Beyond Generic Feedback: Personalized Generative-AI to Support Calibration and Performance

Hassna Akhasbi [1,*], Najib El Kamoun [1], Fatima Lakrami [1], Jean-Luc Gilles [2], Jean Michel Rigo [3], and Emilio Aliss Paredes [4]

[1] Department of Physics and Engineering, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco
[2] Training and Research Unit in Teaching, Learning and Assessment, University of Teacher Education, Haute École Pédagogique du Canton de Vaud, Lausanne, Switzerland
[3] Department Capacity Group Neurosciences, Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium
[4] Department Coordinator of the Center for Academic Development and External Postgraduate Studies (CEDAPE), Faculty of Science and Technology, Universidad Simón I, Patiño, Bolivia
Email: hassna.akhasbi@gmail.com (H.A.); elkamoun.n@ucd.ac.ma (N.E.K.); lakrami.f@ucd.ac.ma (F.L.); jean-luc.gilles@hepl.ch (J.-L.G.); jeanmichel.rigo@uhasselt.be (J.M.R.); emilioaliss@gmail.com (E.A.P.)
*Corresponding author

Manuscript received December 31, 2025; revised February 19, 2026; accepted April 7, 2026; published August 21, 2026

***Abstract*—Generative Artificial Intelligence (GenAI) opens new possibilities for formative assessment, yet most existing systems still provide generic, task-focused comments that say little about how students manage uncertainty. This exploratory pilot study examined the feasibility of an indicator-informed GenAI workflow that combines Degree of Certainty (DC)-based assessment with course-grounded feedback in a Master's-level Artificial Intelligence course. Over one semester, 23 students completed five formative DC-enhanced multiple-choice assessments and one summative assessment. A Python-based engine computed cognitive performance (Percentage of Correct Responses (TRC)) and four calibration indicators: Realism/Calibration (REAL), Overestimation (SUR_EST), Underestimation (SOU_EST), and Net Bias (CENTR). These indicators, together with item-level responses and instructor-approved course documents, were fed into a GPT-4 module through a structured prompt and a canonical JavaScript Object Notation (JSON) template specifying six feedback cases (correct/incorrect × low/medium/high certainty). For each assessment, the system generated item-level explanations and a metacognitive synthesis report, which the instructor reviewed before release. Students also completed a perception questionnaire covering usefulness, clarity, trust, intention to use, cognitive load, and self-efficacy. Results showed higher TRC and REAL, lower SUR_EST and CENTR, and stable SOU_EST within an a priori equivalence margin. Perceptions were neutral to slightly positive, with reliable subscales and associations suggesting that perceived clarity and intention to use were linked to more favorable trajectories in performance and calibration. Given the small sample, single-course context, and absence of a control group, these findings remain hypothesis-generating. They nonetheless support the feasibility of indicator-informed, course-grounded GenAI feedback for fostering more realistic confidence judgments and self-regulated learning in higher education.**

***Keywords*—Generative Artificial Intelligence (GenAI), Large Language Models (LLMs), Degree of Certainty (DC), metacognitive calibration, formative feedback, higher education**

I. Introduction

In higher education, formative feedback is widely recognized as a key driver of learning, particularly when it helps students understand where they stand, why, and what they can do next [1–3]. Beyond correcting errors, feedback can support core processes of Self-Regulated Learning (SRL), particularly when learners are able to interpret and use it productively [4–6]. In engineering education, where students often face complex problems under uncertainty, feedback that addresses not only correctness but also confidence, risk-taking, and decision-making is particularly important for developing calibrated professional judgment [7].

Yet digital feedback often remains generic and task-focused. Many online assessment tools still provide only "correct/incorrect" messages or brief hints, with limited attention to metacognitive dimensions such as confidence monitoring, overestimation or underestimation, and feedback literacy [8, 9]. From an SRL and calibration perspective, this creates a gap between what feedback could support and what is typically available in large engineering courses.

Recent advances in Generative Artificial Intelligence (GenAI) have renewed interest in more personalized feedback. Large Language Models (LLMs) such as GPT can act as AI-supported co-tutors, generating explanations, examples, and study strategies tailored to individual learner inputs [10, 11]. Early empirical work in Science, Technology, Engineering, and Mathematics (STEM) and engineering reports benefits such as on-demand explanations, flexible question answering, and increased engagement, but it also highlights well-documented risks including uneven answer quality, hallucinations, over-reliance, and academic integrity concerns [12–15]. To address these limitations, recent studies recommend combining GenAI with instructor-curated materials, document grounding, carefully engineered prompts, constrained output formats, and teacher oversight [16–18]. LLM-based tools have, for example, been used to generate rubric-based feedback on open-ended responses or to act as conversational tutors with quality-control filters [19, 20]. However, most of these systems remain focused on content-level feedback; relatively few explicitly target metacognitive regulation, feedback literacy, or the integration of confidence into the feedback logic.

In parallel, Degree of Certainty (DC) approaches extend Multiple-Choice (MCQ) testing by asking students to report how sure they are about each answer. Building on Leclercq's work [21], DC-based assessment makes it possible to distinguish assured knowledge, partial knowledge, recognized ignorance, and confident errors, while deriving indicators of cognitive performance and metacognitive calibration [22–24]. When used formatively, DC can support

more realistic self-evaluation and safer decision-making, especially when feedback explicitly links answer accuracy and reported certainty [21, 24]. Within SRL and calibration frameworks, DC therefore provides a concrete means of capturing monitoring judgments and identifying miscalibration.

The present study brings these strands together in an authentic Master's-level Artificial Intelligence course in Networks, Telecommunications, and Cybersecurity. Students completed five formative DC-enhanced multiple-choice assessments and one summative assessment. Their responses were used to compute DC-based performance and calibration indicators, which then informed a GPT-4-based module generating structured, course-grounded feedback reports. For each student and assessment, the reports explain item-level responses in light of both the selected answer and reported DC, and provide a global metacognitive synthesis with an action plan for subsequent study. The feedback design is explicitly grounded in SRL and calibration: the indicators are treated as observable traces of monitoring and judgment, and the GenAI-generated messages are intended to foster awareness of confidence patterns, encourage reflective use of feedback, and support strategic adjustment across assessments.

In this study, performance is summarized by the Percentage of Correct Responses (TRC), while calibration is captured through four DC-based indices: Realism/Calibration (REAL), Overestimation (SUR_EST), Underestimation (SOU_EST), and Net Bias (CENTR). Throughout the paper, confidence refers to the transformed 0–100 m derived from DC. These constructs are detailed in the following section.

Within this framework, we address three research questions:

(1) Do TRC and calibration indicators (REAL, SUR_EST, SOU_EST, CENTR) change across five formative assessments and one summative assessment in this context?
(2) How do students perceive the quality, clarity, and usefulness of DC-structured, GenAI-generated feedback grounded in course materials?
(3) To what extent do objective changes in indicators converge with students' subjective perceptions of the feedback?

Given the small sample size, single-course setting, and absence of a parallel control group, this work should be understood as an exploratory pilot study. Accordingly, it focuses on observed patterns of change and student perceptions rather than on causal claims of effectiveness. More specifically, it examines whether the observed trajectories are consistent with the hypothesis that DC-structured, GenAI-generated feedback may support performance and metacognitive calibration in this context. We expect increases in TRC and REAL, decreases in SUR_EST, and stable or reduced SOU_EST over time, together with positive perceptions of feedback usefulness and clarity.

The remainder of the paper is organized as follows Section II presents the theoretical and empirical background. Section III details the methodological design and GenAI feedback workflow. Section IV reports the quantitative and perceptual results. Section V discusses the implications, limitations, and directions for future research.

II. Literature Review

A. Formative Feedback, Calibration and Self-regulated Learning

Formative feedback is widely recognized as a key driver of learning, particularly when it provides specific information that helps students close the gap between current and desired performance [1, 2]. Contemporary models emphasize not only the corrective function of feedback but also its role in supporting SRL: effective feedback prompts learners to monitor their understanding, reflect on strategies, and adjust their efforts [4, 5]. Within this perspective, feedback operates at multiple levels—task, process, and self-regulation—and its impact depends on how learners interpret and act upon it rather than simply on the provision of comments [6].

Recent work on feedback literacy further highlights that students must learn how to make sense of feedback, manage the emotional demands it may generate, and translate comments into concrete action [9]. In higher education, and especially in engineering programs, students often face complex, ill-structured problems and high-stakes assessments. In such contexts, feedback that explicitly addresses uncertainty, confidence, and decision-making is particularly valuable, as it can help students calibrate their judgment and develop professional habits of mind consistent with SRL models.

B. Certainty-based Marking and Degree of Certainty as a Metacognitive Tool

Certainty-Based Marking (CBM) originates from probabilistic approaches to assessment and decision theory [25, 26]. In this approach, students are asked not only to select an answer but also to state a DC for each response. Pioneering work by Leclercq and colleagues showed that combining response accuracy with DC makes it possible to distinguish assured knowledge, partial knowledge, recognized ignorance, and potentially "dangerous" confident errors [24]. Such information supports a diagnosis richer than correctness alone and can feed a variety of metacognitive indicators, including REAL (alignment between confidence and performance), SUR_EST, SOU_EST, and CENTR.

Several studies have documented the metacognitive benefits of DC-based evaluation. Noël and Leclercq conceptualize metacognition as a cycle of judgment, analysis, and regulation, and argue that attaching a certainty judgment to each answer nudges students to engage in deeper self-evaluation before, during, and after assessments [27]. Longitudinal work shows that repeated use of DC is associated with improvements in REAL and reductions in overconfidence, especially when combined with explicit feedback on confidence-performance discrepancies [27].

