Feedback on learning with generative artificial intelligence in university students

Retroalimentación de aprendizajes con inteligencia artificial generativa en estudiantes universitarios

María Verónica Leiva-Guerrero - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*
Ignacio Araya Zamorano - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*
Rafael Escobar Collins - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*
Francisca Silva Castro - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*

ABSTRACT

Assessment for learning has become increasingly important in university teaching, particularly regarding the feedback process. However, there is still a perception of student dissatisfaction with the quality of feedback provided by faculty, highlighting the need to innovate in feedback strategies. This study aimed to explore the pedagogical and technological relevance of integrating Wilson's Feedback Ladder with generative artificial intelligence, specifically GPT-4o, to strengthen formative feedback in university students. The study was conducted using a qualitative and exploratory approach in two phases. First, a prompt was designed and validated using the Delphi method with the participation of eight experts in assessment and artificial intelligence, applying it to seven state-of-the-art language models. In the second phase, the validated prompt was implemented in two university courses of different nature, Assessment for Learning and Data Structures, integrating automatic feedback into the Moodle platform. The results showed that the experts agreed on the suitability of AI-mediated Wilson's Ladder and highlighted the superior performance of GPT-4o. At the classroom level, students valued the clarity, usefulness, and immediacy of the feedback, although they identified limitations in the tool's lack of contextualization and impersonal tone. It is concluded that the integration of Wilson's Ladder with generative artificial intelligence represents a promising innovation, but one that requires disciplinary adjustments, teacher supervision, and careful attention to the human dimension of feedback in e-learning contexts.

Keywords: feedback; formative evaluation; generative artificial intelligence; ChatGPT; Ladder of Feedback; university students.

RESUMEN

La evaluación para el aprendizaje ha adquirido creciente relevancia en la docencia universitaria, especialmente el proceso de retroalimentación. Sin embargo, persiste una percepción de insatisfacción en estudiantes sobre la calidad de retroalimentación brindada por el profesorado, lo que evidencia la necesidad de innovar en sus estrategias. Esta investigación exploró la pertinencia pedagógica y tecnológica de integrar la Escalera de Retroalimentación de Wilson con inteligencia artificial generativa, específicamente GPT-4o, para fortalecer la retroalimentación formativa en estudiantes universitarios. El estudio se desarrolló bajo un enfoque cualitativo y exploratorio en dos fases. En primer lugar, se diseñó y validó un prompt mediante el método Delphi con la participación de ocho expertos en evaluación e inteligencia artificial, aplicándolo a siete modelos de lenguaje de última generación. En la segunda fase, el prompt validado se implementó en dos cursos universitarios de distinta naturaleza, Evaluación para el Aprendizaje y Estructura de Datos, integrando la retroalimentación automática en la plataforma Moodle. Los resultados mostraron que los expertos coincidieron en la idoneidad de la Escalera de Wilson mediada por IA y destacaron el desempeño superior de GPT-4o. A nivel de aula, los estudiantes valoraron la claridad, utilidad e inmediatez de la retroalimentación, aunque identificaron limitaciones en la falta de contextualización y tono impersonal de la herramienta. Se concluye que la integración de la Escalera de Wilson con inteligencia artificial generativa representa una innovación prometedora, pero que requiere ajustes disciplinares, supervisión docente y resguardo de la dimensión humana en los procesos de retroalimentación en entornos abiertos y a distancia.

Palabras clave: retroalimentación; evaluación formativa; inteligencia artificial generativa; ChatGPT; Escalera de Wilson; estudiantes universitarios.

How to cite: Leiva-Guerrero, M. V., Araya Zamorano, I., Escobar Collins, R., & Silva Castro, F. (2026). Feedback on learning with generative artificial intelligence in university students [Retroalimentación de aprendizajes con inteligencia artificial generativa en estudiantes universitarios]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), 241-265. https://doi.org/10.5944/ried.45547

INTRODUCTION

Assessment for learning with a formative approach has gained increasing relevance at a global level, both in school and university contexts, by centering on continuous support and improvement of the educational process (Black & Wiliam, 2018; Andrade & Brookhart, 2019). Within this framework, feedback is situated as a key component, as it uses evidence of student performance to guide learning improvement (Carless & Boud, 2018).

However, various studies report a persistent dissatisfaction of university students regarding the quality of feedback provided by faculty, especially related to immediacy, specificity and utility (Quezada & Salinas, 2021; Galindo-Domínguez et al., 2023). This situation has been evidenced in research carried out in countries such as Australia, United Kingdom, Mexico and Chile, which suggests this is an issue with international scope.

In several educational systems, feedback practices have not had a significant evolution, which limits its formative impact. Strategies such as Wilson's Feedback Ladder (2013) or the Feedback Panel (Booth et al., 2008) have been proposed as effective alternatives to improve assessment processes, although its incorporation in teaching practices is still limited.

In the face of this scenario, it becomes necessary to renew university assessment strategies, incorporating emerging digital technologies that can support formative feedback (Puertas & Cano, 2024). In particular, Generative Artificial Intelligence (IAGen, for its name in Spanish) offers relevant opportunities to transform feedback in writing and complex thinking tasks (Dai et al., 2023; Al-Azawei et al., 2023).

Among the tools based on this technology we can mention ChatGPT (OpenAI, 2024), which integrates models such as GPT-4o, GPT-4.1 or GPT-4-5 and generates coherent and contextualized texts from prompts designed with pedagogical criteria. Although its incorporation to the educational field is recent, preliminary studies show an increasing use in teaching, learning and assessment, which opens new research lines about its applicability and efficacy (Galindo-Domínguez et al., 2023; García-Peñalvo, 2023).

Within this framework, the present study proposes three main questions: Is Wilson's Feedback Ladder adequate for its application through artificial intelligence? What IAGen tool, programmed with prompts according to these criteria, turns out to be more pertinent to provide feedback in university learning? And how do students perceive feedback received through this strategy applied with GPT-4o?

