STIAEM

Slow Tech Innovation

AI, Ethics and
New Teaching Methods

Dismantling False Controls: Limitations and False-Positive Rates of Anti-Plagiarism Software

Co-funded by the European Union

2025-1-IT02-KA220-SCH-000365213

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them... Project reference code: 2025-1-IT02-KA220-SCH-000365213

Università di Foggia
Studi Umanistici, Lettere, Beni Culturali, Scienze della Formazione

STIAEM

Slow Tech Innovation
AI, Ethics and
New Teaching Methods

Co-funded by
the European Union

2025-1-IT02-KA220-SCH-000365213

eduOpen

CC BY-NC-SA 4.0 International

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.

Università di Foggia
Studi Umanistici, Lettere, Beni Culturali,
Scienze della Formazione

OER 6.2.1: Dismantling False Controls: Limitations and False-Positive Rates of Anti-Plagiarism Software
Reference Module: Module 6 - Classroom Practice and Authentic Assessment
Related Video-Pill: Video 6.2 - Redefining Plagiarism in School
Authoring partner: Esenler District Directorate of National Education - Türkiye
Topic / Thematic Focus: Anti-plagiarism software, false positives and AI-text detection limits

Introduction and Learning Objectives

This resource accompanies, together with OER 6.2.2, video-pill 6.2 (Redefining Plagiarism in School) and constitutes its critical in-depth study. Its subject is a control that promises certainty and instead introduces error: the automatic detection of text generated by artificial intelligence. Faced with a piece of work that seems too good, many teachers entrust their suspicion to a software that returns a percentage. This resource shows why, to a large extent, that path is a dead end.

The text connects to the study by Liang and colleagues (2023) on the *bias* of detectors, to the contributions of Perkins (2023) and of Cotton, Cotton and Shipway (2023) on academic integrity in the era of large language models, and to the UNESCO (2023) guidance on generative AI in education and research. All these sources are used in the form of careful paraphrase, without textual quotations.

By the end of the reading, the teacher will be able to:

- explain why automatic detectors estimate statistical parameters and do not recognise AI;
- describe, in a careful and attributed form, the problem of false positives against non-native-speaker authors;
- place detectors in their exploratory function, excluding them from high-stakes decisions.

1. The case that exposes the crisis: the non-native-speaker student

Let us imagine a situation that is unfortunately not rare. A student writes an essay. She is not a native speaker: her writing is correct, linear, made of short sentences and simple syntax. She submits the work. A software, with a nice red percentage, reports that this text is probably generated by artificial intelligence. From that number an accusation of plagiarism is born, and the student finds herself in the absurd position of having to prove that the text is hers.

Why does that writing seem suspicious? Many detectors confuse a feature of second-language writing with the signature of the machine. Those who write in a language that is not their own tend towards a restricted vocabulary, recurring structures, fewer idiomatic variations; the algorithm reads this regularity as a sign of artificiality. But that text is not artificial: it is the text of someone who is learning to write in a second language.

This brings us to the heart of the problem. A linear and correct text is not proof of automatic generation, and an accusation of plagiarism based on the opaque response of a machine does serious damage to the pedagogical relationship: it establishes a climate of suspicion whose highest price is paid precisely by students with more fragile linguistic profiles. This is the point at which video-pill 6.2 opens its own pathway.

2. Why detectors get it wrong: perplexity and burstiness

Automatic detectors do not recognise AI: they estimate some of its statistical parameters. Two are the main ones. The first is perplexity: how predictable a sequence of words is with respect to a reference language model. The second is lexical variance, which in English is called burstiness: how much the text alternates common and rare words. These are mathematical measures, not proofs of authorship.

The writing of non-native speakers, with a circumscribed vocabulary and recurring structures, tends to have low perplexity: it is, in the statistical sense, predictable. And the algorithm, quite simply, mistakes it for artificial. The misunderstanding is not an isolated incident: it is a structural consequence of the way these tools work, and it is the reason why their response cannot count as an assessment.
Some conclusions follow from this. First: the output of a detector is not a proof, but a probabilistic estimate affected by a measurement error, all the more serious the further the text departs from the standard language of the training data. Second: those scores can be manipulated with very small stylistic changes, and therefore do not serve to unmask those who really cheat. Third: algorithmic surveillance undermines the formative alliance between teacher and student and selectively penalises the most fragile.

Parameter	What it measures	Why it misleads
Perplexity	How predictable a sequence of words is	Second-language writing is regular and therefore appears predictable
Burstiness (lexical variance)	How much the text alternates common and rare words	A restricted vocabulary produces little alternation, read as artificiality
Measurement error	Distance from the standard training language	Those who depart from the standard are exposed to a systematic error

3. The data of Liang et al. (2023): false positives are not random

Video-pill 6.2 recalls a study that deserves to be reported with caution, because its figures are strong and must be attributed precisely. Liang and colleagues, published in the journal Patterns in 2023, tested some of the main commercial and academic detectors. In the tests conducted, the TOEFL essays - written by non-native English speakers - were labelled as generated by AI in more than half of the cases, whereas native speakers' texts recorded negligible false positives.
The datum, reported in this way, should be read with caution as regards the exact figures - which depend on the tests, the tools and the versions considered - but the direction it indicates is clear: the tool does not err at random. It errs systematically, and discriminatorily, to the detriment of those with a non-native linguistic profile. This is the most important aspect for schools: not a physiological margin of error, but an error that concentrates on the most vulnerable.
The practical consequence is clear. These software tools can have, at most, a preliminary exploratory function: an alarm bell to be verified by other means, never a conviction. They are not suited to high-stakes decisions, such as the assignment of a grade, a disciplinary sanction or a formal accusation. Entrusting an opaque number with the power to decide a student's fate means mistaking a probability for a proof.