Empirical evidence across disciplines supports this view. In medical education, DC-enriched multiple-choice assessments help distinguish usable knowledge from potentially dangerous overconfident misconceptions and reveal changes in the distribution of confident-correct, confident-incorrect, and uncertain responses after training [28]. Other work highlights the feasibility and acceptability of DC in authentic settings: students and

instructors generally understand and value the method, while also noting the need for careful explanation of the scale and high-quality item design [29].

More recent syntheses emphasize that DC-based indicators can be used not only to refine summative grading, but also as formative tools for SRL and teaching. Analyses of DC spectra and associated indicators (REAL, SUR_EST, SOU_EST, CENTR) allow instructors to identify profiles such as imprudent, prudent, and realistic learners, and to tailor remediation accordingly [30]. They also provide a basis for dashboards that visualize patterns of confidence and realism over time for both individuals and groups. These findings support the use of DC as a bridge between assessment and metacognitive development, which is central to the present study.

C. Automated Feedback and Online Formative Testing

Parallel to developments in DC, a growing body of work has examined automated feedback and online formative assessment systems. Reviews of computer-based testing and online quizzes in higher education report that frequent low-stakes assessments with immediate feedback can support retention, engagement, and self-regulation, although effects are often modest and conditioned by students' participation patterns [31]. Studies using learning management systems (e.g., Moodle) for repeated practice show positive associations between participation in online quizzes and performance on final exams, but also reveal superficial use and decreasing engagement over time for a substantial proportion of students [32].

Within this line of work, some systems explicitly integrate DC ratings or derived confidence measures into online quizzes. For instance, DC-enhanced assessments in veterinary and medical education have been used to provide immediate feedback on both accuracy and certainty, to prompt reflection on overestimation or underestimation, and to visualize trajectories of calibration across multiple assessments [33, 34]. Other platforms such as DOCIMO combine DC-based items with psychometric indicators (e.g. item-total correlations) to support iterative improvement of question quality and curricular alignment [35]. Although these systems move towards richer, data-informed feedback, they typically rely on rule-based messages (e.g., pre-written hints per item) or simple correctness-based comments.

A recent systematic review of automatic feedback in online learning environments emphasizes similar patterns: automated systems can deliver timely, scalable feedback, but comments tend to be task-focused, generic, and only loosely connected to metacognitive processes or feedback literacy [8]. This motivates the exploration of more expressive feedback generators, such as large language models, while maintaining attention to reliability and pedagogical alignment.

D. GenAI and LLM-based Feedback in Higher Education

The emergence of Large Language Models (LLMs) has prompted a new wave of tools that generate feedback, hints, or explanations on demand. Early commentaries highlight the potential of LLMs to act as conversational co-tutors that offer personalized explanations, examples, and study strategies, while also warning about hallucinations, bias, and over-reliance [17]. Empirical studies in blended and online courses report that access to ChatGPT-like tools can support understanding and motivation, but also raise concerns about academic integrity and the risk of bypassing productive struggle [18]. In engineering education specifically, case studies show that students perceive LLMs as useful assistants for writing and problem solving, yet often struggle to judge the reliability of the generated content [18].

Several recent systems use LLMs specifically to generate feedback on student work. Rubric-based tools leverage LLMs to produce criterion-referenced comments on open-ended responses, essays, or programming tasks, usually anchored in teacher-provided rubrics or example solutions and accompanied by human review [19]. Other work proposes "LLM tutors" whose feedback is automatically evaluated and filtered before being shown to students, in order to control hallucinations and ensure a minimum level of quality [20]. Further studies emphasize the critical role of prompt design and human supervision in maintaining feedback quality: richer prompts, clearer roles, and explicit constraints tend to yield feedback that aligns better with expert judgments and pedagogical goals [36, 37].

These approaches demonstrate that GenAI can substantially reduce the workload involved in producing detailed feedback and can approximate human-like explanations in many cases. However, across these studies, feedback remains primarily at the content level: LLMs comment on correctness, reasoning quality, or alignment with rubrics, but rarely incorporate explicit measures of confidence or calibration into the feedback logic. For example, recent LLM-based feedback systems largely emphasize rubric-/or criteria-aligned comments or quality-filtered tutoring responses, rather than modeling confidence-accuracy calibration as a quantified design target [19, 20, 38]. Metacognitive aims—helping students monitor and adjust their certainty, recognize dangerous overconfidence, or understand patterns of underestimation—are typically mentioned as desirable outcomes but are not embedded as first-class design targets. Likewise, relatively few systems combine LLM-generated feedback with longitudinal visualizations of indicators, thereby enabling both students and teachers to track how performance and calibration evolve over time in authentic STEM and engineering courses.

E. Positioning of the Present Study

The present work sits at the intersection of these strands. Building on DC and CBM research, we adopt TRC and four calibration indicators (REAL, SUR_EST, SOU_EST, CENTR) as key variables for characterizing students' metacognitive profiles in a Master's-level Artificial Intelligence course. These indicators are computed by our analysis pipeline and made available within the Evalviz-Dashboard web application, which can display their evolution across multiple assessments for both students and instructors. In the present study, however, Evalviz primarily serves as the delivery environment for the feedback reports generated by our GenAI module, rather than as an object of evaluation in its own right.

At the same time, we leverage GenAI—specifically GPT-4—to generate personalized, indicator-informed feedback. Prompts specify a pedagogical role for the LLM, provide grounded excerpts from course materials, include

DC-based indicators, and constrain the output to a canonical JSON template with six feedback cases (correct/incorrect × low/medium/high certainty) and a dedicated metacognitive synthesis section. The resulting reports aim not only to explain why answers are correct or incorrect, but also to translate confidence patterns and calibration indicators into student-friendly messages and concrete action plans.

To our knowledge, few existing GenAI systems combine:

- DC-driven calibration indicators as first-class inputs;
- a structured feedback template explicitly tailored to metacognitive profiles rather than solely to content;
- integration into an authentic engineering course with longitudinal DC-enhanced testing.

By articulating these components within a self-regulated learning and calibration framework, this study contributes both a design pattern for "feedback for calibration" and exploratory empirical evidence on how such feedback relates to students' performance, REAL, and perceptions.

III. Materials and Methods

A. Research Context and Participants

This study was conducted in a second-semester course, "Introduction to Artificial Intelligence", within the Master 1 Networks-Telecommunications-Cybersecurity program at the Faculty of Sciences, Chouaib Doukkali University (El Jadida, Morocco). The course combined weekly lectures and hands-on lab sessions. Twenty-three students (12 female, 11 male; aged 22–24) completed all assessments and questionnaires, with no attrition.

Across the semester, students completed low-stakes multiple-choice assessments using Degree of Certainty (DC). Prior to the first assessment, the instructor introduced the DC scale, explained the 0–5 confidence scale with examples, and emphasized that DC was used for formative feedback and research purposes only (i.e., it did not contribute to final grades).

B. Study Design and Timeline

We employed a within-subject repeated-measures design with a single cohort and no parallel control group. Consequently, the analyses document patterns of change over time but do not support causal attribution to any single component of the instructional or feedback ecosystem.

As shown in Fig. 1, students completed five formative DC-based MCQ assessments (T1–T5) aligned with the course chapters (Introduction to AI; Neural Networks; Convolutional Neural Networks; Recurrent Networks and LSTMs; Autoencoders and Variational Autoencoders). A final summative assessment (Ts) assessed students' consolidated understanding using new items covering the same concepts.

This design yielded six repeated measurements of cognitive performance and calibration indicators, complemented by students' perceptions of the feedback after each release.

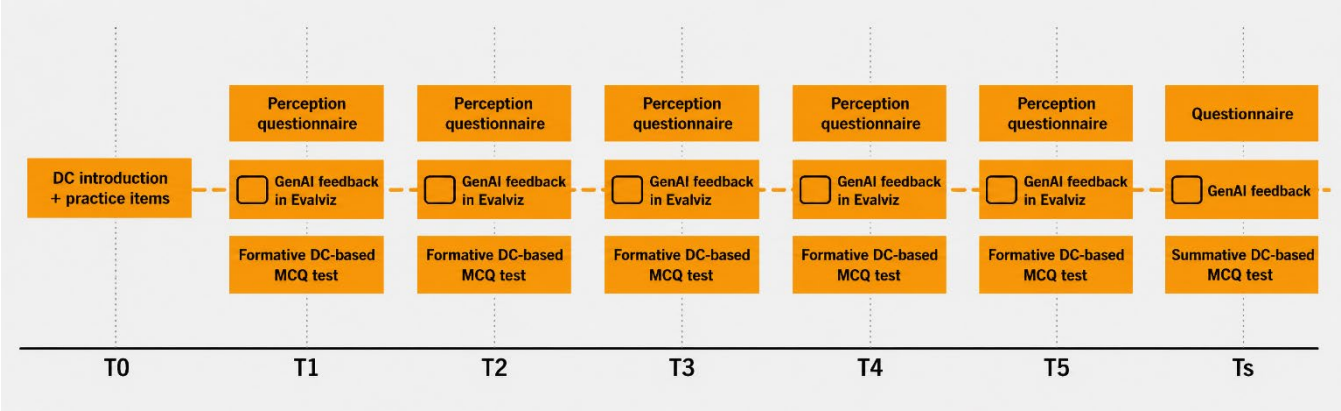

Fig. 1. Study timeline across the semester.