Artificial Intelligence in education

The integration of artificial Intelligence (AI) in the educational field has redefined the traditional teaching framework, becoming a fundamental core of contemporary pedagogical innovation. The role of the teacher, historically centered on content transmission, has experienced a substantial transition towards a role of learning facilitator, thanks to the potentialities of AI to automatize processes, analyze data and customize formative trajectories (Bonales-Daimiel et al., 2025). This reconfiguration allows educators to focus on tasks with greater pedagogical value such as qualitative feedback and differentiated support (García Peñalvo et al., 2024).

In addition, AI has demonstrated to be a strategic ally for the improvement of the educational experience, both in face-to-face and virtual contexts, through the

personalization of resources, automatized evaluation, immediate feedback and real-time monitoring of student performance (López Regalado et al., 2024; Romero Alonso et al., 2025). The implementation of adaptative systems and educational chatbots has allowed not only to optimize teacher efficiency, but also enrich learning processes with student-centered approaches.

Generative Artificial Intelligence

Generative Artificial Intelligence (IAGen) covers techniques and models designed to create new content (text, images, audio, video) that imitates the statistic distribution of large data groups. At the core of many IAGen solutions are the transformers, architectures of attention that, after massive pre-training, learn to preview the net unit (word, pixel, notation) and then generate coherent and contextualized outputs (Vaswani et al., 2017; OpenAI, 2023).

Thanks to fine-tuned algorithms and the incorporation of multiple input and output modalities, these systems can adapt to various tasks – from creative writing to conversational simulations – maintaining fluidity and sense in its answers (Weng et al., 2024).

Since 2024, the LLM ecosystem has diversified with business proposals - GPT-4o y GPT-4.5 (OpenAI, 2024), Claude 3.7 Sonnet (Anthropic, 2024) and Gemini 2.5 (Google DeepMind, 2024)— and from open code —LlaMa 4 (LlaMa, 2024), Mistral (Mistral AI, 2024).

In addition, some models such as DeepSeek R1 (DeepSeek, 2024), stand out by explaining their internal reasoning in the face of complex problems, and xAI Grok 3 (xAI, 2024) dedicates deep "reflection" phases for advanced mathematics and logic.

In education, these tools not only are useful to automatize content, but they help to customize teaching and feedback, adapting to each student and relieving repetitive tasks (work corrections), as long as ethical and methodological aspects are considered. Its responsible integration can significantly strengthen learning (Weng et al., 2024).

Prompts and prompting techniques

In the context of generative artificial intelligence, a prompt is the textual or structure input that is given to a language model to guide their answer. Its design is not trivial: the form, content and format of prompt directly impact quality, precision and utility of the generated output. As a result, the prompt engineering field emerges, dedicated to the creation of specific strategies to build effective prompts according to the task or the educational context (Liu et al., 2023; Reynolds & McDonell, 2021).

Between the main prompting techniques, we can find zero-shot prompting, where the model responds only to a direct instruction with no examples; and few-shot prompting, which incorporates representative examples before the task in order to model the type of expected answer (Brown et al., 2020). These techniques allow the modulation of the model's behavior and better anticipate the type of content that will be created.

Another important technique is chain-of-thought, which guides the model to reason, step by step before offering a final answer, improving its performance in complex logic or structured assessment (Wei et al., 2023). In addition, there is role prompting, that assigns a role explicitly to the model – like an "expert evaluator" or "university professor" – to control the tone, depth and style of feedback (Sahoo et al.,

2024). Finally, structured prompting specifies the required output format, such as a dictionary or a table, which is especially useful when the information generated must be automatically processed by external systems, such as educational platforms (Schulhoff et al., 2025).

In education, prompting techniques allow automatized feedback to be structured in an accurate and pedagogically aligned way. The combined use of role prompting, chain-of-thought and structure prompting guides the model to reason as an expert evaluator, generating organized and coherent responses. This facilitates the integration to educational platforms and fosters a more useful, traceable and learning-centered feedback.

Formative feedback and its mediation with artificial intelligence

Feedback is a pillar of formative learning, conceived not only as error correction, but as an interactive process that promotes reflection, adjusting strategies and active building of knowledge. Winstone et al. (2022) describes it as a dialogic interaction between teacher and student, or between peers, that favors a deep understanding, as long as it is timely, specific and with clear goals. Along the same line, Carless and Boud (2018) underscore that it must be understood as a process of mutual influence and student agency, While Molloy et al. (2020) highlight the need for students to develop competences to interpret and apply it in autonomous contexts.

However, in practice, there are some structural barriers. In Chile, the Ministry of Education (2019) notes that, despite the theoretical recognition of feedback, evaluative practices continue to be mainly unidirectional and corrective, with little emphasis on metacognition and self-regulation. To address this gap, structured approaches are required to strengthen the formative dimension. Within this framework, Wilson's Feedback Ladder (Goodrich, 2011) offers a clear guide with four progressive levels (Clarify: Make questions that point towards solving ambiguities or aspects omitted in the student's work, promoting conceptual understanding; Value: highlight achievements and progress, acknowledging positive aspects of performance in an authentic way; State concerns: identify observed difficulties with a respectful and constructive tone, avoiding judgmental remarks and Suggesting improvements: offering specific orientations that will guide the student towards a deeper understanding and viable improvement actions) aimed at critical thinking and self-regulation. Its value lies on transforming feedback in an empathetic and participative experience that can adapt to different educational contexts, whether they are online or face-to-face.

In this study, Wilson's Ladder is adopted as a conceptual basis to design automatized feedback mediated by artificial intelligence. From this perspective, AI is proposed as a resource to customize and broaden coverage, in line with Wiliam (2011) and Shute and Rahimi (2017) regarding the potential of emerging technologies in evaluation for learning.