4. What it means to measure: probability, not truth

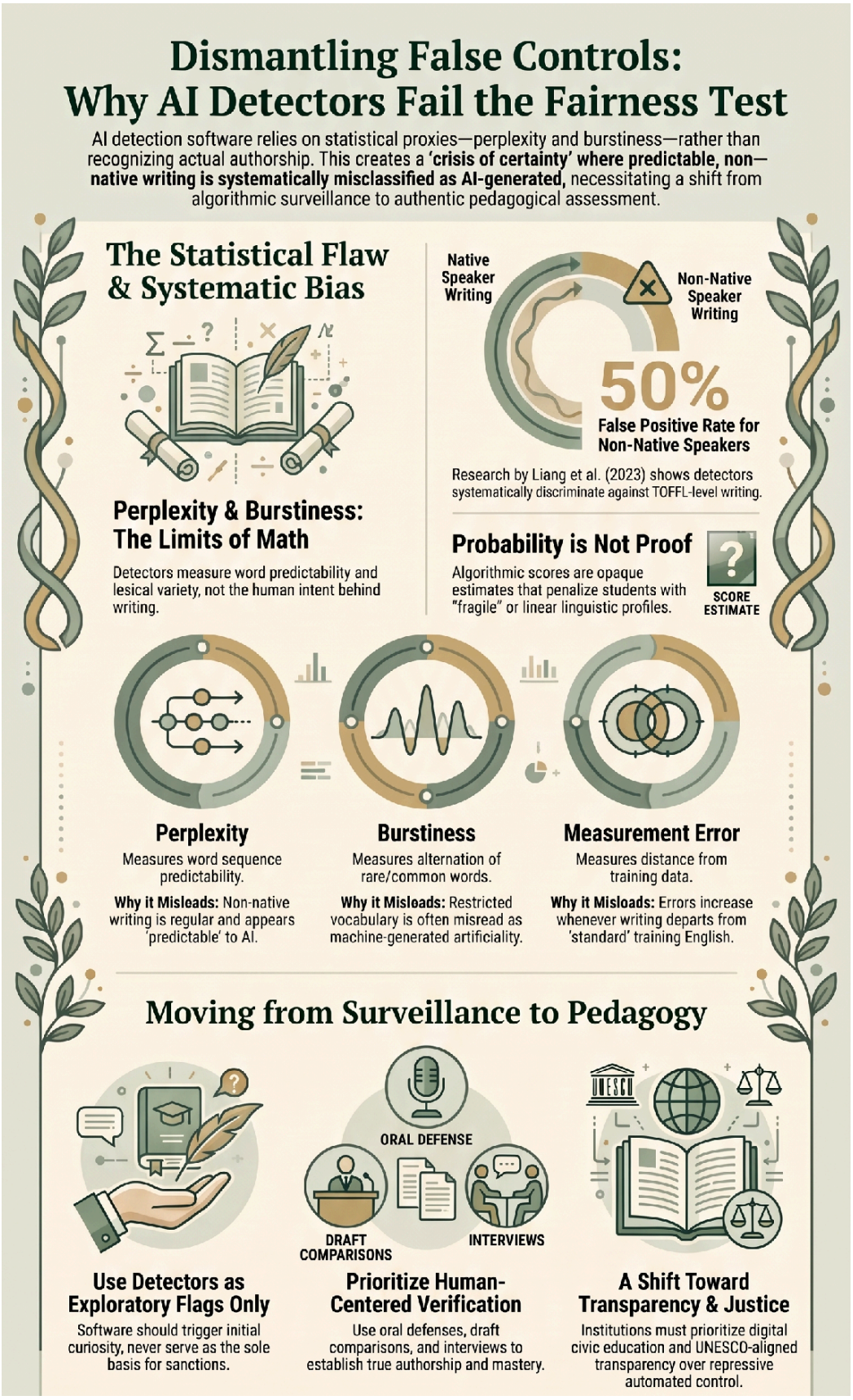
Why is it so difficult to accept that a detector is not a proof? Because the number gives the impression of measurement and measurement gives the impression of objectivity. But what the software produces is not the truth about the authorship of a text: it is an estimate of its statistical similarity to the texts on which it was trained. The difference between the way a tool estimates and what needs to be established is exactly the distinction between probabilistic diagnosis and judgement.
To this is added the question of opacity. Rarely do teacher and student know the data, the objectives and the margins of error of a tool; and the threshold beyond which a percentage

