Jenkins, D., & de Mello-e-Souza Wildermuth, C. (2026). Leading continuous improvement: Generative artificial intelligence and assessment in higher education. *Intersection: A Journal at the Intersection of Assessment and Learning, 7*(5), 92–114. https://doi.org/10.61669/001c.172639

Leading Continuous Improvement: Generative Artificial Intelligence and Assessment in Higher Education

Daniel Jenkins, Ph.D., Cristina de Mello-e-Souza Wildermuth, Ed.D.

Author Note

Daniel Jenkins, https://orcid.org/0000-0002-5895-2587

Cristina de Mello-e-Souza Wildermuth, https://orcid.org/0000-0002-9408-3673

We have no conflicts of interest to disclose.

Intersection: A Journal at the Intersection of Assessment and Learning

Vol 7, Issue 5 Special Issue: The Future of Assessment in an Era of Generative AI

Abstract: Assessment is most powerful when understood as a leadership-driven process of continuous improvement—iterative, evidence-informed, and oriented toward genuine learning rather than compliance. However, the coordination costs of meaningful assessment have historically exceeded most programs' capacity to sustain it. This paper examines how generative artificial intelligence, and in particular retrieval-augmented reasoning environments, can function as leadership-support tools that change the conditions under which assessment becomes practically feasible without displacing the human judgment it requires. Drawing on leadership education and assessment scholarship, we develop a three-level framework—individual and course, program, and institutional—and present six practice-based use cases illustrating AI-supported assessment at each level, including an AI-populated curriculum alignment matrix and an AI-assisted synthesis of external program reviews. We also directly address the tensions practitioners must navigate: data contamination from AI-generated student work, reliability-validity trade-offs, the limitations of AI detection tools, and the faculty professional development gap. Implications for assessment leaders are discussed.

Keywords: *generative artificial intelligence, assessment leadership, continuous improvement, program review, curriculum alignment, higher education*

Introduction

Ask most faculty members about their program's assessment process, and you are likely to hear a familiar story. Evidence exists—somewhere. Someone has the old syllabi, the rubrics used two years ago, and the graduation data that the institutional research office pulls each fall. However, assembling it into a coherent picture of whether the program is actually developing the capacities it claims to develop is an enormous undertaking: a faculty lift that requires hunting down artifacts of learning, negotiating shared time in an already overburdened schedule, and sustaining the kind of collective reflection that genuine improvement demands. The result, too often, is an assessment that functions as compliance rather than inquiry—data collected to satisfy an accreditor, a dean, or a program review cycle rather than to answer the questions that faculty actually care about (Dugan, 2012; Ewell, 1988).

This is not a new observation. Scholarship in leadership education has long argued that assessment is most powerful when understood not as a discrete technical activity but as a leadership-driven process of continuous improvement—iterative, evidence-informed, and oriented toward genuine learning rather than mere accountability (Guthrie & Jenkins, 2018; International Leadership Association [ILA],

2026; Piatt & Woodruff, 2016). Brown and Varghese (2021) make the point with particular directness: without honest and rigorous measurement, institutions cannot make consistent progress in developing leaders. The gap between that aspiration and the reality of how assessment is actually conducted in most programs is one of the field's persistent challenges. It is not primarily a problem of motivation or expertise; it is a problem of process and resources—the time, infrastructure, and coordination that meaningful assessment requires rarely match the demands placed on the faculty and staff who are expected to carry it out.

The rapid integration of generative artificial intelligence (AI) into higher education has raised both familiar questions and genuinely new possibilities. A report of a recent four-person external review team conducting a two-day Academic Program Review of a small, research-intensive doctoral program in leadership studies at a mid-sized public university in the mid-Atlantic region serves as an example. Each reviewer took independent notes across more than a dozen stakeholder meetings. At the end of the second day, rather than spending hours reconciling four sets of notes before the following morning's stakeholder presentation, the team uploaded their collective notes into a persistent AI project environment. The resulting first draft of the closing presentation—and, later, the formal external report—was organized from those human-generated notes, not from the AI's interpretation of institutional documents. The distinction mattered: what the model produced was a synthesis of the review team's own observations organized against the program's self-study and critical questions. The tool did the work that usually falls to one exhausted team member at midnight: the time-consuming synthesis and report drafting that reviewers in Jenkins and Rocco's (2022) study cautioned can easily outrun the time available.

These possibilities arrive alongside legitimate concerns that assessment professionals must take seriously rather than treat as footnotes. The same widespread adoption of generative AI that creates new tools for faculty has also altered the evidence base on which program assessment has historically relied. When students use AI to produce the writing that instructors grade, the course-level data feeding into program assessment may no longer reflect what students actually know and can do. AI-generated student work may produce artificially elevated rates of apparent proficiency—particularly on writing-intensive outcomes such as critical thinking, communication, and the integration of theory and practice—that inflate program assessment results without reflecting genuine learning. Sorting out what assessment data actually means in a post-2022 landscape requires the kind of validity reasoning and contextual judgment that remains the domain of skilled assessment professionals (Brown, 2022; Rosch & Jenkins, 2020).

These possibilities are most tractable when AI is used within a retrieval-augmented reasoning environment—one in which a model reasons across a curated set of institutional documents rather than drawing on general knowledge. When assessment artifacts are compiled in a persistent project environment, the synthesis work that has historically made assessment so expensive can be reallocated to the human judgment that matters most. The result is not that assessment becomes easier in some global sense, but that specific, high-cost bottlenecks—aggregating evidence, detecting patterns, producing first drafts—become feasible for programs that could never have sustained them before (Xia et al., 2024).

This paper examines how generative AI can serve as a leadership-support tool for continuous improvement in higher-education assessment—not by replacing human judgment, but by changing the conditions under which that judgment can be exercised at scale. Drawing on leadership education and assessment scholarship and grounded in practice-based use cases, we develop a three-level framework for AI-supported assessment that operates at the individual and course, program, and institutional levels (Guthrie & Jenkins, 2018; ILA, 2026). At each level, we present concrete examples alongside an honest account of the tensions and cautions practitioners must navigate, and close with implications for assessment leaders seeking to strengthen rather than shortcut the essential work of evidence-based continuous improvement.

A Leadership-Informed Framework for Assessment

Assessment in leadership education is embedded in institutional contexts, shaped by program philosophies, and carried out by people who hold competing demands on their time and attention. Understanding it as a leadership-driven process—rather than a technical or compliance function—reframes both what we expect of it and how we evaluate its effectiveness. This section develops the conceptual scaffold that organizes the paper: a three-level framework for assessment practice, an iterative cycle that operates at each level, and a positioning of generative AI as a resource within that cycle rather than a substitute for human judgment.

Assessment as Continuous Improvement

Ewell's (1988) foundational argument—that assessment provides evidence of whether educational programs achieve their intended purposes—remains as relevant as it was four decades ago. However, subsequent scholarship in leadership education has elaborated on it considerably. Guthrie and Jenkins (2018) situate assessment within a broader framework of leadership learning and program development that emphasizes intentional design, evidence-based practice, and continuous improvement as interconnected commitments rather than sequential activities. The ILA General Principles for Leadership Programs (2026) make the connection explicit: leadership learning and program outcomes are regularly evaluated to ensure continuous quality improvement, and that evaluation is understood as a constitutive part of program leadership rather than an add-on accountability exercise. Piatt and Woodruff (2016) similarly frame assessment as iterative—a cycle of planning, evidence collection, analysis, and action that feeds back into itself across time.

