

DOI: https://doi.org/10.48009/2_iis_2026_108

Detecting AI-Completable Assignments: An Explainable Generative AI Framework for Assessment Governance

Jonathan Hobbs, *Georgia Southwestern University*, jonathan.hobbs@gsw.edu

Abstract

As generative AI becomes more common in student academic work, many traditional assignments can now be completed by general-purpose language models with limited student-specific reasoning, process evidence, or contextual engagement. This paper introduces AI-completable tasks as a practical construct for assessment governance and reframes AI-related risk as an assignment design problem rather than only a post-submission misconduct issue. A faculty-side generative AI evaluator is used to review assignment prompts, assess substitution risk, assign an AI-resistance score, identify structural weaknesses, and recommend targeted revisions. The pilot included 100 assignments from seven faculty members, with 32 assignments completing the full pre/post revision cycle. Mean evaluator-assessed AI-resistance scores increased from 4.16 to 6.59, while Red Zone classifications declined from 62.5% to 12.5%. The findings suggest that modest changes, including contextual grounding, process evidence, student-specific artifacts, and trade-off justification, can reduce evaluator-assessed AI-completable assignments without changing the original learning objectives. The paper contributes a practical framework for helping faculty move AI governance upstream into assignment design.

Keywords: Generative AI, Assessment Governance, AI-Completable Tasks, Explainable AI, Assignment Design, Academic Integrity

Introduction

Generative artificial intelligence has changed how students complete academic work. Large language models such as ChatGPT can produce fluent and technically reasonable responses across essays, programming exercises, documentation, explanations, and applied problem-solving tasks. This creates a practical assessment problem because some assignments that once served as reasonable evidence of learning can now be completed with limited student input. Prior research has shown that ChatGPT can perform effectively on business analytics assignments, raising concerns about academic integrity, skill development, and instructor oversight (Laker & Sena, 2023).

The issue is especially important in information systems, computer science, and related fields, where generative AI is also becoming a normal professional tool for coding, debugging, documentation, and technical explanation. Students often view AI as useful, while faculty remain more cautious about authorship, ethics, and appropriate use (Koohang et al., 2024; Sah et al., 2025). These differences make broad policy statements difficult to apply consistently.

Many institutional responses have focused on detection, prohibition, or acceptable-use rules. These approaches do not fully address the assessment problem. AI-generated work is not traditional plagiarism, and detection tools can produce both false positives and false negatives (Dalalah & Dalalah, 2023). Detection also occurs after submission, leaving faculty to react to possible misuse rather than address assignment risk before release.

This paper introduces the concept of an AI-completable task, defined as an academic task that a generic LLM can complete to a high standard without student-specific context, local course artifacts, or process evidence. It also presents a faculty-side evaluation framework for reviewing and revising assignments before students receive them. Using a multi-faculty pilot of 100 assignments, the paper contributes a practical construct, an explainable evaluator for identifying assignment-level substitution risk, and evidence that targeted revisions can improve evaluator-assessed AI resistance. Figure 1 summarizes the shift from detection-centered governance to design-centered governance by moving AI oversight upstream into assignment review.

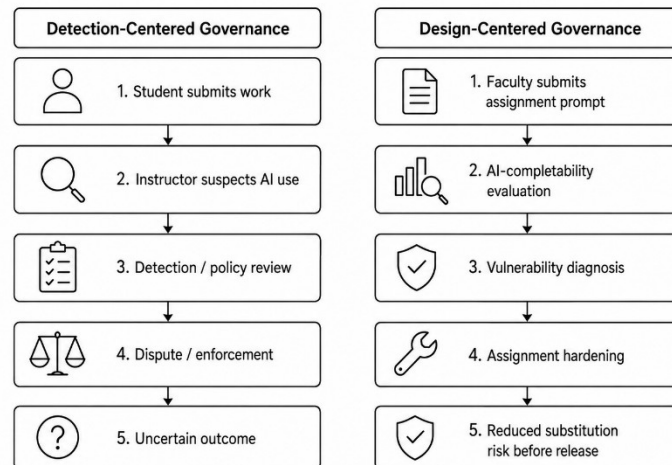


Figure 1. From Detection-Centered to Design-Centered AI Governance

Background and Literature Review

Academic Integrity in the Era of Generative AI

Generative AI is becoming part of both academic and professional work in information systems, computer science, and related fields. Students can use these tools to draft explanations, generate code, debug errors, summarize documentation, and organize technical responses. Prior research has shown that generative AI can support teaching and academic work while also raising concerns about how faculty evaluate learning in AI-enabled environments (Ngo-Ye & Choi, 2024). Laker and Sena (2023) similarly found that ChatGPT could perform effectively on business analytics tasks, creating concerns about academic integrity, foundational skill development, and instructor oversight.