becomes suspicious is almost never made explicit or justified. One entrusts to a system whose criteria are unknown the task of establishing, for a human being, what is authentic. It is an untenable inversion of responsibilities.
There is also a fundamental misunderstanding to be cleared up. The similarity between two texts is not a proof of copying; and, symmetrically, the difference between two texts is not a proof of originality. A text can be almost identical to another through convention, through the use of standardised formulae, through legitimate quotation; and it can be entirely different through the choice of synonyms while nevertheless being derived. Detectors work on surface clues, whereas the pedagogical question concerns authorship and mastery: who thought this text, and what can the person who signs it do. These are different levels, and confusing them is the error that lies at the origin of false positives.
Research on academic integrity insists, for this reason, on transparency and education instead of on control (Perkins, 2023). Understanding how the machine works, in which contexts it is legitimate to use it and how its use is declared is more effective than attempting to repress its use. Understanding, in a word, beats repression: it is the direction that the next resource, OER 6.2.2, develops on the level of policies and attribution.

5. Operational consequences for critical use

What should a teacher do, concretely? Some operational indications follow directly from what precedes. The first: never use the response of a detector as the sole or principal basis of an accusation. The second: if it is consulted, treat it as a clue to be verified with properly pedagogical tools - the oral defence, the comparison of drafts, the interview. The third: make explicit in class the limits of these tools, turning them into a theme of digital civic education.
There is also a consequence of equity, concerning institutional policies. If a detector systematically penalises non-native-speaker students, its reckless use produces harm distributed inequitably. Institutions should therefore govern these tools with caution: limited use, declared purposes, no automation of decisions (Cotton, Cotton & Shipway, 2023; UNESCO, 2023). It is not merely a methodological question, but one of justice.
In the background remains the question that video-pill 6.2 leaves open: if control does not work, what do we put in its place? The answer is not to surrender to the indiscriminate use of AI, but to change the very concept of plagiarism and what we assess. This is the task of the next resource: to shift attention from surveillance to the transparency of attribution and to the redesign of assessment.

Visual Scaffolding



Conclusions and Metacognitive Reflection Prompts

This resource has shown why automatic detectors of generated text cannot found an accusation: they estimate statistical parameters - perplexity and burstiness - and do not recognise AI; they tend to mistake second-language writing for artificiality; and, as documented in careful form by Liang et al. (2023), they produce false positives that concentrate discriminatorily on non-native-speaker authors. The consequence is clear: at most an exploratory function, never high-stakes decisions.

To transfer these core ideas into one's own practice, the following reflection prompts are proposed:

- Have you ever based a suspicion of plagiarism on a percentage? How would your judgement change in the light of the problem of false positives?
- Which pedagogical tool - oral defence, comparison of drafts, interview - would replace, in your practice, the accusation based on the response of a software?
- How would you explain to your students, without technical jargon, that a detector estimates a probability and does not establish a truth?

Technical Metadata Sheet

Metadata Field	Specification / Parameter Value
Identifier	OER 6.2.1
Title	Dismantling False Controls: Limitations and False-Positive Rates of Anti-Plagiarism Software
Creator	Esenler District Directorate of National Education - Türkiye
Description	The resource examines a control that promises certainty and introduces error: the automatic detection of AI-generated text. It explains why detectors estimate statistical parameters and do not recognise AI, and describes, prudently and with attribution, the problem of false positives against non-native-speaker authors (Liang et al., 2023). The practical focus is on placing detectors in an exploratory function and excluding them from high-stakes decisions.
Keywords	AI-text detection; false positives; perplexity; burstiness; non-native speakers; measurement error; high-stakes decisions; academic integrity
Language	EN (English)
Affiliation	Esenler District Directorate of National Education - Türkiye
Contributor	STIAEM Transnational Scientific Board (peer review, Rhodes validation protocol)
Media Format	Text; Infographics
Genre	Training Manual; Educational Guide
URL	
Citation	Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148 Eaton, S. E. (2023). Postplagiarism: The future of academic integrity in the age of artificial intelligence [Academic debate on post-plagiarism]. University of Calgary.

	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 Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching & Learning Practice, 20(2). https://ro.uow.edu.au/jutlp/vol20/iss2/07/ UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693
Educational Level	Secondary School
Associated Module / MOOC Video Lesson	Module 6 - Classroom Practice and Authentic Assessment / Video 6.2 - Redefining Plagiarism in School
Learning Objectives	explain why automatic detectors estimate statistical parameters and do not recognise AI; describe, prudently and with attribution, the problem of false positives against non-native-speaker authors; place detectors in their exploratory function, excluding them from high-stakes decisions.
Rights	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
Date Available	14/09/2026
Date Modified	

References and Webography

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Eaton, S. E. (2023). Postplagiarism: The future of academic integrity in the age of artificial intelligence [Academic debate on post-plagiarism]. University of Calgary.

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

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching & Learning Practice, 20(2). https://ro.uow.edu.au/jutlp/vol20/iss2/07/

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/a