

The AI Teaching Value Alignment Canvas: A Lightweight Framework and Rubrics for Connecting AI-Enhanced Pedagogic Methods and Program Target

Lyu Hai

Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China

Abstract: AI in Education is a rapidly evolving field and program teams often do not have a shared language to discuss the values, evidence, safeguards and scalability of individual learning activities (ILAs) incorporating AI. This study explored the feasibility of the AI Teaching Value Alignment Canvas (ATVAC) which is a one-page instrument that serves as a tool and a rubric to decide whether ILAs integrating AI are suitable for implementation at the program level. The Canvas was developed using design-based instrument development where the authors reviewed the course materials, refined and revised the Canvas based on the program learning outcomes, assessment rubrics and quality assurance materials. An expert panel of faculty members participated in a micro-Delphi method to validate the content and usability of the Canvas. A multi-rater scoring study was conducted in a pilot test to investigate the validity and reliability of using the Canvas to score 12 cases of AI-based teaching practices and found that the percent agreement, weighted Kappa and total-score ICC of the items on the Canvas were higher for more concrete and anchored items (e.g., Problem clarity: $\kappa=0.78$, Safeguards strength: $\kappa=0.76$, Improvement loop: $\kappa=0.73$) and lower for more interpretive items (e.g., Indicator validity: $\kappa=0.41$, Scalability & cost: $\kappa=0.39$). The total-score ICC of the Canvas is 0.79. The Canvas addresses four values which are the learning outcomes, the teacher workload and decision quality, the service experience, and the program quality and consistency. It has feasible and practical evidence indicators which do not require the recording of fine-grained details of the ILAs integrating AI. The Canvas also has clear risk safeguards and governance arrangements that are lightweight. The Canvas will support program teams in making consistent decisions on the adoption of ILAs integrating AI and thus implementing AI-based pedagogies in a responsible manner across a variety of teaching contexts.

Keywords: AI-supported teaching; Program management; Evaluation rubric; Instrument creation; Scalability; Quality control; Agreement between raters

DOI: 10.62639/sspjess03.20260304

1. Introduction

(1) The background

Generative AI is now becoming a standard part of the pedagogical toolkit, available in a variety of formats to support draft feedback, formative feedback, summarizing, speaking practice and knowledge organization^{[11][14][20]}. More advanced AI agents are beginning to undertake multi-step tasks such as information searching, task critique, content editing and learning validation. While the new technology has many real and tangible benefits including, faster feedback for students and faculty and a more human and relevant learning experience, and reduction of routine tasks for faculty^{[23][18]}. These advancements present tangible opportunities, which include accelerated feedback loops, more tailored practice, and reduced repetitive teaching workload, yet they also pose program-

(Manuscript NO.: JIESS-26-4-62A001)

About the Author

Lyu Hai (1982-), Male, Han, Suzhou, China, Postgraduate, Xi'an Jiaotong-Liverpool University, None, Research Interests: Educational Management.

level risks, including inconsistent student experiences across classes, uneven academic integrity controls, unclear accountability attribution, and challenges in generating credible and sustainable evidence without intrusive monitoring^{[1][10][13]}.

Decisions regarding program-level adoption are not the same as the experimentation the teacher has done at the classroom level. A practice may demonstrate its utility in a single class but remain unsuitable for program adoption if it lacks sustainability, proves challenging to verify, exhibits significant variation among instructors, or relies on data collection methods that are impractical or unacceptable^{[2][29]}. These are all factors which need consideration that program managers need keep in mind when deciding which assessment methods will enable teaching teams to determine when a teaching practice including educational AI is ready for adoption by the larger program rather than simply knowing that the practice appears to work in a single classroom^[6].

(2) The gap in the current research

The key challenge is related to how the pedagogical practices implemented using AI are described. Typically, teachers discuss the tools used in the pedagogical practices such as describing the use of an 'agent for feedback' without realizing that they are referring to the learning activity, the validation mechanism to verify the acquisition of the knowledge by the learner and the protection mechanisms to avoid the misuse of the learning activity. Moreover, these descriptions do not allow for comparison between pedagogical practices. The pedagogical practices are often reduced to the discussion of the tools and educational technologies used to implement them, rather than the practices themselves, the types of evidence required to demonstrate that the practices are pedagogically effective, and the governance frameworks required to ensure that the practices are implemented in a way that is educationally effective^{[7][9]}. Teachers often waste too much time discussing the pros and cons of particular tools, rather than focusing on the practices, the quality of the evidence that underpins those practices and the governance arrangements required to ensure that those practices are delivered in an educationally effective way^{[17][12]}.