Artificial Intelligence for educational feedback

The incorporation of AI in educational feedback processes responds to the need to transform traditional practices into more interactive, personalized and effective models. Various research has proposed to innovate in this dimension through the use of AI, highlighting its potential to improve the quality of learning and reduce teacher

assessment workload (Puertas & Cano, 2024; Ossa & Willatt, 2023). The Teaching with AI – Assessment, Feedback and Personalization report, prepared by the European Commission (2023), delves into how AI can positively intervene in assessment and feedback processes from four complementary levels: social, institutional, teachers and students.

In this context, AI offers tools that are capable of adapting feedback to the individual needs of students through the automatized generation of comments, real-time assessment and academic progress follow-up (Ayeni et al., 2024; Holmes et al., 2019). This type of support allows teachers to identify with greater precision the specific difficulties of each student and offer a more focalized and differentiated teaching (Luckin & Holmes, 2016). Likewise, immediate feedback provided by intelligent systems has a positive impact on student motivation, since it offers timely answers that facilitate error comprehension and strengthen key learnings.

In particular, chat tools based on IAGen models have gained relevance as efficient pedagogical assistants. These systems allow the automatization of correction and feedback from previously established criteria, increasing the efficiency and coherence of the assessment process (Carless & Winstone, 2023). For teachers, this means a reduction in the operational workload; for students, an opportunity to receive immediate formative orientation, promoting self-assessment and critical thinking.

The efficiency of these tools has been documented in various empirical studies. Zhang et al. (2024) evaluated the use of ChatGPT to create feedback in programming tasks with students at an introductory Computer Science course. The majority of participants gave a positive value to the clarity and utility of the AI-generated comments, although some expressed they would have preferred feedback with more specific examples. For their part, Jauhiainen and Garagorry Guerra (2024) analyzed the ChatGPT-4 app in its assessment of open answers written by university students. Using a rubric with five criteria – relevance, accuracy, thoroughness, coherence and linguistic correction - they observed that the model offered detailed and consistent evaluations, although they insisted in the need for teacher supervision to guarantee its final accuracy.

Additionally, Baral et al. (2024) compared ChatGPT-4 with other models (like LlaMa and SBERT-Canberra) in the correction of secondary level mathematics answers. From 500 evaluated answers with a common rubric, it was found that ChatGPT-4 reached 92% of agreement with human teacher grading and also created high-quality feedback, evaluated through linguistic metrics and expert judgement. Nonetheless, the authors note that, although AI can approximate to human criteria, its use must be framed within a system that considers expert review and pedagogical reflection.

METHODOLOGY

Design

The study adopted a qualitative approach of exploratory nature, oriented towards designing, validating and implementing a formative feedback prompt automatized through generative artificial Intelligence (IAGen), supported by Wilson's Feedback Ladder (2013). The process was carried out in two phases.

In Phase 1, a prompt was created from a documental review on feedback strategies, and it was validated through the Delphi method (Landeta, 1999). The

procedure included two successive rounds of expert judgement. The first one evaluated the adaptation of the prompt and the four steps of the Ladder (clarify, value, state concerns, suggest) applied to seven representative language models (LLM). In the second phase, a synthesis of results was returned, and re-judgement was requested for agreement. A Likert 0-3 scale anchored in performance was used, and an interquartile ≤ 1 deviation was used as consensus criteria. In addition, Kendall's W coefficient was calculated as a measure of global consistency (acceptability criteria ≥ 0,70).

In Phase 2, the validated prompt was integrated to the Moodle[1] platform in two university courses. Student open answers were processed automatically with GPT-4o to create immediate feedback structured according to Wilson's Ladder. Finally, student perceptions were collected through semi-structured interviews to supplement technical data with student experience.

Procedure

In the first phase, the goal was to ensure the designed prompt generated structured and coherent feedback with the four levels of Wilson's Ladder (2013) before being applied to students. This process included the language model selection, the iterative prompt creation and validation. Seven representative language models were selected for their technical diversity and accessibility: GPT-4o[2], GPT-4o-mini[3] (OpenAI, 2024), Microsoft Copilot, Google Gemini (2024), Claude (Anthropic, 2024), Perplexity AI and LLaMA (2024), accessed from their oficial channels.

The prompt was designed as a structured template with: (i) theoretical context, (ii) question statement, (iii) evaluation criteria and/or expected response, (iv) student answer and (v) explicit instructions to guide feedback following Wilson's steps. Two activities were used during the pilot assessment: an open question based on a case study and a rubric creation task. The answers for each model were analyzed and then evaluated by a panel of eight experts – three assessment experts, three AI experts and two university professors – using a 0 to 3 scale and qualitative comments, in three rounds according to the Delphi approach (Steurer, 2011).

The purpose of the second phase was to implement the validated prompt in two undergraduate courses: Evaluation for learning and Data Structure. The Assessment (written test) was carried out in Moodle through the Coderunner[4] plugin, automatizing the flow between student answers and GPT-4o. Each student answered open questions that, once sent, were processed in real-time generating immediate feedback, structured according to Wilson's Ladder, along to a preliminary score of 0 to 10 based on rubrics.

Students were informed about the experimental nature of the system and the option of an appeal was offered, reviewed manually by faculty, for the sake of equality. In addition, a qualitative analysis was later carried out through semi-structured interviews to volunteer students, enquiring on clarity, utility and reliability of the automatized feedback. Interviews were analyzed using deductive categories (based on Wilson's steps) and emerging subcategories by Mayring (2000). ChatGPT assisted in the analysis systematization, supporting thematic coding and extraction of significant patterns in student testimonies.

Instruments

The following instruments and techniques were used:

1. *Prompt Validation Protocol*

A protocol was designed to validate the performance of the prompt in seven Language models (LLM). This contained feedback generated by each model, organized according to the four steps in Wilson's Ladder (2013): Clarify, Value, State concerns and Suggest improvements. The expert panel assessed each step with an ordinal scale (0-3) and made qualitative observations when the score was equal or less than 1.

