Learning Intelligence

How we teach when the answer is free

A field guide for educators

May 2026

Contents

Learning Intelligence

How we teach when the answer is free

A field guide to assessment, pedagogy, and evidence in the post-output classroom.

In the fall of 2022, a freshman writing instructor at a mid-sized state university could still reasonably believe that the essay sitting in her grading queue was a record of her student's thinking. By the spring of 2026, she cannot. Somewhere between those two semesters, the basic contract of formal education quietly broke.

The break has a date. On November 30, 2022, OpenAI released ChatGPT to the public. Within five days it had a million users. Within two months, 100 million. By the start of 2026, 900 million people were using it weekly, and the question facing every teacher in every classroom on every continent had inverted. It was no longer, *can the student produce this work?* It was, *if the student can produce this work in thirty seconds with a chatbot, what was the work for?*

Three and a half years in, the data is overwhelming. The 2026 Higher Education Policy Institute student survey found that 92% of UK undergraduates now use generative AI for their studies and 88% use it specifically on assessed work, up from 53% the year before. The 2024 Digital Education Council global survey of 3,839 students across 16 countries found that 86% were already using AI regularly, with one in four using it daily. On the faculty side, the Elon University and AAC&U survey of 1,057 college professors released in January 2026 found that 95% expect AI to increase student overreliance, 90% expect it to diminish critical thinking, and 83% expect it to shorten attention spans. The Tyton Partners *Time for Class 2025* report found that 38% of faculty say AI has *increased* their workload, mostly through monitoring for cheating and rebuilding their assessments from scratch, while only 11% say it has decreased theirs.

The instructor with the unreadable essay is not alone. Her problem is now the central operational problem of the entire sector.

What this article is about is the field that is emerging in response to that problem. It does not yet have a settled name, but the most accurate one available is **learning intelligence**: the practice of generating and interpreting trustworthy evidence of how learning is happening, not just what was finally submitted, so that teachers can teach, students can learn, and institutions can certify that something real took place between them.

Learning intelligence is not the same as learning analytics. It is not the same as AI tutoring. It is not the same as plagiarism detection. It borrows from all three, but it points somewhere they do not. It points at the place where

assessment was always pointing, before the post-war research university convinced itself that a stack of essays was a sufficient record of a mind at work: at the *process* of learning itself.

This piece is an attempt to map the field. It walks through what we knew about learning before generative AI made the question urgent again, what the last four years actually did to the classroom, why the assessment crisis is a validity crisis and not a cheating crisis, what AI can and cannot do as a tutor and a colleague, and what a credible model of learning intelligence looks like for the institutions that have to live inside it. It is written for the people who already know that something has to change: K-12 superintendents and deans, provosts and chief academic officers, instructional designers, and the teachers and faculty who are doing the actual work of teaching while the ground shifts under them.

The thesis is simple, and the rest of the article unpacks it. When output becomes cheap, evidence becomes everything. The schools that thrive in the next decade will be the ones that learn how to *see* learning again.

Part I. What we knew before the machines could write

The strange thing about the AI crisis in education is that almost everything we need to solve it is already in the research literature. The science of how people learn has not changed because chatbots got good at writing essays. What has changed is the price of pretending it doesn't matter.

Begin with three findings that nearly every cognitive scientist would put on the same short list.

The first is that **learning requires effortful processing**. Robert and Elizabeth Bjork's work on "desirable difficulties" — the now-classic body of research showing that easier conditions during practice often produce *worse* long-term learning than harder ones — established a principle that has been replicated across decades of memory and cognition studies. Spacing prac-

tice, mixing problem types, generating answers before being told them, and being forced to retrieve information from memory all feel less productive in the moment and produce more durable learning in the long run. Ease is the enemy of encoding.

The second is that **active learning beats passive instruction**, almost everywhere it has been measured. The largest meta-analysis on the question, Freeman and colleagues' 2014 paper in PNAS, pooled 225 studies of undergraduate STEM courses and found that student performance under active learning conditions improved by 0.47 standard deviations on exams and concept inventories, with failure rates 1.5 times higher under traditional lecture. The effect held across every discipline studied. The authors concluded, with unusual force for a meta-analysis, that the results "support active learning as the preferred, empirically validated teaching practice in regular classrooms."

The third is that **feedback is the single most powerful classroom variable we have a strong consensus on**. John Hattie and Helen Timperley's 2007 review in *Review of Educational Research* put the effect size of well-formed feedback on student achievement between 0.70 and 0.79 — extraordinarily large by educational-research standards. The crucial word is *well-formed*. Feedback that tells the student where they were trying to go, where they actually are, and what to do next outperforms feedback that just labels work as good or bad. Praise and grades, by themselves, are weak instructional tools; substantive, forward-looking feedback is a near-miracle.

If you stack these findings together, the picture is clear. Learning is something a person does with effort, in interaction with material and with other people, under conditions where feedback can flow continuously and the student can act on it. Or as Paul Black and Dylan Wiliam wrote in their landmark 1998 monograph *Inside the Black Box*, summarizing the case for formative assessment: "There is a body of firm evidence that formative assessment is an essential component of classroom work and that its development can raise standards of achievement. We know of no other way of raising standards for which such a strong prima facie case can be made."

That was twenty-eight years ago. Most universities still mostly grade the final paper.

There is a fourth, less tidy finding that turns out to matter enormously in the AI era. Learning, even in adults, is regulated by the learner. The body of research on **self-regulated learning** (SRL), built up over decades primarily by Barry Zimmerman, Ernesto Panadero, and their collaborators, treats learning as a cycle: students set goals, plan how to reach them, monitor their own understanding, adjust their strategies when they notice they are off track, and reflect afterward on what worked. Students who do this well outperform students who don't, even controlling for prior achievement. The skills are teachable but rarely taught explicitly. And — this is the part that matters now — when a student offloads cognitive work to an AI, the most important thing they often offload is not the answer. It is the *monitoring*. They stop noticing what they don't understand.

The final piece of the prior consensus is the framework that put a popular language around all of this. In the early 2000s, the U.S.-based Partnership for 21st Century Skills, now hosted by Battelle for Kids, codified what eventually became known as the 4Cs: critical thinking, communication, collaboration, and creativity. The 4Cs are not a research result; they are a synthesis of what employers, educators, and policy bodies converged on as the durable, transferable capabilities every student should develop. They are deliberately not a list of facts. They are a list of *practices*. You can't take a multiple-choice test on creativity. You have to do something creative, in front of someone who can judge it.

This is the framework most schools claim to teach. It is also the framework most schools have the hardest time actually assessing — because the 4Cs are processes, and the conventional assessment infrastructure of higher education is built around products. The Association of American Colleges and Universities' VALUE rubrics, developed by faculty teams across more than two thousand institutions, are the most widely adopted attempt to bridge the gap. There is a VALUE rubric for critical thinking, for creative thinking, for written communication, for oral communication, for teamwork, for inquiry

and analysis, and a dozen others. Each one operationalizes what "good" looks like at progressive levels of attainment, and each one assumes the assessor will be looking at evidence drawn from authentic student work over time, not from a single timed exam.

Here is the synthesis. Before ChatGPT, the research had already told us:

Learning is effortful, social, and continuous. The strongest tool a teacher has is timely, well-formed feedback inside an active task. The skills that matter most for a graduate's life are not facts but practices — critical thinking, communication, collaboration, creativity — and those practices have to be seen and judged over time. Assessment systems that compress all of this into a final product and a number are weak assessments even on their own terms.

What generative AI did was take that weakness and turn it into an emergency.

Part II. The 4Cs, observed not assumed

The 4Cs deserve a closer look, because the entire learning intelligence project depends on whether they can be made *observable*. If "critical thinking" is a vibe, it cannot be evidenced and the field collapses into self-reporting. If it is a sequence of behaviors a person can be seen doing — well, then we can build something.

Take critical thinking first. The VALUE rubric for critical thinking breaks the construct into five observable dimensions: explanation of issues, evidence (selecting and using credible information), influence of context and assumptions, the student's own position (with appropriate complexity), and conclusions and related outcomes. Each of those leaves traces. A student frames a question well or badly. They cite a source or invent one. They acknowledge a counterargument or steamroll past it. They reach a conclusion that follows from their evidence or one that overreaches. In a writing assignment, almost every one of these moves is visible in the draft history if anyone bothers to

look.