(3) The purpose of the research

To bridge this gap, this paper presents the AI Teaching Value Alignment Canvas alongside a corresponding from 0 to 3 point scoring rubric. The Canvas is a one-page instrument that describes a value pathway along 9 fields: problem and context, AI intervention point, educator actions, learner actions, value domains aligned with program priorities, attainable evidence indicators, potential risks and safeguards, governance and sustainability prerequisites and the loop for continuous improvement. The rubric enables the teacher to use the criteria to calibrate each value dimension, compare teaching approaches fairly, and build a bank of concrete examples to help decide which programmes to adopt.

(4) The research questions

This research investigates three key questions:

RQ1: Which value pathways connect AI-supported practices to program priorities?

RQ2: Which evidence indicators are viable without relying on invasive, fine-grained logs?

RQ3: What is the validity of the Canvas and the rubric for various stakeholders in terms of content, usability and reliability?

2. Literature Review

(1) AI-enhanced pedagogy as a design challenge: mechanisms rather than instruments

Educational technology is frequently implemented as a tool in the classroom with little consideration given to

pedagogical implications^{[22][15]}. As a result, Generative AI makes the technology-centered approach highly probable. The Generative AI can produce highly sophisticated and correct responses, which obscure the fact that the students have not constructed a deeper level of understanding^{[25][8]}. A mechanism-oriented approach to program evaluation can help to focus on teacher behavior that restricts the use of technology, student behavior that is associated with learning, and other indicators of learning^[26].

The use of AI agents makes it important to have a clear understanding of how they work. These agents are programmed with certain steps in mind, like getting information, summarizing it, critiquing it, and then revising it, which can affect the work students do and what they expect to happen^{[25][16]}. If different teachers use different agents or set them up in different ways, it could lead to different experiences for students and different standards from one class to another^[25]. By focusing on how these agents work, teachers can have better discussions with their teams about how the agents fit into the workflow, what students are expected to do on their own, how learning is assessed, and how to prevent problems with academic integrity and fairness in education^{[30][21]}. This helps ensure that every student is on the same page and that the use of AI agents supports learning in a consistent and fair way.

(2) Evidence and assessment without detailed learning analytics

Fine-grained interaction logs can support instructional evaluation, yet program-level implementation faces practical and governance challenges, including vendor dependence, ambiguous definitions, heavy compliance burdens, and eroded trust among stakeholders^{[27][3]}. Furthermore, behavioral traces frequently exhibit only a tenuous correspondence with learning assertions^[5]. A pragmatic alternative involves data-minimized assessment leveraging evidence already integrated into teaching and quality assurance procedures^{[28][4]}.

Three categories of evidence are especially viable at the program level: (1) performance evidence aligned with intended learning outcomes, such as rubric-based quality scores, revision improvements, oral verification assessments, and evidence-tracing accuracy; (2) operational indicators that measure process efficiency and rework, including feedback turnaround time, rework rate, frequency of clarification requests; (3) governance artifacts that ensure auditability and standardization, such as moderation notes, calibration records, updates to exemplar banks, and risk safeguard documentation. These forms of evidence are comprehensible, sustainable, and facilitate triangulation without excessive dependence on behavioral records.

(3) Program-level governance and scalable pedagogical innovation

For scalable pedagogical innovation to work, there need to be common norms, methods for calibration and moderation, clear job definitions, ways to reduce risk, and cycles of continuous improvement^[24]. Boundary objects like templates, frameworks, and rubrics make it easier for people with different skills to work together, such as instructors, quality assurance and professional services, leadership, and learning technologists^[19]. A one-page Canvas may be a useful boundary object because it standardizes descriptions, makes evidence and protections clearer, and makes it easier to compare and make decisions without everyone having to agree on one AI tool.

3. Methodology

(1) Research design

This work used a technique for developing instruments that focused on design. The goal was to create a tool-neutral evaluation tool that could be used to compare and choose between different AI systems at the program level, rather than testing a specific AI system.