What distinguishes this framework from more instrumental approaches to assessment is its insistence on the relationship between evidence and improvement. Brown and Varghese (2021) argue that without honest and rigorous measurement, institutions cannot make consistent progress in developing leaders—a claim that implies measurement serves development rather than the reverse. Brown's (2022) Dimensions-Sources-Timeframes framework operationalizes this by identifying three axes along which any assessment approach can be evaluated: the type of outcome being measured (emotional, cognitive, or behavioral), the source of evidence (self-report, observer rating, behavioral trace, or objective outcome), and the timeframe across which evidence is collected. The practical implication is that no single assessment captures the full picture of leadership learning; triangulation across dimensions, sources, and timeframes is what makes assessment evidence-based, credible, and actionable. This principle applies at every level of the framework described below.

Three Levels of Assessment Practice

Assessment in leadership education operates simultaneously at three interconnected levels (Guthrie & Jenkins, 2018; Jenkins & Rocco, 2022; Owen, 2011). At the individual and course level, assessment asks whether students are developing the knowledge, skills, attitudes, and behaviors that specific learning experiences are designed to produce. We treat the individual student and the course as a single level because, in most programs, evidence of individual learning is generated within courses, through the assignments, rubrics, and feedback that faculty design. This level is closest to the day-to-day work of teaching: it includes formative feedback during a course, rubric-based evaluation of written work or performance demonstrations, and reflective exercises that help students connect experience to theory. The central assessment challenge at this level is alignment—ensuring that what is assessed actually reflects what was intended to be learned and taught, rather than conflating, for example, a student's ability to submit work on time with their capacity for critical thinking (Guthrie & Jenkins, 2018; Jenkins, 2016).

At the program level, the assessment asks whether the curriculum, taken as a whole, is producing the outcomes the program claims to develop. This level requires moving from the course as the unit of analysis to the program, and it demands a different kind of evidence: not a single rubric score, but a picture of how outcomes are sequenced, scaffolded, and assessed across multiple courses and experiences. Curriculum alignment matrices—tools that map program learning outcomes to individual courses, distinguishing where outcomes are taught from where student performance is formally assessed—are among the most practical instruments for making this picture visible (Jenkins & Rocco, 2022; Priest et al., 2012). The challenge at this level is coordination: producing and maintaining an accurate map of a curriculum requires the sustained collaboration of faculty across courses, and it is precisely this coordination cost that most programs find prohibitive.

At the institutional level, assessment asks whether *all* programs are fulfilling their obligations to their professions, the institution, and students over time—encompassing program reviews, accreditation, benchmarking against professional standards (Council for the Advancement of Standards [CAS], 2023; ILA, 2026), and longitudinal outcome tracking. Including program review at this level may seem counterintuitive, since a review typically examines a single program. What places it here is not what gets reviewed, but who initiates the review, who receives the findings, and who acts on them. Program reviews are usually prompted by institutional review cycles, accreditation requirements, or senior administrators (Jenkins & Rocco, 2022). They follow institution-wide guidelines, often involve external reviewers, and produce recommendations about resources, priorities, and sometimes program continuation that only the institution can act on. Put simply, the program-level work described above (the alignment matrix, the outcome evidence, the self-study) provides the inputs. At the same time, the review itself serves as a tool for institutional decision-making. Jenkins and Rocco's (2022) comparative case study found that contextual factors—program's theoretical grounding, institutional location, and administrative relationships—were the most critical determinants of review effectiveness. Institutional-level assessment requires a different set of leadership capacities than program-level assessment, including political navigation, stakeholder engagement, and the ability to sustain improvement efforts across organizational boundaries.

Positioning Generative AI Within the Framework

The three-level framework identifies where generative AI adds value by performing three functions that have historically been bottlenecks: synthesizing large bodies of evidence, detecting patterns across complex documents, and drafting structured outputs from human-generated inputs. These functions are most valuable at the program and institutional levels, where the volume of evidence to be synthesized regularly exceeds the time and coordination capacity of the faculty and staff responsible for it. They are valuable at the course level in more targeted ways, primarily for formative feedback and rubric-based assessment. In each case, AI occupies a supporting role within the iterative assessment cycle—contributing to the analysis and use-of-results phases—while leaving the goal-identification, method-selection, and improvement-decision phases to human judgment. As Brown (2022) cautions, the most significant threats to valid assessment are not technical but human—status quo threats that discourage honest evaluation, and belief threats that arise when data contradict educators' assumptions. No AI tool addresses those threats. AI can reduce the logistical barriers that prevent evidence from being assembled and examined in the first place.

Generative AI in Higher Education Assessment: What the Literature Tells Us

The literature on generative AI and higher education assessment has expanded rapidly since 2022, bringing both promise and significant unevenness. Much of the early scholarship was speculative or descriptive—cataloging possible applications of AI tools in educational settings rather than empirically examining outcomes. More recent work has begun to provide empirical grounding. However, the field is still in its early stages, and the gap between what is technically possible and what has been rigorously studied remains wide. This section reviews the most directly relevant scholarship, organized to reflect the three-level framework developed above. It closes by identifying the specific contribution this paper makes to a literature that has attended carefully to the student and teacher levels but has largely neglected the program review and institutional assessment dimensions that are the paper's primary focus.

What the Scoping Review Tells Us

Xia et al.'s (2024) scoping review of how generative AI transforms assessment in higher education provides the most comprehensive and analytically useful synthesis currently available, and its organizing framework maps closely onto the one developed in this paper. Xia et al. examined the literature across three levels of analysis—students, teachers, and institutions—and found that most published scholarship addresses the student and teacher levels. At the same time, institutional-level applications receive comparatively little attention. At the student level, the review identifies both the productive uses of AI (personalized feedback, formative assessment support, adaptive learning pathways) and the integrity challenges that have dominated popular discussion (AI-generated submissions, the difficulty of detecting AI use, the differential impact on students from non-English-speaking backgrounds). At the teacher level, the review documents emerging practices around AI-assisted rubric design, feedback generation, and workload reduction, while flagging concerns about deskilling and the erosion of faculty judgment when AI-generated evaluations are adopted uncritically. The institutional level receives the thinnest treatment in the literature Xia et al. reviewed—a gap that this paper directly addresses by developing a use-case framework for AI-assisted program review and macro-level assessment.

Critical Thinking, Bloom's Taxonomy, and the Course Level

At the course level, the intersection of generative AI and assessment raises particular questions about critical thinking—both as a learning outcome and as an assessment target. The widespread availability of AI writing tools has complicated the longstanding use of written products as primary evidence of higher-order thinking, because AI can now produce text that mimics the surface features of analytical writing without reflecting the cognitive processes the assignment was designed to develop. Lubbe et al. (2025) argue that the higher-order skills of analysis, evaluation, and creation in Krathwohl's revised Bloom's taxonomy are more resistant to AI augmentation than lower-order skills—though not immune—and that this creates a design imperative at the course level: assessments that target higher-order outcomes must increasingly be designed to make the process of thinking visible, not just its products. Scaffolded assignments that require students to submit iterative drafts, annotated sources, or reflective commentary alongside final products are among the approaches leadership educators have identified as more resistant to AI substitution, precisely because they require evidence of the student's own reasoning at multiple points in the process.

