Supplementary

APPENDIX A. INDICATOR DEFINITIONS AND FORMULAS

A. Correct Response Rate (TRC)

The **Correct Response Rate (TRC)** constitutes the basic indicator of cognitive performance. It corresponds to the proportion of items answered correctly by the student out of all items in the test. In studies incorporating degrees of certainty, this indicator is equivalent to the accuracy rate or percentage correct. It measures the level of mastery of the assessed knowledge independently of the degree of certainty associated with the responses.

Formula:

$$TRC = 100\frac{N^+}{N}$$

where N^+ is the number of correct responses and N is the total number of items.

B. Mean Confidence (C_{moy})

The Mean confidence (C_{moy}) corresponds to the average of the degrees of certainty reported by the student across all items, expressed on the certainty scale used in the assessment procedure. This indicator measures the student's average level of subjective confidence at the time of responding to the questions.

Let $DC_i \in \{0, \ldots, 5\}$ denote the certainty level reported for item i. Each DC_i is mapped to a confidence value $c_i \in [0,100]$ using the midpoint of the corresponding probability band. Mean confidence is then computed as:

Formula:

$$C_{moy} = \frac{1}{N}\sum_{i=1}^{N} c_i$$

where N is the number of items and c_i is the transformed confidence associated with item i. This definition is consistent with the certainty-based calibration framework, in which mean declared confidence is compared to observed accuracy to Compute Centration (CENTR) and related indicators.

C. Centration (CENTR)

Centration measures the overall discrepancy between a student's average reported level of certainty and their actual observed performance on the test. It constitutes an indicator of metacognitive calibration bias, making it possible to identify a general tendency toward either overestimation or underestimation of one's knowledge.

Formula:

$$\text{CENTR} = C_{moy} - \text{TRC}$$

where C_{moy} is mean declared certainty and TRC is the correct response rate.

Interpretation:

- CENTR > 0: tendency toward overestimation
- CENTR < 0: tendency toward underestimation
- CENTR ≈ 0: overall balance between certainty and performance

Accordingly, **centration** can be interpreted as the difference between the phenomena of overestimation and underestimation.

D. Realism (REAL)

Realism, also referred to as **calibration**, denotes the degree of correspondence between the degrees of certainty reported by the student and their actual observed performance. In Leclercq's work, it refers to the closeness between the accuracy rates expected on the basis of the declared certainty levels and the accuracy rates actually observed. Realism is considered perfect when, for each certainty level, the observed performance matches the probability associated with that level. Graphically, this situation is represented by points lying on the diagonal of the calibration graph relating certainty to accuracy. Thus, realism does not merely indicate a general tendency toward overestimation or underestimation; more broadly, it reflects the overall quality of the fit between the student's subjective judgment and their objective performance.

Formulas:

Realism is estimated by comparing, for each level of certainty, the observed success rate with the expected probability associated with that certainty level.

(1) Realism for a given certainty level *k*:

$$Real_k = 100 - |T_k - t_k|$$

With:

- T_k = observed success rate for certainty level k
- t_k = target probability associated with that certainty level
- $|\ T_k - t_k\ |$ = difference between observed performance and expected performance

This expression measures the closeness between actual performance and reported certainty.

(2) The student's overall realism

Overall realism corresponds to the weighted average of the realism values calculated for each level of certainty:

$$REAL = \left(\frac{1}{n}\right)\sum_{k=0}^{5} N_k Real_k$$

where:

- N_k: number of responses associated with certainty level k
- n: total number of responses

Interpretation:

A high **REAL** score indicates strong agreement between the student's subjective certainty judgment and their observed objective performance, reflecting good metacognitive calibration. Conversely, a lower score reflects a discrepancy between the student's subjective appraisal of their knowledge and the results actually obtained.

E. Overestimation and Underestimation Rates of Certainty (SUR_EST and SOU_EST)

- Overestimation refers to a student's tendency to assign a higher degree of certainty to their responses than is warranted by their actual observed performance. It

reflects a situation in which the subjective judgment of confidence exceeds the objective reality of success. Within the analysis of metacognitive calibration, overestimation corresponds to a positive centration bias, that is, when the average reported level of certainty exceeds the correct response rate. It therefore constitutes a form of overconfidence, indicating that the student overestimates the validity of their knowledge or responses.

- Underestimation refers to a student's tendency to assign a lower degree of certainty to their responses than is warranted by their actual observed performance. It corresponds to a situation in which the subjective judgment of confidence falls below the objective reality of success. In the analysis of metacognitive calibration, underestimation arises when the average reported level of certainty is lower than the correct response rate. It thus reflects a form of underconfidence, indicating that the student evaluates the validity of their knowledge or responses in an overly cautious or pessimistic manner.

Formula:

Centration measures the overall calibration bias between a student's average reported level of certainty and their actual observed performance. It corresponds to the difference between the average expressed confidence and the actual success rate:

$$CENTR = C_{moy} - TRC$$

This discrepancy may result from two opposite phenomena:

- **overestimation,** when reported confidence exceeds actual performance;
- **underestimation,** when reported confidence is lower than actual performance.

