From Proof of Output to Proof of Agency: Educational Assessment After Generative AI

PIATRA . INSTITUTE

June 2026

Abstract

Generative AI did not create a cheating problem in education; it broke an inference. The take-home essay, the problem set, and the coding assignment were never the thing being measured. They were proxies, and the institution read backward from a good artifact to a competent student. AI severs that link, because a good artifact can now be produced whether or not the student is competent, so the crisis is one of assessment validity and attribution rather than of integrity. This paper models the artifact as a measurement channel from a student's latent competence to an observed score, and asks three computable questions. When does the proxy collapse? Treating substitution as a rational choice, the artifact's validity, the correlation between the recorded score and true competence, falls from about 0.90 under weak AI to near zero under strong AI on a substitutable task, and the collapse is governed by task substitutability: validity crosses one half at a substitutability near 0.68, so low-substitutability work such as supervised and practical assessment keeps a validity near 0.95 while the open take-home loses almost all of it. Can detection save the artifact? No, and the reason is Bayesian: with a detector that flags AI text at a true-positive rate that erodes under cheap evasion and a false-positive rate that does not, at a moderate prevalence the posterior that a flagged student actually substituted is about 0.58 at realistic evasion and below a coin flip beyond it, so roughly 0.42 of accusations fall on the honest while the motivated evade, and enforcement manufactures injustice rather than deterrence. What restores trust? A verification probe, a short viva or a request to modify or explain, measures agency rather than output and is a second channel that substitution cannot fake; because the probe also deters, sampling it on about a third of students restores validity to 0.81 at 6.5 percent of the cost of securely assessing every artifact, and targeting the probe at the suspicious cases reaches 0.95 against 0.66 for random sampling at the same budget. The model separates the crisis from teacher quality: the channel breaks as a structural property of substitutability regardless of how well a course is taught, though poor assignment design accelerates it. The equity costs of the second channel are real and the paper does not minimize them. What AI made expensive is exactly what the artifact was always a proxy for: the agency to explain, reproduce, modify, transfer, and defend the work.

1. The Broken Inference

The discourse calls it a cheating problem, and that framing has cost a year of wasted effort. Cheating is an old offense against a rule, and the response to it is enforcement. What generative AI did is different and worse for the institution: it broke an inference the whole apparatus of grading silently depends on. A take-home essay, a problem set, a program, a lab report was never the object of interest. It was a proxy, evidence from which a grader inferred something unobservable, the student's competence, by reading backward from the quality of the artifact to the quality of the mind that supposedly produced it. The inference ran: this artifact is good, good artifacts are hard to produce without competence, therefore the student is competent. Every step of that was load-bearing, and the middle step is the one AI removed. A good artifact is no longer hard to produce without competence. The arrow from artifact to mind is cut.

This reframing is not a softening. It is a sharpening, because it locates the failure precisely and

tells you which repairs can work. The problem is not that some students break a rule; it is that the artifact has lost validity as a measurement, in the technical sense that validity is the degree to which evidence supports the interpretation placed on a score (Messick, 1989). An essay can still be excellent work and still tell a grader almost nothing about who can do what, because the score it yields no longer correlates with the competence the grader means to certify. The crisis is one of attribution and validity, and detection, the response the framing of cheating recommends, addresses neither; even early integrity-focused accounts of the ChatGPT moment recognized that the problem was reframing assessment rather than merely policing it (Cotton, Cotton and Shipway, 2024).

The rest of this paper treats the artifact as a measurement channel and runs three questions through a model of it. The first asks when the channel collapses, and finds that it collapses as a function of task substitutability rather than uniformly, which is why the answer is to redesign tasks rather than to despair. The second asks whether detection can repair the channel, and finds that it cannot, for a reason that is Bayesian and not technological, so that better detectors would not help. The third asks what does repair it, and finds a second channel, one that measures agency rather than output and that a small, well-aimed sample of can restore at a fraction of the cost the panic assumes. The throughline is the title. Assessment has to move from proving that an output exists to proving that the agency behind it does.

2. When the Proxy Collapses

Let a student have a latent competence c, the quantity the institution wants to measure, and let the artifact yield a score. An honest artifact tracks competence, scoring c up to noise. A substituted artifact, produced with AI, scores $c + s(q - c)$, where q is the quality AI can reach and s is the task's substitutability, the degree to which the tool can stand in for the student. On a fully substitutable task, s near one, the substituted artifact scores near q regardless of c, and the weakest student submits the strongest work; on a non-substitutable task, s near zero, AI cannot help and the artifact still measures the student. Substitution is a choice, and the model lets each student make it rationally: substitute when the reward gain from a higher score exceeds the value of the learning forgone and any expected cost of verification. Cheating has long been understood this way, as an incentive-sensitive decision shaped by opportunity and perceived risk rather than a fixed trait (McCabe, Treviño and Butterfield, 2001). The students with the most to gain are the least competent, because their honest score is lowest and the lift is largest, so substitution is not random across the class but concentrated where the artifact most needs to discriminate.

