



Editorial Article



When Detection Teaches Error: The Case for Critical Artificial Intelligence Literacy in Language Education

Atena Attaran* 

Faculty of Applied Health & Community Studies, Sheridan College, Ontario, Canada

* **Corresponding author:** Atena Attaran, Faculty of Applied Health & Community Studies, Sheridan College, Ontario, Canada.Email: atena.attaran@sheridancollege.ca

ARTICLE INFO

Article History:

Received: 06/06/2026

Revised: 18/07/2026

Accepted: 09/08/2026

Published: 30/09/2026

**Keywords:**

Academic integrity

Artificial intelligence detection

Critical artificial intelligence literacy

English for academic purposes

Postplagiarism

Second language writing

ABSTRACT

Generative artificial intelligence has pushed many post-secondary institutions toward integrity policies in which automated detection plays a central part, even where the course instructor makes the final judgement. This editorial argues that such policies can lead to unfavorable effects on language learning. Drawing on my experience teaching English for academic purposes and English as a second language at the post-secondary level, I describe an emerging practice, deliberate error injection, in which students insert spelling, grammar, and syntax errors into machine-generated text to lower detection scores. I frame the practice within the postplagiarism era, in which hybrid human-AI writing is increasingly ordinary, and within evidence that detection tools are unreliable and biased against non-native writers, consider how the practice may spread through peer networks, and outline why it matters for learners who are still stabilizing their second language. I propose critical artificial intelligence literacy, integrated into language curricula and supported by process-based assessment, as a more defensible response than prohibition or unguided permission, and I identify the research needed to test these claims.

1. Introduction

Higher education has spent the years since the public release of ChatGPT adjusting to a technology that changed faster than its rules could. This is the setting Eaton (2023) calls a postplagiarism era, in which advanced technologies are a normal part of how people teach, learn, and communicate. Policies written to address plagiarism have been amended in haste, and in many institutions the practical result is a reliance on detection software, such as the artificial intelligence writing indicator built into text-matching platforms, as the first line of enforcement. In many programs, the detector does not decide alone: whether a breach of integrity has occurred is judged by the

professor teaching the course, who weighs the report alongside other evidence. That human judgement is a real safeguard, and it is consistent with the advice that a detector's output should never be the sole basis for an accusation (Weber-Wulff et al., 2023). Yet the report still frames what the professor is asked to consider, and students may not know how much weight it will carry. The intention is sound: institutions have a legitimate responsibility to protect the integrity of assessment. My concern is with what this arrangement teaches students, and in particular what it teaches students who are learning to write in a second language.

► *Cite this paper as:* Attaran A. When Detection Teaches Error: The Case for Critical Artificial Intelligence Literacy in Language Education. 2026;5(3): 32-36. DOI: 10.58803/jclr.v5i3.154



The Author(s). Published by Rovedar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

I teach English for academic purposes and English as a second language courses at the post-secondary level, and I see the consequences of detection-centered policy at close range. This editorial is therefore written from a practitioner's point of view. Its central question is what reliance on detection is doing to students, and whether the ethical and pedagogical costs of current approaches have been adequately counted. I argue that they have not, and that language education needs a different response.

2. Writing in a Postplagiarism Era

We now write, teach, and assess in what Eaton (2023) calls a postplagiarism era: a period in which artificial intelligence is part of ordinary life, including how people compose text. Eaton's account rests on several tenets that bear directly on this editorial. Writing co-created by humans and machines is becoming the norm, and trying to locate where the human ends and the machine begins is a pointless exercise. Humans may hand over control of a text to a tool, but they do not hand over responsibility for it. Attribution remains important as a form of intellectual respect and not merely a technical convention. And historical definitions of plagiarism, built around verbatim copying without acknowledgement, can be transcended rather than patched. Eaton (2023) also argues that as tools improve, the chance of accurately telling human from machine text shrinks, which anticipates the detector findings discussed below.

Detection-centered policy quietly assumes the opposite: that a stable line separates human from machine text, and that crossing it is the offence. If the line is neither stable nor the right ethical question, the useful questions change. Does the student take responsibility for what they submit, acknowledge the help they received, and demonstrate their own learning? For language learners, the shift carries a further complication. Eaton's tenet that language barriers will diminish is the most contested of the six, and she clarifies that she meant tools for crossing those barriers will become more available and effective, not that a first language will matter less (Eaton, 2023). That clarification is where language education has most to add. A tool can remove a barrier from a text without removing it from the learner's own repertoire, and when the writing is itself the object of learning, the two can diverge sharply. Deliberate error injection, described below, sits exactly at that point of divergence.

3. Detection with Known Blind Spots

The evidence base on detection tools gives reason for caution. In one of the most comprehensive

independent tests to date, Weber-Wulff et al. (2023) evaluated fourteen tools and concluded that they were neither accurate nor reliable. Their accuracy dropped further when machine-generated text was lightly edited by hand or paraphrased by software, and the authors advised that a detector's output should never be the sole basis for an accusation. They recommended that institutions prioritize prevention and process-focused assessment over detection. It is worth stating fairly that, in their sample, the commercial tool from Turnitin performed best and produced no false accusations, and that the tools have continued to change since that study was conducted in 2023. Even so, the broader finding stands: detection is an unstable foundation for policy. Narrow, purpose-built detectors can reach high accuracy within a single genre, such as graduate admissions letters, but a general and universally effective detector appears to remain beyond current technology (Zhao et al., 2024).