Student Motivation, Self-Direction, and AI-Enhanced Learning

A parallel strand of scholarship examines AI's potential to support student motivation and self-directed learning. Mohamed et al. (2025) found that AI tools positively influenced learning experiences, autonomy, and critical thinking skills across 455 students from four countries, though motivation was not significantly influenced by academic level. Tlili et al.'s (2025) meta-analysis of 85 quantitative studies found a very large overall effect of AI on learning achievement (Hedges' $g = 1.10$, $p < .001$). The caution here is consistent with Brown's (2022) warnings about indirect measures: self-reported improvements in motivation do not by themselves constitute evidence that learning occurred. Leadership educators should ensure that AI-supported activities generate direct evidence of learning—demonstrated performance, behavioral change, integrated application of theory—rather than relying on student satisfaction as a proxy.

Curriculum Integration and the Program Level

At the program level, scholarship on AI integration has focused primarily on curriculum design questions: how institutions can develop coherent, institution-wide approaches to AI literacy and AI-assisted learning. Jha et al. (2025) report on five case studies from an Australian higher education institution examining GenAI integration across disciplines, demonstrating how faculty drew on Technological Pedagogical Content Knowledge to integrate AI tools in ways that supported meaningful learning aligned with content areas and instructional strategies. Their findings highlight a dynamic familiar to most leadership educators: the gap between individual faculty innovation and program-level coherence. Without a deliberate effort to map AI integration to program learning outcomes—asking not just which courses use AI tools but whether their use advances the outcomes the program claims to develop—curriculum integration risks becoming an accumulation of disconnected course-level experiments rather than a coherent program strategy. This is precisely the kind of gap that AI-assisted curriculum alignment matrices, described in Use Case C, are designed to surface.

The Gap This Paper Addresses

The scholarship reviewed above converges on a finding with direct implications for this paper: the literature on generative AI and assessment has been richest at the level closest to the student—course-level feedback, personalized learning, integrity challenges—and thinnest at the level furthest from the student—program review, institutional assessment, cross-unit synthesis. This is partly a function of where AI tools have been most visibly deployed and partly of the longer timescales and greater organizational complexity involved in program and institutional assessment. Jenkins and Rocco's (2022) observation that the leadership program review literature itself was sparse before their comparative case study suggests that the absence of AI-integrated program review scholarship reflects a broader gap in institutional assessment research, not merely a lag in AI adoption. This paper contributes to closing that gap by developing a use-case framework that moves from the course level through the program level to the institutional level, with AI functioning as a leadership-support tool at each level rather than as a replacement for the human leadership that assessment requires.

Three Levels of AI-Supported Assessment Practice

The framework developed above positions generative AI as a resource within the iterative assessment cycle, valuable for synthesis, pattern detection, and structured drafting, but dependent on human judgment for the goal-setting, interpretation, and improvement decisions that give assessment its purpose. This section illustrates what that positioning looks like in practice through six use cases organized across the three levels of the framework, each drawn from actual practice. Together they are offered as worked examples of the framework's possibilities and its limits.

Individual and Course Level

At the course level, the most immediate question is whether generative AI can support assessment of complex outcomes, particularly critical thinking, reflective practice, and applied leadership, without displacing the human judgment rigorous assessment requires. The answer depends significantly on how the tool is used: AI deployed as a consistency check or a scaffolding aid operates very differently from AI deployed as a replacement for faculty evaluation. Two use cases illustrate both the opportunity and the necessary caution.

Use Case A: AI as a Second Evaluative Voice for Student Work

The discovery came, as many do, accidentally. Canvas recently implemented a feature that, when both an assignment description and a rubric are present, generates suggested rubric-based scores for student submissions. The initial reaction—something between alarm and curiosity—is probably familiar to anyone who has encountered AI-generated assessment outputs for the first time. What the feature produced was unexpected: the AI was substantially tougher than a typical end-of-semester human grader. It held the line on rubric criteria without the softening that exhaustion, empathy, and the end-of-term pile invite. Grades that a tired faculty member might nudge upward in the interest of generosity came back lower, and the rationale was traceable directly to rubric language rather than to impressionistic overall judgment.

This observation connects to a broader dynamic worth naming directly. When faculty grade their own students' work—especially high-stakes final assignments submitted during the most demanding week

of the academic calendar—the human tendency to err on the side of the student is both understandable and systematically problematic for program assessment. Grade inflation is a well-documented problem; rising grades account for most of the increase in college graduation rates since 1990 (Denning et al., 2022). Further, the system is unlikely to reward and likely to punish faculty who are strict on grading – faculty evaluations are lower for faculty who grade more harshly (Berezvai et al., 2021). Inflated grades, on the other hand, tell a false story about program outcomes.

AI, by contrast, is not impacted by the negative consequences of lower grades. Provided that the number of assignments per prompt stays within the recommended parameters, it does not experience human fatigue and the consequent decrease in grading quality. Likewise, AI has no ongoing relationship with individual students that could affect its judgment. The result, used judiciously, is a more consistent evaluative voice—not a replacement for faculty judgment, but a useful check on the grade drift that accumulates when human beings do this work alone, at scale, under pressure.

The analytical implications of this asymmetry are worth dwelling on. Scholarship on AI sycophancy has documented that large language models tend to validate and affirm the positions of the person they are interacting with—what Xia et al. (2024) identify as a significant concern at the student level, where AI-generated feedback may reinforce rather than challenge students' existing assumptions. However, when AI is directed to evaluate a third party's work against explicit rubric criteria, the sycophantic dynamic is unlikely to operate in the same way: sycophancy, as Sharma et al. (2024) define and measure it, is a bias toward the preferences of the person the model is conversing with, and in a grading task, that person is the instructor, not the student. This does not make AI-generated rubric scores a substitute for faculty judgment—Brown's (2022) insistence on triangulating across multiple sources applies here as much as anywhere—but it does make AI a potentially useful consistency check, particularly for programs concerned about grade inflation or inter-rater reliability across sections of the same course.

Use Case B: Scaffolded Assignments and Process-Visible Assessment

One productive response to the challenge of AI-generated student work is to redesign assessment tasks so that the process of student thinking becomes the evidence, not just its final product. Scaffolded assignments, those that require students to submit iterative components at multiple points rather than a single culminating artifact, make it considerably more difficult for AI to substitute for student cognition because each stage generates evidence that must be continuous with and responsive to the previous one. A student who submits an annotated outline, a partial draft with instructor feedback incorporated, and a final version with a reflective memo explaining key revisions has produced a documentary trail of their own reasoning that is far more resistant to AI wholesale substitution than a paper submitted in a single file at the deadline.

The connection to assessment validity is direct. When the assignment requires students to show their work iteratively, explaining their interpretive choices, responding to feedback, and documenting how their thinking changes, faculty gain access to evidence of the cognitive processes that leadership education is actually trying to develop: critical thinking, reflective practice, and the integration of theory and experience. Lubbe et al. (2025) make this point precisely: in an AI-augmented environment,

assessment pedagogy must shift toward tasks that target higher-order skills of analysis, evaluation, and creation, requiring demonstrable student reasoning rather than polished output. Scaffolded design is not a workaround for AI; it is a pedagogical improvement that makes evidence of learning more direct, defensible, and useful for program assessment, regardless of whether students have access to AI tools.

