Teacher-AI Interaction in the Selection of Target Texts

Pier Giuseppe Rossi[1,†], Lorella Giannandrea[1,*,†], Francesca Gratani[1,†], David Scaradozzi[2,†] and Laura Screpanti[1,†]

[1] *University of Macerata, Via Luigi Bertelli 1, 62100 MC, Macerata, Italy*
[2] *Università Politecnica delle Marche, Via Brecce Bianche 12, 60131 AN, Ancona, Italy*

Abstract

Even if the importance of feedback in educational training is widely recognized, delivering timely and effective feedback is not always sustainable, especially in educational settings. To address this challenge, this paper presents the use of an Artificial Agent in both the correction process and online feedback delivery. This approach aims to facilitate the delivery of recursive feedback, thus enhancing the overall learning process. The research design investigates three key phases of the assessment process: test preparation, test evaluation and analysis, and recursive feedback delivery. This paper focuses on the second phase, specifically how the artificial agent and the human agent interact in the correction of student assessments, and details the procedure used to select target texts. The evaluation procedure has been developed and tested. The trial was carried out in the 2023/24 academic year using the open answers submitted by 263 first-year students enrolled in the Master's Degree course in Primary Education at the University of Macerata. The adopted methods allowed for reliable data on approximately 80% of the student submissions. Additionally, the system highlights which texts still require further evaluation and indicates the uncertainty of different blocks, identifying those with more reliable evaluations. Although there is still a long way to go and many developments are possible, the results obtained so far are promising for the adoption of models based on recursive interaction between artificial agent and human agent to make the widespread use of feedback in daily university practice more sustainable.

Keywords

Effective Feedback, Artificial Intelligence, Human-AI interaction, Text-based responses, Post-secondary Education, LLMs

1. Introduction

The importance of feedback in educational training is widely recognized [1], [2]. However, in practice, delivering timely and effective feedback is not always sustainable, especially in educational settings. Feedback is most effective when provided shortly after an assessment. Yet, when dealing with large classes of over a hundred students, particularly when assessments involve open-ended responses, the grading process becomes so time-consuming that it conflicts with the demands of teaching. AI (Artificial Intelligence), particularly through Generative AI (GenAI) and Large Language Models (LLMs), has the potential to significantly reshape the classroom experience, impacting not only how feedback is delivered but also how teaching and learning evolve over time. These AI-driven systems can take on routine tasks like grading and feedback, allowing teachers to focus more on complex instructional roles such as mentoring and facilitation [3]-[5]. Over time, this shift could enhance student outcomes by enabling more timely and personalized feedback, fostering greater student autonomy, and encouraging deeper engagement [6]. The use of predictive models in AI, as demonstrated in various studies, shows that real-time data analysis and feedback loops could also enhance formative assessments, allowing students to understand their learning progress dynamically

AIxEDU: 2nd International Workshop on Artificial Intelligence Systems in Education, November 25-28, 2024, Bolzano, Italy

[*] Corresponding author.

[†] These authors contributed equally.

piergiuseppe.rossi@unimc.it (P. G. Rossi); lorella.giannandrea@unimc.it (L. Giannandrea); f.gratani@unimc.it (F. Gratani); d.scaradozzi@univpm.it (D. Scaradozzi); laura.screpanti@unimc.it (L. Screpanti)

https://orcid.org/0000-0001-9801-6307 (P. G. Rossi); https://orcid.org/0000-0002-1169-4795 (L. Giannandrea); https://orcid.org/0000-0003-2974-0101 (F. Gratani); https://orcid.org/0000-0001-9346-2113 (D. Scaradozzi); https://orcid.org/0000-0003-4765-8427 (L. Screpanti)

[7]. These systems can identify gaps in understanding and adapt teaching strategies accordingly, supporting long-term improvements in both student performance and teacher efficiency [6].

The paper presents the strategy to utilize an Artificial Agent (AA) in the correction process and online modes for the delivery of recursive feedback. This approach aims to facilitate the delivery of recursive feedback, thus enhancing the overall learning process.

The work presented in this paper refers to the research carried out in the PRIN AI&F "Artificial Intelligence and Feedback for Effective Learning" project, in collaboration with teams from the Universities of Bari and Padua. It is part of research into the development of systems for the evaluation of open-ended student responses [8]-[9].

