Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 6 n. 3 (2026)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

"Augmented Assessment" and Generative AI: An Empirical Study on Secondary School Teachers. Towards a New "Formative Alliance"?

Maria Vittoria Isidori [1]*, Valentina Matilde Di Michele [1]

[1] University of L'Aquila; mariavittoria.isidori@univaq.it; valentinamatildedimichele@guest.univaq.it
* Correspondence: mariavittoria.isidori@univaq.it

Abstract: This study investigates the evolution of assessment and feedback practices amid the proliferation of Generative AI (GenAI) in education, focusing on strategic direction, critical vigilance and feedback. Using a quasi-experimental design with 170 secondary school teachers-in-training divided into 24 groups, participants used AI to analyze and generate feedback on stimulus texts of varying complexity. Data collected through structured questionnaires were analyzed using a mixed-methods approach. Results indicate that 61.2% of teachers demonstrated intermediate strategic direction, while 38.8% reached an expert level. Regarding critical vigilance, only 22.4% actively corrected AI superficiality, and 63.3% of upper secondary groups experienced an "argumentative depth illusion." Nevertheless, 85.7% achieved optimal human-AI integration in feedback design, with 77.6% preferring Socratic prompting over direct correction. An increased cognitive load was reported by 63.3% of participants. The findings validate the research hypotheses, demonstrating that GenAI shifts professional responsibilities toward higher-order evaluation. Safeguarding the epistemological and pedagogical integrity of formative assessment requires a new "formative alliance" between machine analytical power and human professional judgment.

Keywords: Augmented Assessment; Generative AI; Formative Alliance

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The proliferation of generative artificial intelligence (GenAI) tools such as ChatGPT within educational contexts has significantly reshaped assessment and feedback practices. These technologies redefine the boundaries of thought, writing and educational relationships, particularly in the domain of formative evaluation. Educators confront a critical question: how should assessment and feedback evolve to maintain pedagogical value while leveraging AI's analytical capabilities? This question is not merely technical but profoundly pedagogical, as it touches the very nature of evaluation as a relational and interpretive act (Bower & Torrington, 2024).

The literature highlights growing tension between automation efficiency and depth required for genuine human understanding in evaluative contexts (Baidoo-Anu & Owusu Ansah, 2023; Bower & Torrington, 2024). While AI benefits include personalized feedback and scalable assessment, significant concerns persist

regarding algorithmic bias, epistemological limitations and erosion of professional pedagogical judgment. Regulatory frameworks prohibit assessments based exclusively on automated processes, underscoring necessity for human oversight. This regulatory stance reflects a deeper epistemological principle: assessment is not merely measurement but interpretation, and interpretation requires a subject capable of understanding meaning (Kakungulu, 2024).

Within this contested space, Augmented Assessment emerges as a promising operational model (Kakungulu, 2024), conceptualizing AI not as delegation of judgment but as magnifying lens enhancing teacher expertise in feedback design. This approach integrates AI and learning analytics to observe learning in real-time while combining quantitative data with human sensitivity. The central premise is that algorithms, when appropriately directed through strategic prompting, extract granular information from texts that might escape superficial reading. However, this extraction does not reduce but transforms cognitive load, calling teachers to higher-order critical vigilance and epistemological validation (Isidori et al., 2025).

The theoretical framework of this study draws on three interconnected constructs. The first is strategic direction, understood as the teacher's capacity to formulate pedagogical constraints through prompting, transforming the AI from a generic responder into a situated pedagogical agent. The second is critical vigilance, the ability to identify algorithmic superficialities, exemplified by the argumentative depth illusion, where AI produces apparently sophisticated feedback that lacks genuine comprehension. The third is output processuality, the transformation of raw analysis into Socratic feedback that fosters student engagement and self-regulation.

These constructs are not isolated skills but dimensions of a broader professional posture that we define as the formative alliance: a conscious partnership between machine power and human judgment, in which AI amplifies teacher expertise while the teacher safeguards the meaning-making dimension that resists automation. This alliance is not a simple division of labour but a qualitative reconfiguration of assessment as a pedagogical practice.

The main aim of the work is to empirically verify whether Augmented Assessment can sustain this new formative alliance in secondary school contexts, examining how teachers direct AI, detect its limitations, and transform its outputs into pedagogically meaningful feedback. The study contributes to the growing field of AI-supported formative assessment by offering empirical evidence on the conditions under which AI enhances rather than erodes pedagogical judgment.