Communication is similarly behavioral. The VALUE rubric for written communication asks about context and purpose, content development, genre and disciplinary conventions, sources and evidence, and control of syntax and mechanics. Every one of these is something a reader can see, and every one is something a draft history can reveal as a process. The student who revises for clarity is doing communication. The student who hands in a polished first draft they didn't write may have done communication on the page but not in their head.

Collaboration is where most assessment infrastructure has historically failed. Most universities give group grades. A group grade tells you almost nothing about which student did the collaborating. The VALUE rubric for teamwork instead asks about contributions to team meetings, facilitation of teammates' contributions, individual contributions outside meetings, fostering a constructive team climate, and responding to conflict. Each of these is observable in a peer-evaluation workflow, a shared document's revision log, or a structured peer-review system. The signal exists. We mostly don't capture it.

Creativity is the hardest of the four, and the easiest to game. The OECD's PISA 2022 creative thinking assessment — the first time creativity has been measured at international scale — defines the construct as the capacity to generate diverse and original ideas, *and* to evaluate and improve upon ideas, in open-ended tasks across written, visual, scientific, and social problem-solving domains. The framework is deliberately not about a single output. It is about whether a student can generate alternatives, recognize promising ones, and improve them iteratively. AI can produce a thousand decent ideas in a minute. What it cannot do, yet, is sit inside a learner's head and develop their ability to discriminate between them.

Across all four, the same pattern appears. The construct can be reduced to a number, but doing so destroys most of what makes it valuable. The construct can be observed in process, if the process is captured at sufficient resolution.

And the process is exactly what teachers stopped having time to watch, somewhere around the moment college enrollments crossed twenty million and the average instructor's grading queue stopped being humanly possible to read carefully.

There is a fifth capability that increasingly belongs in this list, even though it is not part of the original 4Cs: **metacognition and self-regulated learning**. This is the practice of monitoring your own understanding, planning your approach, noticing when you're stuck, and choosing a better strategy. In a world where students can outsource almost every other cognitive operation to a machine, the one thing they cannot outsource is knowing whether they have actually understood something. The work of Singh and colleagues at ASIS&T in 2025, which embedded metacognitive prompts into a generative AI search workflow, found that students who were nudged to evaluate the AI's answers — to ask themselves whether what they were reading actually addressed their question — engaged in deeper inquiry and were more discerning about AI responses. The intervention was small. The implication is not. If AI is to be part of how students learn, then helping them notice when it is helping them and when it is fooling them is itself a learning objective.

The 4Cs, plus metacognition, are the right scaffolding for the rest of this argument. They are not abstract. They are practices. They produce evidence. And the practices are exactly what generative AI most threatens to collapse if they are not deliberately preserved.

Part III. A short history of generative AI in the classroom

The story of GenAI in education from the end of 2022 to the spring of 2026 is the story of an entire sector going through the stages of grief in under four years. It is worth walking through, because the policy and product responses still alive today were forged in specific moments, and the moments still shape what is possible.

November 30, 2022 — Release. OpenAI publishes ChatGPT, a free conver-

sational interface on top of GPT-3.5. There is no official "education launch." There doesn't need to be one. Within days, students discover that the chatbot will write a passable essay on almost any topic in any voice, and the news travels through TikTok faster than any administrator can respond.

January 2023 — Panic. New York City Public Schools, then the largest district in the United States, blocks ChatGPT on school networks. Multiple universities follow. Op-eds appear under headlines like "The College Essay is Dead." The first wave of teacher Twitter is a mix of horror, grim humor, and improvisation. Some instructors switch to handwritten in-class essays. Some begin assigning oral defenses. Most do nothing different because the semester is already underway.

February 2023 — Detection arms race. Turnitin announces that AI writing detection is coming to its products. Within months, a parallel industry of "AI humanizers" appears. Students begin running their AI-generated text through second AI tools to bypass detection. The detection market grows. The arms race accelerates.

May 2023 — The first major policy document. The U.S. Department of Education's Office of Educational Technology publishes *Artificial Intelligence and the Future of Teaching and Learning*. Its core message is that AI should "augment human intelligence, not replace it," that teachers must remain central, and that the right response is not bans but informed adoption. The report uses the phrase "humans in the loop" repeatedly. It is the first major institutional signal that the answer to AI in classrooms is not a firewall.

August 2023 — Vanderbilt disables Turnitin's AI detector. In one of the more consequential institutional decisions of the year, Vanderbilt University publicly turns off Turnitin's AI writing detection feature for its entire community. The rationale is unusually candid. Turnitin had claimed a 1% false positive rate, which would have meant roughly 750 of Vanderbilt's 75,000 annual submissions being wrongly flagged. Worse, internal analysis showed that AI detectors were "more likely to label text written by non-native English speakers as AI-written." The university concluded the technology was

not effective enough to be used. Other institutions followed. The detection-first strategy started losing credibility a year before it started losing in court.

September 2023 — UNESCO weighs in. UNESCO publishes the first global guidance document on generative AI in education, calling on governments to regulate use, protect student data, set age limits, and build AI literacy into curricula. The document is mostly cautionary, but it explicitly frames the issue as a curriculum and pedagogy issue, not just a cheating issue.

November 2023 — TEQSA reframes assessment. Australia's Tertiary Education Quality and Standards Agency publishes *Assessment Reform for the Age of Artificial Intelligence*. It is the first national regulator to move the conversation from detection to redesign. TEQSA's premise is that the assurance of learning is the institution's responsibility, and that responsibility cannot be discharged by trying to catch AI use. It has to be discharged by designing assessments that produce evidence AI cannot easily fake.

2024 — Normalization. The Digital Education Council's first global student survey finds 86% of students using AI regularly. Half do not feel "AI ready." OECD releases PISA 2022 creative thinking results, the first international comparable measure of the capability. Singapore tops the rankings; many high-income systems underperform their reading and math scores on creativity, exposing how badly traditional curricula prepare students for open-ended problems. The conversation begins shifting from "how do we ban this" to "what should students actually be able to do."

2025 — The institutional response hardens. EDUCAUSE's 2025 AI Landscape Study finds that 57% of higher-ed institutions now treat AI as a strategic priority, up from 49% the year before, and that 74% are focused on academic integrity, 65% on coursework, and 54% on assessment practices. The same study reveals a "digital AI divide" — well-resourced institutions race ahead; under-resourced ones cannot keep up. Tyton Partners' *Time for Class 2025* finds 38% of faculty reporting increased workload from AI versus only 11% reporting decreased. The biggest workload drivers are monitoring for cheating and rebuilding assessments.

June 2025 — The Kestin RCT. Greg Kestin and colleagues at Harvard publish a randomized controlled trial in *Scientific Reports* in which roughly 180 students in an introductory physics course alternated weekly between traditional in-class active learning and homework using a custom-built AI tutor. The carefully engineered AI tutor — short responses, expert scaffolding, guardrails against hallucination, structured prompting — produced learning gains roughly *twice as large* as the active-learning sessions, on tests administered after both. Students reported higher engagement and motivation as well. The study did not show that ChatGPT-as-such teaches well. It showed that a pedagogically designed AI, built around what we know about learning, can outperform what we previously believed was the gold standard of in-person instruction.

July 29, 2025 — Study Mode. OpenAI launches "Study Mode" in ChatGPT, built in consultation with pedagogy experts from more than forty institutions. Instead of producing direct answers, Study Mode uses guiding questions, the Socratic method, and step-by-step reasoning. It is the first major signal that the platform layer itself recognizes the difference between *answering* a question and *teaching* the person who asked. (OpenAI is also explicit that students can flip back to regular mode at any time. The platform cannot enforce pedagogy. Only an institution can.)

January 2026 — OECD names the problem. The OECD Digital Education Outlook 2026 introduces the phrase that will probably define the policy conversation for the rest of the decade: **false mastery**. The Outlook draws on emerging studies showing that students who practiced math with a generic chatbot performed better in the moment but scored *up to 17% worse* on subsequent closed-book exams than peers who studied alone. The mechanism, OECD argues, is straightforward: when students rely on AI, the metacognitive work that converts answers into understanding never happens. The output looks like learning. The learner has not actually learned.

January 2026 — Faculty hit a wall. The Elon/AAC&U survey of 1,057 faculty publishes. Ninety-five percent expect AI to increase student overreliance. Ninety percent expect it to diminish critical thinking. Eighty-three percent

expect it to shorten attention spans. The headline finding is in the title: *The AI Challenge: How College Faculty Assess the Present and Future of Higher Education in the Age of AI*. The faculty, more than three years in, are not enthusiastic about how this is going.

