Cognitive Outsourcing in Education: Why AI's Real Classroom Crisis Is Verification, Not Cheating

The SIAI Research Editorial

Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

Much of the debate about generative AI in education has been framed around academic integrity, conceived of as rates of detection and compliance with integrity policies. Largely overlooked is a behavioral shift of far greater consequence. It is the proclivity of students and ever more those in instructor roles, to accept AI-generated artifacts as functionally complete without independently vetting them. The paper calls the practice cognitive outsourcing. Drawing on two nationally representative surveys of the United States and the United Kingdom, experimental and quasi-experimental studies on AI-supported learning and labor-economics research on AI's uneven effects across seniority levels, the paper distinguishes cognitive outsourcing from ordinary, often beneficial cognitive offloading. It identifies the conditions under which AI-supported instruction enhances, rather than replaces, durable learning. The analysis shows that prohibition is neither practical nor desirable given current levels of adoption. A more strategic intervention is to steer AI-oriented pedagogy away from substituting for human thought. This approach involves reshaping boundaries of acceptable use at the task level rather than imposing blanket institutional rules and revising assessment design and professional accreditation so that the unassisted skill set remains verifiable at specified points. The evidence further indicates that disparities in access to and prior knowledge of, digital tools and subject matter will exacerbate existing educational inequalities. Evidence from the labor market on declining junior hiring in occupations exposed to AI lends added urgency to this reform agenda. Absent such deliberate interventions, the paper concludes, current AI productivity gains may depend on expertise that the same technology makes harder for new entrants to build.

1 Introduction - Most Students Already Rely on AI Tools for Their Homework

By the middle of 2026, public debate about generative artificial intelligence in classrooms has largely settled into a familiar and comparatively narrow dispute: if students are using chatbots to complete work they were meant to do themselves and whether schools can or should try to stop them. This framing is not wrong so much as insufficient. It treats AI in education as a policing problem, measured in detection rates along with honor-code violations, when the more consequential shift is epistemic rather than disciplinary. The relevant question is no longer only whether an assignment was produced honestly, but whether anyone in the exchange, student or teacher, still reads and checks what the system has produced before acting on it. That behavior, accepting an AI-generated output as functionally finished without subjecting it to independent scrutiny, can usefully be called cognitive outsourcing. It is distinct from ordinary assistance and distinguishing it from assistance is the analytical task this paper sets for itself.

The scale of the underlying behavior is no longer in dispute, even if its interpretation is. In a national poll of more than 1,100 respondents released in July 2026, the education-technology provider Instructure found that ninety percent of college students reported using AI in class at least occasionally, alongside sixty-one percent of instructors.[1] Sixty-five percent of both groups said they worried that AI could seem confident while being wrong, a concern that anticipates much of what follows in this paper.[2] In the United Kingdom, the Higher Education Policy Institute's annual survey of undergraduates recorded a jump in the share of students using generative AI for assessed work from fifty-three percent in 2024 to eighty-eight percent in 2025, with the share reporting no AI use of any kind for coursework falling to twelve percent.[3] The same survey found that ninety-two percent of respondents had used some AI tool during the year, up from sixty-six percent twelve months earlier. That use is not evenly distributed: male students, those on STEM and health courses and more socioeconomically advantaged students were all more likely to use these tools, a disparity that appears consistently across the wider evidence.[4] In the United States, the College Board's ongoing research series on high-school AI use recorded a rise from seventy-nine percent to eighty-four percent of surveyed students reporting generative AI use for schoolwork between January and May of 2025 alone.[5] RAND's American Youth Panel, drawing on a nationally representative sample of young people aged twelve to twenty-nine, found that the share reporting AI use for homework rose from forty-eight percent in May 2025 to sixty-two percent by December of that year, with growth concentrated among middle and high schoolers.[6] The same RAND survey recorded that sixty-seven percent of respondents by the end of 2025 agreed that AI use is harming students' critical thinking, an increase of more than ten percentage points over the prior ten months, indicating that rising use and rising unease are proceeding together rather than one displacing the other.[7]

These figures will continue to rise and there is little reason to expect otherwise. Unlike earlier waves of educational technology, which schools could in principle exclude from their own facilities, generative AI is accessed through devices and accounts that students already control outside any institution's jurisdiction. More importantly, the labor market into which today's students will graduate is beginning to treat familiarity with AI tools as a job-relevant competency in its own right, separate from and sometimes substituted for the underlying

expertise those tools are meant to support. A university's decision to forbid AI use inside its own walls therefore does very little to change how a student engages with these systems everywhere else and it does nothing to prepare that student for an economy in which employers will, correctly or not, expect AI fluency as a baseline. Prohibition, in other words, is not a serious policy option, whatever its appeal as a stopgap.

What has been comparatively neglected in this fast-moving conversation is a precise account of what cognitive outsourcing actually costs and under what conditions it becomes something other than a rational adaptation to a genuinely useful tool. Three distinct problems are folded together in most discussions of AI and student work and separating them is a precondition for saying anything useful about policy. The first is a problem of truth and responsibility: an AI system can produce a plausible-sounding claim that is simply false and the question of who is answerable when a person submits that claim under their own name, whether the person is a student handing in an essay or a professional filing a report, is rarely addressed with any rigor. The second is a problem of comprehension: even when AI output is substantively correct, the person attaching their name to it may understand only a fraction of it well enough to explain, defend, or extend it without the tool present and building that understanding after the fact, a process this paper will call cognitive insourcing, is considerably harder than producing the same understanding from the outset would have been, particularly for someone who lacks the background knowledge to know what to check. The third is a problem specific to the classroom, though not confined to it: when students increasingly use AI to produce their assignments and instructors increasingly use AI to grade them, the pedagogical relationship that assessment is supposed to represent begins to hollow out from both ends simultaneously and it becomes genuinely unclear what is being taught, by whom and to whom.

The argument developed across the sections that follow treats these three problems as related but not identical and it treats the appropriate response to each as falling on a different set of institutional actors. Section one argues that the pedagogical task facing teachers is not to prevent AI use, which is neither feasible nor, used well, desirable, but to redirect what students use it for: away from producing finished answers and toward supporting the thinking that would otherwise have produced those answers unaided. Section two takes up the harder practical question of where usable boundaries between acceptable and unacceptable AI use can actually be drawn and argues that clarity generated at the most rigorous level of an educational system tends to propagate downward in ways that blanket rules issued at any single level cannot achieve alone. Section three follows from both of the preceding arguments to ask what teaching programs, evaluation design and institutional credentialing must change once the premise that AI cannot be excluded is accepted rather than resisted. A final section considers the aggregate consequence of continued inaction, not as rhetorical flourish but as a defensible extrapolation from the labor-market and cognitive-development evidence already on the table.

Two clarifications are worth making before proceeding. First, nothing in this argument treats AI-assisted cognitive offloading as inherently harmful. Human beings have offloaded cognitive work to external aids for as long as they have used calculators, written notes and reference books and much of that offloading measurably improves the work that follows, because it frees limited attention for the parts of a task that actually require judgment. The distinction that matters, developed at length in the sections below, is between offloading the mechanical periphery of a task, such as checking grammar or locating a citation and offloading the intrinsic

cognitive work the task exists to build, such as constructing an argument or diagnosing why a proof fails.[8] Second, the paper deliberately confines itself to what the evidence can support. Where a claim is based on a self-reported survey rather than an objectively measured outcome, or on an experimental setting that may not generalize to ordinary classroom conditions, that distinction is preserved rather than smoothed over, because the policy conclusions that follow depend on getting this distinction right.

