Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 14/01/2026; 1st round notif.: 14/05/2026; New version: 13/06/2026; 2nd round notif.: 24/06/2026;
Camera ready: 04/07/2026; Edition review: 10/07/2026; Available online: 10/07/2026; Published: 10/07/2026;

Which Feedback Is More Effective? A Comparative Study on Teachers' Evaluation of Feedback Models in Essay Writing

Anderson Pinheiro Cavalcanti
Universidade Federal Rural de Pernambuco (UFRPE)
CESAR School
ORCID: 0000-0002-5228-1583
anderson.pinheiro@ufrpe.br

Cleon Xavier
Instituto Federal Goiano – Campus Iporá
ORCID: 0000-0002-7617-5283
cleon.junior@ifgoiano.edu.br

Fabiola Ribeiro
Instituto Federal Goiano – Campus Catalão
ORCID: 0000-0002-3303-9396
fabiola.ribeiro@ifgoiano.edu.br

Weslei Chaleghi de Melo
Universidade Tecnológica Federal do Paraná (UTFPR)
ORCID: 0000-0002-0503-6729
weslei@alunos.utfpr.edu.br

Guilherme Weber
Faculdade UNA Catalão
ORCID: 0000-0002-7502-8070
gweber.gw@gmail.com

Rafael Ferreira Leite de Mello
Universidade Federal Rural de Pernambuco (UFRPE)
CESAR School
ORCID: 0000-0003-3548-9670
rafael.mello@ufrpe.br

Luiz Rodrigues
Universidade Tecnológica Federal do Paraná (UTFPR)
ORCID: 0000-0003-0343-3701
luizrodrigues@utfpr.edu.br

Newarney Torrezão da Costa
Instituto Federal Goiano – Campus Iporá
ORCID: 0000-0002-4954-176X
newarney.costa@ifgoiano.edu.br

Luiz Felipe Bagnhuk Silva
Faculdade de Campina Grande do Sul
ORCID: 0009-0005-2342-2649
felipebagnhuk@gmail.com

Letícia Santana Stacciarini
Instituto Federal Goiano
ORCID: 0000-0002-0038-8645
leticia.stacciarini@gmail.com

Evelyn Vieira
Instituto Federal Goiano
ORCID: 0009-0007-9791-0244
evelyn.vieira@ifgoiano.edu.br

Dragan Gasevic
Monash University
ORCID: 0000-0001-9265-1908
dragan.gasevic@monash.edu

Abstract

Feedback is a key factor in improving the writing skills of students, but providing it at a large scale remains a significant challenge. Large Language Models (LLMs) emerge as a promising solution; however, the pedagogical quality of the automatically generated feedback requires further investigation. This study examined how effective teachers perceived feedback texts generated by an LLM when the prompts were designed to reflect different aspects of feedback theory. To this end, 450 feedback texts were generated for 30 essays and evaluated by five teachers with varying feedback literacy profiles. The results showed that teachers' preferences for feedback models were strongly shaped by their profiles, indicating that there was no single universally "best" feedback model. From a learning analytics perspective, these findings demonstrate how combining theory-informed prompt design with evaluations from diverse teacher profiles can generate actionable insights for the development of scalable, AI-powered feedback systems. The study contributes to learning by showing that the effectiveness of LLM-based feedback depends on personalization not only for students but also for teachers, highlighting new directions for designing adaptive, equitable, and context-sensitive feedback analytics.

Cite as: Cavalcanti, A. P., Rodrigues, L., Xavier, C., Costa, N., Ribeiro, F., Silva, L. F. B., Melo, W. C., Stacciarini, L. S., Weber, G., Vieira, E., Mello, R. F. L., & Gasevic, D. (2026). Which Feedback Is More Effective? A Comparative Study on Teachers' Evaluation of Feedback Models in Essay Writing. Revista Brasileira de Informática na Educação, vol. 34, pp. 1070–1094. https://doi.org/10.5753/rbie.2026.7365.

Keywords: *Educational Feedback; Large Language Models (LLMs); ENEM Essay; Feedback Literacy.*

1 Introduction

The skill of constructing argumentative texts that present clear ideas is an important competence in contemporary professional contexts, and the quality of feedback provided to students is a critical factor in the development of their writing skills (Oliveira et al., 2025). However, offering detailed and personalized feedback in a large-scale scenario presents significant challenges, especially due to the high student-to-teacher ratios and the time required for in-depth grading (Pimentel et al., 2025; Shin et al., 2025; Watts et al., 2023).

To overcome the barriers of time and scale, Automated Essay Scoring (AES) systems emerged as a promising solution (Mello et al., 2024). Recent Brazilian research has investigated AES for essays written in Brazilian Portuguese, including ENEM-style corpora and competencies such as textual cohesion (Monteiro & Barreto, 2022; Oliveira et al., 2023). AES employs computational models that approximate human judgments of writing quality, providing consistent and efficient assessments across large cohorts (Shermis & Burstein, 2003). The Learning Analytics (LA) community has explored approaches for AES. Recent LA studies illustrate both the contextualized design of writing-analytics tools (e.g., AcaWriter) (Shibani et al., 2019) and new human–AI collaborative frameworks that expand AES solutions, aiming to improve feedback quality, explainability, and grader productivity (Xiao et al., 2025). In this context, feedback needs to offer actionable guidance that goes beyond error correction, functioning as a process through which students engage with information from multiple sources to refine their work and learning strategies (Hutt et al., 2024).

Recent advancements in Large Language Models (LLMs), such as GPT and Gemini, have expanded the capacity to generate instant, consistent, and personalized feedback, positioning these models as valuable tools for LA (Shin et al., 2025). Empirical studies demonstrate that LLMs can automate the grading process and augment the effectiveness and efficiency of human graders, enabling productive human–AI collaboration (Xiao et al., 2025). These developments expand AES to a scalable provider of timely and actionable feedback for learners beyond scoring, while highlighting the continuing need to address issues of student trust, pedagogical alignment, and model transparency (Alcock et al., 2024; Mello et al., 2024).

Despite its potential, simple automation does not guarantee the pedagogical quality of feedback (Cavalcanti et al., 2021). Rüdian et al. (2025) demonstrated that students are often unable to distinguish LLM-generated feedback from teacher-written feedback, although their perception deteriorates once the artificial origin of the message is revealed. While this evidence clarifies how the source of the feedback shapes its reception, it leaves unanswered the question of how the pedagogical structure of the message itself, that is, the feedback theory embedded in its generation, influences its perceived effectiveness. In other words, prior work has compared who produces the feedback (human versus AI), but not which theoretical model should guide the AI when producing it. For LA, this distinction matters because grounding AI-generated feedback in consolidated educational theories (Hutt et al., 2024; Rüdian et al., 2025) is what ensures that analytic outputs are not only accurate but also pedagogically meaningful and trusted by learners.

Addressing the challenges of pedagogical quality in AI-powered feedback in LA requires grounding design in established theoretical models such as those by Nicol and Macfarlane-Dick (2006) and Hattie and Timperley (2007). Nicol and Macfarlane-Dick's model (Nicol & Macfarlane-

Dick, 2006) focuses on seven principles that promote self-regulated learning, while Hattie and Timperley's model (Hattie & Timperley, 2007) suggests structuring feedback around three key questions that guide students in their development process. Both theories aim to transform feedback into a tool that empowers the student to understand their goals, evaluate their progress, and identify the next steps (Dawson et al., 2023). Despite this, past research on feedback generation has failed to include these theoretical considerations in how to instruct LLMs (Jansen et al., 2025; Singh et al., 2024; Watts et al., 2023; Xavier et al., 2025). As a result, there is an empirical gap regarding which feedback model is more likely to inform the generation of effective feedback messages in the context of essay writing, a gap that complements source perception studies, such as Rüdian et al. (2025), but is not addressed by them, since here the source is kept constant (a single LLM) and only the theoretical basis of the stimulus varies.

Therefore, the main objective of the current study was to examine the effectiveness of incorporating different feedback theories in the process of generating automatic feedback messages aimed at improving students' writing skills. To guide this investigation, we defined the following primary research question:

> **RQ:** *To what extent does the incorporation of established feedback theories into prompt design improve the perceived effectiveness of LLM-generated feedback messages for essay writing?*

To answer the primary research question, we decomposed it into two secondary research questions (SRQs) that, in a complementary manner, support the answer to the main problem:

> **SRQ1:** *To what extent does the inclusion of feedback theories in prompts improve the perceived effectiveness of feedback messages?*

> **SRQ2:** *To what extent does the perceived effectiveness of different feedback models vary across evaluation criteria?*

To do so, we compared teachers' perceptions of three types of feedback generated by an LLM: naive feedback with no theory associated; feedback structured according to the seven principles of Nicol and Macfarlane-Dick (2006); and feedback based on the guiding questions of Hattie and Timperley (2007). This comparison was designed to identify which approach could be considered most promising for supporting student writing and to consider the implications of embedding theoretically grounded frameworks into LLMs-based systems within LA.