March 2026 — HEPI publishes the new normal. HEPI's 2026 student survey finds that AI use among UK undergraduates is now near universal, with 88% using GenAI for assessed work. The institutional question has fully shifted. It is no longer, *will students use AI?* The answer is yes, all of them, often. The question is, *what are we able to certify about what they learned?*

That is the four-year arc. Panic, ban, detect, regulate, integrate, re-examine. The sector has not landed yet. But the place it is landing on, the one almost every policy body and serious researcher is now pointing toward, is the same place. Not better detection. Not faster grading. Not bigger LMS dashboards. *Better evidence of learning, captured during the learning itself.*

That place is what learning intelligence is for.

Part IV. When output becomes cheap

There is a moment in the life of any measurement system when the thing it measures stops being scarce, and the system stops being useful. The assessment crisis in education is that moment.

For most of the last century, a well-written paper was a reasonable proxy for a literate mind. It took hours of reading, drafting, revising, and re-reading to produce. The labor and the cognition were entangled. To submit the paper was, with allowances for cheating, also to have done the thinking. The artifact was the evidence.

That entanglement is what generative AI breaks. The price of a polished paragraph has collapsed. The price of a coherent five-page argument has collapsed. The price of a passable C+ undergraduate essay has collapsed approximately to zero. Whatever those artifacts used to certify, they no longer

certify in the same way. This is not a moral observation. It is a measurement observation. A thermometer that returns the same number whether or not anything is hot has stopped being a thermometer.

The polite name for this in the assessment literature is the **validity crisis**. The OECD's "false mastery" finding is one way to describe it. The 17% gap between practice performance with a chatbot and unaided exam performance afterward is a direct empirical demonstration that the metric (the practice score) is decoupling from the construct (what the student actually knows). The research on cognitive offloading, much of it published in 2025, makes the mechanism more specific. When students hand cognitive work to AI tools, they engage in less of the self-monitoring and effortful retrieval that produce durable learning. Frequent AI use is now reliably correlated with weaker critical thinking, with cognitive offloading as the statistical mediator. The effect is strongest in younger students and those with less academic experience — the populations whose learning was most fragile to begin with.

The reflex response — find the AI text, punish it, restore the old contract — has been tested at scale and is not working. Vanderbilt's public reasoning for disabling Turnitin's AI detector in 2023 is still the cleanest statement of why. False positive rates that look acceptable in a vendor deck (1%, for example) translate into hundreds or thousands of wrongly accused students at institutional scale. Non-native English speakers are systematically more likely to be flagged. The detection arms race is also asymmetric in time: detectors look at one snapshot of text, while AI generation models update continuously and rapidly outpace the patterns detectors can learn. Multiple recent papers, including Garland's 2026 mathematical framing of the detection problem, argue that text-only one-shot detection is structurally incapable of achieving the fairness properties educational institutions need. Even Turnitin itself has shifted its public messaging, repositioning AI detection as one signal among many rather than a determinant of misconduct.

The detection failure is not an accident. It reflects a deeper truth about the problem. AI text is not, in any robust technical sense, distinguishable from human text. It is a category of writing, and that category is converging on

the same prose features the academy taught us to value: fluency, clarity, organization, conventional grammar, formal register. We trained machines on those features, then asked them to produce text by maximizing those features, and we are now surprised that the resulting text is hard to distinguish from text by students whose teachers also asked them to produce text by maximizing those features. The category boundary was always fuzzy. It is now functionally gone.

What follows from this is uncomfortable but unavoidable. The artifact-only model of assessment was never very strong. It survived because the cost of producing acceptable artifacts was high enough that the artifact and the learning were *in practice* yoked together. Once the yoke is removed, the artifact stops being able to do the assessment job. You cannot certify what a student learned by reading their final paper, in the same way you cannot certify a marathon runner by watching them cross the finish line in a friend's car.

The serious question is what replaces the artifact-only model. The answer cannot be "go back to in-person handwritten exams forever," for two reasons. The first is that handwritten exams have their own validity problems — they tend to measure memorization, performance under stress, and English-language writing speed, none of which are what most courses are trying to certify. The second is that no graduate of any of these institutions will go on to do their actual professional work without AI tools. To grade students exclusively on AI-free performance is to grade them on a skill they will never use again.

The answer cannot be "trust the student," either. Not because students are dishonest. Most are not. But because the question is not about honesty; it is about evidence. Even fully honest AI use — a student who openly used ChatGPT to outline their essay, then wrote it themselves, then asked the AI to suggest improvements, then accepted some and rejected others — produces an artifact that does not, by itself, reveal which parts of the work were done by the student and which were not. The evidence problem is not solved by trust.

The answer has to be a different evidence model. One that does not depend on the final artifact being scarce. One that can absorb the fact that AI is everywhere, in every step of the work, and still produce something an instructor can act on, a student can learn from, and an institution can defend.

That is what learning intelligence is trying to be.

Part V. The two AIs

It is tempting, in the face of the assessment crisis, to conclude that generative AI is bad for learning. Many faculty have concluded exactly that. The Elon/AAC&U numbers — 95% expecting overreliance, 90% expecting weaker critical thinking — describe a faculty population that has seen, up close, what unconstrained AI use looks like.

The evidence is more complicated than the faculty mood. There appear to be two AIs, distinguished not by which model is running but by how the model is used. One AI helps learners. The other replaces them. The same chatbot can do both within the same hour.

The case for AI as a learning amplifier is real and growing. The Kestin et al. randomized controlled trial in *Scientific Reports* found that students using a carefully designed physics AI tutor — short responses, expert scaffolds, explicit step-by-step reasoning, guardrails against giving away answers — learned roughly twice as much per hour as students in an active-learning classroom. The same study found higher self-reported engagement and motivation. These are large effects in real higher-ed settings, not lab experiments. The 2025 meta-analyses on ChatGPT and academic performance, and the 2026 meta-analyses on GenAI and educational outcomes in higher education, point in the same direction on average: AI-supported learning interventions produce significant positive effects on achievement, particularly when the intervention scaffolds the learning process rather than substituting for it.

The case for AI as a learning erosion mechanism is equally real. The OECD Outlook's "false mastery" finding, the cognitive-offloading research, the workload survey showing faculty spending more time policing AI than teaching, the studies showing students who lean heavily on AI tools score worse on subsequent unaided assessments — these are not noise. They describe what happens when AI is used without pedagogical structure. The OECD framing is now widely accepted: GenAI can raise short-term task performance without producing learning, especially when students treat it as a shortcut.

What separates the two cases is whether the AI is being used to do the *learning work* or to do the *output work*.

When AI does the learning work — when it asks the student a question, makes them try, shows them where they went wrong, encourages them to try again, models the reasoning, then steps back — it is doing what good tutors have always done. Bloom's famous "two sigma problem" was about the gap between one-on-one tutoring and conventional instruction. AI tutoring is one of the few interventions in the history of educational technology that has produced effect sizes anywhere near that gap. The Kestin study's roughly two-fold gain over active learning is in the ballpark of what Bloom was describing in 1984 when he proposed the original problem.

When AI does the output work — when the student types the prompt, takes the result, lightly edits it, and submits — the artifact looks the same as it did before, but no learning has happened. The student's neural circuits for thinking through the problem have not been exercised. Their schema for the topic has not been built. Their ability to do the work without the tool has not improved and may have actively decayed. This is what cognitive offloading looks like in practice.

The pedagogical implication is the central practical claim of the entire learning intelligence project: **AI is not a pedagogy. It is an amplifier of whichever pedagogy you bring to it.** Good designs become better. Bad designs become much worse. An assignment that was already a poor test of

student understanding becomes a near-zero test of student understanding when AI is added. An assignment that was already focused on the process of thinking, with scaffolding and feedback and revision, can become much more powerful with AI as the tireless second reader.