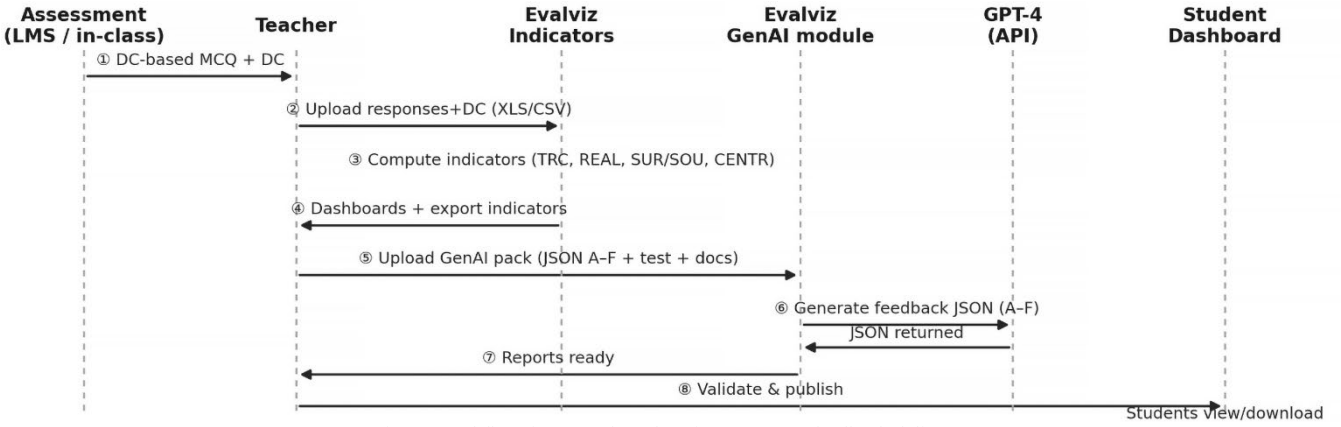

Fig. 2. Workflow from DC-based testing to GenAI feedback delivery.

C. Pedagogical and Analytic Framing

The methodological choices follow a formative assessment and self-regulated learning perspective. DC is treated as an in-task metacognitive judgment of certainty, which can be aggregated into calibration patterns such as REAL,

SUR_EST, and SOU_EST. The feedback is therefore designed to operate at two complementary levels: it provides item-level explanations targeting conceptual understanding, and a global, indicator-based synthesis aimed at fostering metacognitive awareness and guiding subsequent study regulation through concrete recommendations.

This operationalization mirrors the self-regulated learning and calibration framework described in Section II, where DC is regarded as an observable trace of metacognitive monitoring and the derived indicators as explicit targets for regulation-oriented feedback.

D. System and End-to-end Workflow (Evalviz-Dashboard)

The analysis and feedback pipeline was implemented in Evalviz-Dashboard (www.evalviz-dashboard.com), a secure web application used after assessments are administered in the usual course environment (LMS or in-class testing tools).

Fig. 2 provides an overview of the end-to-end pipeline implemented in Evalviz-Dashboard, from DC-based assessment data ingestion to indicator computation, GenAI-assisted feedback generation, instructor validation, and release to students. The following subsections describe each module in detail (Sections E–G).

E. Data Import and DC-based Indicator Computation

After each DC-based test, the instructor exported an Excel response file containing student identifiers, the selected option for each item, the associated Degree of Certainty (DC; 0–5), and the answer key. The file was uploaded to the Results import module of Evalviz-Dashboard. Evalviz parsed the file, matched responses with the answer key and DC ratings, and stored the resulting dataset in an SQLite database on the secure server linked to the assessment identifier and student ID.

A Python-based computation engine then derived, for each student and assessment, a set of performance and calibration indicators computed on a common 0–100 scale (Table 1). In the remainder of the paper, mean confidence is denoted C_{moy}. Calibration indicators follow the certainty-level formulation $k \in \{0, \ldots, 5\}$ and rely on target values t_k derived from the DC probability bands. Full formal definitions, notation, and interpretation are provided in Appendix A.

Indicators were accessible through interactive dashboards (individual trajectories, cohort distributions, item statistics). For this study, all indicators were also exported as a consolidated Excel file (students × indicators), which was used as an input for the feedback generation stage.

Table 1. Formulas and interpretation of TRC and calibration indicators

Indicator	Formula	Interpretation
TRC	$TRC = 100\frac{N^+}{N}$	Higher = higher cognitive performance
C_{moy}	$C_{moy} = \frac{1}{N}\sum_{i=1}^{N} c_i$	Higher = higher reported confidence
CENTR	CENTR = C_{moy}–TRC	>0 = net overestimation, <0 = net underestimation
SUR_EST	$SUR_EST = \frac{(100 - REAL) + CENTR}{2}$	High = harmful overconfidence
SOU_EST	$SOUS_EST = \frac{(100 - REAL) - CENTR}{2}$	High = excessive prudence
REAL	$REAL = \left(\frac{1}{n}\right)\sum_{k=0}^{5} N_k Real_k$	High = good alignment.

Notes: C_{moy} = mean confidence (0–100).

F. GenAI-Based Feedback Generation

The feedback module relies on GPT-4 accessed via the OpenAI API, and is available through a dedicated instructor interface in Evalviz-Dashboard (Fig. 3).

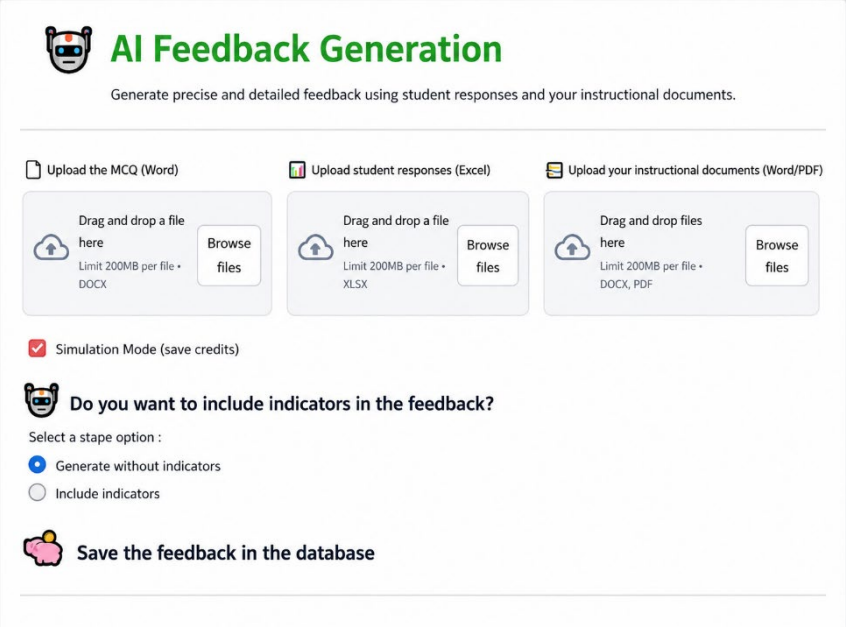

Fig. 3. Instructor interface for AI-based feedback generation in Evalviz-Dashboard.

For each assessment, the instructor provided three elements:

- the MCQ file including the full wording of the questions and the answer key (Word),
- the student response file including identifiers, selected options, and DC values (0–5) for each item (Excel),
- and a set of instructor-approved course documents (Word/PDF) used to ground explanations.

The interface supports feedback generation with or without calibration indicators; in this study, the indicator-enriched mode was used throughout. In this mode, the instructor additionally uploaded the indicator file computed in Section III-E and selected which indicators to include in the prompts (TRC, REAL, SUR_EST, SOU_EST, CENTR). For each student and item, the backend assembled a structured prompt containing the question stem and options, the correct answer and the student's selected option, the item-level DC value, the student's indicator profile for the assessment, and selected excerpts from the course materials.

GPT-4 was instructed to use only the provided information, refrain from introducing external concepts or references, and instantiate a single JSON output matching a canonical template. Item-level feedback followed one of six predefined pedagogical cases (A–F), determined by correctness × DC level. For all analyses and for case assignment, DC was grouped as low (0–1), medium (2–3), and high (4–5). The canonical template encodes a fixed pedagogical structure for each case, including an introductory message, a conceptual explanation, an explicit link to course content, metacognitive reflection on DC, strategy recommendations, encouragement, and key points to remember (Table 2).

A post-processing script converted each JSON output into a human-readable feedback report comprising per-item feedback and a final synthesis interpreting the indicator pattern and proposing a concrete action plan for subsequent study. An example excerpt of a generated report is provided in Appendix C, and full replication materials (verbatim prompt, canonical JSON template, model configuration and inference parameters, and the grounding/document injection procedure) are provided in Supplementary Appendix B.

Table 2. Overview of feedback cases A–F

Case	Correctness	DC level	Main pedagogical goal	Feedback components
A	Correct	High	Consolidation, transfer	Confirmation, advanced explanation, key points, challenge
B	Correct	Medium	Strengthen partial knowledge, build confidence	Confirmation, clarification, mnemonics, recommendations
C	Correct	Low	Reassure, stabilize intuition	Explanation, contrast with wrong options, encouragement
D	Incorrect	High	Repair harmful knowledge, address overconfidence	Alert, corrective explanation, key rules, strategic advice
E	Incorrect	Medium	Complete partial understanding	Explain missing aspect, examples, mnemonics
F	Incorrect	Low	Build knowledge from recognized ignorance	Gentle explanation, simple examples, encouragement

G. Instructor Validation and Student Access

All generated feedback reports were first made available in an instructor-facing view within Evalviz-Dashboard. The instructor reviewed each report, checked conceptual accuracy and alignment with the canonical template, and—when necessary—edited or regenerated specific feedback messages.