The research design investigates three key phases of the assessment process: test preparation, test evaluation and analysis, and recursive feedback delivery. In the first phase, test preparation, the AA collaborates with the human agent (HA) to prepare and organize the test. In the second phase, test evaluation, once the test is completed, it is assessed through an interactive process between the AA and HA. The third phase involves the communication of results to the student, where a recursive interaction between the student and the teacher takes place. After reviewing the teacher's feedback, the student can provide comments, to which the teacher can respond, fostering an iterative dialogue.

This paper focuses on the second phase, specifically how AA and HA interact in the correction of student assessments. The correction process can be structured as a classification task, which involves grouping student submissions into categories, or as a ranking, where submissions are arranged in order based on a specified parameter. In particular, the procedures used to select Target Texts (TTs) will be detailed. TTs are those texts that the AA uses as reference points to compare against student submissions and identify degrees of similarity. These procedures are directly related to the machine learning process, specifically the steps required to train the system.

It is important to note that within the context of this study the AA is not trained on the whole content of the course in a single instance. Instead, each task was designed to focus on specific content areas or particular aspects of a problem. The distinction between what is considered correct or incorrect is not absolute but depends on the depth of study decided for a specific educational context.

2. The AI model

In the initial phase of the experimental process, the goal was to classify student texts. For the classification of open-ended responses, a variant of the BERT (Bidirectional Encoder Representations from Transformers) model was used [10]. This model belongs to the Transformer family, introduced by Vaswani et al. (2017) [11], and is known for its use of the self-attention mechanism, which allows for the generation of contextual word representations by considering both individual meanings and the context in which they appear. The choice of this model, trained on a large corpus of Italian texts, was motivated by the need to handle the linguistic and cultural nuances of the Italian language, including typical syntactic structures and idiomatic expressions. Compared to multilingual or general-purpose models, the selected model (dbmdz/bert-base-italian-cased) offers a more accurate understanding of open-ended responses in Italian, ensuring better performance in the educational context, where answers may contain non-standard phrases or technical terms.

The embeddings for student responses and reference texts are generated through this model. Each response is tokenized using BERT's tokenizer, which splits the text into words or subwords. Dense, high-dimensional numerical vectors, the embeddings, are then computed for each token of the response. The same process is applied to the reference texts (TT), thus allowing the comparison between TTs and students' responses.

Recent studies indicate that embeddings generated by large language models (LLMs) are extremely effective in capturing complex linguistic structures, with BERT demonstrating superiority among lightweight models in terms of both accuracy and computational efficiency [12]. In this context, BERT's ability to accurately represent semantic relationships enables precise evaluation of open-ended responses. It is worth noting that this evaluation, conducted in May 2024, may soon become outdated due to the rapid pace at which new models are being introduced.

Finally, to assess the similarity between the student's response and the TT, cosine similarity is used, a measure that quantifies the closeness of two vectors in high-dimensional space [13].

3. The dataset

To date, the evaluation procedure has been developed and tested with two trials. In the first trial, the construction of the dataset for the procedures that will be illustrated later was carried out using the answers submitted by 263 first-year students enrolled in the Master's degree course in Primary Education, attending the course "Foundations of Teaching and Learning" for the 2023/24 academic year. The assessment consisted of three questions:

Q1. Describe what it means to view activity as a coupling between the subject and the environment.

Q2. The following situations can generate conflict: 1) perturbations in the form of resistance, 2) perturbations in the form of gaps. Explain what each situation means and provide an example.

Q3. Describe appropriation, specifying the triple integration of in-situation, in-corporation, and in-culturation.

The students' responses were collected digitally through an online form. Initially, the task was evaluated by a human evaluator (HA1) using a rubric with a grading scale from 1 to 10. After the first tests, it became evident that in some cases the discrepancies between the Artificial Agent (AA) and the human evaluator were due to subjective assessments by HA1. Consequently, a second evaluator (HA2) was asked to review the evaluations.

At a later stage, a third procedure was adopted: all texts were analyzed separately by three evaluators who had shared a common rubric. The cases where there was no agreement between the three evaluations (which exceeded 35%) were then reviewed to reach a consensus. The fact that the non-concordant evaluations exceeded 35% highlighted the objective difficulty in achieving full agreement, even among human evaluators.

The comparison with the results proposed by the AA led to a deep reflection on correction methodologies and the concept of objectivity. As a result of these considerations, the grading scale was adjusted from 10 levels to 4 levels.