Student and faculty views also differ. Students often see AI as useful, while faculty express greater concern about ethics, authorship, and appropriate use (Koohang et al., 2024; Sah et al., 2025). This makes it difficult to rely on policy alone.

Limits of Detection-Centered AI Governance

Many early responses to generative AI emphasized detection, prohibition, or broad acceptable-use statements. These approaches do not fully address the assessment problem. AI-generated work does not function like traditional plagiarism, and detection tools can produce both false positives and false negatives (Dalalah & Dalalah, 2023).

Detection also occurs after the assignment has been submitted. By that point, the instructor is placed in an enforcement role rather than a design role. This reactive approach does not answer the more important question of whether the assignment itself allowed a generic AI system to replace the intended learning process.

Assessment Design and AI-Resistant Learning

Recent research argues that assessment design must change in response to generative AI. Moorhouse et al. (2023) found that university guidance varies widely in its level of practical detail, while Larsen et al. (2023) recommended contextualized tasks, higher-order reasoning, and formats that are more difficult for generic AI systems to complete.

Authentic assessment provides a useful foundation, but authenticity alone is not enough. A task may use a realistic scenario and still be AI-completable if it lacks student-specific context, process evidence, or local constraints. Kofinas et al. (2025) similarly argued that generative AI requires educators to reconsider how assessment evidence is produced and validated.

Explainable and Faculty-Controlled AI Support

Generative AI can also support faculty as a design tool. In this role, AI can help identify why an assignment is vulnerable and recommend targeted revisions. This aligns with the need for transparent and human-controlled AI systems. Sargent et al. (2024) found that ethical concerns influence perceived trust and risk, reinforcing the importance of systems that make their reasoning visible.

The literature shows that generative AI can complete many academic tasks, that detection is limited, and that assessment design must evolve. What remains underdeveloped is a practical and explainable faculty-facing method for identifying AI-completable assignments before release.

Conceptual Framework: AI-Completable Tasks

Defining AI-Completable Tasks

An AI-completable task is an academic task that a generic LLM can complete to a high standard without student-specific context, local course artifacts, or process evidence. The focus is on task design rather than student intent. The question is not whether a student used AI dishonestly, but whether the assignment allows AI to replace the intended learning process.

This distinction matters because many traditional assignments still appear academically appropriate but rely heavily on generic prompts and final-product grading. In those cases, an LLM may produce a polished response without demonstrating that the student completed the intended analysis, technical work, or decision-making.

AI Assistance Versus AI Substitution

AI assistance occurs when students use AI to support part of the learning process, such as brainstorming, clarification, grammar, or syntax checking. AI substitution occurs when AI can complete the task end-to-end with little student-specific contribution.

Whether AI assists or replaces learning depends on what students must produce, document, explain, and defend. Assignments that require local context, process evidence, verifiable artifacts, or justification make substitution more difficult while still allowing reasonable AI support.

AI-Completeness as a Continuum

AI-completeness should be treated as a continuum. Low AI-completable tasks require local artifacts, observed outputs, process evidence, or decisions tied to a specific context. Moderate tasks allow AI to draft portions of the response but still require student interpretation, evidence, or defense. High AI-completable tasks can be completed with generic prompting and little student-specific input.

This extends authentic assessment by focusing specifically on whether a task requires evidence that a general-purpose AI system cannot easily produce on its own (Kofinas et al., 2025; Wiggins, 1990).

Permission Zones and Design Alignment

The continuum supports three permission zones. Green Zone tasks allow AI assistance but resist full substitution. Yellow Zone tasks permit AI use with disclosure, attribution, revision evidence, or defense. Red Zone tasks should be redesigned because a generic AI system can complete most or all of the work without meaningful student-specific evidence.