Centration therefore represents the net bias resulting from these two opposing tendencies. Accordingly, it can be expressed as the difference between the overestimation and underestimation components:

$$CENTR = SUR\_EST - SOUS\_EST$$

This relation indicates that centration captures the dominant direction of metacognitive bias: a positive value signals a predominance of overestimation, whereas a negative value reflects a predominance of underestimation.

Furthermore, **realism** measures the quality of the fit between subjective certainty judgments and observed objective performance. Discrepancies between confidence and performance may take two opposite forms: **overestimation**, when reported certainty exceeds actual success, and **underestimation**, when it falls below it. These two components represent the different manifestations of calibration bias. Realism therefore corresponds to the proportion of correctly calibrated judgments and can be defined as the complement of the total calibration bias, represented by the sum of overestimation and underestimation. This relation is expressed as follows:

$$REAL = 100 - (SUR\_EST + SOUS\_EST)$$

These two relations make it possible to express analytically the two components of calibration bias. Solving the system leads to the following expressions:

$$SUR_EST = \frac{(100 - REAL) + CENTR}{2}$$

$$SOUS_EST = \frac{(100 - REAL) - CENTR}{2}$$

This formulation therefore makes it possible to decompose metacognitive behavior into three complementary components: a proportion of correctly calibrated judgments (**realism**), a proportion of **overestimation**, and a proportion of **underestimation**, satisfying the relation:

$$REAL + SUR\_EST + SOUS\_EST = 100$$

This decomposition provides a synthetic representation of the degree of fit between the student's subjective self-evaluation and their objective performance.

Interpretation:

- A high SUR_EST value indicates that the student displays excessive confidence relative to their actual performance. It is therefore an indicator of overconfidence bias.
- A high SOU_EST value indicates that the student performs better than would be expected from their reported level of certainty. This may reflect excessive caution, difficulty recognizing one's own mastery, or a form of underconfidence.

APPENDIX B: GENAI REPLICATION PACKAGE

A. Purpose and Scope

This appendix provides the replication package for the GenAI-assisted feedback workflow used in the study. Its purpose is to support transparency and reproducibility by documenting four core aspects of the system:

(1) The full prompt structure and canonical JSON template used to generate feedback,
(2) The exact model version and inference parameters,
(3) The grounding and document selection procedure, and
(4) The instructor review protocol, including quantitative reporting of edits and the most common categories of modifications.

The workflow described here is not based on unconstrained prompting. Instead, it combines deterministic pedagogical case assignment, a canonical JSON template implemented in the system, grounding in instructor-approved course materials, optional integration of performance and metacognitive indicators, and human validation before release. This architecture was designed to ensure that generated reports remain pedagogically structured, aligned with course content, and interpretable for both instructional and research purposes.

B. Full Prompt and Canonical JSON Template

1) Canonical JSON template

The system uses a canonical feedback template in JSON format. This template contains exactly seven top-level keys:

- Case A
- Case B
- Case C
- Case D
- Case E
- Case F
- Global Synthesis

These entries define the pedagogical scaffolding of the workflow.

The item-level feedback is governed by six pedagogical cases determined from the combination of response correctness and Degree of Certainty (DC). In this study, DC values were grouped into three levels:

- low: 0–1
- medium: 2–3
- high: 4–5

The deterministic rules were:

- Case A: correct answer + high certainty (DC ≥ 4) → mastered and confidently assumed knowledge
- Case B: correct answer + medium certainty (DC = 2 or 3) → uncertain success requiring consolidation
- Case C: correct answer + low certainty (DC ≤ 1) → unmastered success, possibly due to guessing
- Case D: incorrect answer + high certainty (DC ≥ 4) → confident error indicating a false representation
- Case F: incorrect answer + medium certainty (DC = 2 or 3) → doubt and confusion requiring clarification
- Case E: incorrect answer + low certainty (DC ≤ 1) → lack of conceptual anchors requiring basic grounding

This distinction is pedagogically central because the expected feedback differs across cases. A correct answer with high certainty is treated as robust understanding and therefore calls for reinforcement and extension. By contrast, an incorrect answer with high certainty is treated as a confident misconception, requiring explicit conceptual correction and metacognitive recalibration. Likewise, a correct answer with low certainty may reflect fragile knowledge or guessing, whereas an incorrect answer with low certainty suggests insufficient conceptual anchors and the need for supportive reconstruction.

Each case contains a structure field that constrains the pedagogical organization of the feedback. The exact fields vary across cases. For example:

- Case A emphasizes confirmation of strong understanding, conceptual extension, and advanced key takeaways;
- Case B emphasizes consolidation, mnemonic support, and practice recommendations;
- Case C emphasizes reassurance, clarification, contrastive examples, and confidence building;
- Case D emphasizes confident error detection, conceptual correction, decision cues, and metacognitive warning;
- Case F emphasizes doubt and conceptual confusion, requiring clarification and discrimination between close concepts;
- Case E emphasizes lack of anchors, reconstruction from basics, and supportive conceptual grounding.

The Global Synthesis section defines the final report-level synthesis based on student indicators such as mastery, realism, overestimation, underestimation, imprudence, and confidence-related indices.