Phase 1: Building the instruments

Iterative drafting and document analysis were used to create the canvas and rubric. Some of the inputs were the

program's learning goals, module descriptions, assessment rubrics, quality assurance documents, and rules about data governance and academic integrity. A purposive sample of AI-supported teaching practices was used to figure out the range of Canvas blocks and rubric anchors. The rubric turned the Canvas into 12 evaluative dimensions that evaluate sustainability, risk governance, safeguards, evidence feasibility, and learning verifiability. It used a 0–3 point anchor scale.

Phase 2: Micro-Delphi/expert panel examination of content validity and usability

A micro-Delphi/expert panel evaluated dimension significance and usability/clarity, providing qualitative feedback on the evidence's viability and identifying any absent or redundant constructs. Ratings and comments helped improve the definitions of dimensions, anchor examples, and governance guidance. The goal was to reduce the amount of subjective interpretation in evaluations of evidence quality and scalability.

Phase 3: Multi-rater scoring research pilot project

To reduce discrepancies in scoring due to information gaps, a pilot multi-rater scoring study utilized the rubric on 12 standardized one-page case briefs corresponding to Canva's core components. Three raters looked at each case on its own and gave short explanations for any extreme or unclear judgments. The study used percent agreement, weighted Kappa per dimension, and ICC for overall scores to analyze the agreement.

The reasoning behind selecting cases (coverage sampling + anchor extremes)

The study used purposive coverage sampling to pick cases that showed typical AI teaching practices at the program level. These included teacher decision support, learner-facing practices, verification-oriented designs, service-experience interventions, and instructional operations/efficiency practices. The study included a high-readiness case (A01) with clear verification, possible evidence bundles, and clear safeguards/governance, and a low-readiness case (B01) with tool use but no verification or governance specification. These cases were chosen on purpose to test discriminant demonstration. To make things easier to compare, all of the cases were put into a common one-page short format that matched the Canvas headers.

(2) Data minimization and ethics

No information that could identify a person was collected. Evidence examples use anonymized artifacts and aggregated indications, which follow the rules of proportionality and data minimization.

4. Result

(1) The framework and evaluation criteria structure

The Value Alignment of AI Teaching Canvas explains how AI-supported practices can be used to create value and shows the difference between the point of AI intervention and the pedagogical mechanism. It requires clear actions and verification artifacts from both the teacher and the learner, which means that using tools cannot be seen as proof of learning.

Table 1: One-page Structure of the AI Teaching Value Alignment Canvas

AI Teaching Value Alignment Canvas (One page structure)		
No.	Factor	Rubrics
1	problem and context	-Course/task/level/assessment link -Instructional problem and why it matters
2	AI intervention point	-Inputs → AI function/constraints → Outputs -What AI cannot do / what must be verified
3	educator actions	-Task constraints, modelling, sampling -Moderation/exemplar use

4	learner actions	-Required verifiable artefacts -Reflection/revision memo
5	value domains aligned with program priorities	-Learning outcomes -Workload & decision quality -Service experience -Programme quality/consistency
6	measurable evidence indicators	-Performance evidence -Operational indicators -Governance artefacts
7	potential risks and corresponding safeguards	-Integrity -Reliability -Fairness -Privacy -Dependency → corresponding controls
8	governance and sustainability	-Light RACI + review cadence + resources/training/cost
9	feedback loop	-What is reviewed, by whom, how often, and how exemplars/ anchors are updated

(2) The content validity and usability

Table 2: Expert-panel ratings of rubric dimensions (importance and usability) and resulting revisions