The remainder of this article is organized as follows. Section 2 presents the theoretical background on educational feedback, feedback literacy, and the use of LLMs to support feedback practices. Section 3 describes the method, including the essay dataset, the feedback generation procedure, the participating teachers, the assessment protocol, and the data analysis plan. Section 4 reports the results for each secondary research question, and Section 5 discusses their implications for the LA community. Finally, Section 6 presents the limitations of the study and directions for future research.

2 Theoretical Background

2.1 Educational Feedback

Feedback is an essential component of the teaching-learning process, conceptualized as "a process through which students make sense of information from various sources and use it to enhance their work or learning strategies" (Carless & Boud, 2018, p. 1315). In its formative function, feedback aims to guide and support the student's continuous development rather than merely certifying a final result. As Hattie and Timperley (2007) highlights, effective feedback works to reduce the discrepancy between the student's current performance and the desired learning goal.

The literature distinguishes between different types of feedback, such as outcome feedback (which only indicates whether the answer is correct or incorrect) and elaborative feedback (which includes explanations, hints, and strategies) (Van Campenhout et al., 2024). Research shows that elaborative feedback is more effective for learning, as it offers a path for correction and improvement. However, providing elaborative and individualized feedback is a task that demands significant time and effort from teachers, especially in large classes, which justifies the search for technological solutions that can assist in this process (Dawson et al., 2023; Van Campenhout et al., 2024).

One of the most influential theoretical approaches is that of Nicol and Macfarlane-Dick (2006), who proposed seven principles of good feedback practices aimed at fostering self-regulated learning. According to the authors, good feedback should: 1) help clarify what constitutes good performance; 2) facilitate the development of self-assessment; 3) provide high-quality information; 4) encourage dialog about learning; 5) encourage positive motivational beliefs and self-esteem; 6) provide opportunities to close the gap between current and desired performance; and 7) provide teachers with information to shape instruction. These principles shift the focus from simply transmitting information to a process that empowers students to reflect on their own learning and become active agents in their development (Nicol & Macfarlane-Dick, 2006; Watts et al., 2023).

Another highly influential model is that of Hattie and Timperley (2007), which conceptualizes feedback as a tool for reducing the discrepancy between current understanding or performance and the desired goal. The model is organized around three essential questions that the learner must be able to answer: Where am I going? (Feed Up), which refers to the clarity of learning goals; How am I doing? (Feed Back), which informs about progress toward these goals; and Where am I going from here? (Feed Forward), which guides the next steps. Feedback, according to the authors, can operate at four distinct levels: task, process, self-regulation, and personal (self). As a result, this feedback model informs students' learning experiences in terms of planning to monitoring, including guidance from what to do (task) to self-improvement (personal).

2.2 Feedback Literacy

For feedback to be effective, both those who provide it and those who receive it need to develop a set of competencies known as feedback literacy. For students, feedback literacy involves the ability to interpret, process, and act on the information received to enhance their learning (Carless & Boud, 2018). Likewise, the teacher's feedback literacy is fundamental, as the development

of the student's literacy largely depends on how the teacher designs and conducts the feedback process (Lee et al., 2023). Carless and Winstone (2023) define teacher feedback literacy as "the knowledge, expertise, and dispositions to design feedback processes in ways which enable student uptake of feedback and seed the development of student feedback literacy".

Based on the concept of feedback literacy, Lee et al. (2023) developed and validated the Feedback Literacy Scale (FLS) for writing teachers, which conceptualizes this competency in a three-factor solution: **Perceived Knowledge** - the teacher's understanding of feedback principles, such as its sources (teacher, peers, self-assessment), effective strategies, and the role of clear assessment criteria; **Values** - the teacher's beliefs, attitudes, and goals regarding feedback, such as viewing it as a shared responsibility between teacher and student, an incentive for revision, and a continuous process aimed at fostering student autonomy; **Perceived Skills** - the teacher's ability to put knowledge and values into practice, such as the skill of avoiding overloading students with too much information, actively engaging them in the process, using technology to enhance practices, and reflecting on their own feedback for continuous improvement.

2.3 LLMs to support Feedback Practice

The use of automated systems to provide feedback to students is a well-established research area, aiming to offer scalable and immediate support for the learning process (Pardo et al., 2019). Some studies have explored the impact and effectiveness of these tools in different knowledge domains, with a growing interest in the application of LLMs to enhance the quality and personalization of this feedback (Cavalcanti et al., 2021).

Previous work has examined how AI can be integrated with learners' contributions to strengthen the feedback process. Singh et al. (2024) investigates a hybrid model in which students and AI collaborate to generate feedback. The study explores the dynamics of this partnership, analyzing how combining AI-generated feedback (GPT-4) with student input (learnersourcing) can create a richer and more effective feedback process. The results indicated that students who reviewed AI prompts produced feedback with greater accuracy, specificity, and usefulness. Despite this, there was no clear preference among students for working with AI support, as a lack of confidence in LLMs and a desire for independent engagement were the main reasons for preferring autonomous writing.

Another important focus in LA has been students' preferences for AI versus human feedback, which shapes how these tools can be adopted at scale. The work of Shin et al. (2025) investigates the preferences of 108 high school students between feedback generated by an LLM (GPT-4) and that provided by peers in a formative peer assessment setting in a mathematics course. The results reveal that the preference for AI feedback was higher in classes with lower engagement and performance, while high-performing classes, which were more aware of AI's potential flaws, preferred human feedback. Individually, students with lower confidence in and perceived value of mathematics also tended to prefer the AI-generated feedback. The study concludes that AI acts as a promising supplement in assessment, especially in scenarios where human feedback is insufficient. However, for it to be effective, its implementation must be carefully adapted to the characteristics of the class and the individual dispositions of the students.

The study presented by Xavier et al. (2025) investigated the use of LLMs to support feedback generation in secondary school chemistry assessments, comparing messages produced exclusively

by teachers with those co-created with a language model. The results indicated that students did not perceive significant differences in the quality of the feedback received and, for the most part, were unable to distinguish its origin. Moreover, the use of LLMs enabled the production of more detailed messages without a substantial increase in grading time, and these messages were largely accepted by teachers. These findings reinforce the potential of LLMs as support for teaching practice, enhancing feedback personalization without compromising pedagogical effectiveness.

Recent studies have also investigated how multimodal evidence can be incorporated into LA systems to broaden the scope of automated feedback. Nguyen and Park (2025) investigated the use of Multimodal LLMs (MLLMs), such as GPT-4o and Claude, to provide automated feedback on sixth-grade science assessments that combine text and graphics. MLLMs were found to be viable for the task, achieving moderate to substantial agreement with human raters in the assessment of scientific reasoning. The performance of the models improved when they were provided with example answers and scores in a few-shot learning process. However, the analysis revealed that the models can underestimate the depth of student responses, generate unsolicited details (hallucinations), and exhibit incorrect numerical reasoning. The findings suggest the feasibility of using MLLMs for real-time feedback but highlight the need for improvements to increase their reliability.

The literature has further highlighted that the effectiveness of LA-driven feedback depends on technical quality and on how learners perceive and engage with it. Another central aspect investigated in the literature is the perception and engagement of students with AI-generated feedback. Rüdian et al. (2025) conducted an experiment in which students could not distinguish feedback generated by an LLM from that created by their teacher. However, the students' perceptions changed drastically to a negative view when they were informed that the feedback was automatically generated, highlighting an emotional and social resistance to the absence of the human element. This finding is corroborated by Jansen et al. (2025), who, in a large-scale study, observed high rates of student non-engagement with feedback (i.e., not making revisions), regardless of whether its origin was human or artificial. This points to a deeper pedagogical challenge that technology alone cannot solve.

In sum, the literature indicates that while LLMs offer the potential for automating feedback at scale, their effective implementation depends on technical accuracy combined with pedagogical structuring and integration into LA systems. For the LA community, the key opportunity lies in human–AI collaboration: using technology to optimize repetitive tasks and generate draft feedback while teachers validate, personalize, and sustain the motivational and relational dimensions of assessment (Jansen et al., 2025; Yan et al., 2024).