2) Canonical JSON template (full structure used)

The workflow uses the following canonical JSON template structure:

```
{
 "Cas_A": {
  "statut": "Correct",
  "dc": "Élevé",
  "structure": {
   "intro": "Bravo pour cette réponse parfaitement ciblée ! En choisissant {STUDENT_ANSWER}, vous démontrez une maîtrise solide...",
   "explication": "L'apprentissage supervisé consiste à apprendre une fonction de prédiction à partir de données étiquetées...",
   "exemple": "Appliquez ces principes sur un jeu de données réel (Iris, Titanic…).",
   "points_cles": [
    "L'apprentissage supervisé exige des labels connus pour l'entraînement.",
    "Il couvre deux sous-types : classification (catégories) et régression (valeurs continues).",
    "Anticipez les défis avancés : surapprentissage, biais-variance, déséquilibres de classes."
   ]
  }
 },
 "Cas_B": {
  "statut": "Correct",
  "dc": "Moyen",
  "structure": {
   "intro": "Très bon choix ! Vous avez bien identifié {STUDENT_ANSWER}, mais votre confiance moyenne (DC {DC}/5)...",
   "explication": "L'apprentissage supervisé s'appuie sur des données étiquetées pour apprendre une fonction de prédiction...",
   "mnemo": "Supervisé ? Cherche la cible !",
   "points_cles": [
    "Supervisé = données étiquetées (label connu).",
    "Deux familles : classification vs régression.",
    "Processus standard : prétraitement → séparation train/test → choix du modèle → entraînement → évaluation."
   ],
   "recommandations": [
    "Faites une fiche comparant supervisé vs non supervisé.",
    "Utilisez scikit-learn avec des jeux comme Iris, Titanic..."
   ]
  }
 },
 "Cas_C": {
  "statut": "Correct",
  "dc": "Faible",
  "structure": {
   "intro": "Bonne réponse, mais votre faible certitude (DC {DC}/5) révèle une hésitation normale à ce stade.",
   "explication": "L'apprentissage supervisé repose sur la relation entrée → sortie...",
   "exemple": [
    "Supervisé : prédire si un client remboursera un prêt.",
    "Non supervisé : regrouper des clients sans label."
   ],
   "points_cles": [
    "Supervisé = features + label.",
    "Classification (catégorie) vs régression (valeur continue).",
    "Schéma : préparation → entraînement → validation → ajustement."
   ],
   "encouragement": "Le fait que vous ayez trouvé la bonne réponse, malgré vos doutes, montre une intuition juste..."
  }
 },
 "Cas_D": {
  "statut": "Incorrect",
  "dc": "Élevé",
  "structure": {
   "alerte": "Erreur confiante ! Vous avez choisi {STUDENT_ANSWER} avec DC {DC}/5, ce qui signale une fausse représentation.",
   "correction": "La proposition correcte était {CORRECT_ANSWER}. Contrairement au non supervisé...",
   "conseil": "Repérez systématiquement si l'énoncé mentionne une 'cible' à prédire.",
   "points_cles": [
    "Non supervisé = découverte de structures sans label.",
    "Supervisé = label connu + prédiction.",
    "Exercice : appliquer la règle 'y a-t-il un label ?' sur des QCM."
   ],
   "recommandation": "Créez une fiche de décision rapide : cible ou non ?",
   "encouragement": "Le fait d'avoir répondu avec assurance signifie que vous investissez activement dans votre raisonnement..."
  }
 },
 "Cas_E": {
```

```
    "statut": "Incorrect",
    "dc": "Faible",
    "structure": {
     "feedback": "Votre hésitation (DC {DC}/5) révèle un manque de repères stables sur le concept.",
     "correction": "La bonne réponse était {CORRECT_ANSWER}. Il est nécessaire de repartir des distinctions de base...",
     "exemple": [
      "Supervisé : prédire si un patient est malade à partir de symptômes connus.",
      "Non supervisé : regrouper des patients selon leurs profils."
     ],
     "points_cles": [
      "Non supervisé = pas de cible → exploration.",
      "Supervisé = cible connue → prédiction.",
      "Exercice : synthétiser ces définitions dans une fiche visuelle."
     ],
     "recommandation": "Reprenez les définitions fondamentales et construisez une fiche d'ancrage conceptuel.",
     "encouragement": "Votre prudence est utile : elle montre que vous ne répondez pas au hasard avec excès de confiance."
    }
   },
   "Cas_F": {
    "statut": "Incorrect",
    "dc": "Moyen",
    "structure": {
     "feedback": "Votre réponse traduit un doute réel et une confusion entre des notions proches.",
     "correction": "La bonne réponse était {CORRECT_ANSWER}. Il faut clarifier ici la différence entre les dimensions du concept...",
     "exemple": [
      "Classification supervisée : prédire si un client va acheter ou non un produit.",
      "Régression supervisée : estimer le revenu annuel d'un client."