There is a third case worth naming, which is the question of whether AI can do the *relational* work of education. Can a chatbot mentor? Can it be empathetic? The honest answer from the research is partial. The 2024 systematic review by Sorin and colleagues in *JMIR* found that large language models can demonstrate elements of cognitive empathy — recognizing emotional content, producing supportive-sounding responses, sometimes outperforming rushed humans on perceived bedside manner. But they do not feel with the learner, and their "empathy" is prompt-sensitive, surface-level, and easily destabilized. A 2024 quasi-experimental study in online higher education found that empathic chatbot feedback was comparable to teacher feedback on learning performance, motivation, and self-regulation in that specific context. But a 2025 study on the "emotional cost of educational chatbots" found that students using a chatbot during an assignment reported significantly lower positive affect than peers who did not. A 2026 study on the "AI empathy choice paradox" found that people generally prefer to receive empathy from humans, even while rating AI-generated empathy as high quality when they receive it.

The synthesis is straightforward. AI can do first-line support, low-stakes encouragement, formative feedback, structured Socratic prompting, and some forms of tutoring at large scale. It cannot do faculty mentorship, the relational trust that underwrites belonging and challenge, the high-stakes advising that shapes a student's life, or the moral seriousness that good teachers bring to difficult conversations with struggling students. To collapse the distinction between these roles is to mistake the substrate for the substance. Empathy is, in the end, a feature of accountability between persons. A chatbot is not accountable to anyone in that sense, and pretending otherwise is the same category mistake that has tripped up every wave of educational technology since the teaching machine.

The implication for learning intelligence is that the question is never *whether* to use AI. It is *for what*. The systems that work treat AI as one instrument in a teacher's hand, capable of expanding the teacher's reach when used carefully and capable of replacing the teacher's effect entirely when used carelessly. The systems that fail treat AI either as an existential threat to be detected and punished or as a magical solution to be deployed and trusted. Both framings are wrong in the same way. Both ignore the actual mechanism of learning.

Part VI. From artifact to process: a new evidence model

If the artifact-only model of assessment is broken, what replaces it? The answer existing in the research literature, well before generative AI made it urgent, is **evidence-centered design**: build assessments around the evidence you actually need to make the claims you want to make about what students can do.

The framework comes most directly from Robert Mislevy and colleagues' work on principled assessment design, dating to the late 1990s. It treats assessment as an argument structure. The argument starts with claims (what we want to say a student knows or can do), specifies the evidence that would support those claims, and only then designs tasks that elicit that evidence. The grade is the conclusion of an argument, not the start of one. This sounds obvious. It is rarely how courses are actually built.

Layered on top of evidence-centered design is the literature on **stealth assessment** — the practice of capturing evidence of learning *during* an authentic activity, rather than interrupting the activity with separate tests. Stealth assessment was developed primarily in the context of educational games and simulations: as a student plays the simulation, the system collects evidence of how they approached problems, what strategies they tried, when they sought help, and how they revised their approach in light of feedback. The evidence is built into the experience. It does not require the student to stop learning in order to be measured.

A third literature is the more recent work on **process data in large-scale assessments**, much of it published by OECD and the major testing organizations. PISA, NAEP, and similar instruments are now routinely capturing not just student answers but the sequences of actions students take to produce those answers: which items they returned to, how long they spent on each step, what tools they used, how they revised. The PISA 2025 "Learning in the Digital World" framework treats iterative knowledge building and effective self-regulation with digital tools as integral parts of the competence being assessed, not as side effects of the task design. The framework's premise is that *how* a student solves a problem is itself a measurable competence, sometimes more diagnostic than whether they solve it.

These literatures converge on the same operational principle, which is the single most important thing to take away from the assessment research of the last twenty-five years: **the more steps you can observe between a student's first encounter with a problem and their final answer, the more confidently you can certify what they learned.**

This is the principle that the AI era forces back into the center of the conversation. The artifact (the final paper, the final exam, the final code submission) is a single point of evidence. Once AI can produce that point cheaply, that point becomes an unreliable estimator of the student's underlying capability. Multi-point evidence is much harder to fake. A student who can produce a draft, revise it in response to peer feedback, defend their argument orally, answer follow-up questions about choices they made in the draft, and reflect afterward on what they learned has produced evidence at five or six distinct points. Faking all six points coherently is not zero-cost. It is also, from the student's perspective, almost the same amount of work as actually learning the material — at which point the incentive structure starts pointing back toward the learning.

The most robust assessment designs in the AI era turn out to be the ones that have been recommended for decades by serious assessment researchers but have rarely been adopted at scale because they are labor-intensive. They include:

Staged writing assignments, where the draft history is itself evidence. Students submit a planning document, an outline, an annotated bibliography, a first draft, a revision memo, and a final draft, with feedback exchanges between each. The grade considers the trajectory, not just the endpoint. AI can be used at any stage and the system still produces signal about what the student actually engaged with.

Inquiry-driven discussion, where the quality of the questions a student asks is itself a measurable thing. The Packback platform's "Curiosity Score," for example, is not a generative AI scoring system; it is an algorithmic measure of the open-endedness, sourcing, and clarity of student-posed questions in a course discussion. A peer-reviewed study published in the *Journal of Computing in Higher Education* on 2,800 long-form assignments found that AI-assisted process feedback improved writing quality and reduced grading workload. The mechanism was process visibility, not output evaluation.

Social annotation, where students mark up shared readings in front of each other before class. The annotation behavior is the evidence. A study published in *Frontiers in Education* on the Perusall platform found that pre-class annotation grades, number of annotations per week, and the ratio of annotations engaged in peer discussion together accounted for **41.8% of the variance in students' weekly post-class essay performance**. That is an extraordinarily strong relationship. Annotation behavior, in that context, is one of the most diagnostic single signals available in undergraduate teaching.

Peer review with calibration, where students assess each other's work using rubrics and are themselves assessed on the quality of their feedback. The peer review is the assessment of collaboration. Done well, it produces evidence about both the reviewed student and the reviewer.

Oral defense or "viva voce" components, where students are asked to explain their work and answer follow-up questions. This is the oldest and most reliable assessment in the academy, the one PhD committees use precisely because they cannot easily certify a dissertation just by reading it. It is also the assessment that AI is most structurally bad at faking, because the student

has to think on their feet in conversation.

Reflective process notes, where students articulate what they tried, what they learned, where they got stuck, and what they would do differently. These are evidence of metacognition. They are also, when done honestly, the cheapest learning intervention any course can add.

Explicit AI use disclosure, where students describe what AI tools they used, for what purposes, and how they evaluated the results. This is the assessment design move that does the most work for the least cost. It does not depend on detection; it does not depend on trust. It simply makes AI use a legible part of the assignment, which both surfaces evidence of student judgment and creates a record that can be assessed on its own terms.

What all of these have in common is that they multiply the number of points where the student is visibly thinking. Each individual point is not necessarily harder to fake than a final essay. But together, they form a pattern that takes much more work to fake than to do, and that produces a much richer record for the instructor to read.

This is the evidence model that learning intelligence systems are being built to support. The systems are not, in their best forms, AI graders. They are AI-augmented evidence captures. They make the process visible. They surface signals to instructors and students. They produce records that can be defended in front of an accreditor or a hearing committee. They do not collapse learning into a number. They are the modern translation of what teachers have always wanted: to see what their students are actually doing as they learn.

What a learning intelligence platform actually looks like

Definitions move quickly to abstraction. It is worth being concrete about the shape such a system actually takes, because most of the products that will be marketed under the "learning intelligence" banner over the next two years will be something else wearing the label. A serious learning intelligence platform sits on three layers, and a buyer can tell whether a product is real by

asking which of the three are present.

Layer 1 — Pedagogy and the constructs. Underneath everything, a platform has to know what it is trying to measure. The 4Cs are the most defensible anchor framework available, and they pair naturally with what a growing number of practitioners are now calling **AI literacy capabilities**: judgment (knowing when to trust an AI output), explanation (knowing how to defend a choice an AI helped produce), coordination (working with AI as one collaborator among many), and agency (deciding when not to use it at all). The pairing matters because it tells the system what to look for. Without an explicit construct layer, every "insight" the system produces is a behavioral metric in search of meaning.

Layer 2 — Assignment types as instruments. The middle layer is the set of authentic activities through which evidence is produced. A credible learning intelligence platform offers more than one assignment type, because no single activity surfaces all of the constructs above. Discussions surface inquiry quality and reasoning chains. Writing assignments surface argument structure, source use, and revision behavior. Close reading and annotation surface comprehension and engagement with text. Peer review surfaces collaboration and the ability to give and receive substantive feedback. Team-based projects surface coordination and equitable contribution. Conversational reasoning and oral defense components surface explanation under pressure. And a newer category — what some platforms now call **AI-integrated assignments** — surfaces AI literacy itself: students prompt, evaluate, accept, reject, and reflect on AI use, and the entire trajectory becomes evidence the instructor can read. The principle is that whoever owns the assignment owns the learning intelligence; the platform is the assignment surface and the evidence layer simultaneously.