The consequence is a collapse, and the model lets us watch it. The artifact's validity, defined as the correlation between the institution's recorded score and true competence over a population of $8{,}000$ students, falls from about 0.90 when AI is weak to about 0.09 when AI is strong and the task is substitutable. Near zero is the literal reading. When the low-competence half of a class submits AI work scoring near the top, the artifact and the competence it was meant to track come apart entirely, and a grader ranking students by their essays is ranking them by who chose to substitute, not by what they know. The validity does not degrade gracefully. It is gone.

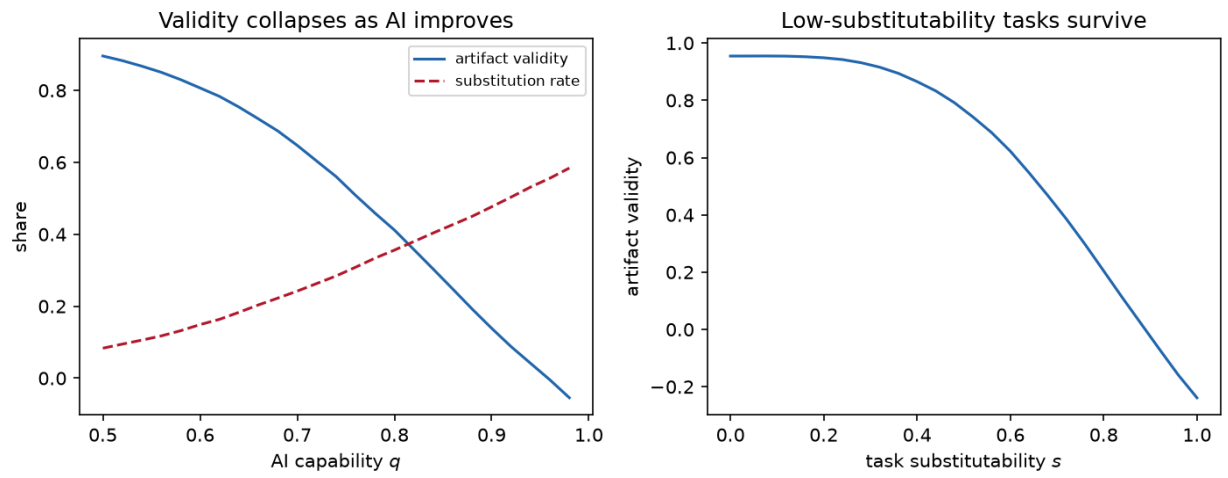

Figure 1: Left: as AI capability q rises on a substitutable task, the artifact's validity (its correlation with true competence) falls from about 0.90 toward zero while the substitution rate climbs, because the students with the most to gain from substituting are the least competent. Right: validity against task substitutability s at full AI capability. The collapse is governed by s: validity crosses one half near $s = 0.68$, so highly substitutable tasks lose almost all validity while low-substitutability tasks, supervised exams, practicals, live performance, keep a validity near 0.95.

What the second panel shows is the part the panic misses. The collapse is not uniform across assessment; it is a function of substitutability, and validity crosses one half near a substitutability of 0.68. Tasks above that line, the unsupervised essay, the take-home problem set, the overnight coding assignment, lose almost everything. Tasks below it, the supervised written exam, the lab practical, the studio critique, the live defense, keep a validity near 0.95 because AI cannot stand in for the student inside them. This is why the instinctive return to in-class, device-free, handwritten assessment is not nostalgia but a correct response to the model: those formats have low substitutability, so their channel never broke. The design lever is substitutability, and it is the lever the rest of the paper turns.

3. Why Detection Is Dominated

The first reflex was to repair the artifact in place by catching the substituted ones, and a market of AI-text detectors appeared to do it. The reflex fails, and it fails for a reason that no improvement in detectors would fix, because the failure is in the arithmetic of base rates rather than in the technology. A detector is a test with a true-positive rate, the share of substituted artifacts it flags, and a false-positive rate, the share of honest artifacts it flags by mistake. Two facts about the setting decide the outcome. The false-positive rate falls on the honest, who are the majority, and it is known to fall unevenly, with non-native English writers flagged far more often (Liang, Yuksekgonul, Mao, Wu and Zou, 2023). And the true-positive rate is not fixed, because substitution can be paraphrased, re-prompted, and run through humanizing tools cheaply, so that text from a determined substituter is hard to detect at all, a fragility the detector literature has documented directly (Sadasivan, Kumar, Balasubramanian, Wang and Feizi, 2023; Weber-Wulff, Anohina-Naumeca,

Bjelobaba et al., 2023).