2 Teach Students to Use AI to Think, Not to Answer Assignments

The pedagogical instinct to treat AI as a threat to be walled off from student work rests on an assumption that is worth stating plainly so that it can be examined: that a student's unaided output is a reliable proxy for what that student has learned. This assumption was already unreliable before generative AI existed, since a hardworking student could always seek help from a parent, a tutor, or a classmate and a finished essay cannot fully distinguish independent understanding from assisted production. Generative AI does not introduce this problem so much as make it dramatically more acute, because the assistance it offers is available instantly, privately and at a level of polish that a parent or classmate rarely matches. A homework prompt that once required a student to have read a short story, formed a view about its theme and expressed that view in writing can now be satisfied by a system that has not read the story in any sense that resembles comprehension but can nonetheless produce prose that is organized, accurate and fluent, leaving the instructor with a finished product that reveals almost nothing about whether the underlying cognitive work occurred.[9]

The research literature on this question increasingly converges on a specific, testable claim rather than a vague worry about laziness: the effect of AI assistance on learning depends on which part of a task's mental effort the AI is asked to absorb. Cognitive demand theory, the dominant framework in educational psychology for describing how working memory constraints shape learning, distinguishes between the difficulty inherent in the material a student is trying to master, which cannot be removed without weakening the learning itself and the difficulty imposed merely by how a task is presented or structured, which can be reduced without any loss. A 2026 report from the University of Technology Sydney's Network for Quality Digital Education, authored by the educational psychologist Jason Lodge and the technology-policy specialist Leslie Loble, formalizes this distinction for the AI era as the difference between beneficial and detrimental cognitive offloading. Beneficial offloading occurs when a student delegates the extraneous parts of a task, checking a citation format or catching an incorrect comma, freeing limited attention for the intrinsic work of constructing an argument or synthesizing evidence. Detrimental offloading, which the report treats as functionally equivalent to what this paper calls cognitive outsourcing, occurs when the student delegates the intrinsic work itself, asking the system to produce the argument rather than to polish one the student has already built.[10]

The empirical case for taking this distinction seriously rather than treating it as a plausible-sounding hypothesis comes from a growing body of controlled research. The most direct evidence concerns what the Sydney report terms the performance paradox: AI assistance can visibly improve a student's performance on the task in front of them while simultaneously degrading the durable learning the task was designed to produce.[11] In a large randomized field experiment involving nearly one thousand high-school mathematics students, published

in the Proceedings of the National Academy of Sciences in 2025, Hamsa Bastani and colleagues found that students given unrestricted access to a generative AI tutor performed well while that access was available, but underperformed students who had worked without AI once the tool was subsequently withdrawn, indicating that the assisted practice sessions had not built the durable skill the practice was meant to build.[12] The same study found that this harm was substantially mitigated, though not eliminated, when the AI tutor was constrained to provide structured hints rather than direct answers, which is direct experimental evidence that the design of the tool, not merely its presence, determines whether offloading is beneficial or detrimental.[13] A separate quasi-experimental study of two hundred and forty students, in which learners were explicitly taught to delegate only lower-order writing tasks such as brainstorming and grammar checking to AI while reserving analysis and evaluation for themselves, found significantly greater gains in critical thinking than were observed in unstructured comparison groups, which is the positive complement to the Bastani finding: offloading managed by explicit instruction produces different outcomes than offloading left to default use.[14]

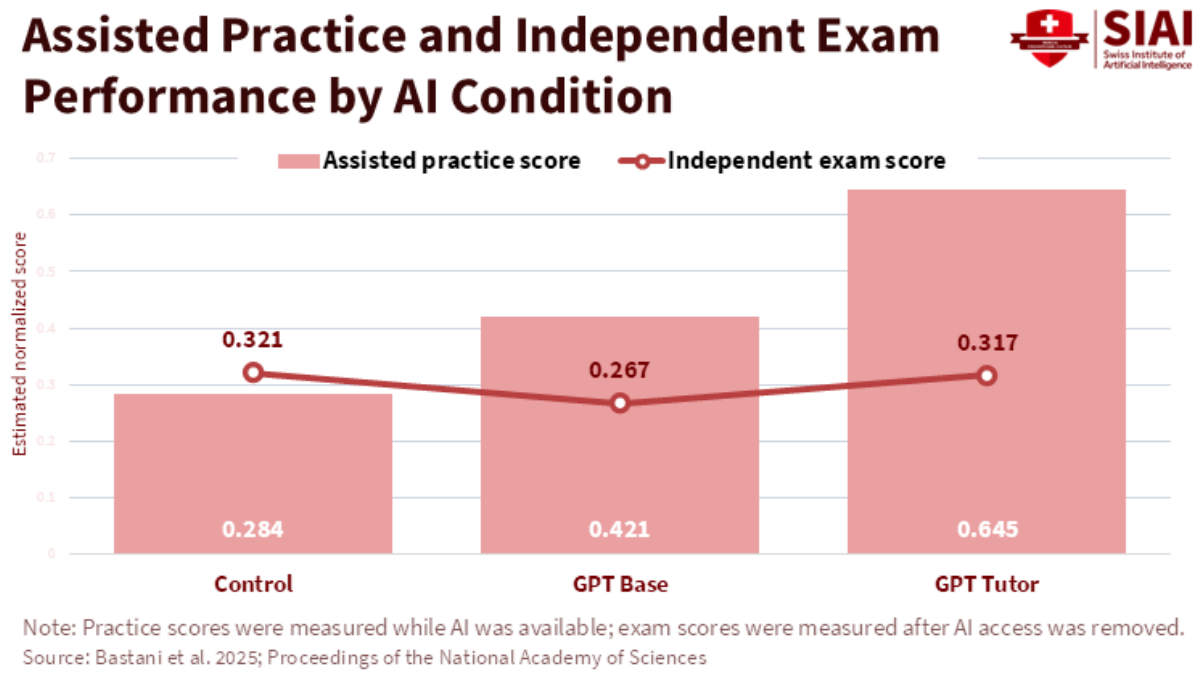

Figure 1: AI assistance substantially raised practice performance, while independent exam performance remained close to or below the control level.

This pattern helps resolve what would otherwise look like a contradiction in the wider research base. Several systematic reviews and meta-analyses published since 2024 report that generative AI has a positive average effect on learning outcomes. This finding circulates widely in more optimistic commentary on AI in education. The Sydney report's reading of this literature, which this paper finds persuasive, is that most of the underlying studies measure short-term, scaffolded performance on a test or assignment completed with the tool still present, rather than durable, independent capability measured after the tool is withdrawn.[15] These are not competing findings about the same thing; they are findings about two different things, performance and learning, that happen to diverge under AI assistance in ways they did not diverge as sharply under earlier forms of technological support. A test score achieved with AI assistance still present says comparatively little about what a student can do

without it and conflating the two measures produces exactly the kind of unearned optimism that a policy response built on accurate evidence needs to avoid.