     ],
     "points_cles": [
      "Supervisé = classification + régression.",
      "Vérifier si l'énoncé inclut une variable cible.",
      "Mnémotechnique : Supervisé ? Cherche la cible. Puis classe ou calcule."
     ],
     "recommandation": "Posez-vous systématiquement la question : catégorie ou valeur numérique ?",
     "encouragement": "Votre hésitation est cohérente avec une compréhension encore partielle, ce qui constitue une bonne base de progression."
    }
   },
 "Global Synthesis": {
   "indicateurs": {
    "TRC": "{TRC}",
    "REAL": "{REAL}",
    "SUR_EST": "{SUR_EST}",
    "SOU_EST": "{SOU_EST}",
    "CENTR": "{CENTR}"
   },
   "structure": {
    "title": "Synthèse Global métacognitive",
    "language": "French",
    "format_rules": [
     "Write the synthesis in clear academic French.",
     "For each indicator, use exactly the following format: indicator label, bullet Value, bullet Interpretation.",
     "After the indicator blocks, add a section titled 'Cross-indicator profile and suggestions'.",
     "Interpret the values pedagogically and supportively.",
     "Connect mastery and calibration indicators in the cross-indicator profile.",
     "End with two complementary lines of work and one concluding synthesis sentence."
    ],
    "indicator_blocks": [
     {
      "indicator": "TRC",
      "label": "TRC – Mastery of the tested concepts",
      "value": "• Value : {TRC} %",
      "interpretation": "• Interpretation: [Interpret the student's level of mastery of the tested concepts. Explain whether the score suggests limited, partial, or solid mastery.]"
     },
     {
      "indicator": "REAL",
      "label": "REAL – Realism of self-evaluation",
      "value": "• Value : {REAL} %",
      "interpretation": "• Interpretation: [Interpret the alignment between the student's confidence and actual performance. Explain whether realism is low, moderate, or high.]"
     },
     {
      "indicator": "SUR_EST",
      "label": "SUR_EST – Overestimation",
      "value": "• Value : {SUR_EST} %",
      "interpretation": "• Interpretation: [Interpret the extent to which the student tends to be confident while being incorrect.]"
     },
     {
      "indicator": "SOU_EST",
      "label": "SOU_EST – Underestimation",
      "value": "• Value : {SOU_EST} %",
      "interpretation": "• Interpretation: [Interpret the extent to which the student tends to doubt correct answers.]"
     },
     {
      "indicator": "CENTR",
      "label": "CENTR – Global bias of confidence",
      "value": "• Value : {CENTR} %",
      "interpretation": "• Interpretation: [Interpret the overall direction of the student's confidence bias. Explain whether the profile tends toward overconfidence, underconfidence, or a balanced confidence profile.]"
     }
    ],
    "cross_indicator_profile": {
     "title": "Cross-indicator profile and suggestions",
     "intro": "Overall, your profile combines:",
     "profile_points": [
      "• [Summarize the level of mastery based on TRC]",
      "• [Summarize the level of realism based on REAL]",
      "• [Summarize the tendency toward overestimation and/or confidence bias based on SUR_EST and CENTR]",
      "• [Summarize the degree of underestimation based on SOU_EST]"
     ],
     "global_interpretation": "This means that [provide an integrated interpretation of the student's metacognitive profile by connecting mastery, realism, overestimation, underestimation, and confidence bias].",
     "recommendations_intro": "To progress, two complementary lines of work are recommended :",
     "recommendation_1": {
      "title": "1. Strengthen conceptual mastery.",
      "actions": [
       "o [Recommend revisiting the items answered incorrectly, especially those answered with high DC.]",
       "o [Recommend identifying the student's original reasoning and comparing it with the correct explanation.]"
      ]
     },
     "recommendation_2": {
      "title": "2. Refine your use of the DC scale.",
      "actions": [
       "o [Recommend reserving the highest certainty levels for answers that can be clearly justified.]",
       "o [Recommend using intermediate DC levels when hesitating and using subsequent feedback to recalibrate judgment.]"
      ]
     },
     "closing": "By combining targeted review of confident errors with a more cautious and calibrated use of high DC values, you can progressively improve both your mastery and your metacognitive realism, while reducing overestimation and confidence bias."
    }
   }
}
```