Put a moderate prevalence of deliberate substitution, here 0.30, through the arithmetic. A naive detector with a true-positive rate of 0.85 and a false-positive rate of 0.08 sounds usable. But the motivated substituters evade, dropping the effective true-positive rate, and at a realistic evasion the posterior that a flagged student actually substituted, the precision of the accusation, falls to about 0.58, barely better than a coin toss, and beyond it below a half. About 0.42 of all accusations land on honest students, and the detector mostly catches the careless few who did not bother to evade while waving the motivated through. The harm is asymmetric in the worst direction: a missed cheat is a small loss to validity, but a false accusation is a catastrophe inflicted on an innocent, and an enforcement regime that produces almost one innocent accusation for every real catch is not a deterrent but a generator of injustice with an adjudication bill attached.

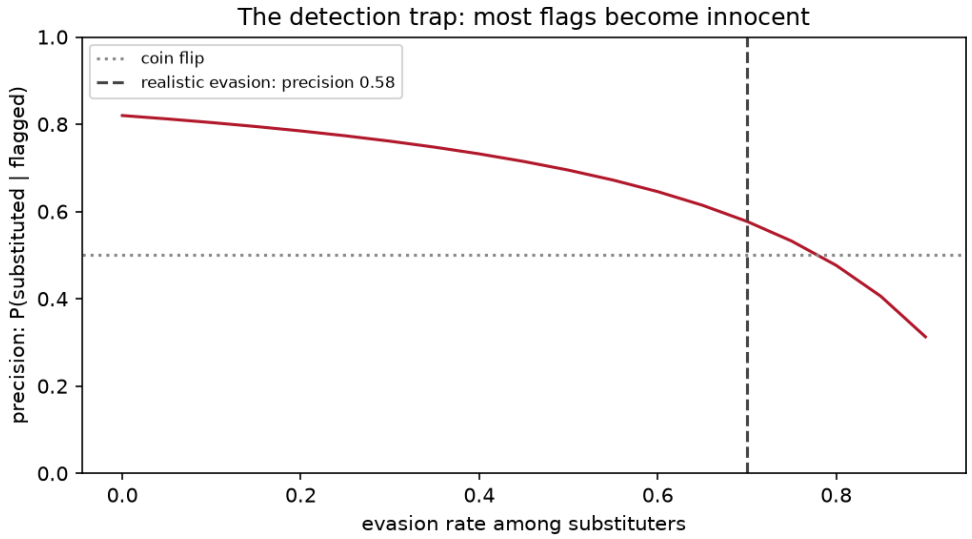

Figure 2: The precision of an accusation, the posterior probability that a flagged student actually substituted, as substituters evade detection. Evasion lowers the effective true-positive rate while the false-positive rate on honest work stays put, so precision falls from a usable level to about 0.58 at realistic evasion (dashed) and through the coin-flip line beyond it. Because the honest are the majority and the motivated evade, the flagged pool fills with innocents while the real substituters pass.

The conclusion is structural. The shape of the problem is the problem, not the tool. Detection is dominated not because today's detectors are immature but because the Bayesian shape, a rare-enough signal, an erodable true-positive rate, and a false-positive rate on a large honest majority, guarantees that flags are mostly wrong exactly when cheating is motivated enough to matter. The institutions that have quietly retired their detectors reached this conclusion empirically. The model says they were right, and that no procurement decision reverses it.

4. The Second Channel

If the artifact channel is broken and cannot be repaired by inspection, the repair has to come from a second channel, independent of the first. The candidate is old and was largely abandoned for reasons of cost: the viva, the oral defense, the request to explain a step, reproduce a result, modify the work under a new constraint, or solve a near-transfer variant on the spot. What these have in common is that they measure agency rather than output. They ask not whether a good artifact exists but whether the student can do the things a competent author of it could do, and that is precisely what a substituter cannot fake, because the agency was never theirs to begin with. In measurement terms the probe is a channel nearly orthogonal to the artifact, so it adds the information the artifact lost. Defending assessment security in a digital world turns on this move from inspecting products to observing process and agency (Dawson, 2021), authentic assessment has been argued to protect integrity and build the agency employers want at once (Sotiriadou, Logan, Daly and Guest, 2020), and authentic and transfer-oriented assessment has argued for the shift on learning grounds for decades (Gulikers, Bastiaens and Kirschner, 2004; Boud and Falchikov, 2006; Perkins and Salomon, 1992).