The mechanism behind this difference has a name in the cognitive-science literature that predates generative AI by decades: desirable difficulties, the finding that durable long-term learning generally requires a degree of effortful struggle that easier, more fluent modes of instruction bypass. A useful non-AI baseline for this mechanism comes from vocabulary research on what is called the generation effect, in which students who must actively generate a word from a cue show significantly better long-term retention than students who are shown the correct word. Generative AI, when used as what the Sydney report calls an answer oracle, is the generation effect's mirror image: it supplies the answer directly and in doing so removes the very retrieval effort that would have built lasting memory of it. Compounding this, AI-generated text carries a further risk documented under the label of fluency: because a model's output is coherent and confidently phrased regardless of whether it is correct, its ease of consumption creates what researchers call an illusion of competence, in which a student mistakes the ease of understanding a polished explanation for having mastered the underlying material. A 2024 study identified a related pattern the authors term metacognitive laziness, in which the convenience of AI assistance measurably reduces students' engagement in the planning, monitoring and revision processes that self-regulated learning depends on, effectively transferring not just the cognitive work of the task but the metacognitive work of managing one's own learning to the tool.[16]

None of this amounts to an argument that AI degrades learning as a matter of technological necessity. The Sydney report's central methodological point, echoed by a 2025 critique published in the Journal of Computer Assisted Learning describing much of the existing research as an effect in search of a cause, is that outcomes are driven by pedagogical design rather than by the medium itself. The practical implication for the assigned angle of this section, that teachers should redirect AI use toward thinking rather than answering, is therefore not a slogan but a specific, evidence-backed instructional strategy with at least three converging strands of support. The first is what the Bastani experiment demonstrates directly: constraining the tool to offer hints and prompts rather than finished solutions preserves most of the assistance value while substantially reducing the learning cost.[17] The second is metacognitive scaffolding, in which structured prompts built into the AI interaction explicitly require students to pause, predict, or justify before proceeding, a design feature that multiple 2025 studies found measurably improved self-regulated learning and depth of inquiry compared with unstructured use of the same underlying model. The third, more speculative though conceptually elegant, reframes the AI's role entirely: instead of positioning it as an answer oracle, some researchers propose engineering it as a deliberately imperfect novice that asks clarifying questions and feigns confusion, forcing the student into the effortful, generative act of explaining a concept aloud, a mechanism long known within educational psychology as the protégé effect.[18]

A separate and more empirically direct source of evidence for the same conclusion comes from watching what students say they actually do when given explicit permission to use AI, rather than inferring their behavior from test scores alone. A qualitative study of thirty-nine undergraduates at a large American research university, conducted within weeks of ChatGPT's public release and published in Scientific Reports in 2025, found that students' documented uses of AI clustered into recognizably different categories with very different

implications for learning. Lower-order uses, chiefly revising and editing already-written text, were the most common, appearing in forty-seven coded instances and students describing this use typically emphasized that the tool sharpened wording without altering their underlying ideas. A smaller but still substantial group of students, however, described what the researchers coded as ghostwriting, in twenty-four instances, where the AI supplied entire paragraphs or arguments that the student had not independently constructed and the degree of reliance within this group varied considerably, from students using AI to overcome momentary writer's block to students who appeared to have delegated most of the substantive thinking. Notably, a meaningful share of students expressed explicit skepticism about the tool's reliability even while using it, reporting that they had caught it inventing citations that did not exist or producing dated information and describing their own practice of verifying AI-generated claims against independent research before incorporating them.[19] This last finding is important because it demonstrates that the distinction between cognitive outsourcing and legitimate assistance is not simply a theoretical construct imposed by researchers from outside; it is a distinction that some students are already drawing for themselves, unprompted, which suggests that explicit instruction in this distinction is building on an existing capacity rather than introducing an entirely foreign discipline.

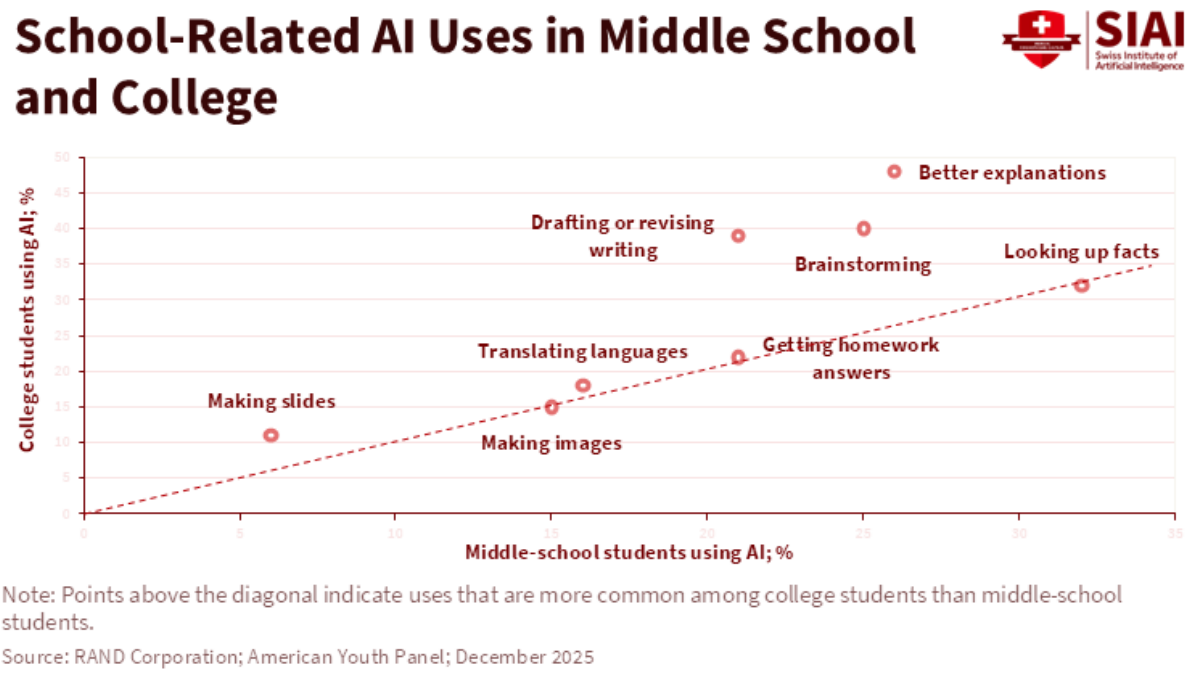

Figure 2: College students reported greater AI use for explanations, brainstorming and writing, while fact-finding and homework-answer use were more similar across the two groups.

A parallel national survey conducted by the University of Southern California's Center for Generative AI and Society reached a structurally similar conclusion using different terminology. Surveying one thousand American college students, the USC researchers distinguished between what they called executive help, defined as seeking a fast solution with minimal effort and instrumental help, defined as using AI to clarify a concept, build a skill, or support independent learning and found that most students defaulted to executive use. The critical finding, however, was not that this default is fixed. Students who reported receiving explicit encouragement from an instructor to use AI thoughtfully were significantly more likely to engage in learning-oriented, instrumental

use rather than executive shortcutting, which is direct survey evidence that faculty guidance, not student disposition alone, shapes which mode of use predominates. As the lead researcher on that project put it in comments accompanying the report, what matters most is ensuring AI use is guided by people with genuine content expertise, rather than leaving students to work out the boundaries on their own. This finding recasts the entire pedagogical question addressed in this section: the goal is not to persuade students to stop wanting fast answers, which is an unrealistic aspiration given ordinary human incentives under academic pressure, but to design assignments and provide guidance that make the instrumental mode of engagement the practical path of least resistance rather than a demanding alternative to the convenient default.[20]