Only after validation did the instructor publish the reports to the feedback database. Reports then became available in each student's space, where students could consult them online and download them as Word documents for later review. Immediately after consulting their feedback, students completed the perception questionnaire described below. Instructor review procedures and audit-based quantitative reporting of interventions/edits are described in Supplementary Appendix B.

H. Instruments and Measures

1) DC scale and transformation to a 0–100 confidence metric

For quantitative analyses, DC ratings were converted to a 0–100 confidence metric by mapping each DC category to the midpoint of its associated probability band (Table 3), following the certainty-based scoring conventions described by Leclercq and Gilles. This transformation allows DC to be interpreted as a probabilistic judgment and ensures that all indicators are expressed on a common 0–100 scale.

Table 3. DC values and associated probability bands

DC	Probability band	Midpoint used (%)
0	0–25%	12.5
1	25–50%	37.5
2	50–70%	60
3	70–85%	77.5
4	85–95%	90
5	95–100%	97.5

Based on item accuracy and this transformed confidence, we computed TRC, REAL, SUR_EST, SOU_EST, and CENTR using the formal definitions reported in Table 1 (see Appendix A for full derivations and definitions). This transformation is grounded in the certainty-based assessment tradition, in which DC is defined as an expression of the respondent's subjective probability that an answer is correct rather than a purely ordinal or verbal judgment.

Accordingly, reference work recommends probabilistic/numerical elicitation to reduce the ambiguity of verbal labels and to enable quantitative calibration indicators such as REAL and CENTR [22, 23, 39]. Because DC categories correspond to probability bands, mapping each band to its midpoint t_k (see Table 3) is a standard operational choice for interval-based certainty scales [40, 41].

Indicator magnitudes can therefore be interpreted in percentage-point units: higher REAL indicates closer agreement between confidence and observed accuracy across certainty levels, whereas CENTR is signed (>0 indicates net overestimation; <0 indicates net underestimation) and should be interpreted cautiously because values near 0 may result from cancellation of opposite biases.

2) Student perception questionnaire

After each feedback release, students completed a short Likert-type questionnaire (1 = strongly disagree, 5 = strongly agree) covering perceived Usefulness (U), Clarity/readability (C), trust in the system/feedback (T), intention to reuse the feedback and perceived impact on study strategies (I), perceived cognitive Load (L), and two optional self-efficacy items (AE). Each construct was measured with 4 items, except self-efficacy (AE), which was measured with 2 items (see Appendix D for full wording). Representative items and subscale codes are reported in Table 4.

Table 4. Questionnaire dimensions (codes, labels, sample items)

Code	Label	Sample items
U	Usefulness	"The feedback helped me understand my mistakes."
C	Clarity	"The feedback was clear and easy to follow."
T	Trust	"I trust the feedback to be accurate and reliable."
I	Intention/Usage	"I will use this feedback to improve my next answers."
L	Cognitive Load	"The feedback was overwhelming or confusing."
AE	Self-efficacy (optional)	"The feedback increased my confidence to avoid similar mistakes."

3) Data analysis

Analyses were pre-specified to align with three research questions. All indicators were analyzed on a common 0–100 scale, enabling direct interpretation as percentage-point differences, with longitudinal change expressed in points per assessment.

For RQ1, we computed descriptive statistics at each assessment and examined pre-post change between T1 and Ts using paired t-tests (or Wilcoxon signed-rank tests when assumptions were violated). We report mean differences with 95% confidence intervals and standardized paired-sample effect sizes (d_z). To leverage the six repeated measurements (T1–T5, Ts), longitudinal change was quantified using student-specific linear trends ("slopes"). For each outcome Y and each student i, we fitted an ordinary least squares regression across time points $t = 1, \ldots, 6$:

$$Y_{it} = \alpha_i + \beta_i t + \varepsilon_{it}$$

where β_i represents the student's average change per assessment (points per assessment). We then tested whether the mean slope differed from zero using a one-sample t-test (with Wilcoxon signed-rank tests as a robustness check) and report mean slopes with 95% confidence intervals.

Because avoiding harmful increases in underestimation was a key pedagogical concern, we complemented these analyses with an equivalence test (TOST) on SOU_EST. The equivalence margin was defined a priori as ±2 points on the 0–100 scale, interpreted as a small and pedagogically negligible change in percentage-point units; sensitivity checks were conducted at ±1 and ±3 points. TOST results are reported using 90% confidence intervals relative to the equivalence bounds.

For RQ2, perception subscales (usefulness, clarity, trust, intention to use, cognitive load, self-efficacy) were summarized descriptively using scores aggregated across the five formative assessments; internal consistency was evaluated using Cronbach's alpha for each subscale (computed per assessment).

For RQ3, we examined convergence between perceptions and objective change using Spearman correlations between students' mean perception scores and (i) pre–post changes in key outcomes (ΔTRC, ΔREAL) and (ii) the longitudinal slope of SUR_EST. We report Spearman's ρ with two-tailed p-values.

Across analyses, we checked assumptions and robustness (normality checks, influence diagnostics, sensitivity analyses). When multiple statistical tests were performed within an indicator family, p-values were adjusted using Holm's procedure. For correlation matrices, significance markers reflect unadjusted two-tailed p-values. All analyses were conducted in R with $\alpha = 0.05$.

I. Ethical Considerations and Data Management

The study complied with institutional guidelines for research in higher education. Participation was voluntary, and students were informed that DC-based assessments and GenAI-generated feedback were purely formative and would not affect final grades; written informed consent was obtained from all participants.

Evalviz-Dashboard is hosted on a secure, password-protected server. Assessment response files, computed indicators, and feedback reports are stored in an SQLite database with access restricted to the course instructor and research team. For analysis, datasets were exported in pseudonymized form: direct identifiers (e.g., names, emails, student numbers) were removed and replaced with study-specific codes, and results were processed and reported at the group level only. Students accessed feedback through their individual accounts, while all reporting in the manuscript uses aggregated statistics. Data are retained solely for research and quality-improvement purposes and are not shared outside the research team.

IV. Results and Discussion

A. Scope of Inference

The present study should be interpreted as an exploratory, single-cohort, within-subject investigation conducted in an authentic course setting. Because the design does not include a parallel control group and because several factors may jointly influence students' trajectories across the semester (e.g., course progression, repeated exposure to the assessment format, practice effects, and feedback exposure), the analyses support primarily descriptive and associative inferences. More specifically, the results allow us to document the feasibility of the proposed workflow and to examine whether the observed patterns are consistent with the hypothesis that DC-structured, GenAI-generated feedback may support performance and metacognitive calibration. However, they do not allow us to establish causal impact, isolate the specific contribution of the GenAI feedback component, or generalize the findings beyond this specific context without appropriate caution.

B. Dataset Description and Analytic Framing

We report two sets of results:

(1) the longitudinal evolution of performance and calibration indicators across six DC-enhanced assessments (RQ1), and
(2) students' perceptions of the indicator-informed GenAI feedback and their associations with changes in performance and calibration indicators (RQ2–RQ3).

Pre-post contrasts compare the first formative assessment (T1) with the summative assessment (Ts). Longitudinal effects are expressed as slopes in points per test on a 0–100 scale (see Table 3 for the 0–5 DC to 0–100 transformation). Slopes correspond to student-specific linear trends estimated across the six time points and summarized at the cohort level (see Section III.H.3 for the model specification).

To support the interpretation of longitudinal trends, Table 5 summarizes key characteristics of each assessment, including the number of items, TRC score distributions, and DC distributions using the study's low/medium/high grouping (low = 0–1, medium = 2–3, high = 4–5). The table highlights differences in assessment length (20 items for T1–T5; 30 items for Ts) and provides context for evaluating comparability across time points.

However, because each assessment targeted a different course chapter, time is partly confounded with content coverage and likely item difficulty. As a result, longitudinal changes in TRC and calibration indicators across T1–Ts may reflect topic- or difficulty-related differences between assessments, in addition to repeated exposure to DC-enhanced testing and feedback reports. We therefore interpret the trajectories as descriptive patterns over successive assessments rather than as evidence of feedback effects.

Table 5. Key characteristics of each assessment (T1–T5, Ts)

Assessment	Topic / chapter	# items	TRC mean (SD)	TRC min–max	DC mean (SD)	DC low	DC med	DC high
T1	Introduction to AI	20	66.09 (12.70)	40.00–85.00	3.77 (1.43)	9.1%	26.4%	64.5%
T2	Neural Networks	20	69.35 (16.67)	40.00–95.00	3.84 (1.32)	6.6%	28.6%	64.8%
T3	Convolutional Neural Networks	20	64.78 (8.19)	40.00–85.00	4.00 (1.26)	4.5%	27.5%	68.0%
T4	Recurrent Networks and LSTMs	20	66.96 (12.95)	40.00–90.00	3.94 (1.22)	3.4%	32.0%	64.5%
T5	Autoencoders and Variational Autoencoders	20	73.04 (15.50)	40.00–100.00	3.69 (1.34)	6.6%	35.7%	57.7%
Ts	Summative consolidation (new items covering same concepts)	30	74.64 (11.71)	53.33–90.00	3.72 (1.29)	5.80%	38.30%	55.90%

Notes. TRC is reported on a 0–100 scale. DC is reported on a 0–5 scale. DC groups: low (0–1), medium (2–3), high (4–5).