The objection to the viva was always cost, and the model's contribution is to show the cost was overestimated, for two reasons. The first is that verification need not be universal. The second is that verification deters. Because a student deciding whether to substitute weighs the chance of facing a probe they would fail, sampling the probe at a rate changes the decision, not only the measurement, and the deterrence is convex: even a modest chance of a viva flips the rational substituter back to honest work, because the expected cost of being caught undefended is high. Sampling the probe on about a third of students restores the artifact's validity from near zero to 0.81, roughly the level of an unaided pre-AI artifact, and at the sampled rate substitution itself falls to about 0.06, because the deterrence does most of the work and the direct measurement of the sampled does the rest. The price is the part that should change the conversation. A five-minute probe sampled on a third of a class costs about 6.5 percent of what it would cost to securely assess every artifact in the old high-touch way. The viva was never unaffordable. It was being imagined as universal when it only ever needed to be sampled.

Sampling can be smarter than random. Because a substituted artifact tends to score far above what a student's prior record would predict, the suspicious gap between the artifact and a noisy estimate of the student's competence is a signal of where to spend the probe. Targeting the same verification budget at the largest such gaps reaches a validity of 0.95, against 0.66 for spending the budget at random, because the probes land where the artifact and the student diverge most rather than on the obviously honest. The design that follows is concrete and cheap: keep the artifacts, because they remain useful work and useful for learning, and attach to them a sampled, partly targeted layer of short agency probes that carry the weight the artifact can no longer bear.

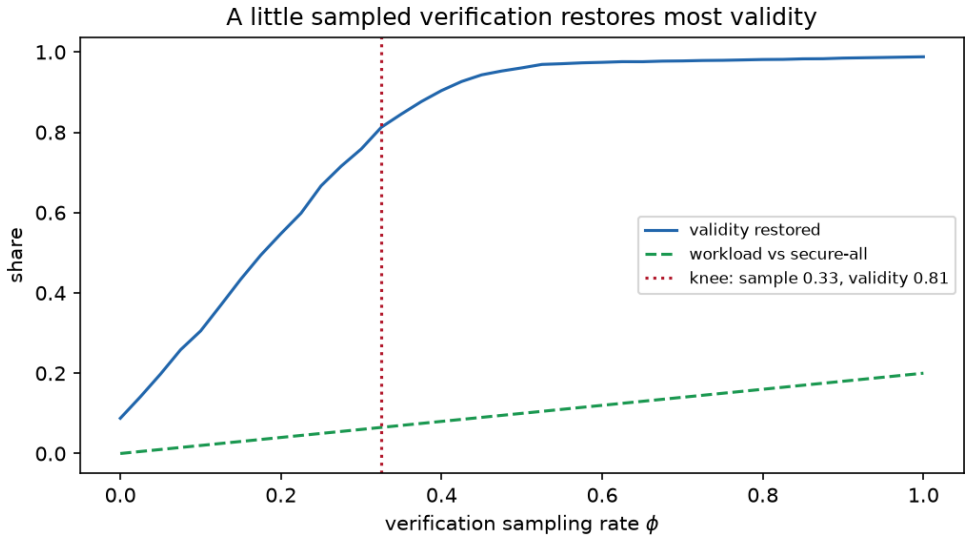

Figure 3: Validity restored against the verification sampling rate ϕ (blue), with the educator workload as a fraction of securely assessing every artifact (green). A little sampled verification restores most of the lost validity, because the probe both measures the sampled student and deters the rest; the knee (dashed) reaches a validity of 0.81 at a sampling rate near a third, for about 6.5 percent of the cost of securing every artifact.

5. What This Says About Teachers

A loud version of the crisis blames teachers: students disengage and outsource to AI because most instructors cannot explain the way a polished online video can, and the assignments they set are lazy. The model lets the claim be separated into a part that is wrong about the cause and a part that is right about a lever. The collapse in section 2 is a property of the channel, of substitutability and AI capability, and it happens regardless of how well the course is taught; a brilliant lecturer who sets an unsupervised, highly substitutable essay watches its validity collapse exactly as a dull one does, because teaching quality and authentication are different variables and the artifact's validity does not depend on the first. So the reduction of the crisis to teacher incompetence is wrong about the cause. The channel did not break because teachers got worse. It broke because a confounder appeared in a measurement that was always more fragile than anyone admitted.