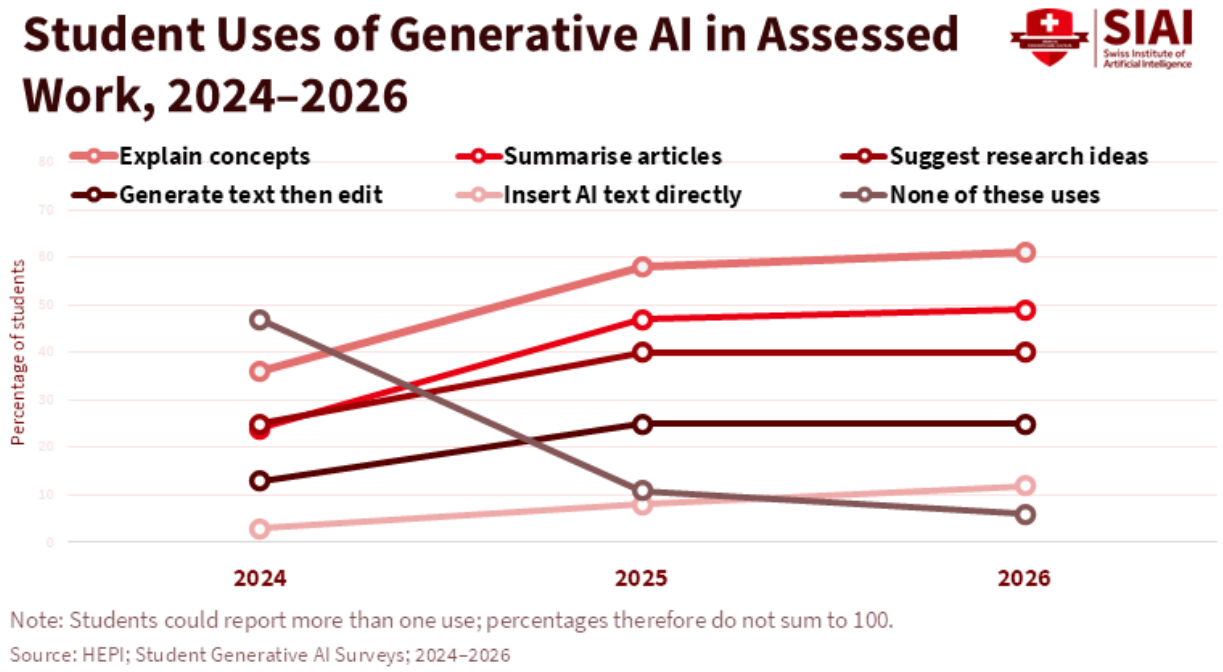

Figure 3: Reported uses expanded rapidly across the period, while the share of students reporting none of these uses fell sharply.

3 Draw a Clear Boundary Between Acceptable AI Assistance and Cognitive Outsourcing

If the pedagogical case in the preceding section establishes that AI use can be redirected rather than merely restricted, it does not by itself answer a harder operational question: how a teacher, a department, or an institution decides, assignment by assignment, where legitimate assistance ends and cognitive outsourcing begins. This question has proven considerably more difficult to resolve in practice than in principle and the difficulty is visible in the data on institutional policy itself. In the same national survey that found ninety percent of college students using AI in class, only a minority of instructors had received any meaningful preparation for making these judgments: just eleven percent reported comprehensive AI training, while forty-one percent reported none at all. A separate 2026 survey of educators in Wisconsin and nationally found that only around a third of respondents said their school or district had a formal AI policy of any kind, meaning that most of the boundary-drawing described in this section is currently happening, if it is happening at all, at the level of

individual teacher discretion rather than institutional design.[21]

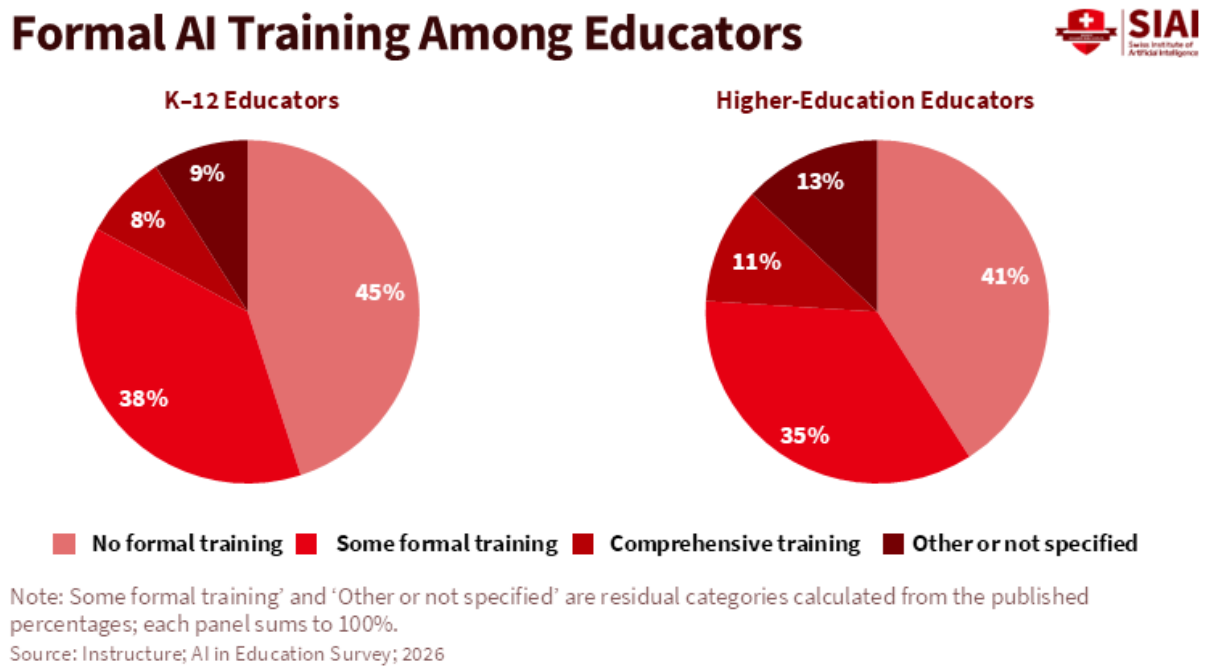

Figure 4: Most educators reported either no formal AI training or only limited preparation, while comprehensive training remained uncommon.

The instinct to resolve this uncertainty through detection software has, on the available evidence, largely failed and understanding why matters for what should replace it. A national survey of sixth- through twelfth-grade teachers found that forty-three percent used AI-detection tools regularly, with a further twenty-seven percent having experimented with them. Still, the underlying accuracy of these tools does not support the confidence with which they are often deployed.[22] One peer-reviewed evaluation of fourteen separate AI-detection systems found false-positive rates as high as fifty percent and false-negative rates as high as one hundred percent depending on the tool, with roughly a fifth of AI-generated text misclassified as human-written even before any deliberate evasion. This figure rose to roughly half once the AI text had been lightly edited and to over seventy percent once it had been paraphrased by a second AI system. A separate study found that detectors falsely flagged non-native English writing as AI-generated at an average rate above sixty percent, a bias with obvious and serious equity implications for any institution serving international or multilingual students. These are not marginal failure rates that better engineering will straightforwardly fix; they reflect a structural problem in trying to statistically distinguish fluent human writing from fluent machine writing once both are genuinely fluent and the practical consequence has been a wave of institutions quietly turning off detection tools they had previously adopted, even as detection vendors continue to market accuracy claims that go largely unverified by any independent standard.[23]

If detection cannot accurately answer where the line sits after the fact, the more promising approach documented in the literature is to specify the line in advance, assignment by assignment, rather than leaving it to be inferred from a single blanket policy. The most developed framework for doing this is the Artificial Intelligence Assessment Scale, first published in 2024 by Mike Perkins, Leon Furze, Jasper Roe and Jason MacVaugh and

subsequently revised in response to feedback from educators worldwide. The framework's central insight, consistent with the cognitive-load distinction developed in the previous section, is that a single institutional rule cannot sensibly govern every assignment, because different assignments are designed to measure different things. An assignment intended to assess a student's independent writing fluency reasonably permits no AI use at all; an assignment intended to assess conceptual synthesis might permit AI for brainstorming provided the student submits original notes and a reflective account of what the tool contributed; and an assignment intended to build critical evaluation skills might deliberately ask students to critique an AI-generated answer and identify what in it is accurate, incomplete, or misleading. By 2026 the scale had been adopted by several hundred institutions and translated into more than thirty languages and it has been formally recognized by Australia's national tertiary quality regulator as an acceptable mechanism for communicating AI expectations to students, which suggests the framework has moved from an academic proposal to genuine operational infrastructure rather than remaining an unused thought experiment.[24]