3 Method

3.1 Context

This study focuses on analyzing feedback messages generated for essays written in the format of the Brazilian National High School Exam (ENEM). ENEM is the examination used in the selection process for higher education in Brazil, and its essay component is considered the most significant element in the final evaluation (Carvalho et al., 2024). The exam requires participants

to produce an argumentative–dissertative text on a socially, culturally, or scientifically relevant topic. Essays are assessed by human graders across five competencies, each scored on a scale from 0 to 200, in intervals of 40 points:

Competency 1 (C1): Demonstrate command of the formal written modality of the Portuguese language.

Competency 2 (C2): Understand the essay's purpose and apply concepts from various areas of knowledge to develop the theme within the structural limits of the argumentative-dissertative text in prose.

Competency 3 (C3): Select, relate, organize, and interpret information, facts, opinions, and arguments in defense of a point of view.

Competency 4 (C4): Demonstrate knowledge of the linguistic mechanisms necessary for the construction of argumentation.

Competency 5 (C5): Elaborate on an intervention proposal for the problem addressed, respecting human rights.

Several studies have proposed solutions for automating the scoring of essays in the ENEM format. For instance, prior research has examined indicators such as the inappropriate use of connectives and excessive word repetition (Oliveira et al., 2025). Among the five ENEM competencies, Competency 4, focused on textual cohesion, has received particular attention in LA research, as it is both one of the most challenging areas for students and a key determinant of clarity and logical progression in writing (Carvalho et al., 2024; Galhardi et al., 2024; Oliveira et al., 2025). However, the provision of feedback to support student improvement is still a gap in the literature.

3.2 Essay Dataset

The corpus used in this research was extracted from Essay-BR (Marinho et al., 2021), a well-established and freely available corpus comprising 4,570 argumentative essays written by Brazilian students in response to 86 distinct prompts (topics), each essay scored by human graders across the five ENEM competencies. Each record in the dataset contains: i) the text of the essay produced by the student; ii) the essay topic; and iii) the supporting text provided to the student as a basis for their writing. This structure allowed the texts to be used to generate feedback.

For this study, we selected a sample of 30 essays through a random draw from the corpus. Proficiency levels were operationalized using the total essay score available in the dataset (0–1000), which corresponds to the sum of the five competency scores assigned by human graders. After the random selection, we inspected the resulting score distribution to verify that the sample was not concentrated at either extreme of the proficiency spectrum, which could bias the feedback generation toward exclusively corrective or exclusively confirmatory messages. Figure 1 presents the distribution of total scores in the final sample, which ranges from 0 to 860 points, with a mean of 419.3 and a median of 440.0, indicating that the sample covers low, medium, and high-proficiency essays. As a consequence of the random sampling procedure, the 30 essays were

distributed across four topics: Traditional communities in Brazil (12), Artificial Intelligence (11), Mental health (6), and Fake news (1). Topic representativeness was not a sampling criterion, since the unit of analysis in this study is the feedback message generated for each competency of each essay, rather than the essay topic itself.

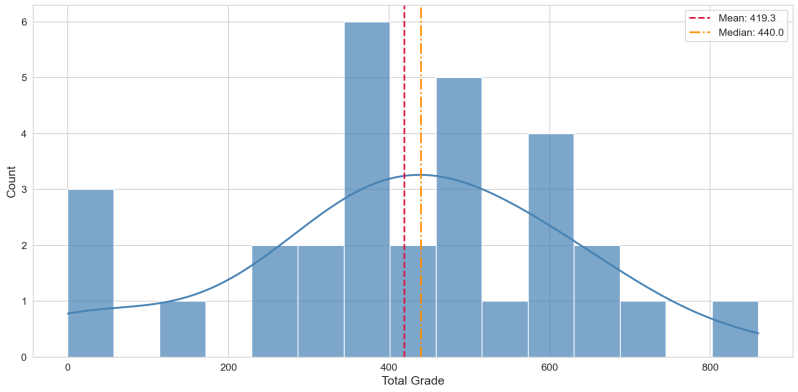

Figure 1: Distribution of the total scores (0–1000) of the 30 sampled essays..

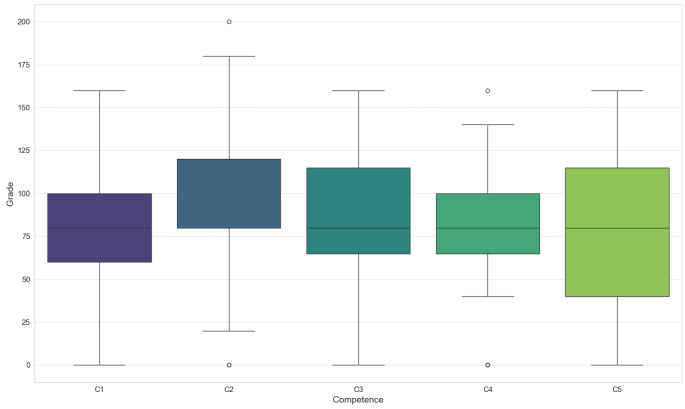

Figure 2: Boxplot with the grades for each competence..

The average text length was 224.10 words, and the average overall grade was 419.3 (out of 1000). The average grades by competence (each with a maximum of 200) were: C1 = 77.33, C2 = 102.00, C3 = 80.67, C4 = 80.00, and C5 = 76.0. Figure 2 presents the box plot showing the distribution of grades for each competence.

We also examined the relationship between essay length and writing performance in the sample. The Pearson correlation between the number of words in the essay and the total score was $r = 0.70$ ($p < 0.001$), indicating a strong positive association: longer essays tended to receive higher scores, consistent with prior evidence from automated essay scoring research (Kucia et al., 2026; Shermis & Burstein, 2003). This association was not used as a sampling criterion.

3.3 Feedback Generation using LLM

The three prompts were developed through a structured, theory-driven procedure. First, the operational instructions of the Nicol and Macfarlane-Dick and Hattie and Timperley prompts were

derived directly from the constructs of each source theory: the seven principles of good feedback practice (Nicol & Macfarlane-Dick, 2006), from which the five principles most aligned with essay-writing goals were retained, and the three guiding questions (Feed Up, Feed Back, Feed Forward) of Hattie and Timperley (2007), respectively. Second, the initial versions of the three prompts were piloted on a set of 10 essays not included in the final sample, in 4 refinement rounds. These pilot rounds were used to (i) calibrate the instruction, limiting each competency's feedback to a maximum of three sentences, (ii) remove generic or repetitive formulations in the generated messages, and (iii) ensure that the three approaches shared an identical base prompt and identical stylistic constraints, so that the only systematic difference among them was the theoretical structure. The outputs of the pilot rounds were reviewed by the authors, and the final versions presented below were frozen before the generation of the experimental corpus.

For each of the 30 selected essays, specific feedback was generated for all five ENEM competencies. A single LLM (Gemini-2.5-flash) was used, but through prompt engineering, we designed three distinct feedback approaches, resulting in 15 feedback texts per essay (5 competencies times 3 approaches). Gemini-2.5-flash was chosen because previous research demonstrated its efficacy for essay assessment tasks (Mello et al., 2025). Below, we present the English translations of the prompts used to generate the feedback messages. Each prompt began with the same base text describing the ENEM competency assessment, which was then tailored according to the feedback approach.

ENEM BASE PROMPT

Evaluate the essay below according to the five criteria established by the ENEM (National High School Exam). Then provide concise, non-generic feedback for each competency, highlighting the overall revision of the essay in that competency. Consider the following criteria used in the ENEM: Competency 1: Demonstrate mastery of the formal written form of the Portuguese language. Competency 2: Understand the essay proposal and apply concepts from various areas of knowledge to develop the topic, within the structural limits of a dissertative-argumentative prose text. Competency 3: Select, relate, organize, and interpret information, facts, opinions, and arguments in defense of a point of view. Competency 4: Demonstrate knowledge of the linguistic mechanisms necessary for constructing an argument. Competency 5: Develop a proposal for intervention to address the problem addressed, respecting human rights.

Naive Feedback: A direct prompt was used to function as a baseline. It asked the LLM to generate corrective and succinct feedback, highlighting the student's main successes and deviations from the assessed competency.

Naive prompt

ENEM BASE PROMPT
Use clear, objective, respectful, and non-generic language. Provide examples whenever possible. The text for each competency should be limited to a maximum of 3 sentences. Consider the steps step by step and clearly justify your choices.

Nicol and Macfarlane-Dick-Based Feedback: The prompt was designed to instruct the LLM to generate feedback incorporating five principles of the Nicol and Macfarlane-Dick (2006) feedback model. These principles were more aligned with the goals of essay writing, such as clarifying good performance, encouraging self-assessment, and providing opportunities to close the learning gap.