Dimension	Importance (Median [IQR])	Usability/Clarity (Median [IQR])	% Important (≥4)	Key feedback (summary)	Revision action (v1.0→v1.1)
Problem clarity	4.5 [4–5]	4.0 [4–4]	88%	Clarify scope and baseline context	Added minimum context fields + examples
Teacher moves alignment	4.0 [4–5]	3.5 [3–4]	75%	Sequence matters	Added stepwise prompts
Learner moves verifiable	5.0 [5–5]	3.5 [3–4]	100%	Must have observable artefacts	Verification artefact checklist + anchor examples
Value pathway clarity	4.0 [4–5]	3.5 [3–4]	75%	Uneven coverage across domains	Require ≥2 value domains stated
Evidence feasibility (no fine logs)	5.0 [4–5]	3.0 [3–4]	88%	Embed in existing processes	Introduced 3-part evidence bundle template
Indicator validity	4.5 [4–5]	3.0 [2–3]	75%	Validity needs examples	Added valid/invalid indicator contrasts
Risk identification	4.5 [4–5]	3.5 [3–4]	88%	Risk categories should be stable	Fixed 5 risk categories (integrity/reliability/fairness/ privacy/dependency)
Safeguards strength	5.0 [5–5]	3.5 [3–4]	100%	Add enforcement details	Added sampling/enforcement/ escalation prompts
Governance clarity (light RACI)	4.0 [4–5]	3.0 [2–3]	75%	RACI feels abstract	Shifted to “minimum viable governance” checklist
Consistency & fairness	4.5 [4–5]	3.0 [3–4]	88%	Show how it standardizes across classes	Added moderation/exemplar cues
Scalability & cost	4.0 [3–4]	3.0 [2–3]	63%	Cost weighting differs by role	Split cost into time/training/ tool-dependence
Improvement loop	4.5 [4–5]	3.5 [3–4]	88%	Specify cadence and responsibility	Added cadence + update outputs
Overall instrument	—	3.6 [3–4]	—	Usable; reduce subjective interpretation	Added anchor examples + minimum case-brief checklist

Analysis: Panel ratings emphasized verifiability, evidence feasibility, and safeguards (importance medians around 5.0), suggesting that programme adoption decisions depend on the assess ability of learning claims and the sustainability of evidence and controls. Usability was less effective for Indicator validity and Scalability & cost, indicating role-based interpretative variations—an initial indication that these dimensions necessitate more robust

anchor examples and standardized case documentation.

(3) The pilot inter-rater agreement

Table 3: Inter-rater agreement for rubric dimensions in the pilot scoring study

Dimension	Mean score across cases (SD)	Percent agreement (%)	Weighted Kappa (95% CI)	Main disagreement reason (summary)
Problem clarity	1.92 (0.74)	86	0.78 (0.66–0.88)	Baseline context missing in some briefs
Teacher moves alignment	1.83 (0.69)	80	0.66 (0.52–0.79)	Sequence/logic of actions unclear
Learner moves verifiable	1.58 (0.83)	83	0.72 (0.59–0.84)	Verification artefacts underspecified
Value pathway clarity	1.71 (0.66)	78	0.64 (0.48–0.77)	Domain coverage uneven
Evidence feasibility	1.54 (0.79)	76	0.60 (0.44–0.74)	Collection effort uncertain
Indicator validity	1.33 (0.77)	66	0.41 (0.22–0.59)	Different interpretations of “valid”
Risk identification	1.46 (0.88)	82	0.70 (0.56–0.82)	Risk priority not explicit
Safeguards strength	1.38 (0.95)	85	0.76 (0.63–0.87)	Enforcement details missing
Governance clarity	1.21 (0.85)	73	0.57 (0.39–0.72)	Ownership/cadence absent
Consistency & fairness	1.29 (0.78)	72	0.55 (0.36–0.70)	Standardisation across classes unclear
Scalability & cost	1.62 (0.71)	64	0.39 (0.19–0.56)	Role-based weighting of time cost
Improvement loop	1.18 (0.92)	84	0.73 (0.59–0.84)	Review cadence/responsibility vague
Total score (0–36)	18.05 (6.40)	—	ICC = 0.79 (0.63–0.89)	Variance driven by evidence/governance clarity