The framework does not require faculty to ban AI or redesign every assignment. It provides a practical way to align AI permissions with the evidence required by the task. Table 1 summarizes the continuum and corresponding governance responses.

Table 1. AI-Completeness Continuum and Permission Spectrum

AI-Completeness Level	Permission Zone	Task Characteristics	Role of AI	Governance Response
Low AI-Completable	Green Zone	Requires student-specific context, local artifacts, process evidence, observed outputs, or decisions tied to a specific course setting.	AI may support brainstorming, clarification, grammar, syntax, or organization, but cannot easily complete the task end-to-end.	Permit AI assistance with normal transparency expectations. Grade the student's evidence, reasoning, and application.
Moderate AI-Completable	Yellow Zone	AI can draft portions of the response, but the assignment still requires student-provided evidence, interpretation, attribution, reflection, or defense.	AI may generate useful starting points, but students must document, adapt, explain, and defend their work.	Allow AI use with disclosure, attribution, revision evidence, or brief written/oral defense. Strengthen weak areas where needed.
High AI-Completable	Red Zone	Generic prompt, product-only grading, no local context, no process evidence, no required justification, and little need for student-specific decision-making.	AI can complete the task end-to-end with minimal prompting and limited risk of failure.	Redesign before release. Add context, constraints, process evidence, proof of execution, trade-off reasoning, or verification.

Faculty-Side AI-Resistance Evaluator

Purpose and Workflow

The Faculty-Side AI-Resistance Evaluator helps instructors identify AI-completable assignments before release. It does not detect misconduct or determine whether a student used AI. Instead, it tests whether a generic LLM could produce a submission that would likely meet the assignment requirements.

Faculty provide the assignment prompt and, when available, the rubric or learning objectives. The evaluator reviews task structure, required evidence, student-specific requirements, and opportunities for AI substitution. It then assigns an AI-resistance score from 0 to 10, explains the score, identifies vulnerabilities, and recommends targeted revisions. Faculty remain responsible for deciding whether those recommendations fit the course and learning objectives (Amershi et al., 2019; Arrieta et al., 2020; Shneiderman, 2020).

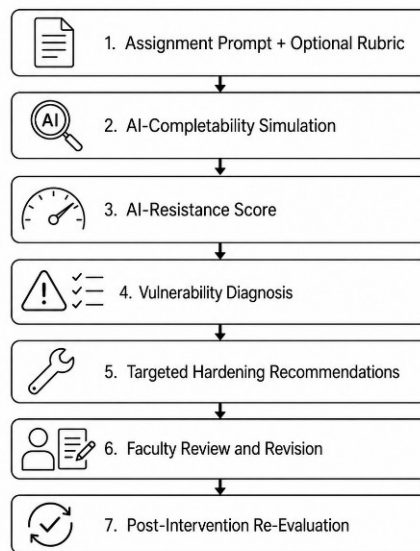


Figure 2. Faculty-Side AI-Resistance Evaluation Workflow

Anchored AI-Resistance Scoring Rubric

To make the evaluator's decision logic more transparent, recurring criteria from the pilot were formalized into five observable dimensions. Each dimension is scored from 0 to 2, producing a total score from 0 to 10. The rubric was developed after the initial pilot to support future replication and independent validation. It was not used to generate the original scores reported in this study.

Table 2. Anchored AI-Resistance Scoring Rubric

Dimension	0 Points	1 Point	2 Points
Contextual Grounding	Task can be answered generically using public information.	Some course or student context is included, but it is optional or weakly connected to the deliverable.	Completion requires student-specific, local, or instructor-controlled information.
Process Evidence	Only the final product is assessed.	Limited process evidence is requested but is not essential to the final evaluation.	Students must provide staged, observable, or documented evidence of how the work was completed.
Artifact Verifiability	No artifacts are required, or artifacts can be easily	Artifacts are required but are only loosely	Artifacts are traceable to actual execution,

	fabricated without task completion.	connected to the student's work.	authenticated activity, or internally consistent records.
Reasoning and Trade-offs	Task emphasizes recall, description, or a single expected answer.	Some explanation or justification is required.	Students must evaluate constraints, compare alternatives, justify decisions, or explain trade-offs.
Verification or Defense	No verification of student understanding is included.	Verification is optional, indirect, or limited to written statements.	The task includes required defense, targeted questioning, spot-checking, or equivalent verification.

