

RETHINKING ASSESSMENT AND MEASUREMENT IN THE AGE OF GENERATIVE AI

Ronald J. Glotzbach

Purdue University / Associate Professor (USA)

Abstract

The rapid integration of Generative Artificial Intelligence (AI) into higher education has intensified longstanding questions about academic integrity, authorship, and the validity of traditional assessment practices. As AI systems become capable of producing essays, programming code, and analytical responses that resemble student work, product-based evaluation models face increasing strain. Detection-based approaches alone offer limited assurance of learning, and in some cases may distract them from deeper pedagogical concerns. This paper argues that assessment and measurement frameworks must evolve to prioritize visible reasoning, verification processes, and structured reflection rather than focusing solely on final artifacts. Rather than positioning generative AI as an external threat to be managed, this paper explores how assessment design can incorporate AI use in ways that preserve rigor while making student thinking more transparent. Drawing on the redesign of a discipline-specific undergraduate assignment, the paper illustrates how requiring students to document AI prompts, test system-generated outputs, identify inaccuracies, and justify revisions transforms AI interaction into assessable evidence of learning. In this model, grading criteria shift toward constraint specification, error detection, conceptual explanation, and reflective decision-making. Students are evaluated not merely on whether a solution works, but on how they engaged with the tool, where they identified limitations, and how they resolved discrepancies between AI output and disciplinary expectations. That reframing aligns assessment with contemporary learning environments in which AI is readily accessible. Instead of attempting to restore pre-AI conditions, educators can design measurement structures that foreground intellectual engagement. By embedding documentation requirements and staged submission processes into assignments, instructors can observe developmental progression and analytical judgment. Such strategies also support fairness and consistency by clarifying expectations and reducing ambiguity around acceptable AI use. The paper further discusses implications for rubric development, formative feedback, and academic standards. When assessment emphasizes reasoning and verification, it becomes possible to maintain expectations of mastery even in AI-rich contexts. That shift also encourages a broader reconsideration of what educational measurement captures: not only knowledge recall or output production, but the processes through which learners interrogate tools, apply constraints, and exercise disciplinary judgment. As generative AI continues to reshape academic work, assessment models must move beyond enforcement toward intentional design. By re-centering evaluation on cognitive processes and documented engagement, educators can respond to technological change without diminishing academic expectations.

Keywords: *Assessment innovation, Generative AI, academic integrity, instructional design.*

1. Introduction

One of the first things many instructors notice once students begin using generative AI regularly is that completed assignments become harder to interpret. In previous semesters, a weak essay, an incomplete programming assignment, or an incorrect problem solution often revealed something meaningful about where a student was struggling. Now, students can generate polished written responses, functional code, and detailed explanations within seconds using conversational AI systems. Sometimes the generated work is accurate. Sometimes it only appears accurate until examined more carefully. Either way, the relationship between the submitted artifact and the student's actual understanding has become less clear. That shift creates growing uncertainty around what instructors are truly measuring when they evaluate student work in AI-supported learning environments (Holmes et al., 2019; Zawacki-Richter et al., 2019).

While many conversations about generative AI in education center around issues of academic integrity, detection, and policies and procedures, many instructors and administrators are looking for ways to operationalize the use of generative AI in the classroom. The issue is therefore not simply whether AI is present, but how its presence changes the visibility of student thinking during assessment. Traditional assignment-based assessments in the classroom typically rely on students demonstrating their critical thinking skills by writing thoughtfully and then revising their work, completing the assignment in accordance with assignment instructions. Yet, as more and more of a completed assignment is generated by an artificial intelligence program after merely receiving topic and parameters from a prompt, the challenge to these traditional assignment-based assessments becomes apparent. Recent discussions of academic integrity in the age of generative AI have similarly argued that traditional assessment approaches are becoming increasingly difficult to sustain without rethinking how student learning is evaluated and documented (Cotton et al., 2024).

For those using assignments as evidence of learning, the traditional grading practices assume a lot of cognitive processes took place as students engaged in the development of the artifact (i.e., creation, revision, and final production). Students plan, analyze, and solve problems while composing, revising, and refining the artifact as they work to meet assignment requirements and expectations. What happens when Generative AI enters the educational ecosystem and allows students to complete assignments and submit them as “perfect” without engaging in any of these processes? Instructors will only know what appears on the surface of the completed work without any insight into the cognitive processes that occurred (or failed to occur) during development. Similar concerns have emerged in broader discussions of AI-supported learning environments, particularly regarding student autonomy, critical evaluation, and the changing role of human judgment in academic work (Popenici & Kerr, 2017; Luckin, 2025).