Nicol prompt

ENEM BASE PROMPT
For each competency, provide objective and clear feedback, taking into account the following best feedback practices proposed by Nicol and McFarlane-Dick (2006): i) Clarify goals and criteria: Clearly explain what is expected for each competency. ii) Promote self-assessment: Show where the text approaches or deviates from the criteria. iii) Provide constructive feedback: Point out strengths/weaknesses and actions for improvement. iv) Encourage self-esteem: Use encouraging language, recognizing achievements. v) Reduce the performance gap: Practical suggestions for improvement for each competency. Use clear, objective, respectful, and non-generic language. Provide examples whenever possible. The text for each competency should be limited to a maximum of 3 sentences. Think through the steps step by step and clearly justify your choices.

Hattie and Timperley-Based Feedback: The prompt was structured so that the LLM generated feedback around the Hattie and Timperley (2007) feedback model's three central questions: Where am I going? (Feed Up), How am I doing? (Feed Back), and Where am I going from here? (Feed Forward).

Hattie prompt

ENEM BASE PROMPT
For each competency, provide structured feedback based on the Hattie and Timperley model: Feed-up: Where am I going? What does this competency require? (clear objective). Feedback: How am I doing? What did the student present in the text? (current diagnosis). Feedback-forward: Where to go next? What can the student do to improve? (next steps). Use clear, objective, respectful, and non-generic language. Provide examples whenever possible. The text for each competency should be limited to a maximum of three sentences. Think through the steps step by step and clearly justify your choices.

Figure 3 shows the distribution of text size for the three prompts used, and for each ENEM competence. Even though the same instruction regarding text length was provided for all three prompts: *"The text for each competency should be limited to a maximum of 3 sentences"*, there was a noticeable difference in text length for each prompt used. The figure suggests that the prompt based on the Hattie and Timperley (2007) feedback model generated longer texts. This may be related to the questions in Hattie's model, which aim to prompt students to reflect on where they are today and how they should act in the next steps of their learning development.

The average text size (in words) generated by the prompts was: 57.06 for Naive, 62.34 for Nicol and Macfarlane-Dick, and 74.78 for Hattie and Timperley. The average text size by competency was follows: C1=60.42, C2=66.48, C3=65.66, C4=65.94, and C5=65.13. Across all three

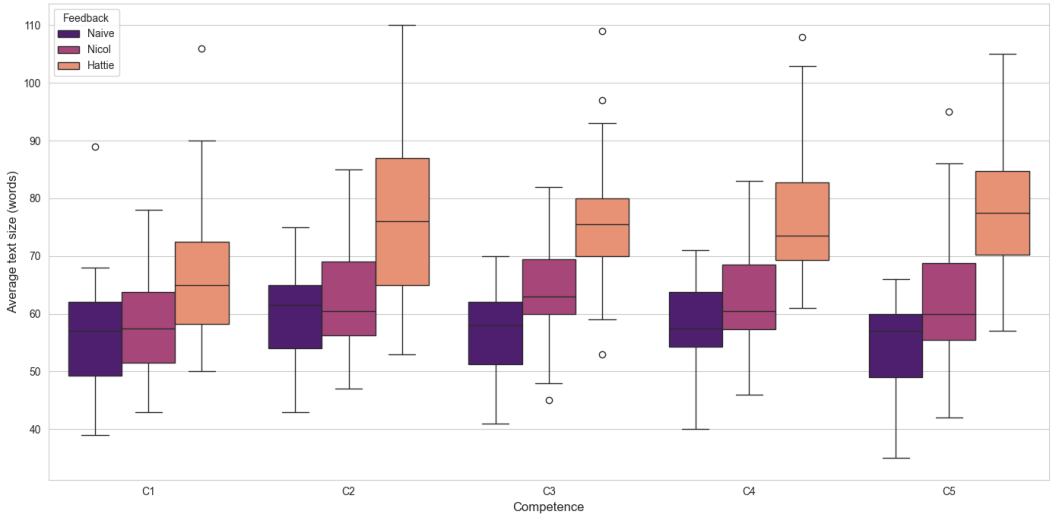

Figure 3: Boxplot with the average text size for each feedback model and competences..

types of feedback, the average text length did not differ significantly between the competencies. However, Figure 3 shows that competency C2 had longer texts for Hattie and Timperley's feedback model, and we assume that this value was aligned with one of the objectives of C2: *"apply concepts from various areas of knowledge to develop the theme"*.

Figure 3 also shows narrower interquartile ranges for competency C3, indicating lower variance in the length of the generated feedback. We hypothesize that this homogeneity is related to the nature of C3, which assesses a single, well-delimited rhetorical operation, selecting, relating, organizing, and interpreting arguments in defense of a point of view, leading the LLM to produce structurally similar diagnoses across essays, regardless of the prompt strategy.

We further verified whether the length of the generated feedback was associated with the quality of the essay being commented on. The correlation between the total number of words in the feedback and the essay's total score was negligible for the Naive ($r = 0.09$) and Nicol and Macfarlane-Dick ($r = 0.13$) prompts, and weak for the Hattie and Timperley prompt ($r = 0.31$). These results indicate that the amount of feedback produced was driven primarily by the structure imposed by each prompt strategy rather than by the proficiency level of the essay, supporting the comparability of the three conditions across the proficiency spectrum.

All feedback messages were generated in a single execution per essay–competency–prompt combination, using the model's default sampling configuration (temperature = 1.0, top-p = 0.95; API snapshot of July/2025). We acknowledge that, given the stochastic nature of LLMs, different executions could produce different surface realizations of the feedback; this design decision and its implications are discussed in Section 6.

3.4 Teachers profile

Five teachers with experience in assessing ENEM essays were recruited for this study. Table 1 presents demographic information that is relevant for interpreting feedback literacy and evaluation results, as it details the teachers' experience.

Table 1: Demographic Information of the Evaluators.

Characteristic	T1	T2	T3	T4	T5
Age range	36-45	26-35	36-45	26-35	26-35
Education	PhD	Postgraduate	PhD	PhD	Master
Teaching Level*	1, 2	0	0, 1, 2, 3	1, 2	0, 1
Years of Experience	> 15 years	> 5 years	> 15 years	> 10 years	> 10 years
ENEM Grading Experience	Excellent (4)	Good (3)	Good (3)	Excellent (4)	Excellent (4)

* (0) Elementary School - (1) High School - (2) Higher Education - (3) Postgraduate

To characterize the profile of the evaluator group and contextualize their choices, we administered the FLS to measure the feedback literacy of each participating teacher. The FLS (Lee et al., 2023) uses a 0–4 Likert scale, where 0 indicates the lowest level of agreement and 4 indicates the highest. Table 2 shows the descriptive statistics for the feedback literacy scale values for the five teachers. The values show that, based on the FLS and taking into account all dimensions, teacher T5 had the highest feedback literacy, followed by teacher T2.

The FLS items were administered through the same online form used to collect the demographic information presented in Table 1; teachers self-reported their level of agreement with each item, and the scores per dimension were computed automatically from these responses.

Table 2: Participating teachers' (T1 - T5) feedback literacy according to the Feedback Literacy Scale (FLS). Values for each dimension range from 0 to 4, where data are shown as Mean (Standard Deviation)..

Dimension	T1	T2	T3	T4	T5
Perceived Knowledge	3.60 (0.52)	3.60 (0.52)	2.30 (0.48)	3.50 (0.53)	3.80 (0.42)
Values	2.58 (0.67)	3.42 (0.51)	2.75 (0.45)	3.00 (0.74)	3.83 (0.39)
Perceived Skills	3.33 (0.98)	3.67 (0.78)	2.58 (0.51)	3.67 (0.49)	4.00 (0.0)
Average	3.11 (0.84)	3.55 (0.61)	2.61 (0.49)	3.38 (0.69)	3.85 (0.36)

3.5 Feedback Assessment by Teachers

After the feedback was generated, it was sent to the five teachers for evaluation through an online questionnaire. For each competency of each of the 30 essays, the three feedback texts (Naive, Nicol and Macfarlane-Dick, and Hattie and Timperley) were presented anonymously. For each set of three feedback texts, the teacher evaluators were asked to answer the following question: *"Which of these three feedback texts do you consider most effective for student development and learning?"* They were required to select only one option, and their choices were recorded for subsequent quantitative analysis to identify the preferred model. Each teacher evaluated 450 feedback messages generated by the LLM, as we provided them with 30 essays that were assessed in terms of five competencies (C1-C5), with each competency containing three different feedback messages (Section 3.3). The only exception was teacher T5, who was unable to complete six out of his 150 assessments (i.e, $< 1\%$) due to logistic issues. To address this issue, we applied literature suggestions for handling missing data, completing these six instances with the teacher's most frequently used answer (i.e., mode) (Robertson & Kaptein, 2016). Hence, our final evaluation set has a total of 750 instances (150 evaluations per teacher times 5 teachers).

