GENERATIVE ARTIFICIAL INTELLIGENCE IN FORMATIVE ASSESSMENT: TRANSFORMING FEEDBACK, STUDENT AGENCY, AND LEARNING OUTCOMES

Hasan Olimjonov
Jizzakh state pedagogical university
hasanolimjonov@jdpu.uz

Abstract. Generative artificial intelligence offers rapid feedback, rubric drafting, and explanatory dialogue, yet speed is not equivalent to formative quality. This conceptual review integrates established formative assessment theory with contemporary guidance on generative AI to propose a Feedback–Interpretation–Action (FIA) cycle for classroom use. The framework distinguishes feedback accuracy, students' interpretation of comments, and demonstrable revision or transfer. It specifies a three-layer checking procedure—disciplinary verification, pedagogical calibration, and fairness review—and offers an illustrative protocol for written assignments. The article argues that teachers should use AI to expand opportunities for formative dialogue while retaining final judgment over consequential assessment. Because no primary student data were collected, the paper presents propositions and a research agenda rather than measured effects. Its main contribution is an auditable instructional design that evaluates whether AI feedback actually enables learners to close a learning gap.

Keywords: generative artificial intelligence; formative assessment; feedback literacy; student agency; assessment validity; pedagogical design

1. Introduction

Formative assessment is a process of eliciting and interpreting evidence so that teachers and learners can modify ongoing instruction. Classic accounts emphasize that feedback is effective only when it helps learners recognize a gap between present work and intended quality and take steps to close it (Sadler, 1989; Black & Wiliam, 1998). Generative AI challenges the practical organization of this process: it can comment on drafts within seconds, hold follow-up conversations, and rewrite explanations, but its judgments may be inconsistent, unjustified, or detached from a rubric.

The pedagogical risk is to equate abundant commentary with useful learning. Students may receive ten suggestions but have little understanding of which matter most. Some may accept fluent advice uncritically, while others may become dependent on rewriting assistance. The purpose of this paper is to articulate conditions under which AI-supported feedback strengthens, rather than bypasses, students' capacity to evaluate and revise their own work.

2. Methodology and Conceptual Foundations

This is a theory-driven integrative review of formative assessment scholarship and selected literature on generative AI in education. The core conceptual anchors are the roles of evidence and instructional adjustment (Black & Wiliam, 1998), learner recognition of quality (Sadler, 1989), the directional questions of effective feedback (Hattie & Timperley, 2007), and self-regulation (Zimmerman, 2002). Risks of bias, inaccurate output, and human oversight were examined through UNESCO (2023) and Kasneci et al. (2023).

The synthesis asked three analytic questions: What information should feedback contain? What must learners do with that information? What quality-control responsibilities remain with educators? This is not a systematic literature search, meta-analysis, intervention, or student survey. The resulting model is a theoretically reasoned proposal, and claims about future outcomes are framed as testable hypotheses.

3. Results: A Feedback–Interpretation–Action Cycle

Stage A, evidence and criteria: before any AI feedback, teachers define the learning objective, a small number of observable success criteria, and an authentic student task. Students submit a first attempt without automated rewriting. This creates evidence of unaided performance and prevents the tool from obscuring the target skill. An AI system may then produce provisional comments that cite exact passages or steps in the student's response, identify strengths, and suggest one or two high-impact revisions.

Stage B, learner interpretation: students translate comments into their own language, identify which suggestion is supported by the rubric, and reject at least one suggestion when justified. This reflective step is essential because feedback that remains only in a machine-generated message cannot demonstrate learning. Peer discussion and teacher questioning can help reveal whether students understand the underlying concept rather than merely follow surface corrections.

Stage C, action and verification: learners submit a revision log explaining what changed, why it changed, and what was not adopted. A brief transfer task assesses whether the improvement carries over to a fresh problem completed without AI. Teachers compare the original response, revision, and transfer performance to decide whether the loop has closed. If understanding remains weak, instruction changes; the system does not simply issue more comments.