There has been a surge of research on the detection of AI-generated content, but even if content is detected as having been generated with the help of AI, that is only half the battle for assessing learning. More important are questions about student behavior when using such technology, what students learn about both the technology and the subject matter as a result of using it, and how students critically evaluate the product of AI. An overemphasis on detection could lead us to overlook these critical questions and divert our attention from assessing learning in a world where AI is increasingly prevalent. The more important question is: Do we still see evidence of reasoning, evidence of reflection, and understanding of key concepts and processes that are discipline specific?

There has been a growing interest in how to assess processes instead of just the end product. To encourage students to make their learning process visible to teachers, assignments can be designed to solicit staged work, documentation of prompts, explanation of how student used the teacher assignment prompt, descriptions of and justifications for any revisions, and verification of certain learning outcomes. In addition, interaction with AI can be incorporated into the portion of the learning that is assessed. Initial efforts to redesign assignments in non-technical and technical courses revealed that having students document the prompt, verify their output, and explain their revisions provided a better sense of their reasoning skills than viewing the end product (Glotzbach, 2026). Thus, the process of learning was given as much weight as the final product.

This paper argues that assessment practices must evolve in learning environments where generative AI is readily accessible. Rather than focusing primarily on detection or attempting to restore pre-AI conditions, educators may need to redesign assignments so that reasoning, verification, and disciplinary judgment become more visible during the learning process itself. The discussion examines how documentation, staged submissions, reflective explanation, and verification activities can function as assessable evidence of student thinking in AI-supported coursework.

2. Assessment and the visibility of student thinking

Assessment in traditional instructional settings generally assumes that some portion of what a student does at the end of a learning effort is a sample and/or record of that learning process. Therefore, the products of student work, such as essays, mathematical calculations, computer programs, data analyses, and other forms of academic activity, serve as indicators of students’ thoughts, understandings, skills, and experiences developed during the learning process. Instructors can observe errors, inconsistencies, incomplete solutions, and unsuccessful attempts to solve problems and interpret them as evidence for what students do and do not know. This is particularly true in disciplines in which the development of the artifact and the artifact itself are fundamentally linked. Generative AI complicates this relationship because students can now produce polished and technically plausible work without engaging fully in the cognitive processes instructors intended the assignment to develop (Holmes et al., 2019).

This shift becomes especially noticeable when assignments are evaluated primarily through product-based grading models. Traditionally, instructors assess the final submission and use it as evidence of mastery, reasoning, or conceptual understanding. But, as technology improves and AI tools become more sophisticated, it's becoming increasingly easy to produce high-quality, impressive products that can fool even the most seasoned teacher into thinking that a student has a deep understanding of a subject. The final artifact alone is not enough to prove student engagement with key concepts of the assignment. By the time an instructor sees the finished product, there may be very little evidence of what students have learned.

Recent conversations around what assessments of student work with AI reveal about what students are learning suggest a growing recognition that a solely surveillance-based assessment model is not sufficient. We used to look at student work and determine if it had been generated by AI or not. However, as we are now designing more assignments labeled “AI free” to supposedly encourage better skills with traditional forms of writing, we are beginning to realize that even these assignments do not demonstrate the type of analytical thinking, thorough revision, concept-based understanding, and other goals we want for our students. On the other hand, work that allows for extended use of AI can produce work that actually demonstrates the type of meaningful intellectual work that students do in order to use these tools well. The central instructional issue is therefore less about the mere presence of AI and more about whether assessment structures make student thinking sufficiently visible.

As teachers integrate AI into their practices, the types of evidence that they might use to measure what students have learned could also change. In process-oriented learning environments, such as those in which students write prompts for an AI and then refine their work before submitting the final product, over the course of a number of models, students provide evidence for skill(s) like forming disciplinary questions, evaluating sources, setting constraints, or making a variety of other decisions that are characteristic of a field. Much of the very best evidence, however, comes from places where students recognize errors, revise because of those errors, and explain why a generated answer is insufficient.

This does not mean that final products no longer matter. Accuracy in the disciplinary content that we teach, accuracy in the way that we use technical language, accuracy in the quality of the final products that students produce – these are all concerns that are just as pressing, and just as difficult to achieve, as they have always been. However, the growing accessibility of generative AI suggests that assessment models focused exclusively on end products may provide increasingly incomplete pictures of student learning. Assessment practices may therefore need to shift toward structures that balance evaluation of completed work with evaluation of the reasoning processes surrounding that work. Rather than treating process documentation as supplemental material, instructors may need to recognize it as central evidence of intellectual engagement in AI-supported coursework (Popenici & Kerr, 2017; Zawacki-Richter et al., 2019).