Validation took place through the Delphi method with an interdisciplinary panel of eight experts. The process included two consecutive rounds, using a 0-3 Likert scale and agreement criteria based on an interquartile (≤ 1) deviation and Kendall's W coefficient. This dynamic allowed the adjustment of the instrument before its application, confirming its coherence with Wilson's Ladder steps and supporting the selection of GPT-4o as a preferred model for the pilot phase.

2. *Written Test for Students (Summative Evaluation)*

The applied evaluations in both courses included open questions designed to activate higher-order cognitive abilities. In "Data Structure", explanations, algorithmic structures and technical justifications were addressed. In "Evaluation from and for learning", questions demanded answers with arguments about pedagogical concepts from observed classes. Questions were accompanied by specific rubrics, which allowed the prompt to be applied coherently.

3. *Structured Prompt*

The central instrument of the study was a prompt designed according to prompt engineering principles (Brown et al., 2020; Wei et al., 2023; Hao et al., 2022; Liu et al., 2023; Schulhoff et al., 2025), comprised of six components: (i) Theoretical context: synthesizes the conceptual and methodological foundations that frame the task, allowing to situate the evaluation within a solid disciplinary and pedagogical framework. (ii) Question statement: lays out the evaluative task with clarity, ensuring coherence with the learning objectives. (iii) Expected evaluation/answer criteria establishes the quality parameters and pertinence of answers, described through an analytical rubric with all their performance levels. (iv) Student Answer contains the production to be evaluated. (v) Instructions for feedback guide the formulation of judgements and comments following the steps in Wilson's Ladder (value, clarifications, concerns, suggestions), with extension limits and constructive tone. (vi) Output format in Python dictionary: organizes the results in the following fields: analysis and score by each evaluation criteria, in addition to the four steps in Wilson's Ladder. It was written in Spanish, using a markdown format to favor the legibility and the role of "expert evaluator" was assigned to ensure technical aspects. After its validation, it was decided to include in point (iii) the complete description of all levels of the analytic rubric and add the point (vi) to facilitate automatization of the answer in the Moodle platform. The validated version is attached as an annex for its replicability. It is important to highlight that points (v) and (vi) can stay unmodified, while the previous sections can be adapted to the corresponding evaluation.

4. *Automatized Feedback System*

An integrated system was developed in Moodle through the Coderunner plugin, that connects data with the GPT-4o model. Each time an answer was sent by a student, a structured prompt was generated and synchronically processed, providing immediate feedback. The output was automatically formatted in Moodle along with the preliminary score. In addition, data was stored for its later analysis. The model was put together with a temperature of 0.1, prioritizing more deterministic answers, consistent with the established evaluation rubric.

Technique

Semi-structured interview

With the goal of exploring student perception of the system, an interview was designed with open questions organized around six dimensions: Wilson's four steps, usability of the system and perceived impact of learning. Before its application, each participant signed an informed consent, protecting the principles of confidentiality, anonymity and voluntary nature. Interviews were recorded on audio with student authorization and completely transcribed for its analysis. The guarantee was that, at all times, the information would only be used for academic purposes, in agreement with current institutional ethical regulations.

The analysis was carried out using Mayring's (2000) qualitative content approach, combining deductive categories (derived from the steps in Wilson's Ladder and the complementary dimensions of usability, impact on learning and critical judgement with recommendations) and inductive, resulting from the information itself. Two coders worked independently in the reading, segmentations and coding of interviews, applying the previously agreed matrix. Then, there was a joint reviewing of results in order to solve discrepancies and consolidate definitive coding, which ensured the reliability of the intercoder and provide more strength to the findings.

The categories were enriched with emerging sub-categories based on participants' experiences, making the distinction between favorable assessments (utility, clarity, motivation, formative guide) and problematic ones (rigidity, de-contextualization, automatism, lack of empathy). The process also had the support of ChatGPT as a supplementary methodological tool, used to organize and contrast textual fragments, explore thematic patterns, and verify the coherence of the category system. Its use was limited to facilitating operational tasks, without replacing the analytical judgement of researchers, and contributed to strengthening the traceability and consistency of the process. Table 1 sums up the categories and subcategories used in the analysis.

Table 1
Categories and subcategories used for the analysis

Category	Favorable Assessment	Problematic Assessment
I. General Experience	- Evaluation of Innovation - Speed and precision of system	- Oddness at automatized evaluation - Absence of human connection

Category	Favorable Assessment	Problematic Assessment
II. Clarification (Ladder)	- Points to useful omissions - Clarifying examples	- Unnecessary corrections - Confusing suggestions
III. Value (Ladder)	- Positive reinforcement generates trust - Stimulates learning	- Emotional impact limited by being AI - Forced or mechanic tone
IV. Concerns (Ladder)	- Constructive and clear criticism - Encouraging improvement	- Redundance or irrelevance in certain criticism
V. Suggestions (Ladder)	- Clarity and applicability - Future improvements	- Suggestions outside of the covered content - Little contextual pertinence
VI. Usability and trust	- Trust in stable technical criteria	- Lack of assessment transparency - Doubts about pedagogical sensitivity
VII. Impact on learning	- Comprehension improvement - Feedback that is useful for study	- Demands perceived as excessive
VIII. Critical judgement and Recommendations	- Willingness to continue to use AI - Contribution to system development	- Need for personalization - Criticism to level adjustment
IX. Perceived autonomy and control (emergent)	- Clarity about expectations - Feedback as guide - AI Strategic adjustments (without explicit conflict)	- Rigid evaluation - Penalization by personal style or creativity - Low motivation due to lack of acknowledgement of alternative valid answers.

Source: Prepared by the authors.

RESULTS

Next, the main results regarding research questions are presented: Is Wilson's Feedback Ladder adequate to be applied through artificial intelligence tools?; Which IAGen tools, with prompts, designed according to Wilson's Ladder, result in being more adequate to provide feedback on learning in university students?; and how do university students perceive the feedback offered through ChatGPT-4o?