4. Quality Assurance and Assessment Integrity

A practical quality gate has three layers. Disciplinary verification asks whether each AI comment is correct, evidence-based, and consistent with the intended answer domain. Pedagogical calibration asks whether suggestions are understandable at the student's level, specific enough to act on, and proportionate to the learning objective. Fairness review asks whether language variety, disability, cultural context, or writing style is being penalized without curricular reason. Teachers retain authority to discard or rewrite outputs.

For example, in an argumentative essay assignment, the approved rubric might specify claim clarity, evidence use, counterargument, and coherence. The teacher provides anonymized excerpts rather than identifiable student records when tool policy permits such use. AI drafts two evidence-linked comments; the student assesses those comments against the rubric and submits a revised paragraph plus a 120-word explanation. The teacher evaluates both the writing and the explanation. This is an illustrative design, not a report of a conducted classroom experiment.

Institutional policy should distinguish low-stakes formative support from consequential grading. Automatically assigning final grades through a general-purpose chatbot raises questions of validity, explainability, bias, and appeals. Where AI helps educators draft feedback, disclosure and verification remain necessary, particularly for children and sensitive work. UNESCO (2023) supports a human-centered, privacy-conscious approach.

5. Discussion and Research Agenda

The proposed cycle reframes AI from an evaluator of learners to a conversational resource for evaluative judgment. This distinction is consistent with assessment theory: feedback is not a product delivered to a passive recipient but information interpreted and used in an instructional process. It also exposes the limitations of metrics such as number of comments, time to first response, or model–teacher score agreement. Such measures can assess operational features, but they cannot establish that learners have developed durable understanding.

Empirical evaluations should pre-register primary outcomes such as blind-rated transfer performance and quality of student revision rationales. Secondary measures might include teacher moderation time,

feedback uptake, students' capacity to identify flawed AI comments, and distributional differences by prior attainment or language proficiency. Studies should compare at least three conditions—teacher-only feedback, unmoderated AI feedback where ethically permitted, and teacher-moderated AI feedback—while maintaining identical tasks and rubrics. Qualitative interviews could examine when learners follow, question, or ignore generated advice.

5. Feedback Design and the Learning Cycle

AI feedback becomes educationally meaningful only when learners can interpret and act on it. A useful workflow begins with a criterion-referenced task and clear success descriptors. Students produce an initial response before consulting any automated suggestion. The teacher may then request feedback that identifies one strength, one priority for improvement, and one question that prompts reconsideration. Crucially, the learner must evaluate the suggestion, decide whether to adopt it, and justify any change. Feedback that simply rewrites the entire response can produce a polished artifact while obscuring the student's understanding. Revision histories and brief oral explanations provide a better indication of whether the learner has gained competence.

An effective feedback cycle also includes a second independent attempt. After reviewing advice on argument structure, a learner might compose a new paragraph on a different prompt without AI access. The teacher then checks transfer rather than merely improved formatting on the original assignment. Students who accept a fluent but unsupported suggestion should receive an opportunity to identify why it is misleading. Such activities turn fallibility into a critical-literacy resource, provided the teacher anticipates likely errors and prevents harmful or discriminatory content from being used uncritically.

6. Assessment Validity, Integrity, and Fairness

The widespread availability of generative AI complicates inferences from take-home submissions. A well-structured essay may reflect subject knowledge, prompt engineering, editorial judgment, or extensive machine drafting. Assessment validity therefore depends on aligning permitted assistance with the construct being measured. If the goal is independent language production, a supervised writing sample may be necessary. If the goal is evidence evaluation, documented critique of an AI-generated claim can itself be an appropriate task. Institutions should communicate these distinctions in advance rather than treating all AI use as equivalent misconduct. Transparency is both more educationally constructive and more defensible than unreliable attempts to infer authorship from textual style.