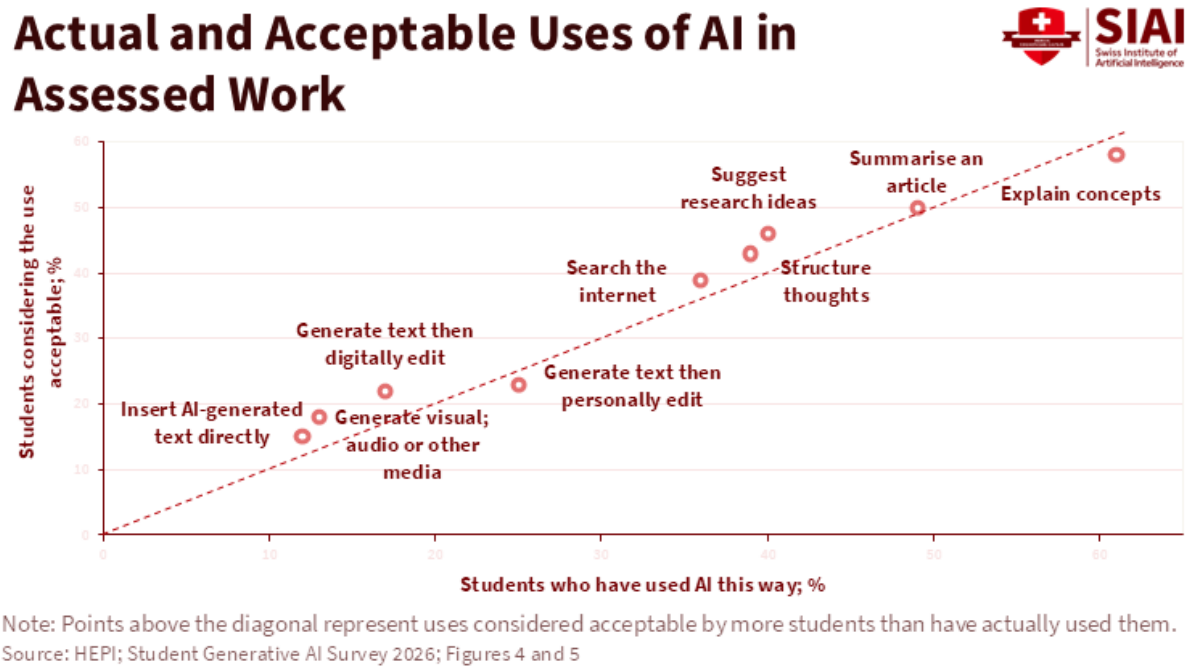

Figure 5: Students generally considered explanatory and organizational uses more acceptable than the direct generation or insertion of assessed content.

The evidence on where this boundary is most consequential in practice, however, complicates any assumption that clarity alone will resolve the basic tension, because the temptation to cross an acceptable-use line appears to scale with frequency of use rather than with any fixed individual disposition toward dishonesty. The largest study of its kind, led by Igor Chirikov of the University of California, Berkeley, in collaboration with researchers at the University of Technology Sydney and Cornell University, surveyed more than ninety-five thousand undergraduates across twenty research-intensive public universities and was published in Science in May 2026. The study found that roughly two-thirds of respondents used generative AI at all, with close to forty percent using it monthly or more often and that at least nine percent of AI-using students admitted to using it in ways they understood to constitute cheating, a figure the authors describe as conservative given the indirect survey

method used to encourage honest self-reporting on a sensitive behavior.[25] The more striking finding concerned the relationship between frequency and misuse: twenty-six percent of daily AI users reported using the tool to cheat, compared with only seven percent of monthly users, a gradient the study's authors describe as a slippery slope rather than a fixed line, in which habitual use erodes the self-regulatory friction that would otherwise prompt a student to ask whether a given use is permitted. Reporting on the study, Chirikov noted that AI policies vary enormously not only between institutions but between individual courses at the same institution, from faculty who permit AI throughout an exam to faculty who ban it outright and that this inconsistency makes self-regulation genuinely difficult even for students who intend to comply, because ordinary tools such as search engines and grammar checkers already embed AI features that blur the line the student is trying to observe.[26]

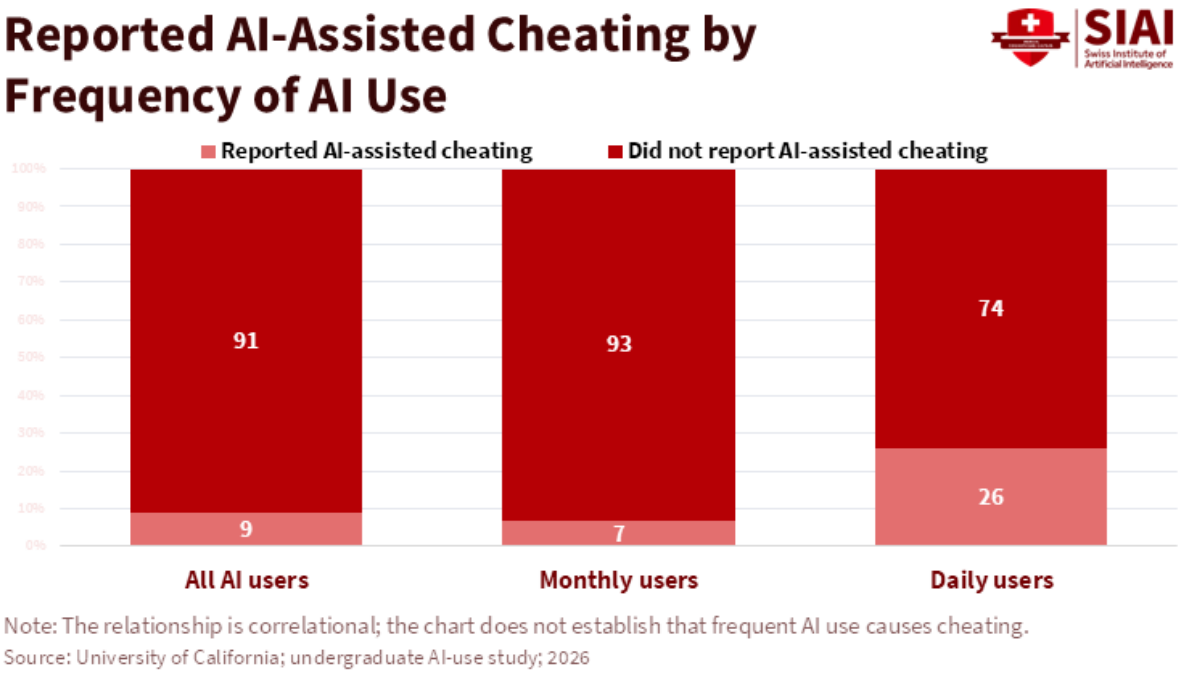

Figure 6: Reported cheating was considerably more common among daily AI users, although the association does not establish that frequent use causes misconduct.