2. Materials and Methods

The study adopts a quasi-experimental, mixed-methods research design incorporating descriptive and correlational elements. While prioritizing controlled experimental tasks over unmonitored observational settings, the methodology balances structured intervention with authentic pedagogical tasks (e.g., evaluating authentic stimulus texts using GenAI platforms). This design ensures ecological validity within the context of professional teacher training while maintaining sufficient experimental control to observe structured human-AI interactions.

Sample. 170 secondary school teachers "in training" (70 lower secondary, 100 upper secondary), organized into 24 groups of approximately 7 individuals. The

sample was selected through convenience sampling within teacher training programmes, ensuring homogeneity in terms of professional status (all "in training") while allowing comparison across school levels. This homogeneity is methodologically significant: it ensures that differences in outcomes reflect task characteristics rather than pre-existing differences in teacher profiles.

Research Hypotheses.

H1 (Strategic Direction): Teachers with greater AI experience formulate richer, more pedagogically constrained prompts.

H2 (Critical Vigilance): AI tends to evaluate abstract thought superficially (argumentative depth illusion); teacher detection indicates epistemological competence essential for valid assessment.

H3 (Feedback Processuality): Effective AI integration improves feedback quality, rendering it more Socratic and fostering student self-regulation.

Procedure and Instruments. Each group used AI tools to analyze and produce prompt on two stimulus texts addressing the concept of individual freedom in the contemporary world. The selection of two texts with different complexity levels served as a methodological device to test the model's adaptability:

Text A: A colloquial text containing an error in the subjunctive mood and weak logical connections. Designed to activate didactic-linguistic vigilance, positioning AI as a formal tutor.

Text B: A seemingly erudite text featuring superficial deployment of the Foucauldian concept of "biopolitics" to trigger the argumentative depth illusion. Designed to activate epistemological vigilance, positioning AI as a conceptual interlocutor requiring an epistemological guarantor.

Texts were assigned by school grade for ecological validity. Both subgroups share homogeneous 'in-training' status; KPI shifts reflect text features, not teacher profiles, supporting task nature as the primary outcome moderator. Dual-text design allows verification of how task nature qualitatively conditions teacher-machine interaction, making human control not only necessary but qualitatively differentiated based on the cognitive challenge posed by the text.

Lower secondary groups (10) analyzed Text A using Brisk Teaching/CoGrader and ChatGPT. Upper secondary groups (14) analyzed Text B using Claude/ChatGPT-4.0 and NotebookLM (Sikri & Nuguri, 2025). Table 1 presents the tasks (Strategic Prompting, Critical Vigilance, Feedback Processuality) and Key Performance Indicators.

Table 1. Key Performance Indicators for the three dimensions of Augmented Assessment

Dimension	KPI	Scale
Strategic Direction	Pedagogical constraints in prompt	0–1 (Basic) 2–3 (Intermediate) 4+ (Expert)
Critical Vigilance	Detection of AI superficiality	0 (not detected) 1 (corrected)
Feedback Processuality	% human modification	<20% 20–50% >80%

Data collected via structured questionnaire (Google Form); mixed-method analysis employed. The questionnaire included both closed items (Likert scales, categorical choices) and open-ended questions, allowing quantification of KPIs and qualitative exploration of teacher reasoning. Mixed-method analysis followed an explanatory sequential design: quantitative results were first computed, then illustrated and deepened through qualitative excerpts from open-ended responses.

Statistical Analysis

Data were analyzed using descriptive statistics (χ^2 tests) and logistic regression to isolate the effects of text complexity from school levels. Analysis of Covariance (ANCOVA) was used to compare cognitive load and prompt constraints across groups while controlling for grade level. Baseline checks confirmed group comparability, while qualitative responses clarified quantitative trends.

3. Results

3.1. Hypothesis 1: Strategic Direction

In total, 61.2% of the sample operated at an intermediate level (2–3 constraints), while 38.8% reached an expert level (4+). Requests focused primarily on syntax/grammar (85.7%), logic/cohesion (79.6%), and citation relevance (57.1%). More experienced teachers consistently formulated richer prompts.

These results confirm H1 and suggest that strategic direction is not a generic digital skill but a pedagogical competence: teachers who understand assessment purposes are better able to translate them into prompt constraints. The predominance of syntax and grammar requests in lower secondary groups reflects the nature of Text A, whereas the attention to logic and cohesion in upper secondary groups reflects the epistemological demands of Text B.