Fairness also requires recognizing unequal access to paid subscriptions, devices, stable connections, and advanced prompting experience. A course that quietly assumes universal access may reward resources rather than competence. Teachers can provide institutionally supported access where appropriate, offer equivalent non-AI pathways, and grade the quality of reasoning against published criteria. Students should not be required to submit private information to third-party tools as a condition of assessment. Rubric moderation and the periodic review of examples across language and ability groups can reveal unintended differences in the quality of feedback.

7. A Proposed Classroom Protocol

A six-step protocol can guide responsible use. First, define the learning target and explicitly state what the student must do independently. Second, identify a constrained role for AI, such as proposing questions rather than giving full solutions. Third, prepare verification criteria and an alternative resource. Fourth, require students to annotate which suggestions were accepted or rejected, and why. Fifth, assess a revised artifact together with an independent transfer task. Sixth, invite reflection on accuracy, confidence, and the effort required to verify advice. This protocol is a pedagogical proposal and should be locally tested before claims about effectiveness are made.

For example, in a lesson on argumentative writing, students may draft an evidence-based claim, ask an approved model for counterarguments, verify all proposed sources, and revise the argument using only credible evidence. The teacher assesses the original claim, the quality of source checking, and the justification for revision. This approach makes feedback visible as a process rather than a hidden replacement of student work. In mathematical reasoning, the parallel exercise could involve evaluating multiple solution strategies, identifying an erroneous step, and explaining a correct method without assistance.

8. Research Priorities and Institutional Safeguards

Future empirical studies should distinguish feedback speed, feedback accuracy, learner uptake, and subsequent achievement. These outcomes are related but not interchangeable. A randomized or quasi-experimental classroom study could compare teacher feedback, teacher-verified AI feedback, and unrestricted model feedback on tasks with common criteria. Investigators should predefine primary measures, examine differences across learner groups, and include delayed assessments. Qualitative analysis of revision decisions would reveal why learners follow or disregard suggestions. No such comparative trial is reported here; the proposed design is a research agenda arising from the conceptual analysis.

At the policy level, schools need guidance on privacy, record retention, acceptable disclosure, accessibility, and routes for contesting a misleading automated recommendation. Teachers should be trained to detect fabricated references, unstable scoring, and culturally narrow assumptions. Administrators should avoid equating automated grading efficiency with educational success. A responsible program can pilot feedback generation for formative use while retaining human accountability for consequential decisions and student support. Ongoing audits should inspect both practical benefits and the possibility that increased automation displaces productive dialogue between teachers and learners.

6. Conclusion

The instructional value of AI feedback depends on verified accuracy, meaningful learner interpretation, and independent action. The FIA cycle offers a transparent design for making those requirements visible. Its practical value lies in protecting assessment validity and student agency while enabling timely support; its effectiveness remains an empirical question for future research.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

Sadler, D. R. (1989). Formative assessment and instructional systems. Instructional Science, 18, 119–144. https://doi.org/10.1007/BF00117714

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410

Saranya, E., Sethuraman, P., Olimjonov, H., Nagendram, S., Ganesh, M., & Sapaev, I. (2026, August). Deep Reinforcement Learning for Cloud Resource Allocation and Energy-Efficient Data Center Operations. In *2026 International Conference on Next-Gen Intelligent Computing and Applications (ICONICA)* (Vol. 1, pp. 1-11). IEEE.

Kannan, N., Kollinal, R. K., Olimjonov, H., Ahammad, S. H., Kanagaraju, M., & Sapaev, I. (2026, August). Intelligent Ensemble Learning for Climate-Aware Agricultural Planning and Resource Allocation. In *2026 International Conference on Next-Gen Intelligent Computing and Applications (ICONICA)* (Vol. 1, pp. 1-10). IEEE.

Guru, K. V., Rajan, C., Olimjonov, H., Lakshmi, N. V., Joseph, A. V., & Nazarova, S. (2026, August). Capsule Network-Based Historical Document Digitization for Intelligent Knowledge Organization and Digital Archiving. In *2026 International Conference on Next-Gen Intelligent Computing and Applications (ICONICA)* (Vol. 1, pp. 1-9). IEEE.