This same study surfaces a second finding directly relevant to where institutional boundary-drawing should be targeted: cheating with AI was not evenly distributed across academic disciplines, with non-STEM students reporting higher rates of AI-assisted cheating than STEM students, which argues against a single university-wide rule and toward the discipline-specific and even course-specific calibration that the assessment-scale literature already recommends. The study also documented a disparity that recurs across nearly every dataset examined in this paper: students from lower-income backgrounds, from underrepresented racial groups and female students were all significantly less likely to use generative AI tools at all, a gap the researchers describe as more concerning than the cheating finding itself, because it implies that unequal access to capable AI tools, driven by the cost of premium subscriptions and the usage limits attached to free tiers, may translate directly into unequal preparation for a labor market that increasingly rewards AI fluency.[27] The Sydney cognitive-offloading report reaches a structurally identical conclusion from a different empirical base, describing what it terms a new metacognitive

equity gap: students who already possess strong domain knowledge and self-regulation skills are positioned to use AI for beneficial offloading and accelerate their learning, while students who lack these foundations are disproportionately exposed to detrimental offloading and the illusion of competence it produces, a dynamic the report likens to the Matthew effect long documented in reading research, in which early advantages compound rather than equalize over time.[28] Any boundary-drawing exercise that treats access to AI as uniform across a student population will therefore systematically misjudge both the benefit and the risk of the tool for different groups within that same classroom.

A further complication, evident from informal faculty discussion as much as from published research, is that professors themselves frequently disagree about where legitimate assistance ends, particularly at the undergraduate level where the line between AI as a research aid and AI as a ghostwriter is genuinely contested even among experienced instructors. This disagreement is not evidence that the boundary cannot be drawn; it is evidence that the boundary cannot be drawn once, generically and then left alone. The most defensible response, consistent with the framework described above, treats the standard set at the most demanding and best-resourced level of the education system, typically graduate and professional programs where the stakes of unverified competence are highest, as a reference point that can meaningfully inform standards further down the system, since students who understand early that unverified AI reliance will not survive scrutiny at the college level have a concrete reason to build the underlying capability in high school rather than deferring the problem. This is not a claim that a single national standard should be imposed top-down; it is a claim that visible, well-justified rigor at the upper end of a system changes the incentives faced by students and teachers throughout the system beneath it, in much the same way that university admissions standards have long shaped instruction in the secondary schools that feed them.

4 Given the New Boundary, Teaching Programs Must Evolve

The preceding two sections establish that AI use should be redirected toward supporting thought rather than replacing it and that the boundary between the two must be set with more granularity than a single institutional rule can provide. Neither conclusion is actionable, however, unless the instructional materials and assessment instruments through which teaching actually happens are redesigned to match. The historical precedent most frequently invoked for this kind of transition, the introduction of the pocket calculator into mathematics education in the 1970s and 1980s, is instructive, but a careful reading of that precedent, developed at length by the Brookings researcher Niam Yaraghi, suggests that generative AI poses a harder version of the same problem rather than a repetition of it and that the policy response therefore needs to be more conservative, not less, than the one mathematics education eventually settled on.

The calculator precedent worked because arithmetic instruction adopted a strict sequencing rule: students mastered manual computation before calculators were introduced and calculators were then treated as instruments of speed rather than substitutes for the underlying competence, which meant a student retained an internal sense of whether a calculator's output was plausible.[29] Generative AI breaks this precedent in three specific ways that matter for how teaching programs must be redesigned. First, a calculator is reliably correct

within its domain, whereas an AI system can produce a confidently wrong answer that only the user's own domain knowledge will catch, meaning a student without that knowledge is worse off than a calculator user, not simply no better off. Second, a calculator executes an operation the student has already decided to perform, whereas an AI system frequently chooses the framing of the problem itself, silently performing the part of the cognitive work, deciding what question is actually being asked, that domain expertise is specifically built by practicing. Third, the convenience differential is enormous by comparison: a calculator saves a student seconds of arithmetic, while a language model can produce an entire essay, memo, or research summary, so the incentive to skip the underlying developmental work scales with the size of the shortcut being offered.[30] These three differences argue for sequencing that is, if anything, stricter than the calculator model achieved, with a correspondingly higher bar for permitting unrestricted use than arithmetic instruction ever required.

Translating this into concrete curricular change requires distinguishing what is properly the responsibility of individual teachers, what belongs to departments and institutions and what requires action from professional and accreditation bodies, because collapsing these levels together produces exactly the vague calls for adaptation and literacy that this paper's brief explicitly warns against. At the level of individual course design, the clearest available example of assessment redesigned specifically to resist naive AI use comes from graduate programs built around applied case analysis rather than recall. One illustrative practice, adopted at a graduate business-and-analytics program built around applied AI coursework, structures final examinations around the requirement that students apply a specific theoretical framework drawn from an assigned text to a scenario the framework was not written to describe, a design that rewards demonstrated command of the underlying theory and penalizes the kind of generic, pattern-matched response that an AI system produces when it has not been given the specific analytical scaffolding the course has spent a semester building.

The general design principle this illustrates, requiring students to apply, rather than merely recall or summarize, a body of theory to a genuinely novel case, is consistent with the wider research finding that critical thinking is not a generic, transferable skill but is instead deeply dependent on the specific domain knowledge a student has actually internalized, which is precisely the knowledge that unrestricted AI access allows a student to bypass building in the first place.[31] The STA501 final examination offers a concrete illustration. The ChatGPT-generated response identifies familiar terms such as measurement error, instrumental variables and endogeneity, whereas the solution key explains the direction of bias, defends the proposed instrument and carries the same theoretical assumptions across linked questions. The comparison shows whether the student can reconstruct and defend the reasoning rather than merely reproduce technically plausible terminology.[32]

At the level of instructor practice more broadly, the evidence reviewed in the first section of this paper points toward assessment that privileges process visibility over final-product polish. Instructors surveyed by RAND and by the Wisconsin study described earlier reported adopting oral components, in-class handwritten work and staged submission of drafts specifically to preserve some direct evidence of independent thinking, changes that do not require new technology so much as a willingness to accept that the finished, polished essay can no longer act as reliable proof of learning on its own. The RAND report frames this shift explicitly in terms of the offloading distinction developed earlier: school and district leaders are advised to identify, assignment by assignment, whether a given use case is likely to produce cognitive offloading, defined there as AI doing the

mental work for the student, or cognitive augmentation, defined as AI prompting the student toward deeper, more independent work and to design instruction around that distinction rather than around a single yes-or-no rule for AI in the classroom overall. One concrete instructional model recommended for this purpose is the flipped classroom, in which students first encounter new material independently, whether or not AI is involved in that first encounter and then complete guided practice during AI-free class time under direct teacher supervision, preserving at least one setting in which independent capability can be directly observed.[33]

At the level of institutional and professional credentialing, the argument developed by Yaraghi is the most fully worked out in the current literature and deserves extended treatment because it identifies a policy lever that individual teachers cannot pull on their own. His proposal is that any credential whose value depends on the holder's underlying capacity - a university degree, a medical license, a bar admission, a professional accounting qualification - should require the credentialing body to certify that the holder has demonstrably developed that capacity without AI assistance at some defined stage of their training, even if AI was used freely elsewhere in the same program.[34] The reasoning behind this proposal rests on empirical findings about where AI genuinely helps and where it does not. In a large field experiment involving over five thousand customer-service agents at a Fortune 500 firm, published in the Quarterly Journal of Economics, access to a generative AI assistant raised average productivity by roughly fourteen to fifteen percent, but this gain was concentrated overwhelmingly among the least experienced workers, whose output improved by around thirty-four percent, while the most experienced agents saw only marginal gains and in some measures a slight decline in quality.[35] A separate, preregistered field experiment conducted with the Boston Consulting Group, involving seven hundred and fifty-eight professional consultants, found a structurally similar pattern that the researchers termed a jagged technological frontier: for tasks that fell within the AI's genuine capability, consultants using it were both faster and produced measurably higher-quality work, but for a task deliberately designed to fall outside that capability, performance among AI-assisted consultants fell by roughly nineteen percentage points relative to the unassisted group, because the consultants, lacking the domain expertise to recognize when the tool had failed, followed its confident but wrong output anyway.[36] The uncomfortable implication of both studies together is that the very expertise required to use AI safely on the hardest, most consequential tasks is the expertise that unrestricted, unstructured AI use during training makes it harder to build in the first place, which is precisely the developmental externality that Yaraghi's credentialing proposal is designed to address.