The total score is interpreted using three permission zones:

- 0-4: Red Zone. A generic LLM can likely complete most or all of the assignment without meaningful student-specific evidence.
- 5-7: Yellow Zone. AI can produce substantial portions of the work, but successful completion still requires some student evidence, interpretation, or defense.
- 8-10: Green Zone. AI may assist, but the assignment requires enough contextual, procedural, or verifiable evidence to prevent straightforward end-to-end substitution.

The numerical score is accompanied by a separate AI-completeness judgment of Yes, Partially, or No. This judgment addresses whether a generic LLM could produce an acceptable submission without access to the student's actual context, artifacts, execution, or decision process.

Vulnerability Categories and Hardening Strategies

The evaluator uses fixed vulnerability categories, including generic prompt framing, product-only evaluation, limited contextual grounding, narrative deliverables without verification, missing process evidence, and limited justification or trade-off analysis.

Recommended hardening strategies include student-specific artifacts, local course materials, annotated evidence, observed outputs, staged work, rejected alternatives, trade-off explanations, and brief defenses. These changes strengthen the evidence required by the task without necessarily changing its topic or learning objective.

Human Oversight and Governance Role

The evaluator supports faculty judgment rather than automating assessment design. Instructors may accept, reject, or modify recommendations based on course level, workload, accessibility, and learning goals. Its main value is moving assignment review before submission rather than relying only on detection or enforcement afterward.

Methodology

Research Design

This study used a design-centered, artifact-based pilot approach. The unit of analysis was the assignment rather than the student submission. Following design science principles, the study identified a practical assessment problem, developed a faculty-facing evaluation artifact, applied it across course assignments, and examined changes in evaluator-assessed AI resistance (Gregor & Hevner, 2013; Hevner et al., 2004; Peffers et al., 2007).

Data Sources

The pilot included 100 assignments contributed by seven faculty members at a public regional university. Most assignments came from undergraduate computer science and information technology courses, with a smaller number from non-computing disciplines, including exercise science.

Thirty-two assignments completed the full pre/post cycle. For these assignments, the data included the original prompt, baseline evaluator output, revised prompt, and post-intervention evaluator output. The remaining assignments contributed baseline vulnerability findings but were not included in paired score comparisons.

The four detailed cases were selected from the complete pairs to represent different computing assignment formats and redesign outcomes. They were chosen as illustrative examples rather than only as the assignments with the largest score increases.

Evaluator Configuration

The evaluator was implemented as a Custom GPT using GPT-4o during Fall 2025 and Spring 2026. Each assignment was reviewed using the same fixed instructions, scoring scale, vulnerability categories, and output structure. Because the Custom GPT interface did not allow direct control of response variability settings, the platform's default settings were used for all evaluations.

Evaluation Procedure

Each original assignment was submitted for baseline analysis. The evaluator assigned a score from 0 to 10 and identified vulnerabilities such as generic prompt framing, product-only evaluation, limited contextual grounding, missing process evidence, and limited justification or trade-off analysis. Each original and revised assignment was scored once using the fixed evaluator instructions; repeat scoring was not conducted to assess within-model consistency.

Faculty then reviewed the findings and revised assignments where the recommendations fit the course and learning objectives. Common revisions included local system details, annotated evidence, expected-versus-actual interpretation, instructor-provided case materials, and decision justification.

The revised assignment was then submitted using the same locked framework. In this study, locked means that the evaluator instructions, original scoring scale, vulnerability categories, and output format remained unchanged across pre- and post-intervention scoring. The five-dimension rubric in Table 2 was formalized after the pilot and was not used to calculate the original reported scores.

Analysis

The analysis was descriptive rather than inferential. It examined changes in evaluator-assessed AI-resistance scores, vulnerability patterns, and permission-zone classifications.

Vulnerability and redesign frequencies were derived from the evaluator's fixed output categories. Assignments could receive more than one category, so the counts were not mutually exclusive. Baseline vulnerability frequencies were calculated across the broader assignment set, while redesign frequencies were calculated across the 32 revised assignments.

The analysis evaluates assignment artifacts and evaluator outputs. It does not measure student learning, misconduct, or behavioral change.