Layer 3 — Synthesis and insight. The top layer takes the signals from individual assignments and rolls them up into views that match the audience. A student should see their own evidence portfolio — what they did, what feedback they received, where they grew, where they are stuck. An instructor should see a per-section view that answers the three questions worth

asking in any teaching moment: who needs help now, what evidence supports that inference, where should I intervene. A department chair or chief academic officer should see aggregated, privacy-preserving rollups of 4C coverage across a course, cohort, program, or institution — the kind of view an accreditor can read and a budget committee can fund against. Most products in the adjacent categories solve at most one of these audiences. The full stack solves all three.

A platform that has only Layer 3 is a dashboard pretending to be intelligence. A platform that has only Layer 2 is a workflow tool. A platform that has only Layer 1 is a framework, not a system. The serious work of the next two years in this category is to assemble all three.

A maturity model for the institution

Institutions, like products, do not arrive at full learning intelligence overnight. A useful way to think about adoption is as three progressive postures an institution can take toward AI in its assessments, each enabled by a different set of platform capabilities.

The first posture is **AI Aware**. The institution recognizes that AI is in every classroom and starts to make its assessment more visible: faculty use the 4Cs explicitly, assignments are designed with process visibility in mind, and basic engagement and course-level signals are available to instructors. The institution is no longer pretending AI is absent. It is not yet doing anything specific about it.

The second posture is **AI Active**. The institution now treats AI use itself as an object of instruction. Assignments include conversational reasoning, co-writing with AI under structured prompts, collaborative and live thinking components, and a metacognitive layer in which students reflect on what they asked the AI and why. The system surfaces cohort-level patterns and a student learning-journey view that crosses individual assignments. Faculty are teaching with AI, not policing it.

The third posture is **AI Native**. AI is embedded throughout, and the

institution measures not only the 4Cs but AI literacy itself. Full learning-journey visibility is available to faculty and administration. AI literacy is benchmarked across cohorts. Predictive engagement and risk signals support intervention before usage stalls. The institution can produce, on demand, defensible evidence of what its graduates can do with AI and without it.

The point of the maturity model is not that every institution should sprint to AI Native. The point is that "AI policy" is not a single decision; it is a position on a continuum, and the right position depends on faculty readiness, student population, governance maturity, and what the institution is trying to certify. A learning intelligence platform that respects this is one that meets the institution where it is and offers a clear path forward.

Part VII. Defining learning intelligence

It is time to be specific.

Learning intelligence is the continuous collection and interpretation of evidence about how learning is happening, used to improve teaching, learning, and assessment.

The definition is one sentence long for a reason. The longer versions tend to obscure what the field is actually for. Learning intelligence is not a technology. It is a practice. The technology exists to enable the practice, the way the microscope exists to enable biology. The relevant question is not "how good is the AI?" but "what evidence about learning is being produced, by whom, for whom, and to what end?"

Several other terms are in the air, and the differences matter:

Learning analytics is the established academic field, defined since 2011 and re-codified in 2025 by SoLAR, the Society for Learning Analytics Research, as "the collection, analysis, interpretation and communication of data about learners and their learning that provides theoretically relevant and

actionable insights to enhance learning and teaching." Learning analytics is the intellectual parent of learning intelligence. The difference is that classical learning analytics has tended toward retrospective dashboards (engagement, time-on-task, click counts) that often sit at the LMS layer and are weakly connected to specific learning constructs. Learning intelligence, as the term is being used now, is more pedagogically opinionated, more assignment-native, and more focused on inferring constructs (critical thinking, communication, etc.) rather than reporting engagement.

AI in education is the broader category that includes everything from AI tutors to administrative chatbots. Learning intelligence is a specific use case within it. Not all AI in education produces evidence of learning; some just produces convenience or automation. The distinguishing question for learning intelligence is whether the system's primary output is *evidence a human can act on*.

Plagiarism and integrity detection has tried to colonize the assessment problem from the integrity side. Detection is a small, narrow, and now-discredited slice of what assessment needs in the AI era. A learning intelligence system may include some integrity signals, but its main job is positive: showing what students did, not just flagging what they might not have done.

Adaptive learning and **personalization** are adjacent fields focused on adjusting content delivery to individual students. These are related but distinct. Adaptive systems are about giving the right next problem; learning intelligence is about understanding what a student's work reveals about their thinking.

Within the learning intelligence space itself, four overlapping variants are emerging:

The first is **instructional intelligence**, mostly K-12 focused, where the AI is embedded in curriculum and lesson delivery. Vendors like Kiddom and Subject have begun marketing under this umbrella. The signal these systems capture is mostly engagement and standards-aligned progress.

The second is **assessment and authorship intelligence**, where systems capture evidence of writing process, draft history, AI use, and revision behavior at the assignment level. Cadmus is the cleanest example. Turnitin's Clarity product, Brisk's writing replay, and others fit here too. The signal is process provenance: how did this artifact come to be?

The third is **institutional learning intelligence**, where LMS and campus platforms aggregate outcomes and engagement data across courses to support program review, accreditation, and student success. Canvas Intelligent Insights, D2L Achievement+, and Anthology Blackboard's analytics suite are examples. The signal is institutional: which courses, programs, and student segments are at risk?

The fourth, and the one with the most upside, is what might be called **process-native higher-ed learning intelligence**: systems that capture evidence of student thinking inside specific high-cognition assignments — discussion, inquiry, writing, peer review, oral defense — and turn that evidence into feedback for students and instructors while it can still change the outcome. Packback's positioning is in this fourth variant. So, in different ways, are FeedbackFruits, Perusall, and Kritik. The category is still being defined.

What these variants share is a set of design principles. The following five, taken together, are what distinguish a learning intelligence system from a generic AI tool or a generic analytics dashboard.

Principle 1: Process over artifact. The primary unit of analysis is what the learner did, not just what they submitted. This is the central commitment. A system that only ingests final submissions cannot do learning intelligence in any robust sense.

Principle 2: Constructs over clicks. Raw events (clicks, time-on-page, post counts) are inputs, not outputs. A learning intelligence system maps events to learning constructs (critical thinking, revision quality, collaboration depth) using transparent evidence models. Engagement metrics are at best weak proxies for learning; a system that surfaces them as if they were

learning measurements is doing engagement theater.

Principle 3: Evidence over scores. The system's primary output should be evidence a human can interpret, not a black-box score. When the system does produce scores, the scores should be defensible: the user should be able to ask "why" and get a useful answer that points back to specific observed behavior.

Principle 4: Transparency over surveillance. Students should know what is being captured, why, and how it will be used. Teachers should be able to see and challenge the system's inferences. Institutional governance — privacy, retention, role-based access, audit logging — is part of the system, not an afterthought. This is not just an ethics requirement. It is a precondition of trust, and trust is a precondition of any assessment that actually changes student behavior.

Principle 5: Human-in-the-loop over autonomous judgment. The system supports instructor judgment; it does not replace it. High-stakes decisions — grades, integrity findings, intervention referrals — remain with humans. The system's job is to make those decisions better-informed, not to make them automatically.

A system that respects all five principles will look quite different from a typical AI-in-education product today. It will collect more signal but display less of it raw. It will be quieter than a dashboard. It will produce reports that look more like a portfolio than a score. It will give students a window into how they are seen, and a way to push back. It will give teachers a way to spend more of their time on the part of teaching that requires their judgment, and less on the parts that don't.

That, in the end, is the test for whether something is learning intelligence or just edtech with AI features: does it help the people in the room understand what is happening between them?

A useful way to ground the abstraction is to walk through how each of the assignment types described earlier maps onto the 4Cs and the AI literacy

capabilities they enable. Discussions evidence critical thinking and communication, and (because students respond to each other) early-stage collaboration. Writing evidences critical thinking, communication, and (through draft history) self-regulation. Close reading and annotation evidence critical thinking and communication in the context of source engagement. Peer review evidences collaboration and a meta-form of critical thinking — the ability to read another person's work the way an instructor would. Team-based work evidences collaboration and creativity in coordination. Conversational reasoning and oral defense evidence explanation under pressure, which is one of the cleanest tests of whether comprehension is genuine. AI-integrated assignments evidence judgment and agency in AI use itself. No single assignment type carries the whole load. The system's coverage of the 4Cs is the sum of the assignments it can host and the constructs it can map them to. A platform that hosts only one of these types — discussion alone, or writing alone — is by definition a partial measurement system, regardless of how good its analytics layer is.