We adopted a forced-choice design, in which teachers were required to select exactly one of the three alternatives, without the option of indicating ties or rejecting all of them. Forced-choice comparison has been validated as a reliable and efficient method for eliciting expert judgments of quality, even when the alternatives are of comparable adequacy (Marcoci et al., 2024), and we adopted it to maximize the discriminative power of the comparison by compelling evaluators to weigh the alternatives against each other rather than rating them independently. The trade-off of this design is that it cannot capture situations in which two or more feedback messages were perceived as equally adequate, or in which none was considered satisfactory; the implications of this choice for the interpretation of the agreement results are discussed in Sections 4.2 and 6.

3.6 Data Analysis

To address SRQ1 and SRQ2, we first conducted a preliminary assessment of inter-rater agreement among the participating teachers regarding which feedback model they perceived as most effective. Since more than two raters were involved, we employed Fleiss' Kappa (Fleiss, 1971), a widely used measure of agreement for categorical data. This step, recommended in methodological research (Cole, 2024; Watson & Petrie, 2010), helped determine whether raters share similar judgments or hold divergent views before analyzing the main research questions. Establishing this level of agreement provided essential context for interpreting our results, clarifying whether raters could be treated as a single combined source of judgment or whether their perspectives needed to be examined separately.

A more granular analysis was conducted to explore patterns of agreement between specific pairs of raters. For this purpose, we calculated Cohen's Kappa (Cohen, 1960) for every possible pair of teachers. Unlike Fleiss' Kappa, which provides a single statistic for the entire group, Cohen's Kappa measures the agreement between two raters while correcting for the possibility of agreement occurring by chance. This pairwise analysis allowed us to identify which teachers, if any, shared similar evaluative criteria and which diverged, providing deeper insight into the variability of the judgments observed in the overall agreement measure.

Building upon the preliminary analysis, we answered SRQ1 and SRQ2 as follows. For SRQ1, we analyzed the number of times each model (Hattie and Timperley, Nicol and Macfarlane-Dick, Naive) was selected as the most effective across all competencies. Given that our data consisted of nominal selections from multiple teachers, we employed frequency counts and Chi-square tests to assess whether the observed differences in selections were statistically significant. Additionally, to account for variability between teachers highlighted in our preliminary inter-rater agreement analysis, we examined selections at both the aggregate level and the individual teacher level. This approach ensured that we captured both overall trends and individual differences in perceived model effectiveness.

For SRQ2, we conducted a competence-level analysis of feedback model selections. Similarly, we used frequency counts and Chi-square tests of independence to test for associations between the feedback model and competence. Furthermore, building on the preliminary inter-rater agreement findings, we analyzed selections separately for each teacher across competences to identify potential individual differences. This dual-level procedure allowed us to determine both the overall influence of competence on teachers' judgments and the consistency of individual teacher preferences.

All analyses were conducted using R[1] within the RStudio environment. We adopted the conventional significance threshold of alpha = 0.05 and relied on base R functions to perform Chi-square tests. The input for these tests consisted of the frequency tables of teachers' selections described above, both at the aggregate and individual teacher levels, ensuring a straightforward and reproducible procedure in answering our SRQs.

4 Results

4.1 Distinctiveness of the Generated Feedback Messages

Before analyzing teachers' choices, we verified whether the three prompt strategies produced textually distinguishable messages, since highly similar alternatives could make the identification of a superior option harder and artificially depress agreement statistics. To this end, we computed the pairwise lexical similarity (TF–IDF cosine) between the three feedback texts generated for each of the 150 essay–competency combinations. The average similarity was 0.30 (SD = 0.08) for the Naive–Nicol pair, 0.33 (SD = 0.08) for the Naive–Hattie pair, and 0.35 (SD = 0.09) for the Nicol–Hattie pair, with maximum values not exceeding 0.68. Considering that the three messages in each set necessarily share diagnostic vocabulary, they comment on the same essay, for the same competency, under identical stylistic constraints, these low-to-moderate values indicate that the prompt strategies produced clearly distinct messages. Two secondary patterns are worth noting: the two theory-grounded models (Nicol and Macfarlane-Dick and Hattie and Timperley) were the most similar pair, as expected given their shared prescriptive structure, and competency C2 presented the lowest similarities across all pairs, suggesting that its more open-ended nature (applying concepts from multiple knowledge areas) gave the strategies greater room to diverge. Therefore, the disagreement among teachers reported in the following section cannot be attributed to the indistinguishability of the alternatives.

4.2 Concordance analysis between the teachers

We found Fleiss' Kappa value of -0.0188 when considering all competences together as an overall agreement among the five teachers. Similarly, even when we investigated the teachers' agreement for each competence, Fleiss' Kappa values were essentially the same (i.e., C1: -0.0071; C2: -0.0319; C3: 0.0064; C4: -0.0616; C5: -0.0648). These results suggest teachers' evaluations were different from each other, highlighting the importance of considering these perspectives in answering our SRQs.

The heatmap graph in Figure 4 shows the agreement between pairs of teachers using Cohen's kappa coefficient. The graph confirms the significant disagreement between teachers in most evaluations and shows reasonable agreement (Cohen's Kappa of 0.36) (Landis & Koch, 1977) between teachers T2 and T5, precisely the teachers who considered Nicol and Macfarlane-Dick's model to be the most effective.

To provide greater visual and interpretive clarity, Figure 5 breaks down the pairwise agreement analysis by competency. The figure shows that the T2–T5 pair, precisely the two teach-

[1] https://www.r-project.org/

Figure 4: Heatmap of Agreement between Pairs of Teachers (Cohen's Kappa).

ers with the highest feedback literacy scores, is the only one to exhibit above-chance agreement across all five competencies, although its strength varies considerably, ranging from weak in C1 ($\kappa = 0.16$) and C4 ($\kappa = 0.07$) to substantial in C3 ($\kappa = 0.78$) and moderate in C2 and C5 ($\kappa = 0.45$ in both). All remaining pairs fluctuate around chance level (between -0.16 and 0.19) in every competency, indicating the absence of shared evaluative criteria outside the T2–T5 dyad. Notably, C3 is both the competency with the highest T2–T5 agreement and the only one for which the overall Fleiss' Kappa was positive, reinforcing that convergence, when it occurs, is concentrated among high-literacy evaluators rather than being a general property of the panel.

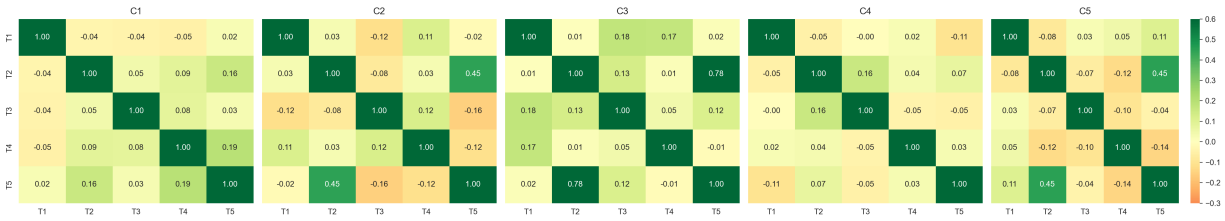

Figure 5: Pairwise agreement between teachers (Cohen's Kappa) for each ENEM competency (C1–C5)..

It is worth noting that the near-zero and negative Kappa values must be interpreted in light of the forced-choice design. When evaluators perceive two or more alternatives as similarly adequate, a plausible scenario given that all three messages were generated by the same LLM under identical stylistic constraints, the forced selection of a single option introduces variability that depresses chance-corrected agreement statistics, even in the absence of fundamentally opposed judgments. Therefore, the low agreement observed here should be read as evidence that no feedback model was unanimously superior, and that teachers' choices were systematically shaped by their individual profiles (as confirmed by the per-teacher analyses in Section 4.3 and Table 3), rather than as evidence of unreliable data.

4.3 SRQ1: Teachers' Opinions on Most Effective Feedback Models

Figure 6 presents teachers' overall evaluations of the most effective feedback models based on LLM-generated feedback messages, considering all competencies together. The figure demonstrates that Hattie's model was selected the most effective 359 times (47.86%), whereas Nicol and Macfarlane-Dick's and Naive models were selected 269 (35.86%) and 122 (16.26%) times, respectively. This difference was statistically significant according to the Chi-squared test (χ^2 = 114.50, df = 2, p-value < 0.001). Hence, these results suggest LLM-generated feedback messages based on Hattie and Timperley's model were the most effective according to teachers.

