] Weber-Wulff, D. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19(1).

[2] Weber-Wulff, D. (2014). False feathers: a perspective on academic plagiarism. Berlin: Springer.

[3] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint.

[4] Perkins, M. (2023). Academic integrity and artificial intelligence. International Journal for Educational Integrity, 19(1).

[5] European Network for Academic Integrity (2024). Position paper on AI detection tools. Brno.

Biography:

Debora Weber-Wulff is a professor at the Hochschule für Technik und Wirtschaft Berlin working on plagiarism, academic integrity and the testing of detection software.

Johannes Krüger is a researcher at the Hochschule für Technik und Wirtschaft Berlin working on the evaluation of text classification systems and their use in academic procedures.