C. Experimental Results (RQ1): Performance and Calibration

1) Global trajectories (T1–Ts)

a) Performance (TRC)

Regarding performance, TRC increased from M = 66.09 (SD = 12.70) at T1 to M = 74.64 (SD = 11.71) at Ts, yielding a mean change of Δ = +8.55 points (95% CI [+0.80, +16.30]). This increase was supported by a significant pre-post comparison, $t(22) = 2.289$, $p = 0.032$, and by the Wilcoxon signed-rank test, $p = 0.048$. In addition, the mean slope estimated across the six assessments indicated an average increase of +1.60 points per assessment (95% CI [+0.42, +2.78]), $t(22) = 2.819$, $p = 0.010$, corresponding to a small-to-moderate standardized change (d_z = 0.48). As shown in Fig. 4, this evolution was not strictly linear across successive assessments, but it followed an overall upward trajectory from T1 to Ts.

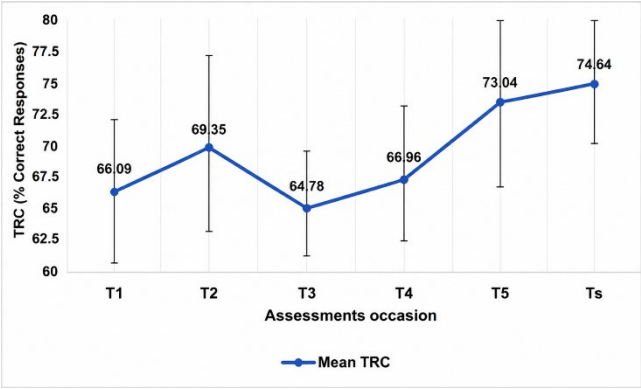

Fig. 4. TRC trajectory from T1 to Ts (0–100 scale).

A higher TRC reflects better mastery of course content, which is consistent with expectations for repeated formative assessment and iterative feedback loops that can support consolidation of learning [2, 4, 31]. At the same time, in this exploratory within-subject design, this progression should be interpreted as potentially reflecting multiple concurrent influences (natural progression through the course, familiarity with the assessment format, practice effects, and differences in topic coverage and item difficulty across assessments), rather than being uniquely attributable to any single component of the intervention.

b) Metacognitive realism (REAL)

On the calibration side, the most pronounced change concerns REAL. Mean REAL rose from M = 74.07 (SD = 8.60) at T1 to M = 84.17 (SD = 6.27) at Ts, a gain of Δ = +10.10 points (95% CI [+5.65, +14.55]). This observed increase was robust ($t(22) = 4.706$, $p < 0.001$; Wilcoxon $p < 0.001$) with a large effect size (d_z = 0.98). As shown in Fig. 5, this evolution was not strictly linear across successive assessments, but it followed an overall upward trajectory from T1 to Ts. The mean slope across the six assessments was +1.79 points per assessment (95% CI [+1.03, +2.55]; $t(22) = 4.898$, $p < 0.001$). Inter-individual variability is summarized in Fig. 6, which displays the distribution of individual slopes (points per assessment), generally positive but heterogeneous (median = +1.24; IQR = +0.80 to +3.20).

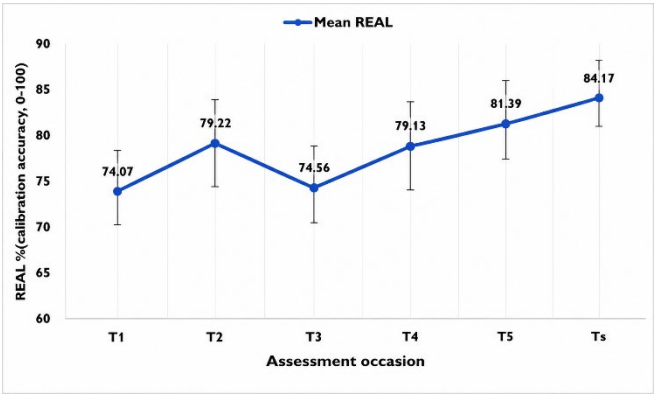

Fig. 5. Mean REAL from T1 to Ts with 95% CI.

In the DC/CBM literature, REAL reflects the alignment between confidence and actual accuracy—that is, metacognitive realism [21–27]. Leclercq [21] formalizes this confidence-accuracy alignment as a form of metacognitive information that can be exploited in assessment, and applied studies report that integrating DC is both feasible and acceptable [29] and that feedback tied to certainty levels can support learning [30]. Our findings are consistent with prior work and document a sustained increase in REAL over

multiple time points in a setting where feedback is not only corrective, but also explicitly structured to highlight episodes of overconfidence and to foster a metacognitive reading of the indicators.

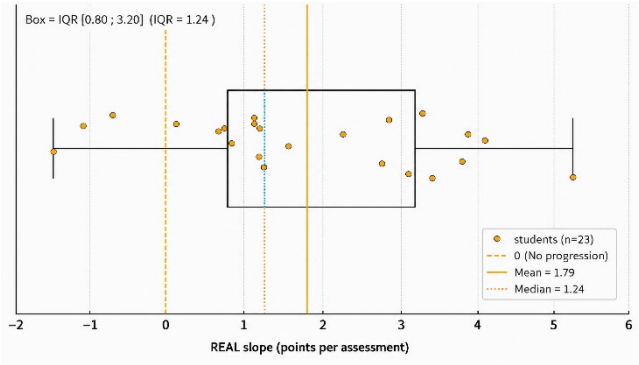

Fig. 6. Distribution of individual REAL slopes across the six assessments.

c) Decomposition of bias: CENTR, SUR_EST, SOU_EST

To clarify how calibration changed, we examined the directional components of miscalibration.

The CENTR index, which captures the overall direction of net bias (C_{moy} – TRC), decreased from M = 16.64 (SD = 15.85) at T1 to M = 8.23 (SD = 9.66) at Ts, a change of $\Delta = -8.41$ points (95% CI [−15.94, −0.88]). This decrease is supported by a significant pre-post comparison ($t(22) = -2.315$, $p = 0.030$; Wilcoxon $p = 0.023$) and by a mean longitudinal slope of −1.77 points per assessment (95% CI [−2.85, −0.68], $p = 0.003$), corresponding to a small-to-moderate standardized pre–post change ($d_z = -0.48$). Because lower CENTR values indicate a weaker overall tendency toward overconfidence, the observed pattern is consistent with a progressively less biased confidence profile (Fig. 7).

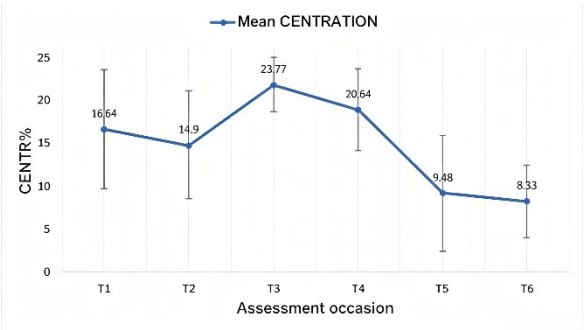

Fig. 7. Longitudinal trajectory of CENTR (T1 to Ts).

The SUR_EST indicator also showed a marked change. Mean SUR_EST declined from M = 21.29 (SD = 11.09) at T1 to M = 12.03 (SD = 6.95) at Ts, a reduction of $\Delta = -9.26$ points (95% CI [−14.70, −3.81]). The pre–post comparison was significant ($t(22) = -3.523$, $p = 0.002$; Wilcoxon $p = 0.002$) with a moderate-to-large standardized change ($d_z = -0.73$). The mean slope across assessments was −1.78 points per assessment (95% CI [−2.61, −0.95], $p < 0.001$) (Fig. 8).

A decline in SUR_EST reflects a reduction in episodes of overconfidence (high certainty on incorrect answers), which is coherent with more finely tuned calibration over time. In DC/CBM studies, reducing overconfidence is often a central aim because such cases are considered particularly "dangerous" for learning—errors that are strongly believed [21–27].

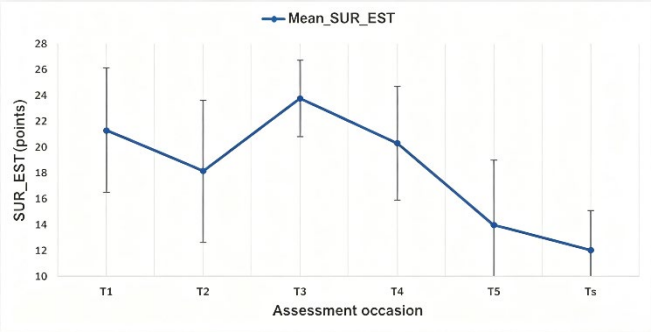

Fig. 8. SUR_EST trajectory from T1 to Ts.

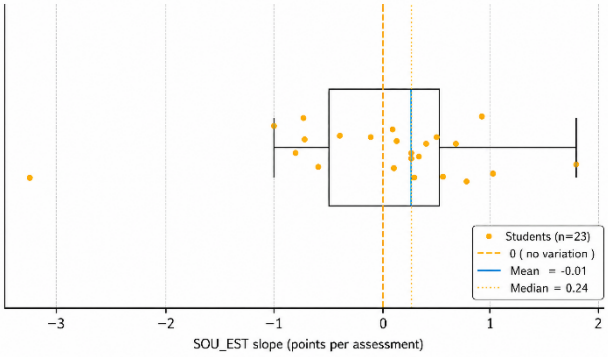

Fig. 9. Individual slopes for SOU_EST (points per assessment).