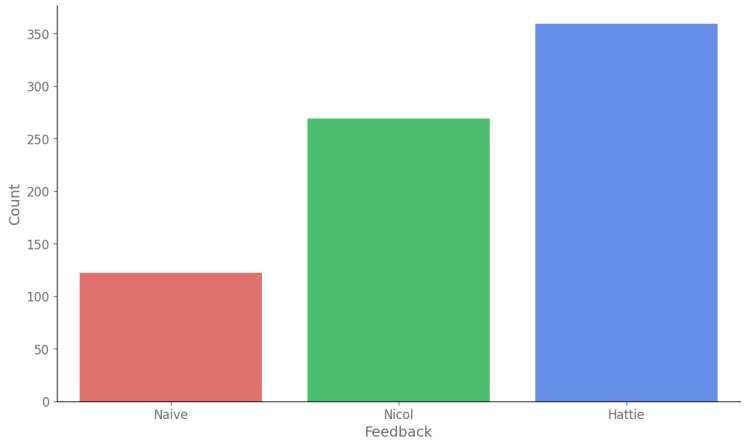

Figure 6: Barplot with the number of times each feedback model was considered the most effective by teachers, including feedback messages for all competences..

Importantly, our concordance analysis suggested that teachers' perceptions were substantially different (Section 4.2). Therefore, Figure 7 expands the previous analysis with insights from each teacher's choices. The figure shows that three of the five teachers considered Hattie and Timperley's model the most effective, although the strength of their preference (i.e., the percentage of times they selected Hattie and Timperley-based feedback) varied. Additionally, the figure also highlights that teachers T2 and T5 diverged, selecting feedback messages based on Nicol and Macfarlane-Dick's model most often. The Chi-squared test of independence revealed a statistically significant association between the feedback model and teachers (χ^2 = 272.61, df = 8, p-value < 0.001). Thereby, whereas the finding that Hattie and Timperley's model was considered the most effective remained, we found that the degree of preference varied among teachers, with a particular case in which Nicol and Macfarlane-Dick's model was considered more effective than that of Hattie and Timperley.

4.4 SRQ2: Competences' Role on Teachers' Opinions on Most Effective Feedback Models

Figure 8 presents teachers' evaluations of the most effective feedback models, based on LLM-generated feedback messages, for each competence individually. Similar to SRQ1's initial findings, the figure suggests that teachers considered feedback messages based on Hattie and Timperley's model to be the most effective. Accordingly, the Chi-squared test of independence revealed a non-statistically significant association between the feedback model and competences (χ^2 = 2.45,

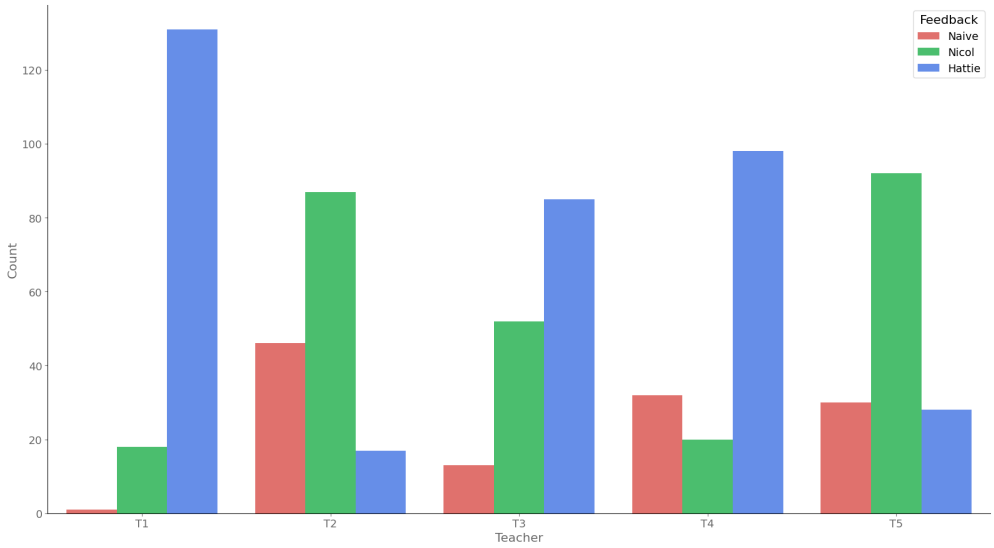

Figure 7: Barplot with the number of times each feedback model was considered the most effective by each teacher, including feedback messages for all competences..

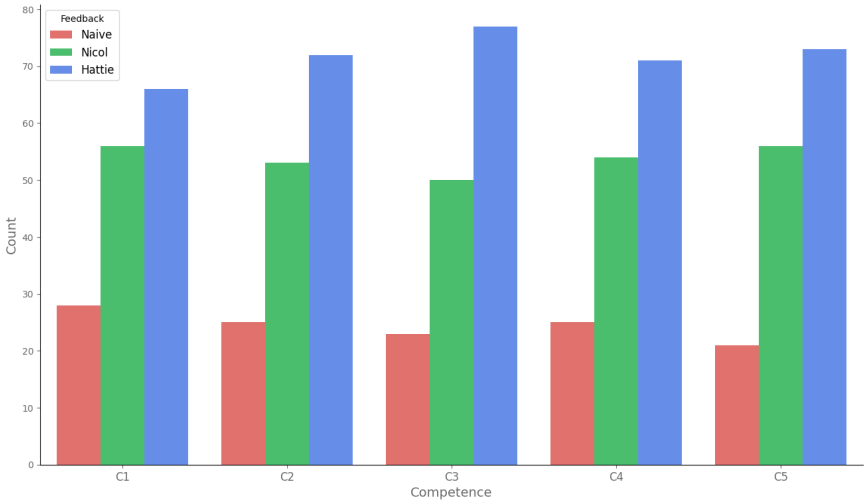

Figure 8: Barplot with the number of times each feedback model was considered the most effective for each competence..

df = 8, p-value = 0.964). Therefore, these findings suggest that teachers' preference for Hattie and Timperley's model as the most effective did not vary across the assessment competencies.

Building upon the concordance analysis (Section 4.2) and the findings of SRQ1's, we expand the previous analysis to compare each teacher's selections within each competence. Table 3 presents the results of this analysis, which corroborate SRQ1's findings. Across all competencies, three of the five teachers (T1, T3, and T4) consistently considered LLM-generated feedback based on Hattie and Timperley's model to be the most effective, although the strength of their preference varied. In contrast, teachers T2 and T5 consistently favored feedback based on Nicol and Macfarlane-Dick's model. Accordingly, the Chi-square independence tests by competence revealed statistically significant differences, further supporting the variation in counts presented in Table 3. As a result, we have evidence that, even when comparing feedback models within each

Table 3: Table of counts of the number of times each teacher selected each feedback model (Nai = Naive; Nic = Nicol; Hat = Hattie) for each competence (C1 to C5). The last row concerns Chi-square independence tests (χ^2 test) for the association between the feedback model and teachers for each competence..

	C1			C2			C3			C4			C5		
Teacher	Nai	Nic	Hat	Nai	Nic	Hat	Nai	Nic	Hat	Nai	Nic	Hat	Nai	Nic	Hat
T1	1	3	26	0	2	28	0	5	25	0	4	26	0	4	26
T2	8	20	2	9	19	2	11	15	4	11	14	5	7	19	4
T3	3	9	18	2	11	17	2	10	18	3	10	17	3	12	15
T4	10	5	15	8	6	16	4	2	24	5	4	21	5	3	22
T5	6	19	5	6	15	9	6	18	6	6	22	2	6	18	6
χ^2 test	χ^2 = 60.714, p < 0.001			χ^2 = 55.666, p < 0.001			χ^2 = 58.680, p < 0.001			χ^2 = 64.441, p < 0.001			χ^2 = 53.008, p < 0.001		

competence, there is a statistically significant association between teacher and feedback choices. It is important to highlight that the two teachers who differ from the others (i.e., T2 and T5 compared to T1, T3, and T4) are those with the highest feedback literacy (see Section 3.4). This shows that the teacher feedback literacy was a much stronger predictor of feedback preference than the nature of the task (i.e., the competency).