3) Deterministic assignment of the pedagogical case

The pedagogical case is assigned deterministically in code, before the LLM is called. The model does not infer the case freely. Instead, the workflow first computes the case from correctness status and DC value, then injects the corresponding case label and case-specific scaffold into the

prompt.

The rules are:

- Correct + DC ≥ 4 → Cas_A
- Correct + DC = 2 or 3 → Cas_B
- Correct + DC ≤ 1 → Cas_C
- Incorrect + DC ≥ 4 → Cas_D
- Incorrect + DC = 2 or 3 → Cas_F
- Incorrect + DC ≤ 1 → Cas_E

This deterministic assignment enhances reproducibility by ensuring that the pedagogical logic is stable and not left to stochastic model interpretation.

4) Full prompt structure used for item-level feedback generation

For each student and each item, the system assembles a prompt containing:

- the question text,
- the answer options,
- the student's selected answer,
- the correct answer,
- the correctness status,
- the Degree of Certainty (DC),
- the detected pedagogical case,
- the corresponding case-specific template section,
- the instructor-approved course documentation excerpt,
- and optional student indicators when this mode is enabled.

A generic reconstruction of the item-level prompt is given below:

```
You are an expert educational feedback assistant specialized in formative assessment,
MCQ pedagogy, metacognition, and confidence-based evaluation.
Your task is to generate ONE personalized feedback for ONE student answer to ONE MCQ.
Pedagogical case: {CASE}
Question:
{QUESTION}
Options:
{OPTIONS}
Student answer:
{STUDENT_ANSWER}
Correct answer:
{CORRECT_ANSWER}
Status:
{STATUS}
Degree of certainty:
{DC}/5
Course documentation:
{COURSE_EXCERPT}
Canonical case structure:
{CASE_TEMPLATE}
Optional indicators:
{INDICATORS}
Instructions:
- Strictly follow the pedagogical logic of the assigned case.
- Rewrite the template into a coherent and natural feedback.
- Keep an academic, supportive, and actionable tone.
- If the answer is incorrect, explain the conceptual error clearly.
- If relevant, explain why the selected distractor may appear plausible.
- Do not label an incorrect answer as partially correct.
- End with key takeaways.
```

5) Prompt used for the global metacognitive synthesis

A separate prompt is used to generate the report-level synthesis when indicators are included. Its purpose is to transform numerical indicators into an interpretable metacognitive profile and action-oriented recommendations.

A generic reconstruction is shown below:

```
You are an expert in metacognitive feedback for higher education students.
Your task is to write a global metacognitive synthesis in English based on the indicators provided below.
The output must be written as a structured report for one student, in a clear, academically rigorous, supportive, and actionable tone.

Indicators:
- TRC = {TRC} %
- REAL = {REAL} %
- SUR_EST = {SUR_EST} %
- SOU_EST = {SOU_EST} %
- CENTR = {CENTR} %
Required output title:
Global metacognitive synthesis
Required output structure:
TRC—Mastery of the tested concepts
• Value: {TRC} %
• Interpretation: [Interpret the student's level of mastery of the tested concepts.]

REAL—Realism of self-evaluation
• Value : {REAL} %
• Interpretation: [Interpret the alignment between confidence and actual performance.]

SUR_EST—Overestimation
• Value : {SUR_EST} %
• Interpretation: [Interpret the extent to which the student tends to be confident while being incorrect.]

SOU_EST—Underestimation
• Value : {SOU_EST} %
• Interpretation: [Interpret the extent to which the student tends to doubt correct answers.]

CENTR—Global bias of confidence
• Value : {CENTR} %
• Interpretation: [Interpret the overall direction of the student's confidence bias.]

Cross-indicator profile and suggestions
Overall, your profile combines:
• [Summarize the level of mastery based on TRC]
• [Summarize the level of realism based on REAL]
• [Summarize the tendency toward overestimation and/or confidence bias based on SUR_EST and CENTR]
• [Summarize the degree of underestimation based on SOU_EST]

This means that [provide an integrated interpretation of the student's metacognitive profile by connecting mastery, realism, overestimation, underestimation, and confidence bias].

To progress, two complementary lines of work are recommended:
1. Strengthen conceptual mastery.
o [Recommend revisiting the items answered incorrectly, especially those answered with high DC.]
o [Recommend identifying the student's original reasoning and comparing it with the correct explanation.]

2. Refine your use of the DC scale.
o [Recommend reserving the highest certainty levels for answers that can be clearly justified.]
o [Recommend using intermediate DC levels when hesitating and using subsequent feedback to recalibrate judgment.]

By combining targeted review of confident errors with a more cautious and calibrated use of high DC values, you can progressively improve both your mastery and your metacognitive realism, while reducing overestimation and confidence bias.

Instructions:
- Write in French only.
- Use exactly the section titles given above.
- For each indicator, use exactly this format: section title, then bullet "Value", then bullet "Interpretation".
- Interpret the values pedagogically, not just descriptively.
- Use supportive but precise wording.
- Connect mastery and calibration indicators in the cross-indicator profile.
- Keep the synthesis coherent, natural, and adapted to a Master's student.
- Do not output JSON.
- Do not add extra sections.
```

- Do not mention these instructions in the output.

C. Exact Model Version and Inference Parameters

1) Model version

The configured model used in the workflow was:

- Model string: GPT-4

2) Inference parameters for item-level feedback generation

The following parameters were used for per-item feedback generation:

- model: GPT-4
- temperature: between 0 and 0.3
- max_tokens: 750
- messages:
  - system role: pedagogical expert / university tutor instruction
  - user role: assembled item-level prompt described in Section B1.5

The low-temperature configuration was selected to favor stable, structure-respecting, and pedagogically consistent outputs rather than creative variation.

3) Inference parameters for global synthesis generation

The following parameters were used for global synthesis generation:

- model: GPT-4
- temperature: between 0 and 0.3
- max_tokens: 700
- messages:
  - system role: metacognitive feedback expert instruction
  - user role: indicator list + synthesis generation instruction

This configuration aimed to produce concise, stable, and interpretable report-level syntheses.