By contrast, SOU_EST (underestimation) remained relatively stable. Means were M = 4.65 (SD = 6.30) at T1 and M = 3.80 (SD = 4.25) at Ts, a small change of $\Delta = -0.85$, which did not reach statistical significance ($t(22) = -0.598$, $p = 0.556$; Wilcoxon $p = 0.903$; $d_z = -0.12$). Given the pedagogical concern of avoiding a shift toward excessive prudence, we complemented this analysis with a TOST (two one-sided tests) equivalence procedure, using an a priori equivalence margin of ±2 points on the 0–100 scale. The hypothesis of an increase beyond this margin was rejected ($p = 0.028$), suggesting that SOU_EST did not increase beyond a pedagogically negligible level. The mean slope across assessments likewise indicated no systematic drift (95% CI [−0.44, +0.42]) (Fig. 9).

Overall, the pattern of higher TRC and REAL, lower CENTR and SUR_EST, and stable SOU_EST is consistent with an evolution toward better content mastery and more accurate, less dangerously overconfident metacognitive judgment, while avoiding a general drift toward under confidence.

Table 6. Pre–post changes in performance and calibration indicators from assessment 1 to the summative assessment

Outcome	T1 Mean (SD)	Ts Mean (SD)	Δ (Ts–T1)	95% CI for Δ	Pre–post test	*p*	Effect size (dz)
TRC	66.09 (12.70)	74.64 (11.71)	+8.55	[0.80, 16.30]	t(22)=2.289	0.032	0.48
REAL	74.07 (8.60)	84.17 (6.27)	+10.10	[5.65, 14.55]	t(22)=4.706	<.001	0.98
CENTR	16.64 (15.85)	8.23 (9.66)	−8.41	[−15.94, −0.88]	t(22)=−2.315	0.030	−0.48
SUR_EST	21.29 (11.09)	12.03 (6.95)	−9.26	[−14.70, −3.81]	t(22)=−3.523	0.002	−0.73
SOU_EST	4.65 (6.30)	3.80 (4.25)	−0.85	[−3.78, 2.09]	t(22)=−0.598	0.556	−0.12

For transparency and ease of reference, Tables 6 and 7 provide the corresponding summary statistics: Table 6 reports the T1–Ts pre–post comparisons (mean differences with 95% CIs and associated tests), and Table 7 summarizes longitudinal trajectories across the six assessments via mean slopes expressed in points per assessment.

Table 7. Longitudinal trends across the six assessment occasions

Outcome	Mean slope (points/assessment)	95% CI	Test vs 0	*p*
TRC	+1.60	[0.42, 2.78]	t(22)=2.819	0.010
REAL	+1.79	[1.03, 2.55]	t(22)=4.898	<0.001
CENTR	−1.77	[−2.85, −0.68]	significant	0.003
SUR_EST	−1.78	[−2.61, −0.95]	significant	<0.001
SOU_EST	≈ 0	[−0.44, 0.42]	non-significant	n.s.

D. Experimental Results (RQ2): Perceptions of Indicator-Informed GenAI Feedback

Students' perceptions were measured with a five-point Likert questionnaire (1–5) covering usefulness, clarity, trust, intention to use feedback, cognitive load, and self-efficacy. Internal consistency was generally satisfactory for Usefulness, Clarity, and Trust ($\alpha \approx 0.73$–0.80, 0.77–0.80, and 0.76–0.80, respectively), and very high for intention and cognitive load ($\alpha \approx 0.96$–0.97), indicating strong internal homogeneity and potential item redundancy; these subscales should therefore be interpreted cautiously and refined in future iterations. Self-efficacy also showed very high reliability ($\alpha \approx 0.91$–0.97). Mean scores aggregated across the five formative assessments were close to the scale midpoint: $U = 3.01$ ($SD = 1.01$), $C = 3.00$ ($SD = 1.04$), $T = 3.01$ ($SD = 1.04$), $I = 2.97$ ($SD = 1.19$), $L = 2.98$ ($SD = 1.18$), and AE = 2.96 ($SD = 1.21$) (see Table 8).

Table 8. Descriptive statistics and reliability for perception subscales (Likert 1–5)

Subscale	M	SD	α (Cronbach)
Usefulness (U)	3.01	1.01	0.73–0.80
Clarity (C)	3.00	1.04	0.77–0.80
Trust (T)	3.01	1.04	0.76–0.80
Intention/Usage (I)	2.97	1.19	0.96–0.97
Cognitive load (L)	2.98	1.18	0.96–0.97
Self-efficacy (AE)	2.96	1.21	0.91–0.97

Notes: Means aggregated over the five formative assessments. Very high α values (≈ 0.97) for I and L indicate strong internal homogeneity and possible item redundancy.

Overall, this "neutral-to-slightly-positive" profile is in line with observations from other automated feedback systems: acceptance is not uniform and depends strongly on perceived clarity, relevance, and students' ability to interpret and reuse comments (i.e., their feedback literacy). In a GenAI context, the observed dispersion also fits with frequently reported issues related to trust and reliability (hallucinations, overconfidence, academic integrity) [10–15, 17].

E. Experimental Results (RQ3): Associations between Perceptions and Objective Changes

To explore how perceived feedback quality relates to objective change, we computed Spearman correlations between students' mean perception scores and their changes (Δ) or slopes on the main indicators. Two statistically significant associations emerged:

- Clarity and ΔTRC: $\rho = +0.44$, $p = 0.037$ (Fig. 10).
- Intention/Usage and the SUR_EST slope: $\rho = -0.45$, $p = 0.030$, indicating that higher intended use is associated with a faster reduction in overestimation (Fig. 11).

A third association was positive but not statistically significant:

- Usefulness and ΔREAL: $\rho = +0.36$, $p = 0.096$ (Fig. 12).

Although these correlations are based on a small sample, they are consistent with the idea that feedback perceived as clear—and intended to be reused—may be more likely to be processed and reinvested by students. This interpretation aligns with formative feedback models oriented toward self-regulation [1–6, 9] and with work on feedback literacy, understood as the capacity to understand, judge, and use feedback productively. Importantly, these associations should be interpreted as correlational within the combined DC-enhanced assessment and feedback workflow and do not identify GenAI-specific effects; they may reflect differences in students' engagement with the overall process.

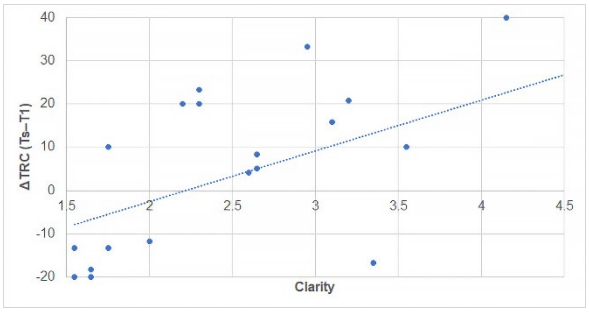

Fig. 10. Relationship between Clarity and ΔTRC (Ts–T1).

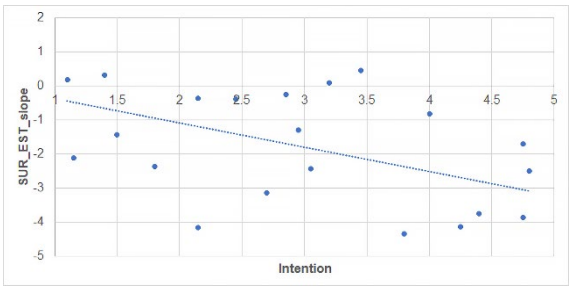

Fig. 11. Relationship between Intention to use feedback and SUR_EST slope.

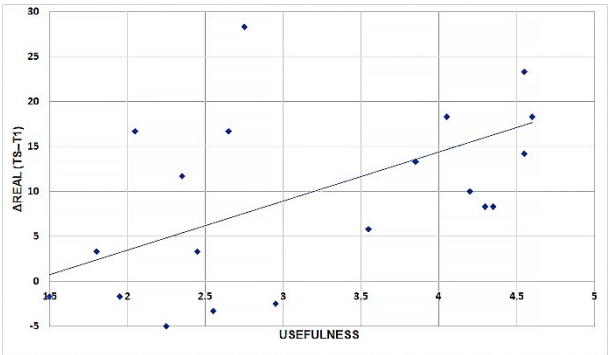

Fig. 12. Relationship between Usefulness and ΔREAL (Ts–T1).

A broader overview is provided in the correlation matrix (Fig. 13), where significant two-tailed Spearman correlations ($p < 0.05$) are marked with an asterisk. Asterisks indicate significance levels based on unadjusted two-tailed p-values. Because the matrix includes multiple comparisons, these associations should be interpreted as exploratory; the multiple-comparison strategy is described in Section III.H.3.

Although based on a small sample ($N = 23$), these moderate correlations suggest a coherent association: students who perceived the feedback as clearer and reported a stronger

intention to use it, tended to show more favorable trajectories in performance and calibration, whereas the usefulness–REAL association was positive but did not reach statistical significance in the current dataset.