3.2. Hypothesis 2: Critical Vigilance

Overall, 75.5% of participants considered AI reliable on target errors, while only 22.4% actively corrected it. The argumentative depth illusion emerged in 63.3% of upper secondary groups, confirming that teacher supervision remains irreplaceable for epistemologically valid assessment.

The low percentage of active correction (22.4%) is particularly significant: it indicates that critical vigilance is not spontaneously activated by AI usage but requires specific epistemological competence. The argumentative depth illusion is not a technical error, but a structural limitation of statistical language processing,

which can reproduce the surface features of erudition without grasping the conceptual relations that give meaning to academic discourse.

3.3. Hypothesis 3: Formative Feedback

Optimal human integration (20–50%) was achieved by 85.7% of the sample, with 77.6% choosing Socratic stimulus over direct correction (22.4%), confirming that AI improves feedback quality and fosters reflective engagement.

This preference for Socratic feedback represents a crucial finding for formative assessment. It suggests that teachers recognize AI's dialogic potential and choose to employ it not as an answer engine, but as a stimulus for student reflection. This orientation aligns directly with the core objectives of formative assessment: promoting self-awareness, reflexivity, and self-regulation.

3.4. Cognitive Load and Professionalism

Regarding cognitive demand, 63.3% of participants reported that AI increased their workload due to the continuous demands of critical vigilance (Table 2).

Table 2. Indicators of increase in cognitive load and professional reconfiguration.

Dimension	Value	Interpretation
Platform Interaction	85.7%	Balance automation-human sensitivity
Algorithmic Indulgence Detected	63.3%	Need for supervision
Socratic Feedback	77.6%	Orientation toward self-awareness

The increase in cognitive load is not a negative outcome but a sign of professional reconfiguration: AI does not reduce teacher work but shifts it toward higher-order evaluation. This finding challenges the common narrative of AI as a workload-reducing technology and reframes it as a technology that redistributes professional attention.

3.5. Differentiated Analysis by Textual Complexity

Text A activated didactic-linguistic vigilance; Text B activated epistemological vigilance. Higher cognitive load in upper secondary (70% vs 55%) confirms vigilance over meaning is more demanding than formal vigilance (Isidori et al., 2025) (Table 3).

Table 3. KPI comparison by textual complexity (Text A vs Text B).

Dimension	Text A	Text B	Synthesis
Strategic Direction	30%	45%	Complex texts require structured prompts
Critical Vigilance	15%	28%	Vigilance shifts

			from form to concept
AI Focus	Syntax 95%	Logic 85% Citations 70%	Formal corrector vs logical analyst
Illusion Detection	N/A	63%	Complexity activates critical thinking
Socratic Feedback	70%	84%	Argumentative texts favor maieutics
Cognitive Load	55%	70%	Complexity increases load

The differentiated analysis confirms the adaptability of the Augmented Assessment model: the same AI tools can be configured as formal tutors or conceptual interlocutors depending on the cognitive challenge posed by the text. This adaptability is not automatic but depends on the teacher's ability to recognize the nature of the task and to direct AI accordingly.

4. Discussion

The findings validate all three hypotheses with significant implications for formative assessment, feedback design and student engagement.

The Argumentative Depth Illusion. This phenomenon reveals AI's constitutive limitation—the gap between statistical language processing and experiential understanding of meaning. AI generates superficially sophisticated feedback but lacks genuine comprehension. This transforms the teacher's role from content evaluator to epistemological guarantor, safeguarding meaningful assessment against algorithmic shallowness. For formative assessment, this means teachers must develop heightened critical vigilance to ensure feedback remains pedagogically sound and fosters genuine learning rather than superficial compliance.

The argumentative depth illusion is not merely a technical flaw but an epistemological phenomenon. It reveals that AI operates on the surface of language, reproducing patterns without accessing the conceptual relations that constitute meaning. This has profound implications for assessment: if AI cannot distinguish between genuine understanding and sophisticated imitation, then the teacher's role becomes indispensable as a guarantor of epistemological validity.

Ontological Transformation in Assessment. The 63.3% cognitive load increase indicates fundamental shift in assessment practice. Far from reducing workload, AI reconfigures professional responsibilities toward higher-order evaluation. The teacher becomes a "critical vigilante," designing reflective assessment experiences where AI serves as tool for deeper inquiry rather than judgment replacement. This transformation has profound implications for assessment literacy and professional development in formative contexts (Crompton & Burke, 2023).

This ontological transformation challenges the instrumental view of AI as a neutral tool. AI is not simply added to existing practices; it reconfigures the very nature of assessment, shifting the teacher's attention from content evaluation to epistemological validation. This shift requires new competences and new forms of professional development.