For language educators, a second finding is more troubling. Liang et al. (2023) showed that widely used detectors frequently misclassified writing by non-native English speakers as machine-generated; more than half of the non-native TOEFL essays in their sample were wrongly flagged, while essays by native-speaking students were classified almost perfectly. The authors linked this to the way many detectors work: text with low linguistic variability, which is common in the writing of learners with a constrained repertoire, is read as machine-like. They also showed that simple prompting strategies that raise the linguistic variability of machine output can bypass the same detectors. Detection logic of this kind, in other words, rewards texts that look irregular and penalises texts that look controlled, and it does so most harshly for the learners we teach.

4. Deliberate Error Injection: An Emerging Practice

What I have observed in my own classrooms, and what I believe deserves systematic study, is a student response to this logic. Technologically confident students appear to be deliberately adding errors to machine-generated text before submitting it: incorrect verb tenses, breaks in subject-verb agreement, misspellings of words they know how to spell, misplaced punctuation, simplified vocabulary, and fractured sentence structures. I call this practice *deliberate error injection*. It is a rational strategy given how detectors are known to behave. A student who understands that flawless prose attracts suspicion will reasonably conclude that imperfect prose is safer. Where a professor makes the final decision, the same logic may extend from the tool to the human reader,

since a draft that looks like a typical learner's draft is less likely to prompt questions from either.

I want to be clear about the status of this claim. It is a practitioner's observation, not a documented prevalence, and I am not aware of published studies that measure how common the practice is. That gap is itself a finding worth reporting, because the behavior is unlikely to be visible in institutional data: a successful evasion looks, by design, like an ordinary learner's draft.

The pedagogical concern is specific to language learning. Ferris (2011) has shown how much second language writers' accuracy development depends on how errors are noticed and treated. A practice in which learners repeatedly and deliberately produce the forms they are trying to abandon runs against that grain. My hypothesis is that the boundary between a strategic error and an internalized one is hard to hold for learners whose knowledge of English is not yet automatic, so that intended errors may leak into unintended ones. This is a hypothesis, not an established effect, and I flag it as a priority for empirical work. What is not speculative is the opportunity cost: time that could be spent building accuracy is spent, instead, engineering inaccuracy.

5. How the Practice Spreads

The reason this matters beyond individual cases is diffusion. Students are densely connected through informal messaging groups organized by cohort, program, or course, and instant messaging is already an established channel for academic communication between and among students and teachers (Nyasulu & Chawinga 2019). A tactic that is cheap, practical, and endorsed by a trusted classmate is well suited to travel through such channels. In my experience, once one student finds that intentional errors lower a detection score, the tip is shared with specifics: which kinds of errors, how many per paragraph, and which tools are least sensitive. Peer endorsement lowers the psychological barrier to adoption in a way that an anonymous online tutorial does not, and a single student's workaround can become a cohort's habit. Again, this account is drawn from practice; how quickly and how widely such tactics spread is an empirical question.

6. Why Prohibition and Permission Both Fall Short

Deliberate error injection is best understood as a response to policy design. A rule that forbids a behavior and enforces the rule through an unreliable detector, without helping students understand why the behavior is harmful, invites evasion and rewards

those who evade most skillfully. Its mirror image, a permissive policy that allows machine assistance without critical framing, leaves students without the concepts needed to tell productive use from harmful use. Neither is fit for purpose where the writing itself is the object of learning. Unlike contract cheating, the use of artificial intelligence tools is not automatically unethical, as Weber-Wulff et al. (2023) note, and this is precisely why a blanket prohibition is a blunt instrument: it fails to distinguish declared, learning-supportive use from concealed substitution. In a postplagiarism era, where hybrid writing is becoming ordinary (Eaton, 2023), that distinction, and not the one between human and machine text, is the one worth defending.

7. Critical Artificial Intelligence Literacy as an Alternative

My proposal is to integrate artificial intelligence literacy, and critical artificial intelligence literacy in particular, into English education curricula. Long and Magerko (2020) define artificial intelligence literacy as a set of competencies that enable people to critically evaluate artificial intelligence technologies, communicate and collaborate with them, and use them as a tool. Ng et al. (2021) organize the field around knowing and understanding, using and applying, evaluating and creating, and ethical considerations. In the language classroom, Kohnke et al. (2023) argue that teachers and learners need digital competencies to use tools such as ChatGPT ethically and effectively. I use *critical* to mark a step beyond operating a tool: the move from knowing how to use it to evaluating the ethical, cognitive, and developmental consequences of choosing to use it in a particular way. Four classroom practices follow from this.