4. Procedures

Students' answers were evaluated using different methods, specifically by comparing them with different TT. The goal in exploring different procedures was to identify an effective strategy for constructing the TTs that will tune the AA and therefore will impact the evaluation of student submissions. This paper describes the process that eventually led to the final procedure, which was reached through sequential refinements over 11 tests.

4.1. Firsts steps of the research

In **test 1**, students' answers were divided into two groups: the first group (Block A) served as TT used to train the AA, while the second group (Block B) consisted of the texts that the AA, once trained, had to evaluate independently. The score assigned by the AA was then compared to the score assigned by the human evaluator (HA). The difference between the evaluations of Block B performed by the HA and the AA is defined as Delta, and this value was used as the main parameter for assessing the effectiveness of the procedure.

In test 1, three trials were conducted: in the first trial, Block A consisted of 50 texts from the total of 263, the second trial included 100 texts, and the third trial included 160 texts. The selection of those texts was random. Table 1 presents the results obtained in trial 3 (Block A used 160 TT) for the three questions.

The best results were obtained when using 160 tasks as TT, with 100 texts being evaluated. The analysis of Q3 provided the most acceptable results (71% of cases showed a maximum deviation of one point from the human evaluations on a scale of 10 levels). This difference was attributed to the nature of the three questions. However, correcting 160 students' answers to achieve this result is not entirely acceptable since it is neither sustainable nor aligned with the desired objective. By further analyzing Q3 to understand where the values were mostly in agreement or disagreement, it was noted that when using a 10-level scale, a difference of 1 or 2 points between evaluators analyzing the same task was very common and difficult to overcome. The first conclusion led to a modification in the number of levels, as previously reported.

Table 1
Results of Test 1: for each question, the table shows the percentage of AA results matching the HA scores (Delta 0), or differing for 1, 2, 3 or 4 points (Delta 1, Delta 2, Delta 3, Delta 4)

	Q1 [%]	**Q2 [%]**	**Q3 [%]**
Delta 0	18.00	17.00	30.00
Delta 1	29.00	21.00	41.00
Delta 2	27.00	20.00	16.00
Delta 3	18.00	13.00	9.00
Delta 4	4.00	16.00	1.00

From test 1 it emerged that data with higher Similarity scores were more reliable than others (see Table 2). Overall, the average value of Delta is 0.61, whereas a Similarity score higher than 0.975 resulted in an average Delta of 0.26. Similarly, the number of evaluations with Delta 0 (i.e., where the AA's evaluation matched that of the HA) was 61% in the total sample, rising to 74% for Similarity scores above 0.975.

Table 2
Results of Test 1 after mapping the 10 levels into 4 levels: distribution of delta score versus similarity score ranges.

Similarity score	**No. of Texts**	**Mean Delta**	**Delta 0 [abs. value]**	**Delta 0 [%]**	**Delta 1 [abs. value]**	**Delta 1 [%]**	**Delta 2 & 3 [abs. value]**
total	100	0,61	61	61%	32	32%	7
0-0,959	12						2
0,96 - 1	88	0,43	56	64%	27	31%	5
0,965 - 1	85	0,37	56	66%	25	29%	4
0,97 - 1	66	0,31	45	68%	20	30%	1
0.975 - 1	43	0,26	32	74%	11	26%	0

Also, when student responses were shorter (i.e., when the character count was lower than the mean value by an amount equivalent to the standard deviation), the AA often overestimated the score. In fact, the instances where there was a Delta of 3 between the AA and the HA were predominantly linked to these shorter responses. Furthermore, some students included examples in their answers, which were rated positively by HA if relevant, even if the prompt did not explicitly request them. However, for the AA, these examples were not only irrelevant but could skew the evaluation. This is because a response with an example and one without, even if similar in content, would not show the same similarity to TT.

To examine the effect of examples, **test 2** used only TTs without examples. This test confirmed the hypothesis, though the difference in Delta was not remarkable (the average Delta decreased from 0.86 to 0.76). Nevertheless, it highlighted the importance of the structure of the TT. The AA does not distinguish between key and minor words that contribute to the overall meaning, particularly in short texts and when the number of available TTs is limited. Redundant words in the TT can affect the similarity measure. This result is evident in Table 3, where providing 20 TTs without examples resulted in a decrease in the average Delta from 0.86 to 0.76.