5 Discussion

The findings of this study highlight how LA contributes to elucidating the complex dynamics in the perceived effectiveness of LLM-generated feedback. It was observed that the notion of "best feedback" was strongly mediated by the feedback literacy of the evaluator. Three main findings emerged: (1) although the model proposed by Hattie and Timperley (2007) was preferred by the majority of teachers, there was no consensus on the most effective model; (2) the disagreement among evaluators was statistically significant and, in some cases, worse than what would be expected by chance; and (3) the preference for a feedback model was explained not by the specific ENEM competence being analyzed, but rather by teachers' feedback literacy (Carless & Winstone, 2023). These findings reinforce the literature on the situated nature of feedback effectiveness (Carless & Boud, 2018; Evans, 2013; Nicol & Macfarlane-Dick, 2006; Sadler, 1989) and suggest that personalization must extend to both learners and educators.

Moreover, our findings suggest that LLM-generated feedback informed by Hattie and Timperley (2007) was more frequently perceived as effective, corroborating previous studies that emphasize the value of structured and directive feedback frameworks (Evans, 2013; Gibbs & Simpson, 2005). Similar to the results of Shin et al. (2025), which showed that student preferences for AI- or peer-generated feedback varied depending on the context, our analysis indicates that effectiveness is not an intrinsic property of the feedback message but emerges from the interaction between theoretical structure and evaluator interpretation. This aligns with research by Xavier et al. (2025) and Rüdian et al. (2025), who found that the perceived usefulness of feedback is contingent on contextual and social factors. Thus, our results contribute to the growing body of evidence that effective LLM-based feedback must integrate pedagogical theory (Carless & Boud, 2018; Hattie & Gan, 2011; Nicol & Macfarlane-Dick, 2006; Sadler, 1989) to ensure both clarity and actionability.

A key contribution of this study is the demonstration that feedback preferences were systematically related to teachers' feedback literacy profiles. Teachers with higher literacy scores, particularly in the "Values" and "Skills" dimensions, tended to prefer the model of Nicol and Macfarlane-Dick (2006), which emphasizes student agency, dialog, and opportunities for self-regulation. This resonates with findings by Lin et al. (2023), who demonstrated the benefits of dialogic feedback for student engagement, and with studies highlighting the role of feedback literacy in mediating uptake (Carless & Winstone, 2023). Conversely, teachers with lower feedback literacy were drawn to the more directive structure of Hattie and Timperley (2007), highlighting prior evidence that prescriptive feedback can serve as a scaffold for both novice learners and novice evaluators (Hattie & Gan, 2011). These patterns highlight that feedback literacy can be a decisive factor in shaping how evaluators perceive LLM-generated feedback.

One relevant result emerging from this study is that there is no "best" feedback model; rather, effectiveness depends on the evaluator's pedagogical profile. This mirrors findings in prior LA research, where variability in engagement has been observed across both students and teachers (Rüdian et al., 2025; Shin et al., 2025; Xavier et al., 2025). The statistically significant disagreement observed in our study shows the limitations of using a single "gold standard" when training or benchmarking AI systems (Gwet, 2014). Instead, variability must be explicitly accounted for (Yun et al., 2025).

From both a practical and theoretical perspective, these results suggest that personalization in feedback systems must extend beyond students to include teachers. LA-based platforms could incorporate mechanisms that allow educators to select or adapt feedback models aligned with their feedback literacy profile (Carless & Boud, 2018; Yun et al., 2025). Such personalization is consistent with recent research advocating for human–AI collaboration in assessment (Hutt et al., 2024; Xavier et al., 2025; Xiao et al., 2025). Theoretically, this reinforces calls to embed established pedagogical frameworks in the design of AI-generated feedback (Hattie & Gan, 2011; Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Sadler, 1989), ensuring that analytic outputs are technically accurate and pedagogically meaningful. Practically, the findings indicate that LLMs should serve as scaffolds that provide multiple theory-informed feedback options (Cavalcanti et al., 2019, 2020; Pardo et al., 2019), which teachers can validate and adapt. In doing so, the LA community can move toward feedback ecosystems that are adaptive, and context-sensitive, advancing both the scalability and quality of educational feedback.

6 Limitations and Future Direction for Research

Despite the methodological rigor employed, this study has limitations that should be considered when interpreting the results, which also open avenues for future research. First, the sample of evaluators was limited to five teachers. Although all were experts with extensive experience, this context restricts the statistical generalization of the findings to the broader population of Portuguese language teachers. The feedback literacy profiles and preferences observed in this group may not be representative of the full diversity of existing pedagogical views.

A further limitation concerns the composition of the essay sample. As a consequence of the random sampling procedure, the topic distribution was uneven, with a single essay on the Fake News topic. Although the unit of analysis in this study was the feedback message generated per

competency rather than the essay topic, future studies should adopt stratified sampling by topic to support thematic representativeness analyses.

Second, the nature of the evaluation task consisted of a forced choice among three pre-generated feedback options. This experimental design measures the evaluators' perception of effectiveness or preference, which is a valuable indicator; however, it does not directly measure the actual impact of this feedback on the student's learning process. Future studies could investigate how students effectively use these different types of feedback to revise their texts, connecting the expert's perception with the learner's practice.

The forced-choice protocol also imposes interpretive limitations on the agreement analysis. Teachers could not indicate that more than one message was equally adequate, nor that none was satisfactory, which may have introduced noise into the agreement statistics (Section 4.2) and prevents conclusions about the absolute pedagogical quality of the generated feedback. The results should therefore be interpreted as relative preferences among the three models. Future designs could combine forced choice with rating scales and explicit "none of the above" options to disentangle relative preference from absolute adequacy.

Furthermore, this study did not include a condition of written feedback by humans, nor did it collect teachers' qualitative perceptions of the usefulness and pedagogical accuracy of the messages generated by the LLM. Comparing the theory-based LLM feedback with feedback produced by expert teachers, and supplementing the choice data with thought verbalization protocols or interviews, would allow future research to assess whether the generated messages reach a pedagogical level comparable to that of experts and identify their main shortcomings.

A third limitation lies in the measurement of feedback literacy, which was based on a self-report instrument (the FLS scale). As noted in the literature, teachers' perceptions of their own competencies may not perfectly align with their classroom practices (Lee et al., 2023). Therefore, although the profiles drawn are fundamental to the analysis in this study, they represent the participants' self-assessment, not a direct observation of their feedback practices.

An additional limitation concerns the generation procedure: each feedback message was produced through a single LLM execution. Since LLMs are stochastic models, the same prompt may yield different messages across runs, and a single execution may not capture the full distribution of outputs of each prompt strategy. Although all three strategies were subject to the same condition, which preserves the fairness of the comparison, future studies should generate multiple completions per prompt and assess the stability of teachers' preferences across executions.

Finally, the results are intrinsically linked to the specific LLM and the prompt engineering used to generate the feedback. The quality, style, and accuracy of the generated texts can vary significantly between different models (e.g., GPT-3.5, GPT-4, Llama) or with different prompt strategies. Furthermore, LLMs have inherent limitations, such as the possibility of "hallucinations" or the generation of false positives, which, although mitigated by expert evaluation in this study, remain factors to consider in feedback automation. Thus, the findings reflect the performance of a specific technological configuration and should not be generalized to all AI systems in LA.

References

Alcock, S., Rienties, B., Aristeidou, M., & Kouadri Mostéfaoui, S. (2024). How do visualizations and automated personalized feedback engage professional learners in a learning analytics dashboard? *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 316–325. https://doi.org/10.1145/3636555.3636886 [GS Search].

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 [GS Search].

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 [GS Search].

Carvalho, R., Lins, L. F., Rodrigues, L., Miranda, P., Oliveira, H., Cordeiro, T., Bittencourt, I. I., Isotani, S., & Mello, R. F. (2024). Exploring NLP and embedding for automatic essay scoring in the Portuguese. *International Conference on Artificial Intelligence in Education*, 228–233. https://doi.org/10.1007/978-3-031-64315-6_18 [GS Search].

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027 [GS Search].

Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., & Gaševic, D. (2020). How good is my feedback? a content analysis of written feedback. *Proceedings of the 10th International Conference on Learning Analytics and Knowledge*. https://doi.org/10.1145/3375462.3375477 [GS Search].

Cavalcanti, A. P., Mello, R. F. L., Rolim, V., André, M., Freitas, F., & Gaševic, D. (2019). An analysis of the use of good feedback practices in online learning courses. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, *2161*, 153–157. https://doi.org/10.1109/ICALT.2019.00061 [GS Search].

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104 [GS Search].

Cole, R. (2024). Inter-rater reliability methods in qualitative case study research. *Sociological Methods & Research*, *53*(4), 1944–1975. https://doi.org/10.1177/00491241231156971 [GS Search].