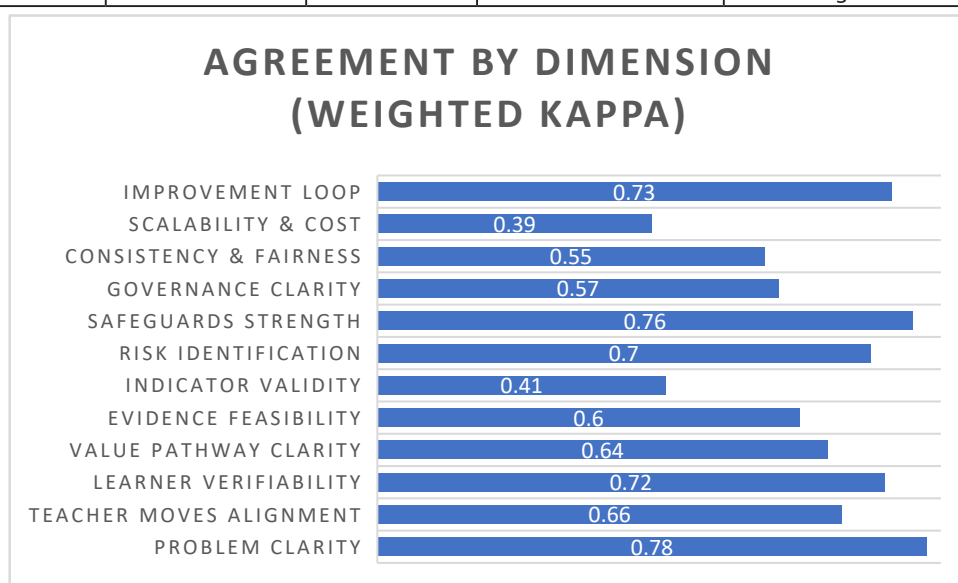


Figure 4: Agreement of dimension (Weighted Kappa)

Analysis: Agreement patterns exhibit a distinct concreteness gradient. Dimensions with a greater number of observable anchors attained higher Kappa values (approximately 0.73–0.78). Additional interpretive dimensions exhibited lower Kappa values (approximately 0.39–0.41), consistent with role-based weighting and the ambiguity regarding what constitutes a valid indicator. Overall total-score reliability was satisfactory (ICC=0.79), indicating that the rubric can support programme-level differentiation of preparedness, even as refinement at the dimension level is prioritized.

(4) Case-set coverage and overall score distribution

To guarantee comprehensive ecological coverage, the pilot set comprised 12 standardized cases encompassing learner-facing practices, verification-oriented designs, teacher decision support, service-experience interventions, and teaching operations. Total readiness scores exhibited a broad distribution ($M=18.05$, $SD=6.41$), characteristic of a

typical program portfolio: numerous practices classified as ‘promising but incomplete’ and a smaller number of more mature exemplars.

Table 5: Total-score distribution across 12 cases

Score band	Count	Case IDs
0–9	1	B01
10–19	7	C05, C06, C07, C08, C09, C10, C11
20–29	3	C02, C03, C04
30–36	1	A01

Table 6: Total readiness scores by case

Rank	Case ID	Brief label	Total score (0–36)
1	A01	Comprehension Verification Chain (Lecture→Summary)	30.35
2	C02	Feedback Triage Assistant (Teacher-in-the-loop)	24.60
3	C03	Speaking Rehearsal + Reflection Prompts	23.80
4	C04	Reading Summary + Evidence Indexing	21.00
5	C10	Writing Revision Coach (Draft→Draft)	18.50
6	C07	Citation & Source Fidelity Checker (Formative)	18.00
7	C06	Peer Review Assistant + Audit	17.00
8	C05	Vocabulary Practice Generator (Curriculum-aligned)	15.90
9	C09	Office Hours Intake Agent (Service)	14.70
10	C11	Translation/Paraphrase Helper (High dependency risk)	13.90
11	C08	Lesson Material Drafting (Teacher)	13.50
12	B01	AI Auto-Summary (Minimal controls)	5.35

Interpretation: The distribution emphasizes the functional role of the Canvas. The majority of practices are not inherently bad but rather incomplete, frequently lacking evidence bundles, safeguards, or proper governance. The instrument facilitates structured improvement pathways rather than simple approve or reject determinations.

(5) The comparison between strong and weak case profile

Table 7: Rubric profiles for two illustrative cases (strong vs weak), averaged across raters (0-3 per dimension)

Dimension	A01 Mean	B01 Mean
Problem clarity	3.00	1.00
Teacher moves alignment	2.67	1.00
Learner moves verifiable	3.00	0.33
Value pathway clarity	2.67	1.00
Evidence feasibility	2.67	0.33
Indicator validity	2.00	0.33
Risk identification	2.67	0.00
Safeguards strength	2.67	0.00
Governance clarity	2.33	0.00
Consistency & fairness	2.33	0.33
Scalability & cost	1.67	1.33
Improvement loop	2.67	0.00
Total (0–36)	30.35	5.35

Table 8: Profile plot (text) for A01 vs B01 (0-3 per dimension)

Dim:	1	2	3	4	5	6	7	8	9	10	11	12
A01:	3	2.7	3	2.7	2.7	2	2.7	2.7	2.3	2.3	1.7	2.7
B01:	1	1	0.3	1	0.3	0	0	0	0	0.3	1.3	0

Analysis: The rubric uses both total scores and diagnostic dimension profiles to measure readiness. A01’s strengths are in important areas of the program, such as verifiable learner actions, achievable evidence collections, safeguards, governance, and cycles of continuous improvement. B01’s higher scalability and cost score means that it will be easy to put into action. But the lack of verification, evidence planning, and safety measures shows why a low

cost doesn't always mean that a program is ready.