Augmented Assessment as Magnifying Lens. AI amplifies teacher expertise in feedback design, reorienting attention toward finer scrutiny while demanding higher-order critical engagement. Teacher vigilance guards the only dimension resisting automation: meaning. In formative assessment, AI enhances—but never replaces—human capacity for understanding student learning and providing feedback that supports self-regulation and deep learning.

The metaphor of the magnifying lens is particularly apt: AI does not replace the teacher's eye but extends its reach, allowing the teacher to see details that would otherwise escape attention. However, the magnifying lens also reveals the limits of automation: meaning is not a detail that can be magnified but a dimension that requires human interpretation.

The Formative Alliance. The predominant choice of Socratic feedback (77.6%) over direct correction confirms teachers recognize AI's dialogic potential for fostering engagement and self-regulation. This maieutic orientation promotes student self-awareness, reflexivity, and deeper engagement—core formative assessment objectives. The new formative alliance must center on conscious partnership between machine power and human judgment, supporting peer and self-assessment while maintaining pedagogical integrity.

The formative alliance is not a simple collaboration but a qualitative reconfiguration of assessment as a relational practice. It requires teachers to develop new competences: strategic direction, critical vigilance, and output processuality. These competences are not generic digital skills but pedagogical competences that must be cultivated through initial and in-service training.

Adaptability Across Contexts. The Augmented Assessment model adapts flexibly to textual complexity. Text A configures AI as formal tutor supporting basic skill development; Text B as conceptual interlocutor requiring epistemological validation. This confirms robustness across diverse educational contexts and assessment tasks, supporting differentiated formative practices.

This adaptability is a crucial finding for the scalability of the model. It suggests that Augmented Assessment is not limited to specific disciplines or levels but can be adapted to different educational contexts, provided that teachers are equipped with the necessary epistemological competences.

Implications for Practice. Critical vigilance requires constant attention and institutional support. Professional development must equip teachers with epistemological competencies for effective AI integration in formative assessment. The formative alliance demands rethinking teacher education, assessment design, and institutional policies to support conscious partnership between human judgment and machine intelligence (Lu et al., 2023).

These implications are not limited to individual teachers but extend to institutional policies. Schools and universities must create conditions for the formative alliance to flourish: time for reflection, spaces for experimentation, and forms of recognition for the new professional competences required by AI integration.

5. Conclusions

The study confirms that GenAI does not replace but reconfigures teacher judgment, calling for higher-order critical vigilance and epistemological validation. Augmented Assessment emerges as an operational model capable of sustaining a new formative alliance between machine power and human judgment. This alliance must be built on conscious partnership, institutional support, and continuous professional development, so that assessment remains pedagogically sound and fosters genuine student self-regulation and deep learning.

The findings suggest that the future of formative assessment lies not in choosing between human and artificial intelligence but in cultivating their alliance. This alliance requires a new professional posture, new competences, and new institutional conditions. It is a pedagogical challenge, not merely a technological one.

References

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bower, M., & Torrington, J. (2024). How should we change teaching and assessment in response to increasingly powerful generative artificial intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*, *29*(12), 15403–15439. https://doi.org/10.1007/s10639-023-12405-0

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), Article 22. https://doi.org/10.1186/s41239-023-00392-8

Isidori, M. V., Muccini, H., Evangelista, C., & Visceglia, D. (2025). When Artificial Intelligence Meets Metacognition: The Experimental Learning Unit Transforming Teaching and Evaluative Practices. *Journal of Inclusive Methodology and Technology in Learning and Teaching*, *5*(3), 1–18. https://www.inclusiveteaching.it/index.php/inclusiveteaching/article/view/406

Kakungulu, S. J. (2024). The Role of Artificial Intelligence in Educational Assessment. *Eurasian Experiment Journal of Scientific and Applied Research*, *5*(2), 44–53.

Lu, H., Wang, Y., Wu, X.-Y., Lu, X.-F., & Liu, X.-B. (2023). Developing and validating the assessment of the skills of preservice teachers in implementing pedagogical model of collaborative problem solving. *Educational Technology Research and Development*, *71*(4), 1799–1819. https://doi.org/10.1007/s11423-023-10250-z

Sikri, D., & Nuguri, V. (2025). From notes to networks – A three-dimensional framework with integrated curation for using Google NotebookLM in health professions education. *Medical Teacher*, *48*(4), 557–561. https://doi.org/10.1080/0142159X.2025.2533408