The part that is right is narrow and worth keeping. Assignment design sets substitutability, and substitutability is the lever, so an instructor who sets work high in substitutability accelerates the collapse and one who sets work low in it slows the collapse. That is a design competence, not a charisma, and it is teachable and cheap, unlike the demand that every instructor perform like an edited, scripted, self-selected video made for already motivated viewers, which is not a fair benchmark for a live classroom of mixed preparation and compulsory attendance. The fair synthesis is that the crisis is structural and the remedy is partly in instructors' hands, but the part in their hands is the design of tasks and the use of sampled verification, not a heroic improvement in explanatory talent that the comparison to online stars was always going to make look like failure.

6. What It Costs the Vulnerable

The second channel is not free of harm, and a paper that recommended it without saying so would be evading the thing it is most likely to get wrong. An oral probe measures agency, but it also measures fluency, confidence, and comfort under interrogation, and those are not evenly distributed. A viva can disadvantage the anxious, the neurodivergent, the non-native speaker, and the student whose competence is real but whose performance under a stranger's questioning is not its equal, in the same way the detector's false positives already fall on non-native writers (Liang, Yuksekgonul, Mao, Wu and Zou, 2023). Process tracking, version histories, prompt logs, and the rest can curdle into surveillance, and a regime that demands a student document every keystroke to prove they are not a machine imposes a burden that lands hardest on those least able to perform legibility on demand. High-touch verification is also resource-bound, and a large class without staff cannot run it at all, which threatens to make the well-funded course assessable and the mass course not.

These are real, and the model's own design mitigates some of them without resolving them. Sampling rather than universal verification means less surveillance, not more, and targeting means the probe falls on the suspicious rather than on everyone, so the median honest student is rarely touched. Multiple probe formats, written reflection, asynchronous code-modification, a choice of defense mode, can widen the channel beyond live oral fluency. None of this makes the second channel costless, and the right posture is to hold the cost in view rather than to pretend the move from artifact to agency is a clean upgrade. It buys validity, and it spends something, and the something it spends is paid disproportionately by students who were already paying the most.

7. What the Model Does Not Settle

The model prices a measurement problem and a verification design, and several large questions sit outside it. It does not measure any real course, detector, or cohort; the 0.90 and the 0.58 and the 6.5 percent are properties of a stipulated channel chosen to be transparent, not estimates of a classroom. It does not claim that AI is bad for education, because the same tool that broke the authentication channel is a genuine instrument of learning, and the argument here is only that an artifact made with it is weak evidence of agency, not that making things with it is wrong; the disclosure frameworks that let students use AI and then defend what they produced are consistent with everything above. And it does not settle whether the credential is worth defending at all. If a degree is mostly a signal sorting workers rather than a record of competence (Spence, 1973; Arrow, 1973; Caplan, 2018), then the collapse of artifact validity is a crisis for the signal, and one response is to let the signal go and verify competence at the point of use, in the work-sample trial and the structured audition, rather than at the point of schooling. The model is neutral on that choice; it only insists that wherever competence is certified, the artifact alone can no longer do it.

What survives all of this is the shift the title names. For a generation the artifact stood in for the agency behind it, cheaply and well enough, and the institution forgot the proxy was a proxy. AI made the substitution visible by making it free, and in doing so it raised the price of the one thing the artifact was always standing in for, the capacity of a particular person to explain, reproduce,

modify, transfer, and defend the work. That capacity is what assessment was always trying to reach and what it can now reach only by asking for it directly. The expensive thing AI exposed is not a flaw in the system. It is the point of it.

References

Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics*, 2(3), 193–216.

Boud, D., and Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, 31(4), 399–413.

Caplan, B. (2018). *The Case Against Education: Why the Education System Is a Waste of Time and Money*. Princeton University Press.

Cotton, D. R. E., Cotton, P. A., and Shipway, J. R. (2024). Chatting and cheating: ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239.

Dawson, P. (2021). *Defending Assessment Security in a Digital World: Preventing E-Cheating and Supporting Academic Integrity in Higher Education*. Routledge.

Gulikers, J. T. M., Bastiaens, T. J., and Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67–86.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779.

McCabe, D. L., Treviño, L. K., and Butterfield, K. D. (2001). Cheating in academic institutions: a decade of research. *Ethics & Behavior*, 11(3), 219–232.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Perkins, D. N., and Salomon, G. (1992). Transfer of learning. In *International Encyclopedia of Education* (2nd ed.). Pergamon Press.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv preprint* arXiv:2303.11156.

Sotiriadou, P., Logan, D., Daly, A., and Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education*, 45(11), 2132–2148.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., and Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, 26.