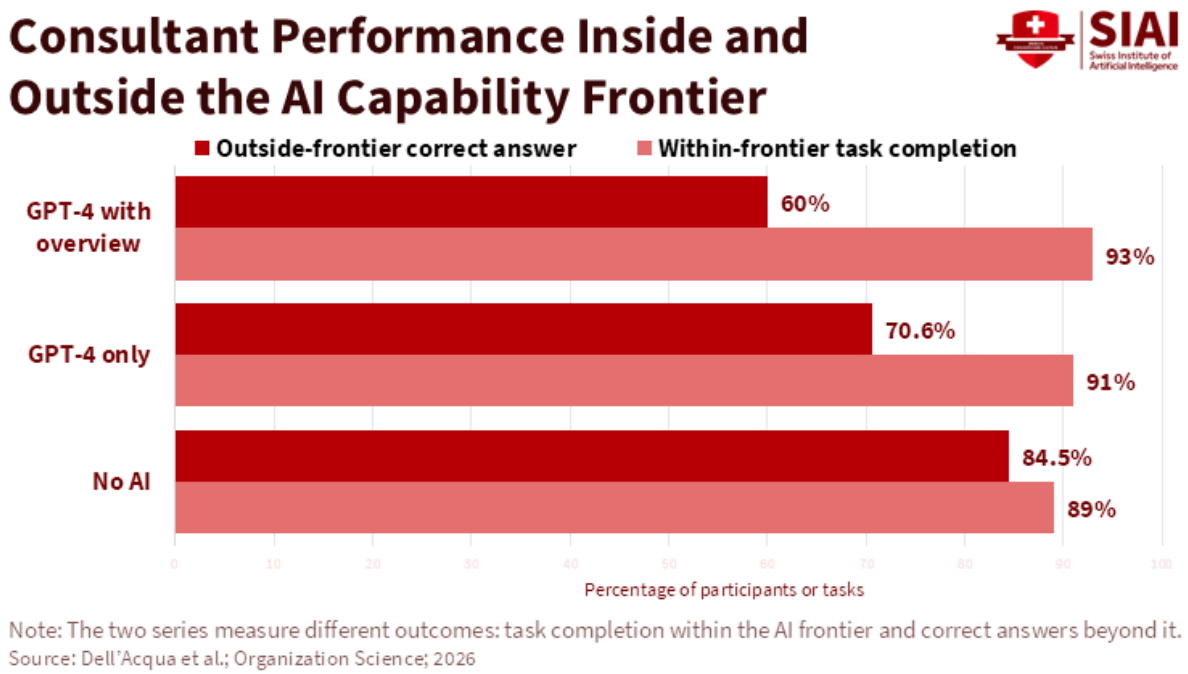

Figure 7: AI improved task completion within its capability frontier but reduced correct performance when users encountered tasks beyond that frontier.

This same body of evidence points toward a labor-market pattern that gives the credentialing argument added urgency rather than remaining a purely academic concern. Using payroll records covering millions of American workers, researchers at Stanford's Digital Economy Lab found a roughly sixteen percent relative decline in employment for workers aged twenty-two to twenty-five specifically in occupations most exposed to generative AI since the technology's general adoption, while employment among older workers in the same occupations remained essentially stable.[37] A separate analysis of resume and job-posting data covering sixty-five million workers across two hundred and eighty thousand firms found that companies actively adopting generative AI sharply reduced hiring at the junior level relative to firms that had not adopted it, a pattern the researchers describe as seniority-biased technological change, driven not by layoffs of existing juniors but by firms simply declining to hire new ones, since AI now performs much of the routine drafting and information-retrieval work that entry-level positions traditionally existed to teach.[38] If firms are hiring fewer juniors precisely because AI has absorbed the developmental work those roles used to provide, then educational institutions cannot treat their own AI-adoption decisions as separable from this labor-market shift; a graduate who has never practiced the underlying skill unaided arrives at a labor market that is simultaneously offering fewer structured opportunities to practice it on the job, compounding rather than offsetting the deficit built during schooling.

A final piece of evidence, concerning research and creative output rather than routine task performance, bears directly on how teaching programs should think about the cumulative effect of encouraging AI-assisted work across an entire cohort rather than in any single classroom. In a controlled experiment published in Science Advances, Anil Doshi and Oliver Hauser found that access to generative AI ideas measurably improved the individually assessed creativity of short stories, with the largest gains concentrated among the least inherently creative writers, effectively narrowing the gap between weaker and stronger writers on an individual basis. At

the same time, the stories produced with AI assistance were measurably more similar to one another than stories produced without it, meaning the same intervention that improved individual output simultaneously reduced the diversity of ideas across the group as a whole. This finding matters for curriculum design because it demonstrates that an assessment regime optimized purely to raise the average quality of individual student work, a reasonable-sounding goal in isolation, can coexist with a systematic narrowing of the range of thinking a cohort produces collectively, which is precisely the kind of second-order, aggregate effect that assessment focused only on individual grades will never detect and that a program-level evaluation, not exclusively a course-level one, is required to notice.[39]

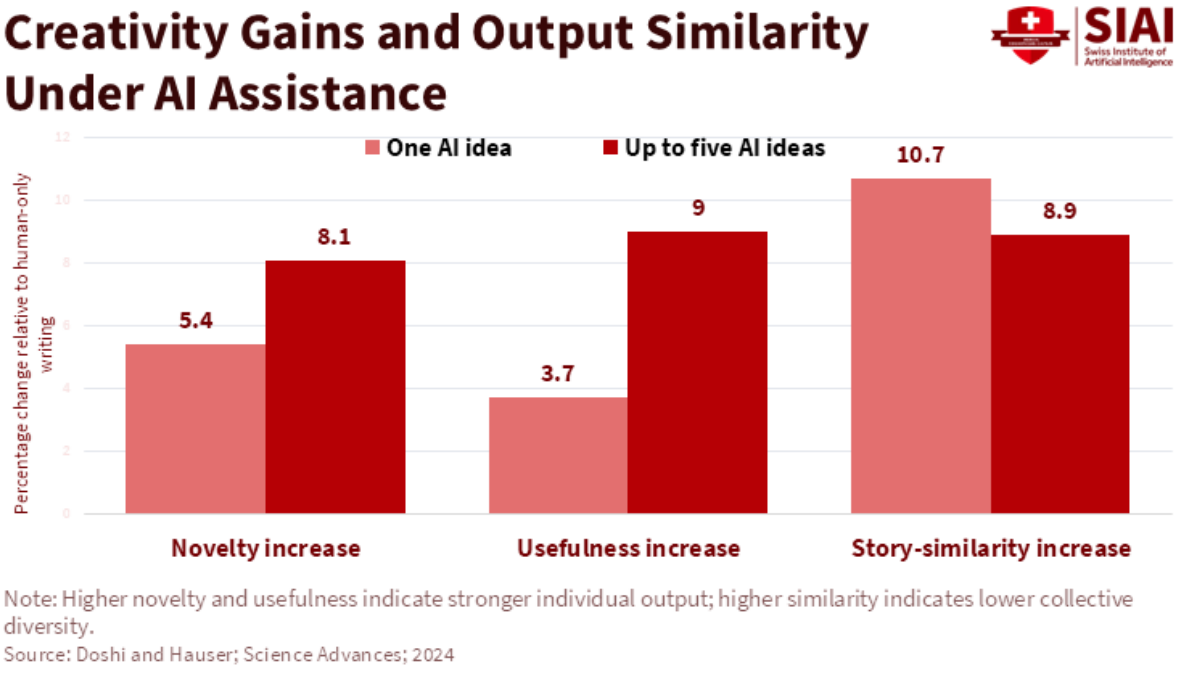

Figure 8: AI-generated ideas improved individual novelty and usefulness while simultaneously increasing similarity across the resulting stories.