The common thread across both course-level use cases is straightforward: assessment tasks designed to make students' reasoning visible generate evidence that AI can support evaluation, but cannot easily replace the production of. Lubbe et al. (2025) argue that this shift—toward tasks targeting the higher-order cognitive processes of analysis, evaluation, and creation—is what assessment pedagogy must pursue regardless of AI. What AI adoption does is make that imperative newly urgent.

Program Level

The program level is where the case for generative AI in assessment practice is most compelling and most clearly grounded in the practical realities that constrain what programs can actually accomplish. The coordination costs of program-level assessment, assembling evidence from multiple courses, maintaining an accurate picture of curriculum alignment, and synthesizing findings across time regularly exceed the capacity of the faculty responsible for it. The three use cases in this section demonstrate how AI can reduce those costs without displacing the interpretive and improvement-oriented work that gives program assessment its value.

Use Case C: AI-Populated Curriculum Alignment Matrix

The curriculum alignment matrix is among the most practical tools for making program-level assessment visible. Yet, many programs aspire to maintain it but find it prohibitively labor-intensive to build and update. The K-State alignment matrix approach, described by Priest et al. (2012) and implemented at Kansas State University's School of Leadership Studies, provides a clean model: program learning outcomes on rows, core courses on columns, with each cell indicating either that student performance in that course is formally used for program-level assessment of the outcome (A) or that students have the opportunity to learn the outcome without formal assessment (X). The matrix makes gaps, redundancies, and assessment concentration immediately legible, patterns that might otherwise require hours of faculty discussion to identify.

To illustrate how generative AI can accelerate this process, we constructed a curriculum alignment matrix for the University of Southern Maine's Bachelor of Science in Leadership and Organizational Studies (LOS) program using Claude (Anthropic), a large language model operating within a persistent Claude Project environment. The program's published learning outcomes and syllabi for five core LOS courses—LOS 300: Organizational Theory, LOS 301: Group Dynamics, LOS 304: Organizational Budgeting and Finance, LOS 350: Leadership, and LOS 440: Organizational Change and Development—were uploaded as project-level knowledge. The K-State matrix was included as a format template. Using a structured multi-turn prompting sequence, the model was directed to read each syllabus against the program learning outcomes and populate the matrix using the A/X/— notation system. The resulting matrix (Table 1) was reviewed by a faculty member with direct knowledge of the program, who confirmed the accuracy of the cell designations.

Table 1

AI-Generated LOS Program Curriculum Alignment Matrix

Learning Outcome	LOS 300	LOS 301	LOS 304	LOS 350	LOS 440
1. Describe and contrast leadership and organizational theories and integrate them with leadership practice	A	X	—	A	A
2. Understand the difference between formal authority and leadership, and that leadership is an activity, process, or practice — not a position	X	A	—	A	A
3. Use critical reflection to continuously improve their practice	A	A	—	A	A
4. Understand the value of collaboration and work collaboratively with others in an effective manner	A	A	—	A	A
5. Identify key issues in organizational and leadership sustainability and implement a project with sustainability factored in	X	X	X	X	A
6. Identify contextual factors that affect competitive advantage in an organization	X	X	X	X	A
7. Explain social responsibility in an array of contexts and develop its applicability	—	—	X	X	X
8. Proficiently analyze leadership situations from a systems perspective and identify parts of the system that could be impacted to achieve positive change	A	A	—	X	A
9. Recognize the value of diversity in all its forms to organizational/community development and adaptation	X	A	—	X	X
10. Identify and apply strategic-level change and strategies to pursue change	X	X	—	X	A
11. Apply innovative and critical thinking as well as complex problem solving	A	A	A	A	A
12. Demonstrate proficient communication skills including written, oral, and computer-mediated interaction	A	A	A	A	A

Note. A = Student performance is used for program-level assessment of this outcome; X = Students have the opportunity to learn this outcome; — = Outcome not addressed in this course. Matrix generated by Claude Sonnet (Anthropic, 2025) using program learning outcomes and five instructor-authored syllabi; reviewed and confirmed by a faculty member with direct program knowledge.

The matrix the model produced in a matter of minutes would have required several hours of faculty coordination to build manually and, more importantly, would have surfaced patterns that manual review might have missed or deferred. Social responsibility (Outcome 7) appears only as an opportunity to learn (X) across all five courses and is never formally assessed. Sustainability and competitive advantage outcomes (5 and 6) are formally assessed only in LOS 440, the final required course, suggesting that students encounter these outcomes throughout the program but are held accountable for them only at the end. Critical thinking and communication (Outcomes 11 and 12) are formally assessed in every course—a distribution that reflects appropriate scaffolding but also raises the question of whether the assessments across courses are coordinated or simply parallel. These are the kinds of observations that should drive faculty conversations about program design, and the AI-populated matrix makes them available at the beginning of those conversations rather than as their conclusion. The full prompting sequence is documented in Appendix A and forms the basis for the AI use disclosure submitted with this manuscript; the key architectural point is that Claude Projects allowed the model to reason across all uploaded documents simultaneously rather than treating each exchange as a new conversation.

Use Case D: AI-Assisted Program-Level Critical Thinking Assessment

The alignment matrix is a structural tool: it shows where outcomes are taught and assessed, but does not evaluate quality. Use Case D extends the matrix into an evaluative function by using the same persistent project environment, with the populated matrix and original syllabi as inputs, to assess a specific outcome: critical thinking. The model was asked to identify relevant assignments across all five syllabi, evaluate whether they provide adequate coverage and developmental scaffolding, and offer a candid assessment of how confident the program should be in students' critical thinking development, directed explicitly to be rigorous rather than affirming, given the sycophancy tendencies documented in the literature (Xia et al., 2024).

The response produced was analytically substantive. The model identified writing-intensive assignments in LOS 300, 301, 350, and 440 as the primary vehicles for critical thinking assessment, noted that LOS 304's assessments were more technically oriented and less likely to generate evidence of critical thinking as leadership educators define it, and flagged a potential scaffolding gap: students encounter critical thinking assignments throughout the program, but the model found no evidence in the syllabi of a deliberate progression in the cognitive complexity demanded, from comprehension-level analysis in early courses to integration and evaluation in later ones. Whether this gap reflects a genuine curriculum design problem or a documentation gap in the syllabi is a question only faculty who know what actually happens in those courses can answer. However, the question itself is now on the table for faculty to examine and debate, rather than remaining implicit in the cumulative weight of individual course decisions made without a program-level view.

Use Case E: Pre- and Post-2022 Comparison and a Validity Check

The two use cases above assume that the student work feeding into the program assessment is a valid reflection of student learning. Use Case E raises the question of whether that assumption holds in a post-2022 environment in which generative AI tools are widely available to students. The logic is straightforward: if an institution can retrieve written assignments from the same course submitted

before the widespread availability of AI writing tools (approximately pre-fall 2022) and after, and evaluate both sets against the same rubric criteria using an AI model, the comparison would provide empirical evidence of whether the apparent quality of student work changed in ways that outpace plausible explanations based on program improvements alone.