Wilson's Feedback Ladder with AI

One hundred percent of the assessments from the expert panel agreed that the strategy of Wilson's Ladder to provide feedback on learning with the use of AI was appropriate, since it offers clarity, value, reflection and suggestions to encourage the learning of students and strengthening the teacher-student relationship when the feedback steps are clearly defined. To this respect, the experts stated:

> It provided a level of description and detail that not only centers on pointing out positive and negative aspects, but goes further, offering more possibilities for reflection, correction and learning improvement through the clarify, value, state concerns and offer suggestions steps (Expert 6).
>
> Wilson's Feedback Ladder strategy is interesting and appropriate and provides a framework for feedback to be empathetic, descriptive and with a focus on contributing to the building of learning, instead of criticism (Expert 2).

More appropriate AI tools according to Wilson's Feedback Ladder

Table 2 presents the averages of assessments provided by experts about feedback generated by different AI tools, evaluated on a scale of 0 (very low) to 3 (very high) according to the four steps in Wilson's Ladder. Each tool was evaluated in terms of clarity, assessment, state concerns, and suggestions, also including its median and standard deviation. According to these data, GPT-4o stands out as the tool with the best general performance, by obtaining consistently higher scores in all criteria and a median of 2.87, along with a low standard deviation (0,18), which reflects both effectiveness and stability in feedback quality.

Table 2

Assessment Averages of Experts on Feedback with AI

AI Tool	Average Feedback Wilson's Ladder Examples 1 and 2					
	Clarify	Value	State Concerns	Offer Suggestions	Mean	Standard Deviation
GPT-4o	2.9	2.75	2.9	2.95	2.87	0.18
GPT-4o mini	2.2	2.1	2.6	2.35	2.30	0.35
Microsoft Copilot	1.55	1.6	1.75	1.95	1.70	0.49
Gemini	1.65	1.45	1.7	1.95	1.67	0.46
Claude	1.6	1.5	1.1	1.35	1.36	0.62
Perplexity	2.45	1.95	2.3	2.5	2.28	0.25
LlaMa 3.2	2	1.9	2	2.3	2.04	0.27

Source: Prepared by the authors.

The agreement between the eight judges when evaluating the different seven tools (GPT-4, Free GPT, Copilot, Gemini, Claude, Perplexity and LlaMa 3.2) was through Kendall's W coefficient, that reached a value of 0.43, which indicates a moderate level according to the judges. The associated meaning test rendered a value of $Chi^2 = 20.67$ with 6 degrees of freedom and a $p < 0.01$, which allows one to conclude that the observed agreement is statistically significant, indicating that the judges showed consistent agreement and statistically validity regarding tool classification according to Wilson's Feedback criteria.

Application of the tool in the Data Structure course

The tool was applied to a total of 107 second-year students. During the test, 846 open answers were processed, each one automatically evaluated through the GPT-4o model, using the previously validated structured prompt.

A voluntary appeal channel was set up through the institutional e-mail, allowing students to request a review of feedback or the assigned score. Thirty-four appeals were received, 27 of those were considered valid, which represents 3.19% of the total answers, a relatively low disconformity rate with the authorized system.

Table 3 presents an analysis of the main patterns of error of the GPT-4o model, on the basis of the total number of processed answers (N = 846). Each identified pattern includes a representative example of generated feedback, an interpretative analysis of the type of error, and the frequency (Freq.) with which it was observed, expressed both in the number of cases and the percentage of the total.

Table 3
Patterns of error observed in responses to the model

Pattern of error	GPT-4o feedback example	Error analysis	Cases	Freq.
Feedback of the model suggests including details that were already present in the student's answer.	"It would be useful to consider a clearer description of the operations that take place on the list, such as the addition of elements at the end and the elimination of the first element".	Feedback suggests that there are still missing details about the operations of adding and eliminating elements on the list. However, the student had already mentioned them in their answer.	31	3.7 %
Feedback of the model asks for details or aspects that were not requested in the question.	"It would be helpful to consider the importance of the hash function and how it affects speech efficiency".	Feedback suggests including additional information (such as the importance of the hash function) that was not requested in the original question. This can confuse the student since it extends the scope of the question and makes them feel their answer was incomplete despite actually answering appropriately to what was asked.	127	15 %
Feedback is unclear or ambiguous, limiting guidance for student improvement.	"Your response does neither address the specific search range nor how to collect elements within this range".	Feedback from AI suggests that the student did not explain about element collection, when in fact they did, but not with the most adequate approach. This can disorient the student since it is not clear if what they wrote was wrong or if they simply had to improve on it.	47	5.6 %

Source: Prepared by the authors.

Application of the tool in the Evaluation for Learning course

The tool was applied to a total of 37 students. During the test, 592 open answers were processed, each one automatically evaluated through the GPT-4o model, using the same structured prompt previously validated.

As in the previous experience, a voluntary appeal channel was activated through the institutional e-mail, so students could request a review of the feedback or the assigned score. In this case, no appeals were received, which suggests a high acceptance of the automatically generated assessments.

Table 4 presents an analysis of the main patterns of error detected in this second experiment, on the basis of the total of processed answers (N = 592).