Table 3. Pilot Study Design Summary

Component	Description
Research Approach	Design-centered, artifact-based pilot study
Unit of Analysis	Assignment prompts and related assessment artifacts
Faculty Participation	Seven faculty members contributed assignments and participated in the evaluation process
Course Context	Primarily undergraduate computer science and information technology courses, with limited non-computing examples for transferability
Pilot Sample	100 assignments, including 32 complete pre/post pairs
Evaluation Tool	Faculty-Side AI-Resistance Evaluator, version 2.0, using fixed instructions, scoring scale, vulnerability categories, and output structure
Analysis Focus	Changes in scores, permission zones, vulnerability patterns, and redesign strategies

Results

Overview of Pilot Findings

100 assignments from seven faculty members were reviewed using the AI-resistance framework. Thirty-two assignments completed the full pre/post revision cycle and were included in the paired analysis. The remaining assignments contributed baseline vulnerability findings but were not included in score comparisons.

Assignments with generic prompts, product-only grading, limited context, and little process evidence generally received lower AI-resistance scores. Revised assignments received higher scores when they added student-specific context, verifiable artifacts, process evidence, expected-versus-actual analysis, or trade-off justification.

Aggregate Results for Complete Pre/Post Pairs

Across the 32 complete pairs, the mean evaluator-assessed AI-resistance score increased from 4.16 to 6.59, while the median increased from 4.00 to 7.00. Red Zone classifications declined from 20 assignments to 4, while 8 revised assignments reached the Green Zone. Thirty assignments received higher post-intervention scores and two remained unchanged.

Table 4 summarizes the paired results. These scores reflect evaluator-assessed changes in assignment design and should not be interpreted as direct evidence of improved learning, reduced misconduct, or changed student behavior.

Table 4. Aggregate Results for 32 Complete Pre/Post Assignment Pairs

Measure	Original assignments	Revised Assignments	Change
Number of assignments	32	32	-
Mean AI-resistance score	4.16	6.59	+2.44
Median AI-resistance score	4.00	7.00	+3.00
Standard deviation	1.27	1.32	-
Minimum-maximum score	2-7	3-8	-
Red Zone, n (%)	20 (62.5%)	4 (12.5%)	-16
Yellow Zone, n (%)	12 (37.5%)	20 (62.5%)	+8
Green Zone, n (%)	0 (0.0%)	8 (25.0%)	+8
Assignments with increased scores	-	30 (93.8%)	-
Assignments with unchanged scores	-	2 (6.3%)	-
Assignments with decreased scores	-	0 (0.0%)	-

Case Selection

Table 5 presents four cases selected from the 32 complete pairs to show variation in assignment type and redesign outcome. The cases include technical execution, certification-based reflection, systems administration, and software engineering. They were selected as illustrative examples rather than only because they produced the largest score increases.

Table 5. Representative Pre-Post AI-Resistance Results

Assignment	Type	Pre	Post	Change	Zone Shift	Primary Design Change
A1	Linux command/concept exercises	3	7	+4	Red to Yellow	System context annotated execution +
A2	GitHub certification reflection	4	6	+2	Yellow strengthened	Artifact-grounded reflection
A3	Linux directories/archiving	5	7	+2	Yellow strengthened	Expected-vs.-actual interpretation
A4	Agile/requirements engineering	2	8	+6	Red to Green	Case artifacts + trade-off reasoning

Representative Assignment-Level Findings

A1: Linux command and concept exercises. The original assignment received a score of 3 because a generic LLM could answer the questions and produce plausible command explanations without actual execution. The revised version required local system details, annotated evidence, observed outputs, and justification of alternatives, increasing the score to 7.

A2: GitHub certification reflection. The original assignment received a score of 4 because the reflection could still be completed with a generic AI-generated response. The revision required references to specific GitHub modules, features, and workflow decisions, increasing the score to 6.

A3: Linux directories and archiving. The original assignment received a score of 5 because AI could predict standard command results and explanations. The revised version added local system context, annotated screenshots, expected-versus-actual analysis, and failure explanation, increasing the score to 7.

A4: Agile and requirements engineering. The original assignment received a score of 2 because students could answer using general software engineering knowledge. The revision added instructor-provided case materials, stakeholder notes, structured requirements artifacts, rejected alternatives, and process evidence, increasing the score to 8.