We offer this not as a completed use case, but as a proposed methodology for programs seeking to assess the validity of their post-2022 outcome data. The practical requirements are modest: a course with a recurring writing assignment, a rubric that can be applied consistently across years, a sample of work products from both periods, and a persistent AI project environment in which the model can evaluate both samples against the same criteria. The comparison would not determine whether individual students used AI, a question fraught with unreliable detection, high false-positive rates, and legal exposure, as documented by Xia et al. (2024). It would instead ask a program-level question: Are the outcomes we are reporting still measuring what we think they are measuring? That question is one of the most important validity challenges facing higher education assessment today, and it is one that AI tools can help programs begin to investigate without requiring the IRB-bound empirical infrastructure of a formal research study.

The comparison is diagnostic, not corrective: it tells a program whether its existing evidence has likely been compromised, but it does not by itself fix the underlying design problem. The natural pairing is with the task-classification approach described below. A program that runs the comparison and finds evidence of contamination has a concrete next step: apply the AI Assessment Scale (AIAS; Perkins et al., 2024) to the assignment in question, and either redesign it toward the AI-resistant, process-visible forms of evidence developed in Use Case B, or, if the task genuinely belongs at a higher permitted-AI level, adjust what the resulting scores are understood to measure. Run together, the two tools answer the contamination problem at both ends: whether the existing evidence can be trusted, and what to do about it if it cannot.

Institutional Level

At the institutional level, the leadership work of assessment expands to encompass program reviews, accreditation processes, cross-unit synthesis, and the sustained organizational conversations through which evidence becomes the basis for strategic decisions. Jenkins and Rocco (2022) found that the most effective program reviewers combined methodological rigor with contextual sensitivity, an attentiveness to the political, cultural, and organizational dynamics that shape how evidence is received and acted upon. Generative AI cannot supply that sensitivity, but it can reduce the logistical burden of the synthesis work that precedes it.

Use Case F: AI-Assisted Synthesis in External Program Review

The program review use case introduced earlier deserves fuller treatment here. Although the review focused on a single program, it was an institutional undertaking in every sense that matters for this framework. It was part of the university's regular Academic Program Review cycle, coordinated through the provost's office and conducted in accordance with institution-wide guidelines. The review team presented its preliminary findings to senior administrators, including academic affairs leadership and the college and graduate school deans, and the program was required to develop an action plan in

response. Several of the critical questions the review addressed, including recruitment, graduate funding, and revenue, can only be answered by decisions made above the program level. During a two-day Academic Program Review of a small, research-intensive doctoral program in leadership studies at a mid-sized public university in the mid-Atlantic region, a four-person external review team conducted more than a dozen structured interviews with faculty, students, administrators, and advisory board members. Each reviewer took independent notes throughout the visit. At the end of the second day, three reviewers' notes, several thousand words of observations, direct quotes, and analytical impressions, were uploaded into a persistent Claude Project environment that also contained the program's self-study documents, the review team's charge and deliverables, and a set of critical questions identified in the self-study as the primary foci for the review. A fourth reviewer, who had left the visit a day early, submitted notes separately; in a follow-up step, the model was asked to identify what those notes added beyond the developing synthesis and integrate the two into a single, non-redundant account.

The model was asked to synthesize the team's notes against the self-study and critical questions, identifying areas of convergence across reviewers, areas where observations added nuance to the self-study's account, and areas where on-site evidence diverged from what the self-study had anticipated. The resulting synthesis served as the first draft of the closing stakeholder presentation and, later, the formal external report. Two aspects distinguished this from asking an AI to summarize a document: the inputs were the review team's own observational notes rather than existing institutional text, and the persistent project environment allowed the model to maintain context across the full corpus simultaneously—producing a synthesis responsive to the review's specific frame rather than a generic summary.

The resulting first draft required substantial revision by the team; the political navigation, interpretive judgment, and calibration of tone that Jenkins and Rocco (2022) identify as hallmarks of effective review practice are not functions AI can perform. Nevertheless, the draft provided a structured scaffold that allowed the team to work efficiently, redirecting the hours typically spent on synthesis toward the judgment-intensive work of deciding what to emphasize, how to frame difficult findings constructively, and how to sequence recommendations for institutional reception. In this sense, AI did not conduct the review; it made space for the reviewers to do so more effectively.

Tensions and Cautions: What Human Judgment Must Hold

The use cases above are optimistic in their framing, and deliberately so. But genuine utility does not mean uncritical adoption, and a responsible account of what generative AI can do for assessment practice must attend to what it cannot do, what it risks distorting, and where algorithmic efficiency substituted for human judgment produces outcomes that are actively harmful to the assessment enterprise. The most fundamental of these concerns begins with an assumption that optimistic use cases tend to take for granted: that the student work being assessed reflects what students actually know and can do. In a pre-2022 environment, that assumption was imperfect but defensible; in a post-2022 environment in which AI tools capable of producing polished academic prose are freely available to students at every level, it is an assumption that programs can no longer make without deliberate investigation.

The Contamination Problem

The contamination problem, as it bears on program assessment, is this: if students use AI to generate assignments that faculty use to evaluate learning outcomes, then the grades those assignments receive are not evidence of student learning; they are evidence of AI output quality. Moreover, when a program aggregates those grades to report, say, that 92% of students demonstrated proficiency in critical thinking, it is reporting a finding that may have no relationship to its students' actual cognitive development. The circularity is uncomfortable but important: AI-assisted program assessment tools are being asked to evaluate the quality of work that AI tools may have produced. A model grading a model's output and finding it proficient is not assessment; it is a hall of mirrors.

What makes this particularly difficult is that the contamination is invisible in the aggregate data that program assessment typically uses. A rise in mean rubric scores on a critical thinking assignment between 2021 and 2024 could reflect genuine program improvement, a change in student population, grade inflation by faculty, or widespread AI use—and nothing in the scores themselves distinguishes among these explanations. The pre- and post-2022 comparison methodology proposed in Use Case E is one approach to surfacing this ambiguity. However, the more fundamental implication is that programs relying on written artifacts as their primary evidence of learning outcome achievement need to interrogate whether those artifacts still constitute valid evidence—and build that interrogation into their assessment cycles rather than deferring it until an accreditor or a faculty committee raises the question.

One way to operationalize that interrogation is to stop treating AI involvement as a single, binary contamination risk and instead classify assessment tasks by the type and degree of AI use they permit. The AIAS (Perkins et al., 2024) offers this kind of framework, specifying levels ranging from no AI use through AI-assisted stages of the process to full AI collaboration, each tied to the task it is designed to measure. Applied at the program level, the AIAS reframes the contamination question from an unanswerable forensic one—did this particular student use AI on this particular assignment?—into a design question a program can act on: for outcomes where the evidence must demonstrate the student's own reasoning, tasks should be classified and redesigned toward the lower end of the scale, using the structural safeguards described in Use Case B; for outcomes where AI-assisted work is a legitimate part of professional practice, higher-level AI use can be permitted without undermining the validity of what is being measured. What types of learning are being assessed determines how much AI involvement the assessment evidence can tolerate before it stops measuring that learning at all.

Reliability Versus Validity: A Critical Tradeoff

A related concern involves the reliability-validity tradeoff. Reliability—the consistency with which an assessment produces the same result when applied to equivalent work—is an area where AI has the potential to outperform overburdened human faculty. Humans get tired, so assignments graded first may receive higher grades than those graded last (Wang et al., 2024). Moreover, while LLMs have their own version of the “fatigue” problem (one should not, for example, feed 30 essays to an LLM in the same chat box), the fix is fairly straightforward – each submission can be graded in its own pass rather than combined. Furthermore, it is possible to have multiple LLMs review the same batch, which would not be feasible for human graders.