4) Parameters left at defaults

Unless explicitly configured otherwise, the following parameters were left at API defaults:

- top_p: default
- presence_penalty: default
- frequency_penalty: default
- seed: not used
- n: single completion (1)

D. Grounding and Document Selection

1) Input sources used for grounding

The workflow uses instructor-provided materials as grounding sources. Instructors upload:

- the MCQ file (Word), containing question text, options, and answer key;
- the student response file (Excel), containing selected options and DC values;
- pedagogical documents (Word or PDF), used as the grounding basis;
- an indicator file containing student-level performance and metacognitive indicators.

2) Text extraction procedure

Text is extracted from uploaded pedagogical documents as follows:

- Word documents: all non-empty paragraphs are concatenated;
- PDF documents: text is extracted page by page and concatenated.

This produces a unified textual grounding base that can be injected into prompts.

3) Document selection and injection strategy

The implementation used in the study relies on a concatenation-based grounding strategy, not on retrieval-based document selection. All instructor-approved course documents are converted to text and merged into a single grounding string.

This grounding text is then injected into the item-level prompt as a course documentation excerpt. In practice, the injected excerpt is truncated to the initial portion of the concatenated documentation so as to control prompt length.

This design prioritizes:

- simplicity of implementation,
- instructor control over the pedagogical corpus,
- and consistent use of instructor-approved materials.

4) Limitation of the grounding procedure

Because the system relies on concatenation and truncation rather than item-specific retrieval, not every prompt necessarily contains the most directly relevant passage for the target concept. As a result, some outputs may be broader or more generic than those produced by a retrieval-augmented design.

This limitation is mitigated by instructor review before release.

E. Instructor Review and Quantitative Reporting of Edits

1) Human-in-the-loop review protocol

After generation, the instructor reviews the feedback before it is made available to students. This review checks:

- compliance with the expected pedagogical case structure,
- consistency between student answer, correctness status, and DC value,
- alignment with instructor-approved course content,
- coherence of displayed indicators when indicators are enabled,
- clarity, usefulness, and academic tone of the feedback.

After review, the report may be:

- approved as generated,
- edited locally by the instructor,
- then re-uploaded and stored for student access.

2) Stabilization before deployment

Before use, repeated feasibility and stabilization testing was conducted. These tests aimed to verify:

- correct extraction of item and response data,
- correct mapping of answer choice and DC into the report,
- deterministic assignment of pedagogical cases A–F,
- structural compliance of outputs with the canonical template,
- consistency of the synthesis section when indicators were enabled.

This phase informed several code adjustments and repeated internal checks before deployment in authentic classroom conditions.

3) Logging limitation

The platform did not systematically archive paired "raw generated draft" and "final released version" logs for every report. Therefore, a fully automated retrospective computation of edit rates across the complete dataset cannot be reconstructed from system traces alone.

This limitation should be stated explicitly for transparency.

4) Quantitative reporting of review and edits

The workflow supports quantitative reporting at two levels.

Review coverage by assessment

- For the first assessment, the full cohort was reviewed to validate the end-to-end pipeline under authentic conditions.
  - After stabilization, subsequent assessments were subjected to instructor spot-checking before release, generally covering approximately 20–30% of reports per assessment.

Audit-based intervention statistics

- In the absence of exhaustive draft-to-final logs, edit statistics can be established through a post hoc audit of a defined subset of reviewed reports from the stabilization and deployment period.
- For this subset, each report is coded as:
  - validated without modification, or
  - edited before release.

5) Edit taxonomy

Instructor edits are coded into one or more of the following categories:

1. **Content accuracy / course alignment**
   conceptual correction or improved alignment with course wording.
2. **Template compliance**
   missing, incomplete, or misplaced sections relative to the expected case structure.
3. **Input consistency correction**
   mismatch between selected option, correctness status, DC value, and generated feedback.
4. **Indicator consistency correction**
   mismatch between displayed indicators and uploaded values.
5. **Clarity / wording**
   rephrasing for readability, precision, or fluency.
6. **Tone / hedging**
   adjustment of assertiveness, supportiveness, or academic tone.

Because one report may contain more than one type of modification, category frequencies are not expected to sum to 100%.

F. Reproducibility Statement

The GenAI-assisted feedback workflow used in this study is not based on free-form prompting alone. It combines:

- deterministic assignment of one of six pedagogical cases,
- a strict canonical JSON template,
- case-specific prompt construction,
- grounding in instructor-approved course documents,
- optional integration of student-level performance and metacognitive indicators,
- and instructor validation prior to release.

This combination explains the structural regularity of the generated feedback, both at the item level and in the global metacognitive synthesis. It also provides a clearer basis for replication than a simple model-and-prompt description, because the workflow depends not only on the language model, but also on the pedagogical scaffolding, the deterministic case logic, the grounding documents, and the review procedure.

APPENDIX C. EXAMPLE OF A GENAI-GENERATED FEEDBACK REPORT (EXCERPT)

The excerpt below is adapted from an actual feedback report produced for one student in the study. It illustrates (a) item-level feedback combining content and metacognitive comments, and (b) a global synthesis based on DC-derived indicators.

A. Example of Item-level Feedback

Question 2. Which activation function is best avoided when training a recurrent neural network?