Across the four cases, higher scores were associated with stronger evidence requirements rather than changes to the original learning objectives. The largest improvement occurred when revisions combined instructor-controlled materials, process evidence, and decision justification.

Common Vulnerability Patterns

Across the broader baseline assignment set, vulnerabilities were coded using the evaluator's fixed diagnostic categories. Because assignments could contain multiple vulnerabilities, category totals were not mutually exclusive. Table 6 shows the most frequently identified structural weaknesses.

Table 6. Common Vulnerability Patterns Across the Baseline Assignment Set

Vulnerability Category	Assignments identified, n (%)	Interpretation
Generic prompt framing	73 (73%)	The task could be answered using general knowledge without course-specific evidence.
Product-only evaluation	68 (68%)	Grading emphasized the final response or artifact rather than how it was produced.
Limited contextual grounding	65 (65%)	The task did not require student-specific, local, or instructor-controlled information.
Limited process evidence	61 (61%)	Students were not required to document intermediate work, execution, or revision.
Narrative without verification	57 (57%)	Written explanations could be generated without demonstrating actual task completion.
Limited justification or trade-off analysis	52 (52%)	Students were not required to compare alternatives or explain decisions.
No interactive or authenticity check	44 (44%)	The task lacked a defense, spot-check, live demonstration, or comparable verification method.

Generic prompt framing was the most common weakness, followed by product-only evaluation, limited contextual grounding, and missing process evidence. These findings suggest that AI-completeness usually resulted from several overlapping design weaknesses rather than a single feature.

Common Redesign Patterns

The 32 revised assignments were also coded for the hardening strategies used. More than one strategy could appear in the same assignment.

Table 7. Common Redesign Patterns Across 32 Revised Assignments

Redesign pattern	Assignments using strategy, n (%)	Primary purpose
Student-specific or local context	25 (78.1%)	Prevents fully generic responses by connecting work to the student's system, case, or experience.
Annotated process evidence	23 (71.9%)	Requires students to explain outputs, artifacts, or intermediate work.
Expected-versus-actual analysis	19 (59.4%)	Requires comparison between predicted and observed results.
Rejected alternatives or trade-offs	18 (56.3%)	Moves the task from description toward decision-making and justification.
Instructor-provided case materials	14 (43.8%)	Grounds responses in controlled or non-public information.
Brief defense or spot-check	11 (34.4%)	Provides an additional check of student understanding or authorship.

Student-specific context and annotated process evidence were the most common redesign strategies. Expected-versus-actual analysis and trade-off reasoning were also used frequently. Brief defenses and spot-checks appeared less often, suggesting that faculty generally preferred changes that fit within existing assignment and grading workflows.

Discussion

From Detection to Design

The findings support moving AI governance earlier in the assessment process. Detection-centered approaches place faculty in a reactive role after work has been submitted, while a design-centered approach focuses on whether the assignment itself allows AI to replace the intended learning process. This is especially important because AI detection remains imperfect and can raise fairness concerns when used as evidence of misconduct (Dalalah & Dalalah, 2023).

AI-Completable Tasks as a Practical Signal

The AI-completable task construct gives faculty a practical way to identify assignment-level risk. Many vulnerable assignments were not poorly designed in a traditional sense. Their weakness was that they relied on generic prompts, final-product grading, and limited evidence of how the work was completed. The Red, Yellow, and Green zones provide a clear way to connect those design risks with appropriate AI-use expectations.

What Reduced AI-Completeness

Higher evaluator-assessed scores were most often associated with student-specific context, process evidence, verifiable artifacts, expected-versus-actual analysis, and decision justification. These changes strengthened the evidence of student work without requiring faculty to replace the original learning objectives. They also align with authentic assessment by emphasizing applied reasoning and evidence of performance rather than final products alone (Kofinas et al., 2025; Wiggins, 1990).

Explainable and Faculty-Controlled AI Support

The evaluator's value is not limited to its score. It explains why an assignment may be AI-completable and connects each weakness to a possible redesign strategy. Faculty remain responsible for deciding whether the recommendations fit the course, workload, and learning goals. This supports a human-controlled approach in which AI assists professional judgment rather than replacing it (Arrieta et al., 2020; Shneiderman, 2020).