Table 4
Patterns of error observed in responses to the model

Pattern of error	GPT-4o Feedback example	Error Analysis	Cases	Freq.
Feedback of the model asks for details or aspects that were not requested in the question.	"What specific examples of formative evaluation did you observe during class?"	Feedback does not render pertinent neither regarding the student's answer nor the case stated in the question, which does not reflect an explicit application of the principles of formative evaluation.	3	0.5 %
Feedback has conceptual errors or incorrect statements.	The answer correctly highlights the pedagogical function of assessment and its formative approach, which is fundamental in a context of language learning. In addition, it adequately identifies the use of hetero evaluation, which shows an understanding of the roles of assessing agents in the class.	Feedback suggests that the student's answer is correct and adequate; however, it does not address all aspects required by the question. This inconsistency can generate confusion, especially if the maximum score is not assigned and feedback does not clearly specify which elements should be improved.	4	0.7 %
Feedback uses negative or rigid concepts that do not agree with constructive feedback.	The answer lacks a discussion on how standardized tests could have an impact…	The word "lack" was replaced for not being appropriate enough in the context of constructive feedback, since it could result in being rigid or negative. In its place, a more formative and improvement-oriented statement was used, such as "maybe you could incorporate…" or "it is still to be addressed…".	26	4.4 %

Feedback is unclear or ambiguous, limiting the student's orientation to improvement.	The answer correctly highlights the formative nature of the classroom assessment.	A double evaluation when using terms such as "highlights" and "correctly"; however, it does not provide constructive feedback from the perspective of the achieved goal. The answer does not specify which aspect of the formative assessment is valued in the classroom context, which limits its usefulness to guide improvements.	12	2.0 %

Source: Prepared by the authors

Student Perception on Received Feedback

From this category system, the systematic analysis of the interviews was performed, coding student answers in relation to defined categories and subcategories. This coding allowed for the detection of patterns of sense, thematic recurrences and contrast of individual values, as well as relevant nuances in the way in which participants interpreted automatized feedback. The analysis was carried out combining an inductive reading of the discourse with the structured classification by dimensions, which facilitated a holistic and nuanced understanding of the student experience. Each testimony was interpreted in the light of its internal coherence, its relationship with the categories of the analytical framework and its connection with the rest of the cases.

In general terms, students assign positive value to the experience of receiving automatized feedback, especially highlighting its immediacy, clarity and innovation in comparison to teacher traditional feedback. The surprise due to the use of artificial intelligence is recurring: "it is surprising that it reviews your assessment, since as a student, I am used to having this feedback come from the teacher (student 3). This innovation is seen as a significant change in the evaluative practice, perceived as new and promising. However, this technical assessment coexists with emotional nuances: some describe the experience as "a little stressful" (students 1 and 8), because "it makes you think that you have to be more prepared and be more accurate" (students 1,4,6). Together, students recognize the functional value of the system, but they also notice the absence of the "human factor" (students 1,2,5,7 and 8) and a certain affective distance in the feedback style.

Regarding the clarification step, participants appreciate that AI is capable of identifying omissions and ambiguities that they hadn't noticed themselves. It is valued, for example, when the AI indicated that "there are still details of the complete process of pointer adjustments that need to be listed so that the new node integrates correctly to the list" (student 7). These observations are perceived as useful and formative, but they are not exempt from criticism. Some students question the fact that "they correct me and explain something that I already wrote down in my answer" (student 10), which can generate confusion and a sensation of unfair assessment. This tension between structure and subjectivity is reinforced when the "AI becomes more critical when one freely provides an answer" (student 9) which questions the interpretative flexibility of the system in the face of non-literal but valid answers.

In relation to the value step, many students acknowledge the positive reinforcement as an incentive to continue learning. Affirmations such as "highlight what you did well and invites you to continue to study" (student 6) or "confirms that what I know is not so wrong" (student 2) show that recognition, even when automatized, can take a meaningful pedagogical function. However, its emotional impact is limited to some: "if the comments came from a professor, I would probably feel more acknowledged" (student 4). This ambivalence reveals that, although reinforcement validates knowledge and guides the student, it lacks the emotional charge and the empathy that a human teacher normally contributes, which poses ethical considerations about the quality of the pedagogical connection and interaction mediated by AI.

The concerns step shows that AI criticism is, in general, well received thanks to its clear and respectful tone. Student 1 points out that "this criticism is constructive, which reaffirms its mistakes", while others comment that "I felt well criticized and to the point, straightforward" (student 3). Nonetheless, there are questions that arise when AI demands additional explanations that were not requested by the statement. For example, "although I received all the points, I believe that the part where it mentions including examples is a little unnecessary" (student 9). These observations suggest that feedback would be more effective if the level of demands was adjusted to the content that was actually worked with and the explicit evaluation criteria.

Regarding the suggestions step, students consider that AI offers recommendations that are "clear, concise and easy to apply", which contributes to their autonomous learning. "It helped me to consider it for future occasions" (student 10), one student comments, while another highlights that "it emphasizes the writing order when explaining my algorithm" (student 3). However, the utility of these suggestions depends on the degree of alignment with the context of the course. Some criticize that "we hadn't covered temporal complexity, and this confused the feedback" (student 5), showing that the system could lose efficacy if operating disconnected from the scheduled topics. In spite of this, the majority values structure and clarity of the suggestions as a valuable resource for progressive improvement.

Finally, in relation to the impact on learning, students agree that the use of artificial intelligence has strengthened their own formative process, significantly contributing to reflection about their own learning. Immediate feedback was valued for its ability to activate previous knowledge, "remember details I used to forget" (student 6) and allowing for timely review: "feedback the moment" (students 2); in addition, it was described as "useful to reflect about my teaching-learning process" (student 4).

Some participants recognize the contribution of this tool as a strategic resource: "it made me improve in topics I had not gone into depth about" (student 8) and "it helped me to identify important elements I had not considered before" (student 1). These perceptions reflect an active appropriation of automatized feedback as support for the development of reflective and metacognitive competences.

On the other hand, regarding critical judgement and recommendations, we can observe a proactive and analytical attitude. Students not only value the use of AI, but also expressed concrete proposals to improve its functioning. Among them, reorganizing the steps of Wilson's Ladder, placing value before clarifying, incorporating a discussion section in order to explore contents further, and previously train AI with course materials, and avoid unjustified penalties, not deduct points due to the lack of examples if they are not required or are part of the evaluation criteria.

These proposals show that students do not think of themselves as a passive receptor of the system, but as a reflective agent, committed to perfecting and pedagogical instrument that they consider useful, but that they want to become more contextualized, fair and aligned with the formative goals of the course.