Options:

A. Softmax
B. Sigmoid
C. Hyperbolic tangent (tanh)
D. ReLU

Your answer: A
Status: Incorrect
Degree of Certainty (DC): 4/5

Feedback (excerpt)

Response feedback

You selected *Softmax* with a high level of certainty (DC 4/5). This choice reveals a confusion between the role of activation functions inside the network and the role of the output function.

The correct answer was: **D—ReLU**.

In recurrent neural networks, ReLU is generally discouraged as the activation function inside recurrent cells, because it can lead to zero gradients ("dying ReLU") and block learning for certain sequences.

Key concept

In RNNs (and their variants LSTM, GRU), it is generally preferable to use bounded activation functions such as tanh or sigmoid, which help stabilize backpropagation through time.

- *Softmax* is usually used in the output layer for classification.
- *ReLU* is effective in feedforward networks, but more problematic in sequential architectures for modeling internal dynamics.

Metacognitive reflection

Your high degree of certainty shows that you were confident in your reasoning, even though the answer was incorrect. This is a typical case of **overestimation**: a mismatch between expressed confidence and actual performance. Identifying this kind of confident error is important for adjusting your learning strategies.

Advice for improvement

I recommend creating a comparative table of activation functions (**ReLU, tanh, sigmoid, Softmax**) and, for each one, specifying:

- its main role (internal layer vs. output layer),
- its behavior with respect to the gradient,

- the architectures where it is most appropriate (MLP, CNN, RNN/LSTM/GRU).

Referring back to this table when solving exercises will help you choose the appropriate activation function more systematically.

Key points to remember

- In RNNs, tanh or sigmoid are usually preferred for internal states.
- ReLU should generally be avoided in recurrent cells because of the risk of zero gradients.
- Softmax is reserved for the output layer in classification tasks.

B. Global Metacognitive Synthesis (Indicators Excerpt)

TRC—Mastery of the tested concepts

- **Value :** 40.00 %
- **Interpretation:** This score indicates limited mastery of the concepts assessed in this test. A substantial proportion of items were answered incorrectly, which suggests that several core notions still need to be consolidated.

REAL—Realism of self-evaluation

- **Value :** 65.38 %
- **Interpretation:** This level points to a **moderate realism**: in many cases your confidence is aligned with your actual performance, but important discrepancies remain, especially on items where you were confident and wrong.

SUR_EST—Overestimation

- **Value:** 34.50 %
- **Interpretation:** This value is **high**. It indicates frequent situations where you feel sure of your answer while it is in fact incorrect. This pattern makes it harder to detect your own misconceptions and can delay remediation.

SOU_EST—Underestimation

- **Value:** 0.12 %
- **Interpretation:** This value is **very low**, which is positive: you rarely doubt your correct answers. In other words, you almost never underestimate your knowledge when you are actually right.

CENTR—Global bias of confidence

- **Value:** 34.38%
- **Interpretation:** This index summarizes the **overall direction** of your bias and confirms a strong tendency toward **over-confidence**. Your confidence is globally "over-centered" relative to your actual performance.

Cross-indicator profile and suggestions

Overall, your profile combines:

- **Low mastery** (TRC = 40 %) on this test,
- **Moderate realism** (REAL ≈ 65 %),
- **Strong overestimation and over-centering** (SUR_EST and CENTR high),
- **Almost no underestimation** (SOU_EST ≈ 0 %).

This means that you *trust yourself a lot* on many items, including when your answers are incorrect. To progress, two complementary lines of work are recommended :

1. **Strengthen conceptual mastery.**
   - Revisit the items you got wrong, especially those answered with high DC.
   - For each one, identify precisely what you had in mind and compare it with the correct explanation.
2. **Refine your use of the DC scale.**
   - Reserve the highest certainty levels (4–5) for cases where you can justify your answer with a clear argument or example.
   - When you hesitate between two options, deliberately choose an intermediate DC (2–3) and use the subsequent feedback to recalibrate your judgment.

By combining targeted review of confident errors with a more cautious use of high DC values, you can progressively increase both your TRC and your REAL, while reducing SUR_EST and CENTR toward a more balanced and accurate confidence profile.

APPENDIX D: STUDENT PERCEPTION QUESTIONNAIRE: FULL ITEM WORDING AND RESPONSE SCALE

This appendix reports the full wording of the student perception questionnaire used to assess learners' perceptions of the AI-generated feedback. The instrument comprised six subscales: usefulness (U), clarity/readability/actionability (C), trust (T), intention to use / reuse (I), cognitive load / comfort (L), and self-efficacy (AE). Most constructs were measured with four items; self-efficacy was measured with two items. Table A1 provides the complete list of questionnaire items, organized by subscale and item code, in order to clarify how each dimension of students' perceptions was operationalized.