However, reliability without validity is worse than useless: it produces consistent measurements of the wrong thing. Brown (2022) identifies this in his discussion of Campbell's Law—the tendency for any high-stakes measurable indicator to be gamed in ways that decohere it from the construct it was designed to measure. In an AI-saturated environment, the risk is that programs develop highly reliable assessment systems that consistently measure students' proficiency at using AI rather than their leadership learning. The solution is to design assessment tasks that are resistant to this dynamic—a challenge that requires human expertise in both AI capabilities and leadership learning theory.

AI Detection, Academic Integrity, and Legal Exposure

The instinct to detect and address AI-generated student work is understandable, and many faculty members have acted on it—only to encounter institutional resistance grounded in legitimate legal concern. The experience of raising an AI integrity case to an associate dean and being told to leave it alone is not uncommon: the advice, when it comes, typically reflects awareness of litigation risk rather than indifference to academic integrity. At least one high-profile case—a federal lawsuit alleging that a university's disciplinary action, triggered by an AI detection tool flag, discriminated against a student as a non-native English speaker—has made administrators across higher education reluctant to allow those tools to serve as the evidentiary basis for misconduct findings (Doe v. Yale University, 2025). The concern is well-founded: commercially available AI detection tools exhibit significant false-positive rates, which disproportionately affect submissions by non-native English speakers and students from certain demographic backgrounds (Liang et al., 2023; Xia et al., 2024). This pattern creates both ethical and legal exposure for institutions that treat detector outputs as definitive.

The practical result is a deeply uncomfortable institutional position: faculty who are confident, based on their knowledge of a student's prior work and the stylistic signatures of AI output, that a submission was AI-generated, but who cannot act on that judgment through formal channels because the available tools cannot meet the evidentiary standard for academic misconduct. This impasse is unlikely to resolve quickly; AI writing continues to improve in ways that make detection harder, not easier, and the legal landscape around institutional action remains unsettled. The implication for program assessment is one of the strongest arguments for the program-level, aggregate approach developed in this paper: rather than attempting to identify and adjudicate individual cases of AI use—a task for which neither the tools nor the institutional frameworks are currently adequate—programs can ask validity questions at the level of the outcome data itself, using approaches like Use Case E to surface whether their assessment evidence has been systematically affected by AI, without ever accusing any individual student.

Sycophancy as a Context-Dependent Risk

The sycophancy of large language models—their tendency to validate and affirm the person they are conversing with—presents differently depending on how AI is used in assessment. At the course level, when students use AI for formative feedback on their own work, the model will tend to praise their approach and suggest only minor adjustments, producing the educational equivalent of a yes-man; research has found that computer-generated flattery affects self-perception comparably to human flattery (Xia et al., 2024), meaning students may develop inflated pictures of their own development. At the program level, sycophancy becomes a risk when faculty use AI to evaluate their own program

outcomes: a model prompted to assess critical thinking results may produce a more favorable analysis than warranted. The mitigation, explicitly prompting the model to be rigorous and critical, as in Use Case D, is imperfect but meaningful and reflects an emerging practice of adversarial prompting that assessment leaders should build into their AI-assisted evaluation workflows.

The Faculty Professional Development Gap

Perhaps the most practically significant caution in this paper is also the least technically complex: the faculty and staff who would most benefit from AI-assisted assessment tools are often the least prepared to use them effectively. Xia et al. (2024) identify teacher professional development in AI literacy as one of the most critical preconditions for productive AI integration in assessment; without it, faculty are likely to either avoid the tools entirely or adopt them uncritically, producing the reliability-without-validity problem described above. Jha et al. (2025) similarly found that effective AI integration depended on faculty having developed technological pedagogical content knowledge—an understanding of how AI tools interact with specific disciplinary content and learning goals, not just how to operate them. Investment in AI-assisted assessment tools must be paired with investment in the professional development that enables their use with the judgment that effective assessment requires.

Implications for Assessment Leaders

Build Before You Prompt: Compile the Assessment Artifact Base First

The framework and use cases in this paper point to four practical priorities for faculty, assessment coordinators, and program leaders as they integrate generative AI into their practice. The first and most foundational: before deploying AI, compile the document base that enables AI-assisted analysis. The program-level use cases in this paper depended on the availability of program learning outcomes, course syllabi, prior alignment matrices, and accreditation or program review frameworks in a single, accessible location. Most programs have all these documents; very few have them compiled in a form that can be uploaded into a persistent project environment and reasoned across simultaneously. Building that base—an assessment artifact repository—is a one-time leadership investment that pays dividends across multiple assessment cycles; a persistent AI project environment (such as Claude Projects) is the appropriate architecture, maintaining uploaded documents across sessions and allowing the model to reason across the full corpus rather than treating each query as independent. Programs that build and maintain this repository will find that the logistical barriers that have historically made program-level assessment so expensive begin to dissolve.

Use AI for Synthesis; Reserve Human Judgment for Interpretation

The framework developed in this paper positions AI as most valuable in the analysis phase of the iterative assessment cycle—specifically, in the synthesis and pattern-detection work that precede interpretation and decision-making for improvement. This positioning should be reflected in how assessment leaders design their workflows: AI produces the first-pass synthesis, the populated matrix, the draft summary of program review notes; humans examine those outputs critically, correct errors and misreadings, and make the interpretive and improvement judgments that assessment is ultimately for. The temptation to use AI's outputs directly—to accept the alignment matrix as accurate without

faculty review, or to present the model's critical thinking assessment as the program's official finding—should be resisted for the reasons developed in the Tensions and Cautions section. Appendix A provides a model for documenting AI-assisted synthesis that is transparent about the human review step, where the output derives its validity.

Design Toward Process Visibility at the Course Level

The course-level implications of this paper point in a different direction from the program and institutional implications: rather than using AI to analyze existing assessment data, they call for redesigning assessment tasks to generate better data in the first place. Assignments that require students to make their reasoning processes visible—through iterative drafts, annotated sources, reflective commentaries, or recorded explanations—generate evidence of learning that is both more resistant to AI substitution and more useful for the kind of developmental, formative assessment that leadership educators have long prioritized. Lubbe et al. (2025) provide a useful framework for this redesign work, arguing that assessment pedagogy in an AI era should orient toward the higher-order cognitive skills of Krathwohl's revised Bloom's taxonomy—analysis, evaluation, and creation—and design tasks that require students to demonstrate those processes in forms that AI cannot easily produce without the student's own thinking occurring.

Invest in Faculty AI Assessment Literacy

None of the implications above can be achieved without sustained investment in faculty professional development. The gap between what AI tools can do for assessment and what faculty are currently equipped to ask of them is real and consequential, and it will not close through occasional workshops or technology demonstrations. What Xia et al. (2024) call assessment literacy in an AI era—understanding how to design assessments that are appropriate for an AI-augmented environment, how to use AI tools for synthesis without surrendering interpretive authority, and how to communicate with students about the role of AI in their learning and evaluation—is a form of professional expertise that requires deliberate cultivation. For programs and institutions considering AI-assisted assessment, the professional development investment should precede rather than follow the deployment of the technology. Faculty who understand the conceptual framework described in this paper—who can articulate why AI belongs in the analysis phase of the assessment cycle rather than the judgment phase—will use these tools more effectively and more ethically than faculty who are simply handed access to a new platform and expected to figure it out.