Taken together, these three levels of change, redesigned individual assessments, instructor practices that preserve visible evidence of independent thinking and institutional credentialing that certifies unaided capability at defined checkpoints, describe a coherent response to the argument developed across this paper, but they share a common precondition that deserves explicit statement: none of them can be implemented by teachers who do not themselves understand how these systems work, where they fail and what structured versus unstructured use actually looks like in practice. Investment in that specific form of educator preparation, distinct from generic technology training, is likely the single highest-leverage policy lever available to the institutions this section has been discussing and it is also, on the current evidence, the one most consistently underfunded relative to its importance.[40]

5 Conclusion - What Happens If Education Does Not Act Now?

The evidence assembled here points toward a specific and avoidable failure mode rather than a vague technological anxiety. If the redirection of AI use described in the first section, the calibrated boundary-drawing

described in the second and the credentialing and curricular reform described in the third are not undertaken, the population entering the workforce over the coming decade will disproportionately consist of graduates whose credentials certify capability that was never independently built, because the developmental work those credentials are supposed to represent was quietly outsourced during the years it should have been practiced. This is not primarily a story about individual laziness; it is a story about an education system and a labor market that, left to their current incentives, jointly remove the conditions under which expertise has always been formed, replacing effortful practice with fluent, confident output that looks like competence without reliably being competence. Firms are already hiring fewer junior workers precisely because AI performs the routine work that once trained them, narrowing the pipeline that would produce the next generation of senior judgment at the same moment schools are declining to certify that unaided judgment was ever built in the first place. The resulting productivity gains are real but borrowed, sustained by an expert generation whose capability predates these tools and unlikely to persist once that generation retires without deliberate institutional effort to replace what it is quietly spending down. The appropriate response is neither to prohibit AI, which the adoption data show is no longer a live option, nor to defer to it uncritically, but to insist, through assessment design, credentialing standards and teacher preparation, that the distinction between having used a tool and having built the capacity it assists remains visible, checkable and consequential, before the institutions responsible for drawing that distinction lose the capacity to draw it at all.

References

[1, 2, 21] Spitalniak, L. (2026) '90% of students use AI in the classroom, Instructure poll finds', *Higher Ed Dive*, 21 July.

[3, 4] Freeman, J. (2025) *Student Generative AI Survey 2025.* HEPI Policy Note 61. Oxford: Higher Education Policy Institute, in partnership with Kortext.

[5] College Board (2025) *New Research: Majority of High School Students Use Generative AI for Schoolwork.* New York: College Board Newsroom.

[6, 7, 33] Schwartz, H.L. and Diliberti, M.K. (2026) *More Students Use AI for Homework, and More Believe It Harms Critical Thinking: Selected Findings from the American Youth Panel.* RR-A4742-1. Santa Monica, CA: RAND Corporation.

[8, 10, 11, 15, 16, 18, 28] Lodge, J.M. and Loble, L. (2026) *Artificial Intelligence, Cognitive Offloading and Implications for Education.* Sydney: University of Technology Sydney, Network for Quality Digital Education.

[9, 21, 22, 23, 33] DeJager, B. (2026) '84% of students use AI for homework. Only 3 in 10 schools have rules for it', *Fortune*, 16 July. Republished from The Conversation.

[12, 13, 17] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö. and Mariman, R. (2025) 'Generative AI without guardrails can harm learning: Evidence from high school mathematics', *Proceedings of the*

National Academy of Sciences, 122(26), e2422633122.

[14] Hong, H., Vate-U-Lan, P. and Viriyavejakul, C. (2025) 'Cognitive offload instruction with generative AI: A quasi-experimental study on critical thinking gains in English writing', *Forum for Linguistic Studies*, 7(7), pp. 325–334.

[16] Duplice, J. (2025) 'Generation and L2 vocabulary learning', *Vocabulary Learning and Instruction*, 14(1), 102482.

[16] Fan, Y. et al. (2024) 'Beware of metacognitive laziness', *British Journal of Educational Technology*, 56(2), pp. 489–530.

[17] Weidlich, J., Gašević, D., Drachsler, H. and Kirschner, P. (2025) 'ChatGPT in education: An effect in search of a cause', *Journal of Computer Assisted Learning*, 41(4), e70105.

[19] Black, R.W. and Tomlinson, B. (2025) 'University students describe how they adopt AI for writing and research in a general education course', *Scientific Reports*, 15, 8799.

[20] McQuiston, P. (2025) 'AI is changing how students learn, or avoid learning', *USC Today*, 18 September.

[24] Perkins, M., Furze, L., Roe, J. and MacVaugh, J. (2024) 'The Artificial Intelligence Assessment Scale: A framework for ethical integration of generative AI in educational assessment', *Journal of University Teaching and Learning Practice*, 21(6), pp. 49–66.

[25, 26, 27] Kapoor, M.L. (2026) 'The largest study of AI use by undergrads is in, revealing disparities in access, and in cheating', *University of California News*, 28 May.

[29, 30, 34, 40] Yaraghi, N. (2026a) 'Repaying the inheritance: How education and research policy can address AI's borrowed expertise', *Brookings Institution*, 20 July.

[31] Willingham, D.T. (2019) 'How to teach critical thinking', *NSW Department of Education.*

[31] Tricot, A. and Sweller, J. (2014) 'Domain-specific knowledge and why teaching generic skills does not work', *Educational Psychology Review*, 26(2), pp. 265–283.

[31, 32] Swiss Institute of Artificial Intelligence (2021) *STA501: Data-Based Decision Making: Final Examination 2021.* SIAI institutional teaching material.

[31, 32] Swiss Institute of Artificial Intelligence (n.d.) *STA501 Final 2021 Solution Key for External Audiences' Reference.* SIAI institutional teaching material.

[35] Brynjolfsson, E., Li, D. and Raymond, L.R. (2025) 'Generative AI at work', *The Quarterly Journal of Economics*, 140(2), pp. 889–942.

[36] Dell'Acqua, F., McFowland III, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K.C., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K.R. (2023) *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality.*

Harvard Business School Working Paper No. 24-013.

[37] Brynjolfsson, E., Chandar, B. and Chen, R. (2025) 'Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence', *Stanford Digital Economy Lab Working Paper*, August, revised November.

[38] Hosseini Maasoum, S.M. and Lichtinger, G. (2025) 'Generative AI as seniority-biased technological change: Evidence from U.S. résumé and job-posting data', *SSRN Working Paper*, August.

[39] Doshi, A.R. and Hauser, O.P. (2024) 'Generative AI enhances individual creativity but reduces the collective diversity of novel content', *Science Advances*, 10(28), eadn5290.