Table A1. Full wording of the student perception questionnaire items, organized by subscale and item code

Subscale	Code	Full item wording
Usefulness	U1	The feedback I received helped me better understand my mistakes.
Usefulness	U2	The feedback is relevant for improving my learning.
Usefulness	U3	The feedback provided information that I can use directly.
Usefulness	U4	I consider this feedback important for my progress.
Clarity / Readability / Actionability	C1	The feedback was written in a clear and understandable way.
Clarity / Readability / Actionability	C2	The feedback was sufficiently detailed and specific.
Clarity / Readability / Actionability	C3	The feedback indicated concrete actions I could take to improve.
Clarity / Readability / Actionability	C4	Thanks to the feedback, I know exactly what I need to correct or improve.
Trust	T1	I trust the reliability of the feedback I received.
Trust	T2	The feedback corresponds to my actual performance.
Trust	T3	The feedback accurately reflects my strengths and weaknesses.
Trust	T4	The feedback seemed fair and credible to me.
Intention to use / reuse	I1	I reread the feedback carefully.
Intention to use / reuse	I2	I applied at least one recommendation from the feedback.
Intention to use / reuse	I3	The feedback encouraged me to correct my mistakes or practice more.
Intention to use / reuse	I4	The feedback motivated me to adjust my learning strategy.
Cognitive load / Comfort	L1	The feedback was easy to understand.
Cognitive load / Comfort	L2	The feedback did not seem too long or too complex.
Cognitive load / Comfort	L3	The feedback did not discourage or stress me.
Cognitive load / Comfort	L4	Reading this feedback was a positive experience.
Self-efficacy	AE1	After reading the feedback, I feel more capable of avoiding this type of mistake in the future.

Self-efficacy	AE2	The feedback strengthened my confidence in my ability to succeed.

Note: U = usefulness; C = clarity/readability/actionability; T = trust; I = intention to use / reuse; L = cognitive load / comfort; AE = self-efficacy.

Response scale. All items were rated on a 5-point Likert scale: 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; 5 = strongly agree.

APPENDIX E. FEATURE COMPARISON WITH REPRESENTATIVE CBM/DC AND LLM-BASED FEEDBACK SYSTEMS

Appendix E provides a concise feature comparison between Evalviz-Dashboard and representative lines of work in (i) certainty-based assessment (CBM/DC) and (ii) recent LLM-based feedback assistants. The goal is not an exhaustive review, but a structured positioning along design-relevant dimensions for this manuscript: whether calibration is explicitly modeled, what inputs the system uses, how feedback is generated and constrained, how grounding is handled, what level of instructor oversight is involved, and whether longitudinal indicators are supported. Table A2 summarizes this comparison across selected systems and lines of work, highlighting the distinctive combination of certainty-based calibration, course-grounded GenAI feedback, instructor validation, and longitudinal dashboards implemented in Evalviz-Dashboard.

TABLE A2. Feature comparison of Evalviz-Dashboard with certainty-based assessment and LLM-based feedback systems

System / Line of work	Primary input	GenAI/LLM feedback generation?	Calibration / DC explicitly modeled ?	Feedback target	Output constraints	Grounding	Human oversight	Longitudinal indicator/dashboard
MOHICAN "check-up" edumetric + metacognitive indices [23]	MCQ responses + certainty/ confidence information (CBM/DC-style)	**No** (non-GenAI; analytic framework)	Yes (confidence–accuracy indicators; metacognitive indices)	Diagnostic measurement and interpretation of knowledge/calibration; may inform formative remediation	Rule-based scoring and index computation	Test/course content used in the check-up	Instructor-led interpretation	Indices can be computed; not necessarily translated into structured narrative feedback
DOCIMO / CGQTS-based standardized assessment workflow [35]	Test blueprint / table of specifications, item bank, student responses, answer key, and degree of certainty (DC)	**No**	**Yes**	Response accuracy, score interpretation, and certainty-aware assessment feedback	Rule-based, structured feedback/ output; not LLM-generated	Grounded in the test blueprint, item bank, predefined scoring rules, and student response + DC data	High	Assessment results/ statistics available
LEAP (LLM-based formative assistant) [38]	Student work + teacher-designed prompts	**Yes**	**No** (not a quantified calibration target)	SRL-oriented scaffolds; content/process guidance	Prompt-structured, but output varies	Instructor-provided materials / prompt context	Emphasizes human-in-the-loop design	Not a core focus
FreeText (LLM feedback on open-ended responses) [19]	Open-ended student response + rubric/criteria	**Yes**	**No**	Rubric-aligned formative feedback	Rubric/criteria-guided; structured feedback components	Instructor-specified criteria (held-out rubric)	Instructor controls rubric/criteria; oversight possible	Not a core focus
Dean of LLM Tutors (LLM feedback evaluation/ filtering) [20]	Candidate LLM feedback outputs	**Yes**	**No**	Quality assurance of LLM-generated feedback	Automated evaluation/ filtering across dimensions	Depends on upstream system	Automated "dean" evaluator (and possible human review)	Not a core focus
Evalviz-Dashboard (this work)	MCQ + **DC (0–5)** + indicator profile + course docs	**Yes**	**Yes** (TRC, Cmoy, REAL, CENTR, SUR_EST, SOU_EST…)	**Calibration-aware formative feedback** (item-level + global synthesis)	**Canonical JSON template (A–F + synthesis)**	**Course-grounded** (instructor-approved documents)	Instructor validation before release	**Yes** (trajectories, cohort distributions; indicators reused in feedback)