3. Process-oriented assessment in AI-supported coursework

One response to the growing difficulty of interpreting student artifacts is to redesign assignments so that reasoning processes become more visible throughout completion of the work. Rather than evaluating only final submissions, instructors can require documentation, staged development, verification, and reflective explanation. In this model, assessment shifts from evaluating only finished products toward evaluating how those products were developed, tested, and revised.

Early redesign efforts involving structured AI integration within technical coursework demonstrated that requiring students to document their interaction with AI systems produced a very different kind of assessable evidence than traditional submission models (Glottzbach, 2026). In addition to completing all of the assignments, students were required to share their prompts, initial AI-generated outputs, revised outputs, and a rationale for any changes made. After grading and reviewing all of the assignments, one thing was more apparent than any other: it was how the students' decisions made after receiving the initial outputs revealed whether or not they had developed sufficient understanding of their discipline. Some of the students recognized quickly the faulty structure of their initial outputs and spent considerable time and thought to revise and complete the assignments more effectively. The opposite was true of some other students who hurried to accept and turn in their outputs with many faults and failed to evaluate them critically.

In addition to feedback on design ideas, verification of the generated content was a huge aspect of the assignments. At first glance, many of the AI-generated content often looked extremely convincing, but upon closer reading had many errors, such as syntactic errors, incomplete ideas, missing constraints, and/or inconsistencies. The students who generally did the best on the technical assignments were those who tested their generated content rigorously, and used the resulting set of problems as additional material to further solidify and expand upon their own emerging understanding of the concepts. But some of the most valuable learning moments came when the generated content did fail – allowing students to debug, re-trace their

steps back to the fundamental concepts, and explain to themselves where things went wrong, and why what they asked the AI for was not what they received. Productive struggle remained part of the learning process, even though AI had become involved in generating the initial output (Kapur, 2008).

In addition to these new ways to use AI to assess student work, seeing student prompts provided to the AI to receive a response gave yet another perspective on how students were thinking about the problem, what they envisioned as the task's boundaries, and how they could communicate their intentions and goals to a non-human problem solver. Moreover, seeing how students developed prompts exposed another aspect of student reasoning that the instructor could assess. Prompting itself became part of the intellectual work of the assignment rather than an invisible step occurring outside the instructor's view.

In addition to these features, a staged submission format for assignments provided increased visibility into student learning. Rather than submitting a finished assignment at the end of the learning process, the assignment format described here allowed for multiple points at which students could submit work for evaluation of student engagement and understanding. Some of these included first drafts on which students had done some work and then subsequent submissions after having made revisions, as well as students' logs from the testing and iteration process, their written reflections on the assignment, and the final version of the assignment. This format provided an opportunity for students to learn in significant ways and for instructors to see that learning, thus shifting the focus from a static end product to a developmental process of learning, application, formation of point of view or argument, and use of evidence from appropriate sources.

In documenting student work from a processing perspective, it is easy to overlook the accuracy and completeness of what has been produced from the perspective of the discipline. Even as students achieve high levels of achievement by using external aids, such as AI, and drawing on content knowledge, the documentation can show how their work relates to their emerging understanding of that content. For example, what looks from the product perspective like minimal understanding of a problem may reveal significant knowledge of that problem when viewed through the lens of prompt history, revisions, and verification activities. Rather than asking if students are using AI, we should be paying more attention to how they are using AI and what level of cognition is evident in that use.

As AI tools continue to evolve, assignments that make reasoning processes visible may become increasingly important across disciplines. The broader instructional challenge is no longer simply determining whether AI was involved in producing student work. Instead, educators may need to focus more directly on how assessment structures capture evidence of judgment, revision, verification, and analytical reasoning in learning environments where AI interaction is becoming routine.

4. Implications for assessment and measurement

As generative AI tools are incorporated into academic environments, the traditional ways of assessing learning outcomes are being forced to adapt. In many academic disciplines, instructors assess student learning based on the products that students create and complete throughout their studies. Whether it be an essay, project, lab report, programming assignment, or solution to qualitative or quantitative problems, the end products serve as indicators of what students have learned. As generative AI tools become more prevalent in academic environments, however, the issue of assessing learning outcomes when the same end products can be produced by a machine complicates this approach, indicating that a shift in assessment models is necessary. That shift does not necessarily diminish the importance of final products, but it does suggest that assessment models built entirely around artifact evaluation may provide increasingly incomplete pictures of learning.