Conclusion

The important work of leadership education both begins and ends with assessment—and generative AI does not change that fundamental claim. What changes are the conditions under which the leadership work of assessment becomes practically possible. The argument of this paper has been neither that AI makes assessment easy nor that it renders established assessment practice obsolete. It is, rather, that a specific set of high-cost tasks that have historically constrained what programs can accomplish at the program and institutional levels—synthesis across heterogeneous evidence, pattern detection in large document corpora, first-draft production from human-generated inputs—can now be redistributed to tools that perform them quickly and without fatigue, freeing human judgment to concentrate on the interpretive, relational, and improvement-oriented work that no algorithm can perform.

The three-level framework developed in this paper—individual and course, program, and institutional—provides a structure for understanding where AI adds value and where it does not. At the course level, the most important implication is not what AI can assess but what it demands of assessment design: tasks that make students' reasoning visible, that target higher-order cognitive processes, and that generate evidence of learning rather than evidence of AI proficiency. At the program level, persistent-context AI environments make curriculum alignment matrices, program-level outcome assessments, and longitudinal validity comparisons feasible for programs that could never have sustained the coordination costs required by those analyses. At the institutional level, AI-assisted synthesis of reviewer notes and self-study documents accelerates the drafting of program reviews without displacing the contextual sensitivity and political judgment that Jenkins and Rocco (2022) identify as the hallmarks of review practice that produce real improvement.

The tensions and cautions documented above are not reasons to defer adoption; they are reasons to adopt thoughtfully. The contamination problem is that AI may evaluate AI-generated student work and deem it proficient, posing a genuine threat to validity that programs must investigate rather than assume away. The reliability-validity tradeoff, the limitations of AI detection tools, the context-dependence of sycophancy, and the faculty professional development gap are not peripheral concerns; they are the conditions under which AI-assisted assessment either produces genuine insight or the appearance of insight without the substance.

Three directions for empirical follow-up are particularly urgent. First, validity studies of AI-rubric grading at the program level would provide the evidentiary foundation that practitioner use cases like those in this paper cannot supply. Second, the pre- and post-2022 comparison methodology outlined in Use Case E deserves systematic application across multiple programs and disciplines, yielding longitudinal validity data that assessment leaders need to understand how their post-2022 outcome evidence reflects. Third, longitudinal studies tracking whether AI-assisted evidence use leads to program improvements would evaluate whether the optimistic case made in this paper holds up against the standard set by Brown and Varghese (2021) for the field: does the assessment produce genuine, consistent progress in developing leaders?

Generative AI, in the end, is most useful to assessment leaders who already know what a good assessment looks like and who understand the differences between evidence and compliance, between reliability and validity, and between synthesis and judgment. For those leaders, it offers something genuinely new: the practical capacity to do at the program and institutional levels what they have always known should be done. The gap between aspiration and practice in higher education assessment has persisted for decades, not because faculty do not care about improvement, but because assembling and examining evidence has been too expensive for the time and resources most programs have. If generative AI can close even part of that gap, if it can produce the alignment matrix that clarifies a curriculum's hidden structure, or the synthesis that turns four reviewers' notes into a first draft by morning, then it will have served the purpose assessment has always served: helping educators see clearly enough to do better.

References

Anthropic. (2025). *Claude Sonnet* [Large language model]. https://www.anthropic.com/claude

Berezvai, Z., Lukáts, G. D., & Molontay, R. (2021). Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assessment & Evaluation in Higher Education, 46*(5), 793–808. https://doi.org/10.1080/02602938.2020.1821866

Brown, R. P. (2022). *Measuring the mist: A practical guide for discovering what really works in leader development.* Storied Publishing.

Brown, R. P., & Varghese, A. H. (2021). *Measuring the mist: Leadership program evaluation at the Doerr Institute for New Leaders.* Rice University. https://doerr.rice.edu/measuring-mist

Council for the Advancement of Standards in Higher Education. (2023). *CAS professional standards for higher education* (11th ed.).

Denning, J. T., Eide, E. R., Mumford, K. J., Patterson, R. W., & Warnick, M. (2022). Why have college completion rates increased? *American Economic Journal: Applied Economics, 14*(3), 1–29. https://doi.org/10.1257/app.20200525

Doe v. Yale University, No. 3:25-cv-00159 (D. Conn. filed Feb. 3, 2025).

Dugan, J. P. (2012). Exploring local to global leadership education assessment. *New Directions for Student Services*, *2012*(140), 89–101. https://doi.org/10.1002/ss.20034

Ewell, P. T. (1988). Implementing assessment: Some organizational issues. *New Directions for Institutional Research, 1988*(59), 15–28. https://doi.org/10.1002/ir.37019885904

Guthrie, K. L., & Jenkins, D. M. (2018). *The role of leadership educators: Transforming learning.* Information Age Publishing.

International Leadership Association. (2026). *General principles for leadership programs* (2nd ed.). https://ilaglobalnetwork.org/about/advancement-of-leadership-programs/

Jenkins, D. (2016). *Relation of leadership educators' perceived learning goals and assessment strategy use* [Paper presentation]. Association of Leadership Educators 26th Annual Conference, Sacramento, CA, United States. http://www.leadershipeducators.org/resources/Documents/Annual%20Conference%20Proceedings%202016.pdf

Jenkins, D. M., & Rocco, M. L. (2022). Conducting leadership program reviews in U.S. higher education: A comparative case study. *Journal of Leadership Education, 21*(3), 1–18. https://doi.org/10.12806/V21/I3/R6

Jha, M., Jha, S., Holmes, A. M., Smith, B., Murphy, B., Kansal, M., & Pidgeon, D. (2025). Integrating generative artificial intelligence across the curriculum in higher education: Multi-disciplinary case studies. *Journal of Education, Innovation, and Communication, 7*(2). https://doi.org/10.34097/jeicom-7-2-1

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Lubbe, A., Marais, E., & Kruger, D. (2025). Cultivating independent thinkers: The triad of artificial intelligence, Bloom's taxonomy and critical thinking in assessment pedagogy. *Education and Information Technologies, 30,* 17589–17622. https://doi.org/10.1007/s10639-025-13476-x

Mohamed, A. M., Shaaban, T. S., Bakry, S. H., Guillén-Gámez, F. D., & Strzelecki, A. (2025). Empowering the faculty of education students: Applying AI's potential for motivating and enhancing learning. *Innovative Higher Education, 50,* 587–609. https://doi.org/10.1007/s10755-024-09747-z

Owen, J. E. (2011). Assessment and evaluation. In S. R. Komives, J. P. Dugan, J. E. Owen, C. Slack, & W. Wagner (Eds.), *The handbook for student leadership development* (2nd ed., pp. 177–202). Jossey-Bass.