DISCUSSION AND CONCLUSIONS

The results of this study show the pedagogical potential of integrating automatized feedback mediated by generative artificial intelligence, in particular through GPT-4o structured through Wilson's Ladder. This proposal is aligned with the need to transform feedback practices in university teaching, which students tend to perceive as late, superficial and non-specific (Carless & Boud, 2018; Winstone et al., 2022; Quezada & Salinas, 2021). In open and distant contexts, where the immediacy and clarity are critical factors, the IAGen demonstrates that is it capable to reduce response times and deliver structured and focussed comments, strengthening the formative dimension of evaluation (Puertas & Cano, 2024; Holmes et al., 2019).

Wilson's Ladder offered a useful scaffolding to organize feedback in the four steps (clarify, value, state concerns and suggest), which allowed the generation of more understandable, specific and improvement-oriented observations, even in an automatized format. Students highlighted this experience because of its immediacy, clarity and innovation, recognizing it as a tool capable of supplementing the teaching role. Nonetheless, they also mentioned tensions associated to feelings of stress, greater demand on the preparation for answers and the absence of the human factor which offers closeness and warmth. These nuances strengthen the idea that technology cannot replace the teacher but must be conceived as a supplementary support to pedagogical interaction. This interpretation coincides with Valenzuela and Pérez (2025) who underscore the irreplaceability of teachers, in terms of human interaction and personalized support, remembering that artificial intelligence can build valuable resources, but always subordinate to the relational and affective dimension that is typical of teaching.

A relevant finding was the student observation regarding the clarify step. In current practice, the tool does not allow the student to contribute additional information before receiving comments, which reduces the formative reach of this step. To include an interactive mechanism, such as a chat of clarifying window, could enrich the personalization of feedback, aligning comments with real needs of students (Luckin & Holmes, 2016; López Regalado et al., 2024). This possibility gains greater importance in massive contexts or distance education, where interaction spaces tend to be limited.

In the technological dimension, validation by experts and the positive perception of students agree with what is proposed by Zhang et al. (2024) and Baral et al. (2024), who highlight that GPT can generate coherent and functional feedback as long as prompts are carefully designed. This aspect is critical in scenarios that require standardization and scalability (Carless & Winstone, 2023). However, as stated by Jauhiainen and Garagorry Guerra (2024), limitations linked to empathy, contextualization and affective sensibility persist, which reasserts the need for teacher supervision to ensure pedagogical pertinence and formative impact.

The study also shows that students from different fields and academic levels (second and fourth year) valued automatized feedback in a similar way, which suggests that Wilson's Ladder, applied with generative AI, becomes an adaptable and replicable

framework in various contexts. However, its implementation requires incorporating ethical criteria, teacher reviews and rigorous prompt design that will adequately guide the AI production.

Finally, there are challenges that emerge that invite further exploration in future research, in particular regarding the analysis of algorithmic biases, evaluation of ethical risks connected to privacy and equality, and also the need to guarantee the technical sustainability of these tools. It is also critical to explore its adaptation to various academic disciplines, ensuring pedagogical relevance and its alignment with higher education quality standards. Along this line, it is recommended that progress is made towards a responsible and contextualized integration of artificial intelligence in university evaluation processes, in such a way that its technical benefits supplement teacher pedagogical mediation. It is pertinent for the future to expand the reach of the studies to different disciplines and explore new dimensions of its impact in education, considering variables such as argumentative quality, problem resolution, critical thinking, student autonomy, and the transformation of evaluation practices. Likewise, it is suggested that longitudinal research is developed in order to analyze the sustained effect of IAGen use in learning and teaching practice. Simultaneously, the need emerges to strengthen teacher training in prompt design, ethical reflection and critical evaluation of the use of these technologies, with the purpose of ensuring a more conscious and formative implementation.

NOTES

1. Moodle is an open code learning management platform (LMS, because of its name in English) widely used in educational contexts for creation, delivery and follow up of online courses.
2. Version gpt-4o-2024-08-06
3. Version gpt-4o-mini-2024-07-18
4. CodeRunner is a Moodle plugin that allows automatic evaluation of information coding. It integrates with an execution engine (for example, Python) that processes answers in real time, creating feedback and immediate scores according to preset evaluation criteria.

REFERENCES

Al-Azawei, A., Abdullah, A. A., Mohammed, M. K., & Abod, Z. A. (2023). Predicting online learning success based on learners' perceptions: The integration of the information system success model and the security triangle framework. *International Review of Research in Open and Distributed Learning, 24*(2), 7295. https://doi.org/10.19173/irrodl.v24i2.6895

Andrade, H. L., & Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice, 26*(1), 103-117. https://doi.org/10.1080/0969594X.2019.1571992

Anthropic. (2024). *Claude 3.7* [Large language model]. https://www.anthropic.com

Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B., & Adewusi, O. E. (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews, 18*(2), 261-271. https://doi.org/10.30574/gscarr.2024.18.2.0062

Baral, S., Worden, E., Lim, W.-C., Luo, Z., Santorelli, C., Gurung, A., & Heffernan, N. (2024). *Automated feedback in math education: A comparative analysis of LLMs for open-ended responses*. arXiv. https://doi.org/10.48550/arXiv.2411.08910

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Bonales-Daimiel, G., Martínez-Estrella, E. C., & Sierra-Sánchez, J. (2025). Evolución del perfil docente y surgimiento de nuevos roles profesionales en la era de la inteligencia artificial (IA). *Pixel-Bit. Revista de Medios y Educación, 73*, Article 3. https://doi.org/10.12795/pixelbit.109085

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed.). University of Chicago Press. https://doi.org/10.7208/chicago/9780226062648.001.0001

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education, 28*(1), 150-163. https://doi.org/10.1080/13562517.2020.1782372

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). *Can large language models provide feedback to students? A case study on ChatGPT*. EdArXiv. https://doi.org/10.35542/osf.io/hcgzj