Practical Limitations

Assignment hardening may increase student or faculty workload when it adds annotated evidence, staged submissions, or brief defenses. These requirements should be proportionate to the learning objective and course level. Accessible alternatives may include text-based logs instead of screenshots, written rather than oral defenses, and flexible artifact formats. The pilot did not formally measure workload or accessibility outcomes, which should be examined in future studies.

For institutions, AI-completable tasks provide a shared language for reviewing assignment risk before relying on detection or broad policy statements. For assessment research, the construct can also be tested against faculty judgments, student performance, model completion success, and assessment-validity outcomes.

Limitations and Future Research

This study evaluated assignment artifacts and evaluator outputs rather than student submissions or learning outcomes. The findings therefore show changes in evaluator-assessed AI-completable, not direct evidence of improved learning, reduced misconduct, or changed student behavior.

A second limitation is that the same evaluator identified vulnerabilities, suggested revisions, and scored the revised assignments. Although fixed instructions, categories, and output structures reduced procedural variation, they did not eliminate the possibility that the evaluator favored revisions aligned with its own recommendations. The anchored rubric developed after the pilot provides a clearer basis for future independent validation.

The sample was also drawn primarily from computer science and information technology courses at one public regional university. Additional testing is needed across disciplines, course levels, institutions, faculty raters, and AI models.

Future research should compare evaluator scores with independent faculty ratings, student testing, and model completion success. It should also examine workload, accessibility, assessment validity, and whether lower AI-completable is associated with stronger student engagement or performance. A further direction is the development of semi-agentic workflows that can evaluate, revise, and re-evaluate assignments while keeping faculty approval in the process.

Conclusion

Generative AI has changed what many assignments can reasonably demonstrate about student learning. Tasks that once appeared appropriate may now be completed by a general-purpose LLM with limited

student-specific reasoning, context, or process evidence. This paper argues that the challenge should be addressed not only through detection or misconduct policy, but also through assignment design.

The AI-completable task construct gives faculty a practical way to identify when AI may replace the intended learning process. The pilot findings suggest that evaluator-assessed AI resistance can improve through focused changes such as local context, verifiable artifacts, process evidence, expected-versus-actual analysis, and trade-off justification.

The broader contribution is a shift from reacting to possible AI use after submission to reviewing assignment risk before release. In this approach, generative AI serves not only as a challenge to assessment integrity, but also as a tool for improving assessment design.

References

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, 21(2), 100822. <https://doi.org/10.1016/j.ijme.2023.100822>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.5555/2017212.2017217>
- Kofinas, A. K., Tsay, C. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, 56(6), 2522–2549. <https://doi.org/10.1111/bjet.13585>
- Koohang, A., Sargent, C. S., & Svanadze, S. (2024). Students' perceptions of benefits and opportunities of artificial intelligence (AI). *Issues in Information Systems*, 25(2), 438–450. https://doi.org/10.48009/2_iis_2024_134
- Laker, L. F., & Sena, M. (2023). Accuracy and detection of student use of ChatGPT in business analytics courses. *Issues in Information Systems*, 24(2), 153–163. https://doi.org/10.48009/2_iis_2023_113
- Larsen, S. K., Narayan, J., & Kapur, M. (2023). Creating large language model resistant exams: Guidelines and strategies. *arXiv*. <https://doi.org/10.48550/arXiv.2304.12203>

- Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, 5, 100151. <https://doi.org/10.1016/j.caeo.2023.100151>
- Ngo-Ye, T. L., & Choi, J. (2024). Practical use cases of generative AI in teaching and academic professional work: An exploratory study. *Issues in Information Systems*, 25(1), 434–449. https://doi.org/10.48009/1_iis_2024_135
- Peppers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Sah, R., Hagemaster, C., Adhikari, A., Lee, A., & Sun, N. (2025). Generative AI in higher education: Student and faculty perspectives on use, ethics, and impact. *Issues in Information Systems*, 26(2), 373–386. https://doi.org/10.48009/2_iis_129
- Sargent, C. S., Koohang, A., Floyd, K., & Kilburn, R. (2024). Artificial intelligence: Ethical concerns, trust, and risk. *Issues in Information Systems*, 25(2), 71–83. https://doi.org/10.48009/2_iis_2024_106
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, 2(2). <https://doi.org/10.7275/ffb1-mm19>