Part VIII. How to teach with AI

The categorical answer to "how should I teach with AI" is "it depends," which is true and useless. The operational answer, drawn from the research evidence assembled so far and from the practices of the institutions that are getting this right, is more specific.

Start with what you are trying to certify. Most courses, if pressed, can name three to five intended learning outcomes. The first question for AI redesign is whether those outcomes are still meaningful in a world where AI can produce most of the artifacts that previously evidenced them. If "the student can write a coherent five-page essay on a literary text" is the outcome, that outcome is now a weak proxy for what the course probably actually cares about, which is something more like "the student can read closely, generate a thesis from evidence, defend it against an alternative reading, and revise

based on critique." The redesign starts by sharpening the outcome.

Make AI use explicit and bounded. The single most leveraged move any instructor can make right now is to write an AI use policy into each assignment that specifies what AI tools may be used, for what stages, and with what disclosure. The policy can be permissive ("any AI tool, any stage, with a disclosure paragraph") or restrictive ("no AI tools for this exam"). What it cannot be is implicit. Implicit policies turn every assignment into a guessing game for students and a detection puzzle for instructors. Explicit policies turn AI use into a piece of evidence the instructor can read.

Stage the work. Almost any high-cognition assignment can be broken into stages with checkpoints. A research paper can have a topic proposal, an annotated bibliography, an outline, a draft, a peer review exchange, a revision memo, and a final draft. Each stage is short. Each stage is its own piece of evidence. The grade can be weighted toward the final product or distributed across the process; either works. The point is to multiply the points of evidence. AI cannot easily fake six stages of legible thinking across two months.

Add a metacognitive layer. The cheapest way to convert AI use from passive offloading into active learning is to require students to reflect on it. "What did you ask the AI? What did you accept? What did you reject? Why?" These three questions, asked in a short reflective paragraph attached to every AI-permitted assignment, do an enormous amount of work. They surface the student's judgment, which is the thing the AI cannot do for them. They produce a record that can be assessed. And they nudge students into the metacognitive practice that the Singh et al. study found was associated with deeper inquiry and better evaluation of AI responses.

Reintroduce the voice. Oral components have been falling out of higher education for decades because they don't scale. They scale fine in small classes, and they should be brought back. A five-minute oral defense after a major paper produces more evidence about whether the student understood their argument than another five pages of writing would. AI is structurally bad at fake-defending an argument it generated. A student who wrote their paper

genuinely can usually defend it. A student who didn't, usually can't.

Use AI as a feedback amplifier, not a grading machine. The role generative AI is best at in the classroom is the role of patient, tireless, immediate first reader. AI feedback on a draft — what is the thesis, what is the strongest evidence, what is the weakest evidence, what is missing, what would a hostile reader push back on — is often genuinely useful, and it can be delivered at 2 a.m. when the student is working. What AI is bad at is making the final grading judgment that affects the student's transcript and life. The Kestin et al. RCT's effect size came from the AI doing the tutoring; the grading remained with the instructor. That division of labor is the one most likely to produce both better learning and better learning evidence.

Calibrate peer review. Peer review is one of the most underused tools in higher education. It is also one of the highest-leverage. Done badly, it is busywork that students don't trust. Done well, it produces evidence of collaboration, evidence of communication, evidence of critical reading, and a powerful learning experience for the reviewer. The trick is calibration: train students on what good feedback looks like, give them rubrics, and assess them on the quality of their reviews as well as on their own work. The peer-assessment literature (much of it from Christopher Topping and colleagues) has shown for decades that calibrated peer review produces inter-rater reliability comparable to instructor grading on many tasks, and the AI era makes the case for it stronger, not weaker.

Audit the artifact, but lightly. Detection should not be the center of an assessment strategy, but selective integrity checks still have a role, especially on high-stakes assignments. The role is similar to the role of random tax audits: the existence of occasional careful scrutiny shapes behavior even when most submissions are never deeply audited. The audit cannot rely on automated AI detection alone, given what we now know about false positives. It can rely on inconsistencies between the artifact and the rest of the evidence: a polished final draft from a student whose drafts and discussion contributions all suggested they could not yet write that well is a signal to talk to the student, not to file a misconduct case.

Build in the boring stuff. Cumulative exposure to material matters. Spaced practice matters. Retrieval practice matters. AI does not eliminate these findings; if anything, AI raises their importance, because frictionless answer-generation makes it more tempting to skip the encoding work that turns information into memory. Quick, low-stakes retrieval quizzes, spaced over weeks, are the simplest and most evidence-based way to ensure that students are actually learning the substantive material, not just producing artifacts about it. The quizzes do not need to be high-stakes; they need to be frequent and to require effortful recall.

Talk about it openly with students. This is the move faculty most often skip and most often regret. Students are not the enemy. Most of them want to learn the things their courses are nominally trying to teach. Many of them are confused, often legitimately, about what is and is not allowed. A short, honest conversation at the start of a course about why the assignments are designed the way they are, what the instructor is trying to certify, what the AI policy is, and what the consequences of misuse will be does more to shape student behavior than any technical countermeasure. It also models the kind of professional thinking the course is supposed to develop.

These moves are not exotic. They are mostly old. What is new is the urgency. Before generative AI, instructors who used these practices were doing exceptional teaching. After generative AI, instructors who don't use them are doing weak assessment, regardless of whether they realize it. The bar has moved.

The good news is that the bar has moved toward the kind of teaching most teachers came into the profession wanting to do. Less grading of artifacts that may or may not represent student thought. More conversation. More feedback. More visible learning. The AI era is a forcing function on a transition the research literature has been quietly recommending for thirty years.

Part IX. The institutional layer

What individual faculty can do, by themselves, has limits. The transition learning intelligence describes is also an institutional transition, and the institutions that get this right will get it right at the system level. The ones that don't will keep treating each new wave of AI capability as a discrete crisis to be managed by the academic affairs office.

The institutional response has so far been mixed. EDUCAUSE's 2025 AI Landscape Study found that 57% of higher-ed institutions now consider AI a strategic priority, up from 49% the previous year. The proportion with AI Acceptable Use Policies climbed from 23% to 39% in a single year. And yet only 9% reported that their cybersecurity and privacy policies were adequate for AI-related risks. The infrastructure is catching up to the urgency, slowly. Most provosts and CIOs would say it is not catching up fast enough.

What should institutions actually do? Five priorities, in roughly the order they typically need to be addressed.

One: get clarity on what the institution wants to certify. Most colleges have learning outcomes documents. Most of those documents are aspirational and rarely consulted. The first step toward a defensible AI-era assessment strategy is to take those outcomes seriously enough to ask, for each one, what evidence the institution is actually generating, and whether that evidence is robust to AI. This is a faculty-governance conversation as much as it is an administrative one. It cannot be outsourced.

Two: align assessment design at the program level. Individual courses can do good assessment in isolation, but a degree certifies what the *program* did, not what each course did. Programs need to map outcomes to courses, identify where each outcome is taught, practiced, and assessed, and ensure that the cumulative evidence supports the credential. This is not new — the accreditation literature has been pushing this for decades — but generative AI has made it suddenly more urgent. A program where every course relies on take-home essays as its primary assessment is now structurally vulnerable. A program that distributes assessment across written, oral, project-based,

and applied work is much more robust.

Three: fund the infrastructure for process evidence. Capturing process evidence is more expensive than capturing artifact evidence. Draft histories take storage. Peer review systems take licenses. Oral defenses take time. Reflection components take rubrics. The institutions that succeed will treat this as a capital expense, not an operating burden on individual instructors. The institutions that fail will quietly require faculty to do this work in addition to everything else they were already doing, at which point most faculty will quite reasonably refuse, and the assessment redesign will not happen.

Four: build the data governance before you need it. Every learning intelligence system is also a student data system, and student data is regulated under FERPA in the United States, GDPR in Europe, and equivalent regimes elsewhere. The questions that matter — what data are collected, how long they are retained, who can see them, what inferences can be made from them, how students can challenge those inferences — should be answered before the platforms are bought, not after. The Digital Education Council's 2024 global survey found that 80% of students said AI in universities was not fully meeting expectations, with 60% specifically worrying about the fairness of AI evaluations and 70% citing privacy as a major concern. These are not abstract policy questions. They are trust prerequisites. A system students do not trust will not produce honest evidence, because students will hide their actual learning behavior from it.