DeepSeek. (2024). *DeepSeek R1* [Generative AI model]. https://www.deepseek.com/

European Commission. (2023). *Teaching with AI – Assessment, feedback and personalisation. Briefing report No. 7 (European Digital Education Hub)*. Erasmus+ Programme. https://resitve.sio.si/wp-content/uploads/sites/7/2023/11/AI-squad-output_briefing-report-7.pdf

Galindo-Domínguez, H., Delgado, N., Losada, D., & Etxabe, J. M. (2023). An analysis of the use of artificial intelligence in education in Spain: The in-service teacher's perspective. *Journal of Digital Learning in Teacher Education, 40*(1), 41-56. https://doi.org/10.1080/21532974.2023.2284726

García-Peñalvo, F. J. (2023). La percepción de la inteligencia artificial en contextos educativos tras el lanzamiento de ChatGPT: ¿Disrupción o pánico? *Education in the Knowledge Society (EKS), 24*, e31279-e31279. https://doi.org/10.14201/eks.31279

García Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia, 27*(1), 9-39. https://doi.org/10.5944/ried.27.1.37716

Goodrich, H. (2011). Ladder of Feedback [Adaptation based on Ron Berger's Ladder of Feedback]. In A. Goodrich (Ed.), *Protocols in the classroom*. Harvard Project Zero.

Google DeepMind. (2024). *Gemini 2.5* [Multimodal AI model]. https://deepmind.google/

Hao, Y., Sun, Y., Dong, L., Han, Z., Gu, Y., & Wei, F. (2022). *Structured prompting: Scaling in-context learning to 1,000 examples*. arXiv. https://doi.org/10.48550/arXiv.2212.06713

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 1-18.

https://doi.org/10.1080/14703297.2024.2422337

Landeta, J. (1999). *El método Delphi: Una técnica de previsión para la incertidumbre*. Ariel.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1-35. https://doi.org/10.1145/3560815

LLaMa. (2024). *LLaMA4* [Large language model]. https://www.llama.com/

López Regalado, O., Núñez-Rojas, N., López Gil, O. R., & Sánchez-Rodríguez, J. (2024). El análisis del uso de la inteligencia artificial en la educación universitaria: Una revisión sistemática. *Pixel-Bit. Revista de Medios y Educación, 70*, 97-122. https://doi.org/10.12795/pixelbit.106336

Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/IntelligenceUnleashedSPANISH.pdf

Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research, 1*(2), 1-10. https://doi.org/10.17169/fqs-1.2.1089

Ministry of Education of Chile. (2019). *Orientaciones para la implementación del Decreto 67/2018 de evaluación, calificación y promoción*. MINEDUC. https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/14279/orientaciones%20decreto%2067.pdf

Mistral AI. (2024). *Mistral* [Large language model]. https://mistral.ai/

Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education, 45*(4), 527-540. https://doi.org/10.1080/02602938.2019.1667955

OpenAI. (2023). *Best practices for prompt engineering with the OpenAI API*. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api

OpenAI. (2024). *GPT-4o, GPT-4.5 model cards*. https://openai.com

Ossa, C., & Willatt, C. (2023). Uso de inteligencia artificial generativa para retroalimentar escritura académica en procesos de formación inicial docente. *European Journal of Education and Psychology, 16*(2), 1-16. https://doi.org/10.32457/ejep.v16i2.2412

Puertas, E., & Cano, E. (2024). ¿Puede la inteligencia artificial proporcionar un feedback más sostenible? *Digital Education Review, 45*(1), 50-58.

Quezada, S., & Salinas, C. (2021). Modelo de retroalimentación para el aprendizaje: Una propuesta basada en la revisión de literatura. *Revista Mexicana de Investigación Educativa, 26*(88), 225-251.

Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). https://doi.org/10.1145/3411763.3451760

Romero Alonso, R., Araya Carvajal, K., & Reyes Acevedo, N. (2025). Rol de la inteligencia artificial en la personalización de la educación a distancia: Una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia, 28*(1), 9-36. https://doi.org/10.5944/ried.28.1.41538

Sahoo, S. S., Plasek, J. M., Xu, H., Uzuner, Ö., Cohen, T., Yetisgen, M., & Wang, Y. (2024). Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association, 31*(9), 2114-2124. https://doi.org/10.1093/jamia/ocae074

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). *The prompt report: A systematic survey of prompt*

engineering techniques. arXiv. https://doi.org/10.48550/arXiv.2406.06608

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*(1), 1-19. https://doi.org/10.1111/jcal.12172

Steurer, J. (2011). The Delphi method: An efficient procedure to generate knowledge. *Skeletal Radiology, 40*, 959-961. https://doi.org/10.1007/s00256-011-1145-z

Valenzuela Caico, R., & Pérez Carvajal, A. (2025). Inteligencia artificial en educación superior: ¿Un reemplazo para los profesores o una herramienta de apoyo? *Revista Iberoamericana de Investigación en Educación, (9)*. https://doi.org/10.58663/riied.vi9.221

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://doi.org/10.48550/arXiv.2201.11903

Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology, 40*(6), 37-55. https://doi.org/10.14742/ajet.9540

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wilson, D. (2013). *Ladder of feedback* [Working paper]. Project Zero, Harvard Graduate School of Education. https://pz.harvard.edu/resources/ladder-of-feedback

Winstone, N. E., Boud, D., Dawson, P., & Heron, M. (2022). From feedback-as-information to feedback-as-process: A linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education, 47*(2), 213-230. https://doi.org/10.1080/02602938.2021.1902467

xAI. (2024). *Grok 3* [Large language model]. https://grok.com/

Zhang, Z., Dong, Z., Shi, Y., Price, T., Matsuda, N., & Xu, D. (2024). Students' perceptions and preferences of generative artificial intelligence feedback for programming. In *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23250–23258. https://doi.org/10.1609/aaai.v38i21.30372

Date of reception: 1 June 2025
Date of acceptance: 2 September 2025
Date of approval for layout: 1 October 2025
Date of publication in OnlineFirst: 17 October 2025
Date of publication: 1 January 2026