5. Discussion

(1) Theoretical and methodological contribution

This research recontextualizes AI-supported instruction as a manageable practice rather than merely a tool deployment. The common problem of treating AI fluency as learning is solved by using a one-page Canvas and a scored rubric that clearly explains how things work, how to check them, and how to protect them. As shown in Figures 2 and 3, stakeholder validation and pilot reliability can be used to help with iterative refinement. The panel determines the most important constructs (verification, evidence feasibility, safeguards), and the agreement results show where anchors and documentation standards need to be improved (indicator validity, scalability/cost). The concreteness gradient in Figure 4 provides a useful design tip, which shows dimensions that need contextual judgment will stay noisier unless the instrument gives stronger examples, clearer anchor criteria, and fewer case-brief requirements. Meanwhile, the instrument's total-score reliability (ICC=0.79) shows that it can already help with program-level comparison and prioritization.

(2) Practical implication for teaching staffs

The score distribution shows that most practices are in the middle range. This means that program governance should not be seen as supporting or opposing the use of AI. Instead, it should be seen as helping practices get ready for the program by finishing evidence bundles, safeguards, and governance elements. The Canvas allows for three workflows: (1) building an exemplar bank by standardizing the documentation of mechanisms, verification, evidence, safeguards, and governance; (2) calibration and fairness through multi-rater scoring sessions that lower cross-class variability; and (3) conditional adoption based on diagnostic profiles (e.g., strong learning value but weak safeguards) instead of all-or-nothing judgments.

(3) The design principle of data-minimized evaluation

A data-minimized approach is both feasible and acceptable, as evidenced by panel ratings and pilot results. This method simultaneously generates interpretable evidence that is associated with quality processes and learning claims, while simultaneously reducing vendor dependence and surveillance concerns. In future iterations, the anchors for indicator-validity ($\kappa=0.41$) and scalability/cost ($\kappa=0.39$) should be strengthened by providing examples of valid and invalid indicators, aligning evidence expectations with existing assessment rubrics and moderation routines, and clarifying role expectations for resource trade-offs.

6. Future work and limitations

This paper presents preliminary validation utilizing a limited panel and a constrained experimental case set. The outcomes are dependent on the thoroughness of the case briefs. Future research should (a) assess the instrument across various institutions and programs, (b) perform longitudinal implementation studies to determine if the Canvas diminishes practice variability and improves program consistency, and (c) evaluate specific refinements for low-agreement dimensions through supplementary anchor exemplars, rater training, and standardized case-brief checklists.

7. Conclusion

As AI-supported pedagogy becomes more common, especially in the context of more complex AI-agent

workflows^{[11][14]}, program teams need a simple tool that clearly shows evidence, safeguards, and governance, as well as links practice design to program priorities. Descriptions that focus on tools are not enough for program decisions because they do not explain how they work or how to check them, and they do not include the controls and ownership that are needed for responsible, scalable adoption^{[17][12]}.

This study developed and preliminarily validated the AI Teaching Value Alignment Canvas and its corresponding rubric. The instrument defines AI-supported practices as value pathways that set AI intervention apart from what teachers and students do. Additionally, it includes a data-minimized evidence strategy that focuses on practical performance, operational, and governance indicators, as well as risk safeguards and light governance.

Illustrative stakeholder ratings and pilot scoring results suggest that the instrument can be used in different roles and can tell the difference between different levels of readiness in ways that are important to the program. The overall total-score reliability allows for program-level comparison, and the patterns of agreement show that there are different areas that need to be improved, especially indicator validity and scalability. By iteratively strengthening anchor exemplars and standardizing case documentation, the Canvas can help programs turn scattered AI teaching innovations into program assets that can be audited, compared, and improved. This method makes it possible to use these new ideas in a wide range of teaching settings in a responsible and scalable way.