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice, 21*(6). https://doi.org/10.53761/q3azde36

Piatt, K. A., & Woodruff, T. R. (2016). Developing a comprehensive assessment plan. *New Directions for Student Leadership, 2016*(151), 19–34. https://doi.org/10.1002/yd.20198

Priest, K., Friedel, C., & O'Dell, I. (2012, July 8–11). An invitation to innovation in leadership program assessment [Roundtable]. In *2012 ALE conference proceedings* (pp. 384–385). Association of Leadership Educators. http://www.leadershipeducators.org/resources/Documents/Annual%20Conference%20Proceedings%202012.pdf

Rosch, D. M., & Jenkins, D. M. (2020). What do we know about formal leadership courses and their effects? *New Directions for Student Leadership, 2020*(168), 31–41. https://doi.org/10.1002/yd.20406

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024).* https://openreview.net/forum?id=tvhaxkMKAn

Tlili, A., Saqer, K., Salha, S., & Huang, R. (2025). Investigating the effect of artificial intelligence in education (AIEd) on learning achievement: A meta-analysis and research synthesis. *Information Development, 41*(3), 825–842. https://doi.org/10.1177/02666669241304407

Wang, Z., Pei, J., & Li, J. (2024). *70 million Canvas grading records reveal widespread sequential bias and system-induced surname initial disparity*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4603146

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, Article 40. https://doi.org/10.1186/s41239-024-00468-z

About the Authors

Daniel Jenkins, Professor, Leadership & Organizational Studies, University of Southern Maine daniel.m.jenkins@maine.edu

Cristina de Mello-e-Souza Wildermuth, Professor of Leadership, Organizational Leadership, Barry University, cwildermuth@barry.edu

Appendix A

Prompting Documentation for Use Cases C and D

To illustrate the analytic potential of generative AI in program-level assessment, the authors employed Claude Sonnet (Anthropic, 2025), a large language model, to construct a curriculum alignment matrix for the University of Southern Maine Bachelor of Science in Leadership and Organizational Studies (LOS) program. The following materials were uploaded as persistent project-level knowledge within a Claude Project environment: (a) the published LOS program learning outcomes retrieved from the program's institutional website, (b) instructor-authored syllabi for five core LOS courses (LOS 300; LOS 301; LOS 304; LOS 350, and LOS 440), and (c) a model alignment matrix from Kansas State University's School of Leadership Studies, included to orient the model to the target format and notation conventions. The matrix was generated through the prompting sequence reproduced below, in which the model was directed to map each program learning outcome against each course using a three-value notation system (A = formally assessed; X = opportunity to learn; — = not addressed), modeled on the approach described by Priest et al. (2012). The model produced cell designations by independently reading and cross-referencing each syllabus against the stated outcomes; no cell values were pre-assigned or suggested by the authors.

The resulting matrix was reviewed by a faculty member with direct knowledge of the LOS program, who confirmed the accuracy of cell designations. In generating the matrix, the model flagged a divergence from an earlier illustrative version of the same matrix that had circulated in prior program materials, particularly regarding LOS 304. The faculty reviewer confirmed the syllabus-derived matrix presented here as authoritative, on the basis that it reflected a closer, document-grounded reading of that course's current design. Cell designations reflect the model's reading of stated course objectives, assignment descriptions, and grading structures as documented in the syllabi; they do not constitute independently verified empirical evidence of student learning or instructor intent, and may not fully capture the enacted curriculum as it unfolds in practice. This process is offered as a replicable use case rather than a substitute for faculty-driven assessment; the human review step is treated as essential to the validity of the output.

Prompt 1 (material inventory): "I am working on an academic paper about the use of generative AI in higher education assessment. As part of that paper, I want to demonstrate how a faculty member or assessment leader can use a Claude Project to conduct program-level assessment using a curriculum alignment matrix approach. I have uploaded the following materials to this project: the University of Southern Maine's Bachelor of Science in Leadership and Organizational Studies (LOS) program learning outcomes; syllabi for the core LOS courses; a model alignment matrix from Kansas State University's Leadership Studies program review. Please begin by reviewing all uploaded materials and confirming what you can see. Then wait for my next instruction." The model returned a structured inventory of the uploaded materials without performing analysis.

Prompt 2 (matrix construction): "Using the USM LOS program learning outcomes and the syllabi I have provided, please build a curriculum alignment matrix modeled on the K-State example. Structure the matrix as follows: Rows = each USM LOS program learning outcome; Columns = each uploaded course; Cell values: A = Student performance in this course is formally used for program-level assessment of

this outcome; X = Students have the opportunity to learn and practice this outcome in this course; — = This outcome is not addressed in this course. After the matrix, provide a 2–3 paragraph narrative that: (1) identifies any program-level gaps where an outcome appears infrequently or only at the 'X' level across all courses; (2) notes any outcomes that appear over-concentrated in a single course; and (3) flags any outcomes where assessment ('A') is sparse relative to learning opportunities ('X')." The model produced the matrix reported in Table 1 and the accompanying narrative, and independently flagged the LOS 304 divergence described above.

Prompt 3 (critical thinking assessment): "Now, using the alignment matrix you just built and the syllabi I provided, please conduct a program-level assessment of critical thinking as a student learning outcome in the LOS program. Specifically: 1. Identify all assignments across the uploaded syllabi that are likely measuring or developing critical thinking; 2. Evaluate whether those assignments, taken together, provide sufficient coverage and scaffolding of critical thinking across the program (introductory → developing → advanced); 3. Note any gaps, redundancies, or developmental sequencing concerns; 4. Provide a brief, candid overall assessment: Based on these course materials alone, how confident should the program be that students are genuinely developing critical thinking capacity by the time they complete the required LOS courses? Please be rigorous and honest rather than affirming — I need this assessment to be useful for genuine program improvement." The model's response is synthesized in the Use Case D discussion in the main text.

Appendix B

Prompting Documentation for Use Case F

Before the synthesis work began, the following materials were available as project knowledge within a persistent Claude Project: the program's self-study report, the review team's charge and deliverables, the site visit itinerary, and relevant professional-standards documents, including the ILA General Principles for Leadership Programs. Three reviewers' independent notes from the site visit, several thousand words of observations, direct quotes, and analytical impressions, were uploaded together at the close of the visit. The model was asked to synthesize these notes against the self-study and the critical questions identified as the review's primary foci, and produced a first-pass synthesis organized around emerging themes.

A fourth reviewer had left the visit a day early and submitted notes separately. In a follow-up step, the model was asked to review that reviewer's notes against the developing synthesis, identify what they added that was not already reflected, and integrate the two into a single, non-redundant account. This synthesis served as the basis for the first draft of the closing stakeholder presentation, which was presented the following morning.

The formal external report underwent a separate, substantive revision session roughly 11 days after the site visit, in which specific recommendation language was expanded, tightened, and more explicitly grounded in cited evidence at the authors' direction. One instruction from that phase illustrates the level of control exercised over the model's output when revising an already-drafted deliverable: "You previously generated the attached PowerPoint. I am also attaching an updated report document containing expanded and revised content. Your task: 1. Compare the updated report to the existing slides. 2. Update ONLY the slides where content has materially changed. 3. Preserve: Slide order / Slide titles / Formatting / Visual layout / Tone and audience framing. 4. Do NOT regenerate the entire deck. 5. Do NOT expand slide count unless absolutely necessary. […] Be surgical and minimal." The model revised eight of fourteen slides, left the remaining six untouched, and the reviewers verified the result against the source report before use.

As with Use Cases C and D, all outputs in this sequence were reviewed and revised by the review team before use in any deliverable; the AI-generated drafts themselves are not reproduced here or elsewhere in this manuscript, consistent with the anonymity this use case requires.