Make the limits of detection discussable. Students should hear directly that detectors are fallible and biased toward misreading second language writing (Liang et al., 2023; Weber-Wulff et al., 2023). Paradoxically, honesty about the tools removes some of the mystique that makes evasion attractive, and it shows students that the institution is willing to reason with them.

Teach the linguistic cost of camouflage. Learners can compare machine output with their own drafts, examine what makes writing accurate, fluent, or natural, and discuss why deliberately degrading a text works against the accuracy they are trying to acquire. The aim is to treat errors as evidence of learning to be examined and revised, not as a disguise.

Assess process, not only product. Weber-Wulff et al. (2023) recommend focusing written assessment

on the development of student skills. Staged drafts, annotated revisions, and short reflective commentaries make learning visible and make detection scores unnecessary as the primary evidence.

Use declared-use frameworks and treat flags as conversation starters. Where machine assistance is permitted, students should state what they used it for and why. Where a detector raises a concern, the result should open a conversation with the student and not close a case. Because the professor usually makes the final decision, this is also where instructor judgment matters most: a flag is one piece of evidence to weigh alongside drafts, revisions, and the student's own account. This follows two principles from Eaton (2023): that people may relinquish control of a text but not responsibility for it, and that attribution remains important.

The ethical justification for this approach is that students have a right to understand the consequences of the choices they make about their own learning. Critical literacy treats them as ethical agents who can make informed decisions, rather than as rule-followers to be monitored and penalized. In language education that posture is also the more pedagogically productive one. It also fits the postplagiarism premise that what remains constant is human responsibility, not our ability to police the boundary between human and machine text (Eaton, 2023).

8. Challenges and Limitations

This proposal has real costs. It requires instructors who are themselves critically literate about artificial intelligence, a capacity that cannot be assumed and that institutions must support through professional development. It requires curriculum time in programs where contact hours are already tightly allocated. It also requires institutions to treat artificial intelligence ethics as part of language and writing education and not as a compliance exercise. There are further limits to what I can claim. The argument rests on a practitioner's observations, so the prevalence and spread of deliberate error injection remain undocumented. I also cannot say how instructors weigh detector reports when they make integrity decisions, which is likely to vary across programs. The detection studies I cite are largely based on tool versions from 2023 and on English-language texts, and detectors have continued to change. Finally, whether critical artificial intelligence literacy actually reduces evasion or protects language development is an empirical question that this editorial cannot answer.

These limits point to a research agenda: surveys and interviews to establish whether and how students use error injection; longitudinal writing data to test whether deliberate errors affect accuracy development; studies of how instructors weigh

detector reports in integrity decisions; comparative studies of detection-centered and literacy-centered policies; and evaluations of professional development for language instructors.

9. Conclusion

Deliberate error injection illustrates an overlooked consequence of detection-centered integrity policy: students may be incentivized to produce inaccurate language on purpose, which is especially concerning where language development is itself the learning objective. Institutions are right to protect academic integrity, but in a postplagiarism era (Eaton, 2023), policies that rely mainly on detection and enforcement risk generating behaviors that undermine learning. A more effective and more defensible approach is to give students the knowledge to evaluate their own use of artificial intelligence and to understand its consequences for their academic and linguistic development. Integrating critical artificial intelligence literacy into language education is one way to build a culture of informed, responsible, and reflective engagement with these technologies. Language educators are well placed to lead that work, and the field should now test it.

Declarations

Authors' contributions

Atena Attaran conceived the argument, drafted and revised the manuscript, and approved the final version. The author agrees to be accountable for all aspects of the work.

Acknowledgement

None.

Competing interests

This editorial was externally peer-reviewed.

Ethical consideration

The author confirms that ethical issues, including plagiarism, double submission, and data originality, were observed. This editorial reports the author's professional perspective, involved no human participants or student data, and required no ethical approval.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Statement of artificial intelligence use

The intellectual direction, question selection, and scholarly framing reflect the author's own academic judgment. Claude (Anthropic; model Claude Sonnet

5.5, accessed through the claude.ai web interface in September 2026) was used to refine the language.

References

1. Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity*, 19(1), Article 23. <https://doi.org/10.1007/s40979-023-00144-1>
2. Ferris, D. R. (2011). *Treatment of error in second language student writing* (2nd ed.). University of Michigan Press.
3. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550. <https://doi.org/10.1177/00336882231162868>
4. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779. <https://doi.org/10.1016/j.patter.2023.100779>
5. Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. <https://doi.org/10.1145/3313831.3376727>
6. Nyasulu, C., Chawinga, W. (2019) Using the decomposed theory of planned behaviour to understand university students' adoption of WhatsApp in learning. *E-Learning and Digital Media*, 16(5). <https://doi.org/10.1177/2042753019835906>
7. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, Article 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
8. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
9. Zhao, Y., Borelli, A., Martinez, F., Xue, H., & Weiss, G. M. (2024). Admissions in the age of AI: Detecting AI-generated application materials in higher education. *Scientific Reports*, 14, Article 26411. <https://doi.org/10.1038/s41598-024-77847-z>