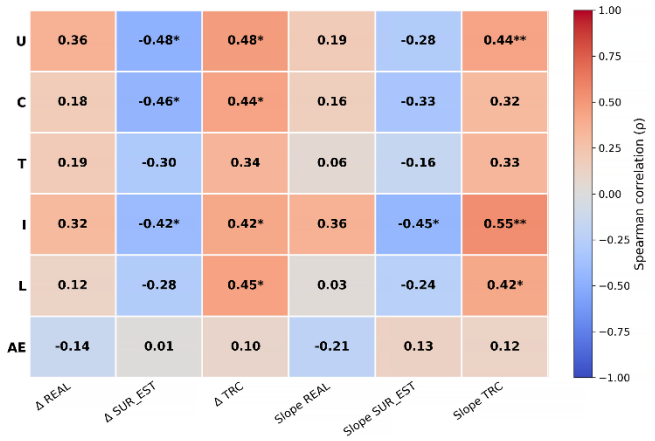

Fig. 13. Spearman correlations between perception dimensions and objective change indicators (ΔREAL, ΔSUR_EST, ΔTRC, slopes).

F. Comparison with Related Work

Overall, our findings sit at the intersection of three strands of work: Certainty-Based Marking and Degree of Certainty (CBM/DC), formative feedback and self-regulated learning, and LLM-based feedback assistants. A distinctive feature of our design is that calibration indicators are not only computed, but also explicitly integrated into the feedback logic.

First, the empirical pattern we observe—higher TRC, higher REAL, lower SUR_EST and CENTR, and stable SOU_EST—is coherent with the rationale of CBM/DC approaches. In these frameworks, combining accuracy and certainty serves to surface "dangerous" strongly believed errors and to support metacognitive calibration. Our trajectories are consistent with recurrent DC/CBM findings showing that repeated exposure to confidence–performance discrepancies, combined with feedback, can be associated with increases in REAL and reductions in overconfidence in STEM and health education [21–27].

Second, the design is aligned with formative feedback and SRL models, which emphasize that feedback is most effective when it supports monitoring, interpretation, and strategy adjustment, rather than simply providing correct answers [1–6, 9]. In our case, the key step is the move from indicators to action: instead of only displaying indicators on a dashboard, the system transforms item-level DC information and student-level indicators into a structured narrative report that combines diagnosis, metacognitive interpretation, and concrete recommendations for future study.

Third, the contribution speaks directly to recent GenAI and LLM-based feedback systems. Prior studies show that tools such as ChatGPT can support learning and cognitive skills through rapid, interactive assistance [14], and that combining LLMs with active pedagogies can be associated with changes in performance and engagement [42]. Other work on LLM-based formative feedback assistants underlines the value of structured comments and SRL-oriented scaffolds [38], while systems for open-ended responses stress the importance of teacher-guided prompting in producing actionable feedback [19]. However, these initiatives generally remain content-centric: they rarely treat calibration (alignment between confidence and accuracy) as an explicit design target or rely on quantified calibration profiles. Appendix E provides a concise feature-level comparison with representative CBM/DC approaches and recent LLM-based feedback assistants (inputs, calibration modeling, grounding, output constraints, instructor oversight, and longitudinal analytics).

In this respect, our approach adds a calibration-aware layer to LLM use. The feedback module relies on carefully engineered prompts [43] and a canonical JSON template that encodes the six DC cases and the intended pedagogical functions of each section of the feedback. In response to well-documented concerns about LLM reliability (hallucinations, inconsistency), we combine document grounding and constrained generation, in line with retrieval-augmented strategies for improving factual fidelity [16] and with recent efforts to evaluate and filter LLM feedback quality [20]. Where CBM/DC approaches instrument calibration without always translating it into structured guidance, and where many LLM assistants generate rich but weakly calibrated feedback, our module aims to integrate both: An LLM used not as a generic explanation generator, but as a calibration-aware formative feedback engine aimed at helping students understand and regulate their learning strategies more deliberately.

Importantly, this design does not disentangle familiarization with DC from the potential contribution of the feedback reports, nor does it isolate GenAI-specific effects. Accordingly, we position our contribution as a feedback-augmented DC workflow: DC captures calibration-relevant signals, and the GenAI module operationalizes these signals into structured, course-grounded feedback (item-level A–F cases plus an indicator-based synthesis) under instructor validation, without implying that observed calibration changes are attributable to GenAI alone.

G. General Implications and Future Work

Taken together, the results indicate a consistent pattern: TRC and REAL increased, CENTR and SUR_EST decreased, and SOU_EST remained stable. At the perceptual level, two dimensions—clarity and intention to use—were associated with more favorable trajectories, whereas usefulness showed a positive but non-significant trend in relation to REAL. In this exploratory single-cohort pilot, these observations are compatible with the hypothesis that an indicator-informed, DC-structured feedback workflow may support both performance and metacognitive calibration, while also underscoring the need for stronger evidence.

Several limitations temper this interpretation. The study relies on a small sample ($N = 23$) from a single Master's-level AI course and uses a within-subject design without a parallel control group. Consequently, we cannot disentangle the respective contributions of DC-enhanced testing, the feedback reports, and natural course progression. The observed changes should therefore be interpreted as co-occurring with the intervention rather than as uniquely caused by any single component. In addition, we did not systematically track students' engagement with the reports (e.g., time on page, depth of reading, or integration into study

strategies), and some questionnaire subscales (notably Intention/Usage and Cognitive Load) showed very high internal consistency, suggesting potential item redundancy and calling for refinement in future iterations. Finally, feedback literacy was discussed as a relevant lens, but it was not directly measured in this pilot. Moreover, participants had not previously been exposed to indicator-enriched feedback reports combining item-level explanations with cognitive and metacognitive indicators (e.g., REAL, over-/underestimation) and their interpretation. This novelty may have influenced both perceived usefulness/clarity and uptake, and should be addressed in future work through baseline measurement of feedback literacy and brief orientation activities to support interpretation of the indicators.

Despite these constraints, the study illustrates how DC-based indicators and LLM-assisted feedback can be integrated into a coherent, calibration-oriented design. From a practical perspective, this suggests that it may be useful to (1) embed DC and calibration indicators in regular assessment, (2) translate these indicators into structured, metacognitively explicit feedback reports, and (3) support students in interpreting and acting on these reports, for instance through brief in-class debriefings or guided reflection activities. Future research should replicate and extend this work with larger and more diverse cohorts, explicit comparison conditions (e.g., DC without feedback reports; feedback without DC; human-only feedback), and richer traces of feedback use. Such studies would help clarify the mechanisms through which a feedback-augmented DC workflow contributes—or fails to contribute—to changes in performance, calibration, and feedback literacy, and could inform refinements of both the indicators and the feedback templates.

H. DC Familiarization versus Feedback Effects

An important internal-validity consideration is that calibration changes may partly reflect familiarization with the DC technique itself rather than the feedback reports. Repeated certainty rating can function as a metacognitive training signal by making certainty judgments salient and encouraging monitoring across successive assessments. Because DC practice and feedback were delivered as a combined routine in our design, we cannot isolate the incremental contribution of the GenAI-generated reports beyond course progression and repeated DC use.

Within these constraints, the GenAI component primarily serves an interpretive and scaffolding function: it translates item-level correctness and DC into six pedagogical cases (A–F) and provides an individualized synthesis of calibration indicators to help students make sense of their confidence-accuracy profile. We therefore interpret the observed trends as co-occurring with a feedback-augmented DC workflow, rather than as evidence of a GenAI-specific effect.

V. Conclusion

This exploratory pilot study examined the feasibility of coupling indicator-informed generative AI feedback with certainty-based assessment in an authentic Master's-level engineering course. By combining Degree of Certainty (DC) ratings with course-grounded GPT-4 prompts and a structured feedback template, we implemented a workflow in which each assessment yielded both performance scores and calibration indicators, together with personalized feedback reports.

The observed patterns—higher TRC, higher REAL, lower SUR_EST and CENTR, and stable SOU_EST—are consistent with the hypothesis that such a design may support more accurate confidence judgments over time.

Questionnaire data indicate that students generally perceived the reports as moderately useful and clear. In addition, clarity and intention to use showed significant associations with more favorable trajectories in performance and calibration, whereas the usefulness–REAL association was positive but not statistically significant. These convergences remain tentative and should be interpreted as associative rather than causal, but they suggest a potentially meaningful link between how feedback is experienced and the evolution of confidence-accuracy profiles.

From a practical perspective, the approach remains lightweight for instructors: DC-enhanced assessments are administered in the usual course environment, indicators are computed automatically in Evalviz-Dashboard, and feedback is generated through a constrained, document-grounded LLM pipeline before being reviewed and released via the student dashboard. This makes it feasible to provide individual, calibration-aware feedback at the cohort scale without specialized technical expertise.

Given the small sample size, the single-course context, and the absence of a control group, the findings should be interpreted with caution and viewed as hypothesis-generating rather than conclusive. Future work with larger and more diverse cohorts, explicit comparison conditions (e.g., DC without GenAI, GenAI without indicators), and richer traces of feedback use will be needed to clarify the conditions under which indicator-informed, course-grounded GenAI feedback may support performance, calibration indicators, and self-regulated learning.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

H.A. was primarily responsible for writing the manuscript, guiding the experimental design, and analyzing the results N.E. supervised the study, provided technical research input, and collected the data. F.L. offered supervisory guidance and contributed to manuscript editing. J.L.G. provided comments on the feedback-generation process and contributed to basic statistical analyses. J.M.R. and E.A.P. assisted with manuscript revisions and editing. All authors reviewed and approved the final version of the manuscript.