References

- [1] Akhasbi, H., El Kamoun, N., Lakrami, F., Gilles, J. L., Rigo, J. M., & Aliss, E. (2025). Personalizing formative feedback through generative artificial intelligence: An experiment in higher education. In *EDULEARN25 Proceedings* (pp. 9338-9347). IATED.
- [2] Alfaleh, M. (2026). Sustainable AI-Driven Assessment in Higher Education: A Systematic Review of Fairness, Transparency, Pedagogical Innovation, and Governance. *Sustainability*, 18(2), 785.
- [3] Anaroua, F. I., Li, Q., Tang, Y., & Liu, H. P. (2025). AI-driven formative assessment and adaptive learning in data-science education: Evaluating an LLM-powered virtual teaching assistant. *arXiv preprint arXiv:2509.20369*.
- [4] Akindipe, D., & Lalude, O. M. (2025). Balancing public safety and individual rights in AI-driven public administration in Nigeria. In *AI Driven Tools for Sustainable Public Administration* (pp. 53-80). IGI Global Scientific Publishing.
- [5] Bakar, S., & Tapsoba, R. (2026). Artificial Intelligence in Classroom Teaching: Prospects, Challenges and Framework for Responsibly Orchestrated Mediation. *Social Science Chronicle*, 6, 1-21.
- [6] Chou, C. M., Shen, T. C., Shen, T. C., & Shen, C. H. (2025). Teachers' adoption of AI-supported teaching behavior and its influencing factors: Using structural equation modeling. *Journal of Computers in Education*, 12(3), 853-896.
- [7] Chen, J., & Alibakhshi, G. (2026). AI-Powered Applications' Effects on English Language Learners' Cognitive, Metacognitive, and Resource Management Strategies, and Language Achievement. *Journal of Computer Assisted Learning*, 42(1), e70171.
- [8] Chandafa, M. J., & Huang, F. (2025). Artificial Intelligence Enhanced Instruction: Reflection on Teachers and Students' Perceptions and Practices. *International Journal of Technology in Education*, 8(4), 1101-1128.
- [9] Cefa, B., Macgilchrist, F., ElGamal, H., Bai, J. Y., Zawacki-Richter, O., & Loglo, F. S. (2025). Responses to the initial hype: ChatGPT supporting teaching, learning, and scholarship?. *Open Praxis*, 17(2), 227-250.
- [10] Hargreaves, S. (2025). ChatGPT and other generative AI tools as university teaching aids. *Legal Education Review*, 35(1), 55-74.
- [11] Habiyambere, J. D., & Niyibizi, O. (2025). Assessing the Influence of AI-Supported Learning Methods on Instructional Quality in Higher Education. *Educational Journal of Technology and Innovation*, 2(1).
- [12] Ibrahim, H. (2025). A Suggested Framework for Integrating Artificial Intelligence in Foreign Language Education: Implications for EFL in the Middle East. *English Language and Literature in Education Journal*, 3(2), 53-69.
- [13] Jiang, L., & Lai, C. (2025). How did the generative artificial intelligence-assisted digital multimodal composing process facilitate the production of quality digital multimodal compositions: toward a process-genre integrated model. *Tesol Quarterly*, 59, S52-S85.
- [14] Kaushica, J. K. (2026). Technology-Integrated Teaching in Higher Education: An Empirical Study on AI-Supported Learning, Digital Tools, and LMS Innovations. *Sri Lankan Journal of Applied Sciences*, 4(2), 36-44.
- [15] Khan, I., Tufail, M., Naseer, M. U., & Shahzad, N. (2025). Investigating how AI tools can assist teachers in refining their pedagogy and enhancing student learning experiences. *The Critical Review of Social Sciences Studies*, 3(2), 383-396.
- [16] Khoza, N. G., & Van Der Walt, F. (2025, December). A systematic review on AI-enhanced pedagogies in higher education in the Global South. In *Frontiers in Education* (Vol. 10, p. 1667884). Frontiers Media SA.
- [17] Kim, J., Detrick, R., Yu, S., Song, Y., Bol, L., & Li, N. (2025). Socially shared regulation of learning and artificial intelligence:

- Opportunities to support socially shared regulation. *Education and Information Technologies*, 30(9), 11483-11521.
- [18]Leiva-Guerrero, M. V., Araya Zamorano, I., Escobar Collins, R., & Silva Castro, F. (2026). Feedback on learning with generative artificial intelligence in university students. *RIED: Revista Iberoamericana de Educación a Distancia*, 29(1).
- [19]Norton, J. (2026). Designing and Assessing AI-Enabled Programs in the Helping Professions: Leadership, Learning Design, and Accreditation. In *AI-Powered Learning in Social Work, Counseling, and Psychology* (pp. 197-226). IGI Global Scientific Publishing.
- [20]Otto, F., Kling, N., Schumann, C. A., & Tittmann, C. (2023). Conceptual approach to an AI-based adaptive study support system for individualized higher education. *International Journal of Advanced Corporate Learning (IJAC)*, 16(2), 69-80.
- [21]Peng, X., Sun, K., Shan, X., & Zhang, J. (2026). Empowerment or dependency? A systematic review of the impacts of intelligent assessment and generative AI on learners' self-beliefs and cognitive agency in music education. *Frontiers in Psychology*, 17, 1776445.
- [22]Peinemann, K., & Peinemann, M. (2025). Potentials and Limitations Of Digital Education: Creating Personalization, Collaboration, And Equity In The Age Of Ai-Enhanced Learning Environments. In *ICERI2025 Proceedings* (pp. 2300-2310). IATED.
- [23]Prompiengchai, S., Narreddy, C., & Joordens, S. (2025). A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI. *arXiv preprint arXiv:2505.23405*.
- [24]Sangwa, S., Mutabazi, P., & Muvunyi, J. B. (2025). AI-Enabled Framework for Program and Course Design in Higher Education.
- [25]Sidorkin, A. M. (2026). *AI-Enhanced Pedagogies: Rethinking Learning, Curriculum, and Human Potential in the Age of Intelligent Machines*. Taylor & Francis.
- [26]Sulolipu, A. A., Syata, W. M., Ampa, A. T., Ram, S. W., Rosmayanti, V., Sabillah, B. M., & Thayeb, A. M. D. R. (2025). *The AI-Enhanced Educator: Multidisciplinary Approaches to Smart Teaching*. CV Eureka Media Aksara.
- [27]Wu, Z. (2025). *Designing and Evaluating Fine-Grained Interactive Practice Tools for Novice Programming Learners* (Doctoral dissertation).
- [28]Yusmaneli, Y., Fathurrahmad, F., & Raihani, A. (2025). A hybrid "probabilistic scaffold" model for enhancing reading literacy, numeracy, and AI literacy among Indonesian primary school students. *Design Journal*, 3(2), 78-98.
- [29]Yan, L., Suleman Abdullah Alwabel, A., & Mohamad, U. H. (2025). AI-Powered Education: Transforming Teacher-Student Interactions and Advancing Sustainable Learning Practices. *European Journal of Education*, 60(4), e70351.
- [30]Zang, Z. (2025), Investigating the Intrinsic Mechanisms by which AI-driven Educational Games Enhance Student Cognition.

Appendix: Pilot case set overview (12 cases)

Case ID	Brief label	Coverage category	Total readiness score (0–36)
A01	Comprehension Verification Chain (Lecture→Summary)	Learner-facing + verification	30.35
B01	AI Auto-Summary from Notes (Minimal controls)	Learner-facing (tool-only)	5.35
C02	Feedback Triage Assistant (Teacher-in-the-loop)	Workload/decision support	24.60
C03	Speaking Rehearsal + Reflection Prompts	Practice + metacognition	23.80
C04	Reading Summary + Evidence Indexing	Learner-facing + verification	21.00
C05	Vocabulary Practice Generator (Curriculum-aligned)	Practice + assessment support	15.90
C06	Peer Review Assistant + Audit Sampling	Peer learning + quality control	17.00
C07	Citation & Source Fidelity Checker (Formative)	Integrity + learning support	18.00
C08	Lesson Material Drafting for Teachers	Teaching operations/efficiency	13.50
C09	Office Hours Intake Agent (Service)	Service experience/triage	14.70
C10	Writing Revision Coach (Draft→Draft)	Writing development	18.50
C11	Translation/Paraphrase Helper (High dependency risk)	Support (risk-prone)	13.90