The ability to process information and complete tasks using computer-assisted systems requires a change in assessment criteria as well. In addition to using disciplinary correct solutions to grade student output, teachers must also use criteria such as verification, explanations, revisions, and evidence-based decisions and reasoning. In other words, in addition to correctly completing problems using AI-supported systems, students should receive credit for the processes they employed to reach the final answer. This means that students should receive credit for identifying errors, making revisions based on strong reasoning and evidence, constraining the output of AI systems to meet particular objectives, and explaining the quality of their revisions. Teachers must recognize that students are learning critical thinking skills to process information generated outside of a student's control and are, therefore, developing a unique set of competencies. With AI generators becoming increasingly sophisticated and convincing, it is becoming more important for teachers to understand the difference between student-generated content and AI-generated content.

This way of organizing assignments and assessing student work also affects how formative assessment and feedback will take place. Students' work can offer a window into their growing abilities, revealing the thinking that went into responses to prompts, the revisions they made along the way, and their sense of how they verified what they read. Teachers will see feedback on assignments evolve from a focus on accuracy of final product to a deeper evaluation of students' reasoning and intellectual decisions made throughout the assignment. Educators will be better able to identify student misconceptions, examine patterns in their decision making, and see whether students are relying on AI to complete assignments. Assessments become more holistic, giving teachers the ability to see how students are growing in ability to understand complex topics throughout the assignment.

Students often struggle to understand the boundaries of using outside AI assistance and academic dishonesty. Process-oriented assessments aim to mitigate some of this confusion. Assignments can be structured to clearly incorporate key expectations into assessments, by documenting prompts, submissions over time, and explanations of use of AI for both process and product. This structure enables assessment of students' use of AI for discipline-specific reasoning rather than simply determining if students are using AI.

There are broader questions about what educational measurement captures when generative AI enters the scene and transforms the learning environment in which students are educated. While measures of recall, production, and accuracy may still be relevant in the near term, the evidence suggests that going forward measures will also focus on the student's ability to critically evaluate the product produced by the AI; use it as the basis for revising and improving one's own work, recognizing limitations of AI generated content; and then support claims made with evidence and sound argument.

This does not mean that AI should replace disciplinary rigor or lower academic standards. If anything, AI-supported learning environments may require even greater emphasis on analytical reasoning and conceptual understanding. Students who cannot evaluate generated outputs critically are often unable to recognize inaccuracies, inconsistencies, or misleading information. The presence of AI therefore increases the importance of disciplinary judgment rather than diminishing it. Assessment models that foreground reasoning, verification, and reflection may provide one way of preserving rigor while still acknowledging the realities of contemporary learning environments (Holmes et al., 2019; Luckin, 2025).

5. Conclusion

Generative AI is reshaping many of the assumptions that have traditionally guided assessment and measurement in higher education. As students gain access to systems capable of producing polished written responses, technical solutions, and analytical content, final products alone may no longer provide sufficient evidence of student understanding. This paper argued that assessment practices in AI-supported learning environments may need to place greater emphasis on reasoning, verification, revision, and reflective explanation. Rather than focusing primarily on detection or attempting to recreate pre-AI conditions, instructors can redesign assignments so that cognitive processes become more visible throughout the completion of the work itself. Process-oriented assessment does not reduce academic rigor. Instead, it may increase the importance of analytical reasoning, disciplinary judgment, and conceptual understanding. As generative AI continues to influence academic work, the challenge for educators will increasingly involve designing assessment structures that still allow meaningful evidence of learning to remain visible.

References

- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239.
- Glottbach, R. J. (2026). Embedding AI Literacy within Technical Coursework: An Instructional Design Approach. *Proceedings of the 37th international conference of the Society for Information Technology and Teacher Education (SITE 2026)* (pp. 1203-1207). Philadelphia, PA.
- Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education. In M. Hildebrandt, J. W. B. G. Peters, & B. van den Hoven (Eds.), *Data ethics: Building trust* (pp. 619–653). Globethics.
- Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379–424.
- Luckin, R. (2025). Nurturing human intelligence in the age of AI: rethinking education for the future. *Development and Learning in Organizations: An International Journal*, 39(1), 1-4.
- Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 1–13.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, 16(1), 1–27.