Five: invest in faculty development that respects faculty time. EDUCAUSE found that "training for faculty" (63%) topped the list of AI-related strategic priorities at most institutions. The training that works is not generic AI literacy. It is discipline-specific assessment redesign, run by colleagues who teach similar courses, with concrete examples, model assignments, and time to revise actual syllabi. The institutions that have invested seriously in this — pairing instructional designers with faculty cohorts, offering course buyouts for redesign, supporting communities of practice — are the ones seeing the most substantive curriculum change. The institutions that have offered a webinar and a policy document are mostly still where they were three years

ago.

There is a sixth priority that is harder to name but more important than any of the others: **the institution has to decide what business it is in.** The transactional model of higher education — pay tuition, complete assignments, receive credential — has been weakening for years, and AI accelerates the weakening. If the credential is going to mean anything in 2030, it will need to mean that the institution can credibly say what the holder of it can do. That means the institution needs to know what its graduates can do. Which means it needs to have evidence. Which means it needs to invest in producing that evidence, deliberately, at scale, across its curriculum.

This is the institutional version of the personal argument: when output becomes cheap, evidence becomes everything. The institutions that act on this will be in a strong position. The ones that don't will find that their credentials are slowly losing their power to certify anything an employer or a graduate school cares about, and that nobody can quite say when it happened.

Part X. Where to look in the field

A category is more real when you can see who is building inside it. The market for learning intelligence and AI-era assessment is still forming, and the categories below are more useful than vendor names — most companies fit primarily inside one of them, and the test for any product is what it actually does, not what it markets itself as.

The five categories that follow were derived by reading the field across higher education and K–12 vendor materials, peer-reviewed efficacy research where it exists, and the institutional buying patterns visible in EDUCAUSE, Tyton Partners, and HEPI data. Each category solves a different piece of the assessment-after-AI problem. Each has a place in a well-considered assessment architecture. None of them, on its own, is yet a complete answer.

The capability matrix

The matrix below is the fastest way to read the market. Categories run down the rows; the seven capabilities a learning intelligence platform needs to deliver run across the columns. A buyer can read it in under a minute and immediately see what each category gives them and what it leaves on the table.

Category	Process evidence	Construct mapping (4Cs)	Multiple assignment types	Faculty-readable	Accreditor-defensible	Institutional rollups	LMS integration
Process-native LI	Yes	Yes	Yes	Yes	Partial	Partial	Yes
Integrity-first & detection	Partial	No	No	Partial	Partial	Partial	Yes
LMS-native analytics	No	Partial	No	Partial	Yes	Yes	Yes
AI tutoring	Partial	No	Partial	No	No	No	Partial

Category	Process evidence	Construct mapping (4Cs)	Multiple assignment types	Faculty-readable	Accreditor-defensible	Institutional rollups	LMS integration
K-12 instructional	Partial	Partial	Partial	Yes	No	Partial	Partial

Legend: Yes Full capability · Partial Partial · No Not in the category

The five categories, explained

01 · The newest — Process-native learning intelligence. Products in this category capture evidence of student thinking inside specific high-cognition assignments — discussion, writing, peer review, oral defense, AI-integrated work — and map those signals to learning constructs like the 4Cs and AI literacy. They are pedagogically opinionated by design and are the only category architecturally aligned with the "process over artifact" thesis the rest of the field is now converging on. *Strength*: produce defensible evidence that learning happened, before the final artifact is submitted. *Gap*: cross-program institutional reporting at the scale of LMS analytics is still expanding. Most platforms stop at the assignment or course level. *When to look*: when the question is "how do we know learning happened" and the artifact alone is no longer trustworthy. *Representative players*: Packback, Cadmus, FeedbackFruits, Perusall, Kritik.

02 · The incumbent — Integrity-first and detection. The established category for academic-misconduct workflow. Products here detect similarity, identify likely AI-generated text, capture authorship transparency, and provide audit trails for hearing processes. The category leaders are now pivoting toward process visibility (notably Turnitin's Clarity product), recog-

nizing that detection alone is structurally limited. *Strength*: surface signals that a piece of work may not be the student's own; maintain the integrity infrastructure accreditors still expect. *Gap*: cannot prove learning occurred (only flag where it may not have). Vanderbilt's 2023 decision to disable Turnitin's AI detector on fairness grounds was an early signal of waning institutional confidence; non-native English speakers continue to face 2–3× higher false-positive rates. *When to look*: when the institution still needs an integrity workflow for high-stakes assessments, with full awareness of false-positive risk and an explicit policy that detection alone is not sufficient evidence for misconduct. *Representative players*: Turnitin, Originality.AI, GPTZero, Copyleaks.

03 · The institutional layer — LMS-native analytics. The reporting infrastructure where provost- and CIO-level conversations already happen. Products here aggregate course-level data — submissions, grades, page views, login frequency — into dashboards that flag at-risk students and report on course health and outcome trends. They are where institutional accreditation reporting is already structured. *Strength*: provide the institutional view; native data flow from existing LMS adoption; already trusted by accreditors for outcome reporting. *Gap*: engagement counts are not learning measurements; participation and login frequency are weak proxies for whether anything has been learned. The 24–48 hour data lag also makes them retrospective by design, not real-time intervention tools. *When to look*: for the institutional reporting layer, not as a substitute for assignment-level evidence. Best deployed as the destination that ingests process evidence from elsewhere. *Representative players*: Canvas Intelligent Insights, D2L Achievement+ and Lumi, Anthology Analytics for Learn.

04 · The fastest-growing — AI tutoring. The category propelled by both consumer interest and pedagogically engineered systems. Products here scaffold individual study sessions, generate explanations, provide practice problems, and deliver immediate formative feedback. Some are extraordinarily effective when carefully designed; others are general-purpose chatbots in a study skin. *Strength*: personalize learning at scale. The Kestin et al. 2025 RCT

in *Scientific Reports* showed a carefully designed AI tutor producing learning gains roughly twice those of in-class active learning. *Gap*: AI tutoring is an input to learning, not a measurement of it. A student who learned a topic through a tutor still needs an assessment surface that captures whether the learning held without the tool. *When to look*: for personalized study support, supplemental instruction, and outcomes-aligned remediation — not for grading, certification, or program-level assurance. *Representative players*: Khan Academy and Khanmigo, ChatGPT Study Mode, Squirrel AI, Carnegie Learning.

05 · The K-12 wedge — K-12 instructional intelligence. The category that has most actively claimed the "learning intelligence" phrase, generally inside K-12 curriculum and lesson workflows. Products here align lessons to standards, deliver classroom-level visibility to teachers, and provide AI productivity tools for routine instructional tasks. The K-12 context — younger students, more standardization, less faculty autonomy — is structurally different from higher ed, and tools built primarily for it often do not generalize directly upward. *Strength*: standardize AI use across schools and districts; reduce teacher workload on lesson planning and feedback generation; strongest district-level adoption motion in the field. *Gap*: most are not yet evidence-architected for cross-program assurance in the way higher education will require. Strong in the classroom, weaker in the program-review or accreditation layer. *When to look*: for K-12 specifically, especially districts standardizing AI policy and seeking teacher-productivity gains. *Representative players*: Kiddom, Subject, MagicSchool, Brisk.

Convergence is happening

The five categories overlap more than they used to. Integrity vendors are adding process-visibility features. LMS vendors are partnering with AI providers to embed tutoring and feedback directly. Process-native platforms are extending upward into the institutional reporting layer. K-12 vendors are eyeing higher ed.

The convergence is not random. Every category is moving toward the same

destination: a system that captures evidence of how learning is happening, maps it to defensible constructs, and produces interpretable reports for the people who need them. That destination is what this guide has been calling learning intelligence throughout. The categories represent different starting points, not different end states.

This matters for the buyer because it means a vendor's category of origin tells you what they will do best in 2026 — but the maturity of their other layers tells you what they will be able to do in 2028. Integrity vendors moving into process visibility, LMS vendors moving into construct mapping, and process-native vendors moving into institutional rollups are all making the same bet on the same destination. The question is which ones will actually arrive.