Dawson, S., Pardo, A., Salehian Kia, F., & Panadero, E. (2023). An integrated model of feedback and assessment: From fine grained to holistic programmatic review. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 579–584. https://doi.org/10.1145/3576050.3576074 [GS Search].

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of educational research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350 [GS Search].

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378. https://doi.org/10.1037/h0031619 [GS Search].

Galhardi, L., Herculano, M. F., Rodrigues, L., Miranda, P., Oliveira, H., Cordeiro, T., Bittencourt, I. I., Isotani, S., & Mello, R. F. (2024). Contextual features for automatic essay scoring in Portuguese. *International Conference on Artificial Intelligence in Education*, 270–282. https://doi.org/10.1007/978-3-031-64315-6_23 [GS Search].

Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning [Link]. *Learning and teaching in higher education*, (1), 3–31. [GS Search].

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC. [GS Search].

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In *Handbook of research on learning and instruction* (pp. 263–285). Routledge. https://doi.org/10.4324/9780203839089-22 [GS Search].

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487 [GS Search].

Hutt, S., DePiro, A., Wang, J., Rhodes, S., Baker, R. S., Hieb, G., Sethuraman, S., Ocumpaugh, J., & Mills, C. (2024). Feedback on feedback: Comparing classic natural language processing and generative AI to evaluate peer feedback. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 55–65. https://doi.org/10.1145/3636555.3636850 [GS Search].

Jansen, T., Horbach, A., & Meyer, J. (2025). Feedback from generative AI: Correlates of student engagement in text revision from 655 classes from primary and secondary school. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 831–836. https://doi.org/10.1145/3706468.3706494 [GS Search].

Kucia, F. J., Chakraborty, A., & WrĂłblewska, A. (2026). LLM essay scoring under holistic and analytic rubrics: Prompt effects and bias. *arXiv preprint arXiv:2604.00259*. https://doi.org/10.48550/arXiv.2604.00259 [GS Search].

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310 [GS Search].

Lee, I., Karaca, M., & Inan, S. (2023). The development and validation of a scale on l2 writing teacher feedback literacy. *Assessing Writing*, *57*, 100743. https://doi.org/10.1016/j.asw.2023.100743 [GS Search].

Lin, J., Dai, W., Lim, L.-A., Tsai, Y.-S., Mello, R. F., Khosravi, H., Gasevic, D., & Chen, G. (2023). Learner-centred analytics of feedback content in higher education. *LAK23: 13th international learning analytics and knowledge conference*, 100–110. https://doi.org/10.1145/3576050.3576064 [GS Search].

Marcoci, A., Webb, M. E., Rowe, L., Barnett, A., Primoratz, T., Kruger, A., Karvetski, C. W., Stone, B., Diamond, M. L., Saletta, M., van Gelder, T., Tetlock, P. E., & Dennis, S. (2024). Validating a forced-choice method for eliciting quality-of-reasoning judgments. *Behavior Research Methods*, *56*(5), 4958–4973. https://doi.org/10.3758/s13428-023-02234-x [GS Search].

Marinho, J. C., Anchiêta, R. T., & Moura, R. S. (2021). Essay-BR: a Brazilian Corpus of Essays. *arXiv preprint arXiv:2105.09081*. https://doi.org/10.5753/dsw.2021.17414 [GS Search].

Mello, R. F., Anthony, L., Lobo, J., Ribeiro, F. G. C., Xavier, C., Costa, N. T., Gasevic, D., & Rodrigues, L. (2025). Empowering equitable learning with LLMs: Enhancing writing skills in low-resource contexts. *European Conference on Technology Enhanced Learning*, 183–197. https://doi.org/10.1007/978-3-032-03870-8_13 [GS Search].

Mello, R. F., Freitas, E., Cabral, L., Pereira, F. D., Rodrigues, L., Rakovic, M., Raniel, J., & Gaševic, D. (2024). Words of wisdom: A journey through the realm of natural language processing for learning analytics–a systematic literature review. *Journal of Learning Analytics*, *11*(3), 82–105. https://doi.org/10.18608/jla.2024.8403 [GS Search].

Monteiro, E., & Barreto, R. (2022). Um método baseado na teoria da resposta ao item para avaliação e feedback automático no contexto do ENEM. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 255–266. https://doi.org/10.5753/rbie.2021.29.0.746 [GS Search].

Nguyen, H., & Park, S. (2025). Providing automated feedback on formative science assessments: Uses of multimodal large language models. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 803–809. https://doi.org/10.1145/3706468.3706480 [GS Search].

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090 [GS Search].

Oliveira, H., Mello, R. F., Miranda, P., Alexandre, B., Cordeiro, T., Bittencourt, I. I., & Isotani, S. (2023). Classificação ou regressão? Avaliando coesão textual em redações no contexto do ENEM. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 1226–1237. https://doi.org/10.5753/sbie.2023.234516 [GS Search].

Oliveira, H., Mello, R. F., Miranda, P., Batista, H., Silva Filho, M. W., Cordeiro, T., Bittencourt, I. I., & Isotani, S. (2025). A benchmark dataset of narrative student essays with multi-competency grades for automatic essay scoring in Brazilian Portuguese. *Data in Brief*, *60*, 111526. https://doi.org/10.1016/j.dib.2025.111526 [GS Search].

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138. https://doi.org/10.1111/bjet.12592 [GS Search].

Pimentel, E. P., Silva, L. D. L., Takeuti, R. H., & Braga, J. C. (2025). Abordagem com LLM para geração automatizada de feedback no ensino de programação de computadores. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 590–602. https://doi.org/10.5753/sbie.2025.12543 [GS Search].

Robertson, J., & Kaptein, M. (2016). An introduction to modern statistical methods in hci. In *Modern statistical methods for hci* (pp. 1–14). Springer. https://doi.org/10.1007/978-3-319-26633-6_1 [GS Search].

Rüdian, S., Podelo, J., Kužílek, J., & Pinkwart, N. (2025). Feedback on feedback: Student's perceptions for feedback from teachers and few-shot LLMs. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 82–92. https://doi.org/10.1145/3706468.3706479 [GS Search].

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714 [GS Search].

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge. [GS Search].

Shibani, A., Knight, S., & Buckingham Shum, S. (2019). Contextualizable learning analytics design: A generic model and writing analytics evaluations. *Proceedings of the 9th international conference on learning analytics & knowledge*, 210–219. https://doi.org/10.1145/3303772.3303785 [GS Search].

Shin, I., Hwang, S. B., Yoo, Y. J., Bae, S., & Kim, R. Y. (2025). Comparing student preferences for AI-generated and peer-generated feedback in AI-driven formative peer assessment. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 159–169. https://doi.org/10.1145/3706468.3706488 [GS Search].

Singh, A., Brooks, C., Wang, X., Li, W., Kim, J., & Wilson, D. (2024). Bridging learnersourcing and AI: Exploring the dynamics of student-AI collaborative feedback generation. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 742–748. https://doi.org/10.1145/3636555.3636853 [GS Search].

Van Campenhout, R., Kimball, M., Clark, M., Dittel, J. S., Jerome, B., & Johnson, B. G. (2024). An investigation of automatically generated feedback on student behavior and learning. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 850–856. https://doi.org/10.1145/3636555.3636901 [GS Search].

Watson, P., & Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology*, *73*(9), 1167–1179. https://doi.org/10.1016/j.theriogenology.2010.01.003 [GS Search].

Watts, F. M., Dood, A. J., & Shultz, G. V. (2023). Automated, content-focused feedback for a writing-to-learn assignment in an undergraduate organic chemistry course. *LAK23: 13th international learning analytics and knowledge conference*, 531–537. https://doi.org/10.1145/3576050.3576053 [GS Search].

Xavier, C., Costa, N. T., Valdo, A. K., Alves, G., Rodrigues, L., Rodrigues, L. F., Silva, M., Neto, R., Falcão, T. P., Gasevic, D., & Mello, R. F. (2025). Human teacher vs. LLM-generated feedback in secondary education: A comparative study on student perceptions. *European Conference on Technology Enhanced Learning*, 534–548. https://doi.org/10.1007/978-3-032-03870-8_36 [GS Search].

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025). Human-AI collaborative essay scoring: A dual-process framework with LLMs. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 293–305. https://doi.org/10.1145/3706468.3706507 [GS Search].

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370 [GS Search].

Yun, J., Nie, A., Brunskill, E., & Demszky, D. (2025). Exploring the benefit of customizing feedback interventions for educators and students with offline contextual multi-armed bandits. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 944–949. https://doi.org/10.1145/3706468.3706551 [GS Search].