References

[1] J. Hattie and H. Timperley, "The power of feedback," *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007. doi: 10.3102/003465430298487

[2] V. J. Shute, "Focus on formative feedback," *Rev. Educ. Res.*, vol. 78, no. 1, pp. 153–189, 2008. doi: 10.3102/0034654307313795

[3] C. Bosc-Miné, "Characteristics and functions of feedback in learning," *L'Année Psychologique*, vol. 114, no. 2, pp. 315–353, 2014. doi: 10.4074/S000350331400205X (in French)

[4] D. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," *Studies in Higher Education*, 31, pp. 199–218, 2006. doi: 10.1080/03075070600572090

[5] D. L. Butler and P. H. Winne, "Feedback and self-regulated learning: A theoretical synthesis," *Rev. Educ. Res.*, vol. 65, no. 3, pp. 245–281, 1995. doi: 10.3102/00346543065003245

[6] S. M. Brookhart, *How to Give Effective Feedback to Your Students*, 2nd ed. Alexandria, VA, USA: ASCD, 2017.

[7] D. H. Jonassen, "Toward a design theory of problem solving," *Educ. Technol. Res. Dev.*, vol. 48, no. 4, pp. 63–85, 2000.

[8] A. P. Cavalcanti, A. Barbosa, R. Carvalho, F. Freitas, Y.-S. Tsai, D. Gašević, and R. F. Mello, "Automatic feedback in online learning environments: A systematic literature review," *Comput. Educ.*, vol. 191, 104641, 2022. doi: 10.1016/j.compedu.2022.104641

[9] D. Carless and D. Boud, "The development of student feedback literacy: Enabling uptake of feedback," *Assess. Eval. High. Educ.*, vol. 43, no. 8, pp. 1315–1325, 2018. doi: 10.1080/02602938.2018.1463354

[10] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, "Vision, challenges, roles and research issues of artificial intelligence in education," *Comput. Educ.: Artif. Intell.*, vol. 1, 100001, 2020. doi: 10.1016/j.caeai.2020.100001

[11] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms," *Comput. Educ.: Artif. Intell.*, vol. 2, 100020, 2021. doi: 10.1016/j.caeai.2021.100020

[12] J. Bravo and C. Cruz-Bohorquez, "Generative AI in engineering education: Opportunities and risks," *Int. J. Eng. Educ.*, 2024.

[13] L. M. Sánchez-Ruiz, S. Moll-López, A. Nuñez-Pérez, J. A. Moraño-Fernández, and E. Vega-Fleitas, "ChatGPT challenges blended learning methodologies in engineering education: A case study in mathematics," *Appl. Sci.*, vol. 13, no. 10, 6039, 2023. doi: 10.3390/app13106039

[14] H. B. Essel, D. Vlachopoulos, A. B. Essuman, and J. O. Amankwa, "ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational Large Language Models (LLMs)," *Comput. Educ.: Artif. Intell.*, vol. 6, 100198, 2024. doi: 10.1016/j.caeai.2023.100198

[15] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer *et al*., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, 102274, 2023. doi: 10.1016/j.lindif.2023.102274

[16] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," arXiv preprint, arXiv:2005.11401, 2020.

[17] T. Rasul *et al.*, "The role of ChatGPT in higher education: Benefits, challenges, and future research directions," *J. Appl. Learn. Teach.*, vol. 6, no. 1, pp. 41–56, 2023. doi: 10.37074/jalt.2023.6.1.29

[18] D. De Silva, N. Mills, M. El-Ayoubi, M. Manic, and D. Alahakoon, "ChatGPT and generative AI guidelines for addressing academic integrity and augmenting pre-existing chatbots," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, 2023, pp. 1–6. doi: 10.1109/ICIT58465.2023.10143123

[19] J. K. Matelsky, F. Parodi, T. Liu, R. D. Lange, and K. P. Kording, "A large language model-assisted education tool to provide feedback on open-ended responses," arXiv preprint, arXiv:2308.02439, 2023.

[20] K. Qian, Y. Cheng, R. Guan, W. Dai, F. Jin, K. Yang *et al.*, "Dean of LLM tutors: Exploring comprehensive and automated evaluation of LLM-generated educational feedback via LLM feedback evaluators," arXiv preprint, arXiv:2508.05952, 2025.

[21] D. Leclercq, "Confidence marking: Its use in testing," *Evaluation in Education*, vol. 6, no. 2, pp. 161–287, 1982. doi: 10.1016/0191-765X(82)90011-8

[22] J.-L. Gilles. (1996). Use of degrees of certainty and realism norms in examination and self-assessment situations at FA.P.S.E.–ULG. [Online]. Available: http://hdl.handle.net/20.500.12162/4809 (in French)

[23] J.-L. Gilles, "Spectral qualities of standardized university tests: Development of edumetric indices for analyzing the spectral quality of assessments of university students' learning outcomes and application to the MOHICAN Check Up '99 tests," Ph.D. dissertation, Univ. de Liège, 2002. (in French)

[24] D. Leclercq and M. Poumay, "Edumetric analyses and metacognitive indices applied to the questions of the 10 MOHICAN check-ups," Liège, Belgium: Éd. Univ. de Liège, 2003. (in French)

[25] B. De Finetti, "Methods of discriminating levels of partial knowledge concerning a test item," *British Journal of Mathematical and Statistical Psychology*, vol. 18, pp. 87–123, 1965. doi: 10.1111/j.2044-8317.1965.tb00695.x

[26] E. H. Shuford, A. Albert, and H. E. Massengill, "Admissible probability measurement procedures," *Psychometrika*, vol. 31, no. 2, pp. 125–145, 1966. doi: 10.1007/BF02289503

[27] D. Leclercq and B. Noel, "How can cognitive and metacognitive abilities be developed?" in *Research and Actions in Support of University Student Success: Twenty Years of Collaboration within the "Success" Commission of the Interuniversity Council of the French Community*, Bruxelles: CIUF, 2011, pp. 55–60. (in French)

[28] D. A. Curtis, L. S. Lind, C. K. Boscardin, M. Dellinges, "Does student confidence on multiple-choice question assessments provide useful information?" *Med Educ.*, vol. 47, no. 6, pp. 578–84, 2013. doi: 10.1111/medu.12147

[29] C. Monod-Broca, J. Bouix, and D. Leclercq, "Feasibility and acceptability of assessment using degrees of certainty," *Médecine et Armées*, vol. 49, no. 3, pp. 21–36, 2023. doi: 10.17184/eac.7900. (in French)

[30] H. Akhasbi, N. El-Kamoun, F. Lakrami, J. L. Gilles, J. M. Rigo, and E. Aliss, "Remote formative assessment with multiple-choice questions and levels of certainty: The impact of feedback on learning," in *Proc. ICERI2023*, 2023, pp. 3709–3719. doi: 10.21125/iceri.2023.0948

[31] H. L. Roediger III and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," *Psychological Science*, vol. 17, no. 3, pp. 249–255, 2006. doi: 10.1111/j.1467-9280.2006.01693.x

[32] C. Redondo, C. Garcin, J. Pironom, and O. Thuilier, "Value of online practice tests on Moodle: A formative assessment system in the context of distance university teaching," *Evaluer. Journal International de Recherche en Education et Formation*, vol. 7, no. 1, pp. 41–70, 2022. (in French)

[33] A.R. Gardner-Medwin, and M. Gahan. (2019). Formative and summative confidence-based assessment. Figshare. [Online]. Available: https://hdl.handle.net/2134/1910

[34] C. H. Houzé-Cerfon, D. Lauque, and S. Charpentier, "Integrating the degree of certainty into the assessment of emergency medicine students' knowledge," *Ann. Fr. Med. Urgence*, vol. 6, pp. 389–394, 2016. doi: 10.1007/s13341-016-0685-1 (in French)

[35] J. L. Gilles, B. R. Pignolet, and E. Trisconi. (2019). Collective development of two large-scale examinations using the Construction and quality management cycle for standardized tests (CGQTS cycle) and the online Docimo platform. [Online]. Available: https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Lehre/Veranstaltungen_2019/Workshop4_HEPVaud.pdf (in French)

[36] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Proc. 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[37] H. Akhasbi, N. El-Kamoun, F. Lakrami, J.-L. Gilles, J.-M. Rigo, and E. Aliss, "Personalizing formative feedback through generative artificial intelligence: An experiment in higher education," in *Proc. EDULEARN25*, 2025, pp. 9338–9347.

[38] S. Steinert, K. E. Avila, S. Ruzika *et al.*, "Harnessing large language models to develop research-based learning assistants for formative feedback," *Smart Learn. Environ.*, vol. 11, 62, 2024. doi: 10.1186/s40561-024-00354-1

[39] D. Leclercq, "A meta-analysis of degrees of certainty expressed in words," *Journal International de Recherche en Education et Formation*, vol. 2, no. 3, 2017. (in French)

[40] D. Leclercq. (2020). Degrees of certainty: Spectral analyses of correct answers. Université de Liège, ORBi. [Online]. Available: http://hdl.handle.net/2268/257334 (in French)

[41] O. Prosperi, "Realism with degrees of certainty," *Mesure et Évaluation en Éducation*, vol. 38, no. 1, pp. 121–140, 2015. (in French)

[42] T. Al-Shloul *et al.*, "Role of activity-based learning and ChatGPT on students' performance in education," *Comput. Educ.: Artif. Intell.*, vol. 6, 100219, 2024. doi: 10.1016/j.caeai.2024.100219

[43] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," arXiv preprint, arXiv:2402.07927, 2024.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).