Eight questions to ask any vendor

The most useful thing a buyer can take into a vendor evaluation is not a competitor matrix but a short list of architectural questions. The following hold up regardless of which vendor is sitting across the table.

1. Does the product capture process evidence, or only final artifacts?
2. Can you map signals to specific learning constructs (such as the 4Cs and AI literacy), or only to engagement counts?
3. Can a faculty member see *why* a student was flagged or scored, in terms of observable behavior?
4. Can a student see their own evidence, and contest incorrect inferences?
5. Does the platform host multiple assignment types, or only one?
6. Is the data architecture defensible to an accreditor, with rollups by course, cohort, and program?
7. What student data is captured, how long is it retained, and who has access?
8. Is any high-stakes judgment — grades, integrity findings, intervention referrals — held by humans?

The vendor that can answer all eight questions clearly, in plain language, without sales evasion, is the vendor worth talking to longer. The vendor that

responds to half of them by talking about how innovative their AI is — that's the answer to the question you actually asked.

Part XI. Open questions and limits

Honesty requires acknowledging what the field of learning intelligence does not yet know, and what its risks are.

The research base is still young. Most of the strongest empirical studies on AI in higher education are short-term, often in specific disciplines (often language learning or introductory STEM), often with carefully engineered interventions that may not generalize to the chatbot a typical student uses on a typical Tuesday night. The 2025 meta-analyses are encouraging but not yet definitive. The Kestin RCT is the strongest single piece of evidence we have for AI as a learning amplifier, and it is one study, in one course, with a custom-built tutor. Larger and longer trials are coming, but the field's current claims should be held with appropriate humility.

The process-data and stealth-assessment literatures are robust as frameworks but uneven as implementations. Most existing edtech "process data" is engagement counting in disguise. The hard work of mapping events to constructs, validating those mappings, and demonstrating that the resulting inferences are fair across student populations is mostly still ahead of us. The OECD process-data work has been explicit that valid process measurement requires investment in psychometric validation, fairness audits, and transparent scoring logic, none of which most current edtech products invest in seriously.

Affective AI — the use of emotion recognition, facial expression analysis, sentiment analysis, or other "feeling-aware" technologies — is not ready for high-stakes educational use. Multiple recent systematic reviews, including Shingjergji and colleagues' 2026 review of 96 studies, have found that affective computing in education tends to study engagement, confusion, and frustration from facial expression CNNs, often without real classroom validation

and almost always without serious ethical analysis. Privacy concerns are large. Accuracy in real classrooms is unproven. The safe use case is opt-in, low-stakes, instructor-facing support; the unsafe use case is anything resembling automatic grading of "engagement" or "attention" from a webcam. A learning intelligence system that ventures into affective data should do so cautiously, transparently, and with strong governance.

Equity outcomes are unclear. There is real evidence in both directions. A within-subject writing experiment by Tukachinsky Forster and colleagues in 2025 found that all students benefited from AI assistance but less-skilled writers benefited more, suggesting AI could narrow some performance gaps. The Digital Education Council and HEPI surveys, on the other hand, show socioeconomic divides in usage patterns, with students from higher-income backgrounds and well-resourced institutions accessing better tools, more support, and clearer policies. The 2026 Elon/AAC&U faculty survey found that faculty broadly expect AI to widen digital inequities. Whether AI is a leveler or an amplifier of existing inequalities is probably context-dependent, and learning intelligence systems should be designed to monitor and audit their own equity effects rather than assume them.

Privacy and proportionality are the chronic risks. Capturing process evidence at high fidelity creates real surveillance risks. Keystroke logging, screen recording, full prompt capture, network analysis, and behavioral inference all sit on a continuum from useful to oppressive, and the line moves depending on context, student age, jurisdiction, and what the data are used for. The safe defaults are data minimization (collect only what you need for the specific claim you are trying to make), bounded retention (don't keep data longer than required), redaction (strip PII from anything not strictly needed), role-based access (instructors see what instructors need; administrators see what administrators need), and auditable model use (log what inferences are being drawn from what data, and let users see those logs). A learning intelligence system that does not respect these defaults will produce backlash that may delay the entire field.

The category itself is contested. "Learning intelligence" is already being used

by multiple companies, including 1EdTech (as "predictive and prescriptive learning intelligence"), Kiddom ("Learning Intelligence Technology"), Subject ("built-in learning intelligence"), and Brisk (under the related label "curriculum intelligence"). The phrase has not been claimed by any single body, and probably will not be. The thing that matters is not who owns the phrase but who builds the practice. The practice can succeed under several names. What cannot succeed is treating the phrase itself as a moat.

And finally: the assessment crisis is not the only crisis. Higher education in 2026 is also navigating enrollment declines, financial pressure, political scrutiny, accreditation reform, the unbundling of credentials, and a changing labor market. Learning intelligence is part of the response to one of those forces — the validity crisis in assessment — but it does not solve any of the others. The institutions that treat AI as the whole crisis will misallocate. The institutions that treat AI as the forcing function on a much bigger transition will probably do better.

Part XII. Closing

There is a temptation, in a moment like this, to be either apocalyptic or utopian. Either the universities are ending and the AI is winning, or the AI is liberating and the old gatekeepers should get out of the way. Both moods miss what is actually happening.

What is actually happening is that a measurement system whose limits had been quietly tolerated for decades has now broken in public. The old contract — submit the artifact, receive the grade, accumulate the credential — relied on a scarcity that has been quietly removed. Faculty are not wrong that something has been lost. Students are not wrong that something has been gained. Both are responding to real features of the situation.

The way out is not nostalgia and not utopia. The way out is to take what we have always known about how people learn, and what we have always known about how to assess what they have learned, and finally to build the

infrastructure that takes those things seriously. The research has been telling us for thirty years that learning is a process, that feedback is the most powerful intervention we have, that the 4Cs are practices, that authentic assessment is the only assessment that produces durable evidence. We did not act on it at scale because the artifact-only model was good enough. It is no longer good enough.

Learning intelligence is the name being attached, for now, to the work of building the new system. Whether the term sticks is less important than whether the practice does. The practice is captured in a few principles that have been the through-line of this entire piece. Watch the process, not just the product. Map signals to constructs, not to clicks. Treat evidence as something the human reads, not something the system decides. Be transparent about what is captured and why. Keep humans in the loop for any decision that matters.

A teacher in 2026 who designs an assignment that produces six points of legible thinking, gives students explicit AI use guidelines and a metacognitive reflection prompt, calibrates peer review, requires a short oral defense, and uses AI as a tireless feedback amplifier between drafts is not doing something exotic. They are doing the kind of teaching that the research literature has recommended since before most of their current students were born. The difference is that, in 2026, they are also doing the only kind of teaching whose evidence still holds up.

The institutions that fund this work, that align their assessment infrastructure to it, that protect their faculty's time to do it, and that build the data governance to support it without slipping into surveillance, will be the institutions whose credentials still mean something at the end of the decade. The institutions that don't will continue to graduate students whose transcripts certify achievements the institutions can no longer credibly verify.

The post-output classroom is here. It has been here, in fragments, for years. What learning intelligence is for is to assemble those fragments into a system the next generation of students can actually be learners inside.

That is the work. It will take a decade. It is the most interesting work in education right now.

This article synthesizes peer-reviewed research, major institutional reports, and current sector survey data published between 2014 and early 2026. Specific studies cited inline include the Freeman et al. 2014 PNAS meta-analysis on active learning, Hattie and Timperley's 2007 Review of Educational Research feedback paper, Black and Wiliam's 1998 Inside the Black Box, the 2026 Elon/AAC&U faculty survey, the 2026 HEPI student survey, Tyton Partners' Time for Class 2025, the 2025 EDUCAUSE AI Landscape Study, the Kestin et al. 2025 Harvard AI tutoring RCT in Scientific Reports, the OECD Digital Education Outlook 2026, the Digital Education Council 2024 Global Student AI Survey, PISA 2022 creative thinking results, the AAC&U VALUE Rubrics, the Partnership for 21st Century Learning Framework, the Perusall annotation-and-performance study in Frontiers in Education, OpenAI's Study Mode launch announcement, Vanderbilt's 2023 statement on disabling AI detection, UNESCO's 2023 Guidance for Generative AI in Education and Research, TEQSA's 2023 Assessment Reform principles, the U.S. Department of Education's 2023 AI report, Gerlich's 2025 study on AI tools and cognitive offloading in Societies, and the SoLAR 2025 updated definition of learning analytics.