Table 3
Results of Test 2

	No. instances	Mean Delta
Text with examples (N=20)	243	0,86
Text without examples (N=20)	243	0,76

After the initial tests, it was decided to use a grading scale with four levels instead of the original ten. Additionally, a new correction of all tasks was carried out by HA, as previously described, with increasing focus on the selection of TT.

In **test 3** the TTs were no longer randomly selected as happened in test 1. Instead, concise texts with few redundancies and clearly fitting into one of the four levels were chosen. Sets of 16, 24, and 48 TTs were provided, and each group contained an equal number of texts from levels 1, 2, 3, and 4. The results obtained from 247 instances from the dataset (i.e., 263 total tasks minus the 16 used as TTs in the first trial of test 3) showed an increased agreement between HA and AA agreement compared to previous tests (see Table 4). Additionally, there was no substantial difference between the test using 16 TTs and the one using 48 TTs (see Table 5). In fact, the 16 TTs in the first trial were much more homogeneous and consistent than the 48.

Table 4
Results of Test 3

	Similarity range	No. instances	Mean Delta	Delta 0 [abs. value]	Delta 0 [%]	Delta 1 [abs. value]	Delta 1 [%]
16 TT	Total	245	0,82	96	39	120	49
	0,975 – 1	46	0,65	21	46	21	46
	4th quartile	62	0,45	36	58		
24 TT	Total	237	0,67	86	36	121	52
	0,975 – 1	51	0,65	20	39	31	61
	4th quartile	59	0,58	31	52		
48 TT	Total	213	0,6	89	42	106	50
	0,975 – 1	57	0,55	27	47	30	53
	4th quartile	54	0,5	27	50		

Table 5
Results of Test 3: difference between 16 TT and 48 TT on the whole dataset

	No. instances	Mean Delta	Mean similarity	Delta 0	Delta 1	Delta 2	Delta 3	Delta 0 [%]	Delta 1 [%]
16 TT	246	0,676	0,961	105	119	21	2	43	48
48 TT	213	0,664	0,961	90	106	18	0	42	49

Moreover, once the results were ordered and divided into four quartiles according to the similarity score values, it could be observed that where the similarity measure was higher, the results agreed more with the score assigned by HA and thus could be considered substantially more reliable (see Table 6). The findings also indicated greater sustainability, as 16 TTs were used instead of the 160 used in test 1, while achieving similar results, with an average Delta that was not much larger (moving from an average Delta of 0.61 in test 1 with 160 TTs to an average Delta of 0.82 with 16 TTs and an average Delta of 0.60 with 48 TTs) (see Table 4). From this point onward, the focus shifted away from the Similarity threshold (e.g., 0.975) and attention was instead directed to quartile-based analysis.

Table 6

Results of Test 3: difference between 16 TT and 48 TT on the 4th quartile

	No. instances	Mean Delta	Delta 0	Delta 1	Delta 2	Delta 3	Delta 0 [%]	Delta 1 [%]
16 TT	62	0,468	36	24	1	1	58	39
48 TT	54	0,426	31	12	0	0	58	42

An additional finding emerged from test 3. As shown in Table 7 the AA better identify the highest-quality submissions, particularly those in level 4, where it correctly identifies 72% of the submissions without error. Furthermore, a further attempt using only the 12 level-4 TTs yielded results very similar to those obtained in the trial where 48 TTs were used (12 from level 1, 12 from level 2, 12 from level 3, and 12 from level 4). This further confirms the central role of level-4 TTs. Additionally, as shown in Tables 7 and 8, the system discriminates more easily between submissions from levels 1 and 4.

Table 7

Results of Test 3: distribution of delta score depending on the 4 levels

	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
Delta 0	6	8	36	**55**	38%	21%	31%	**72%**
Delta 1	7	15	77	20	44%	38%	66%	26%
Delta 2	1	16	3	1	6%	41%	3%	1%
Delta 3	2	0	0	0	12%	0%	0%	0%
total	16	39	116	76				

Table 8

Results of Test 3: distribution of delta score depending on the 4 levels

	Level 1	Level 2	Level 3	Level 4
1st quartile	75%	46%	15%	8%
2nd quartile	25%	25%	31%	12%
3rd quartile	0%	21%	28%	28%
4th quartile	5%	8%	26%	42%
	100%	100%	100%	100%

4.2. Target texts using textbooks or teachers' notes

The previous tests raised the following question: how can the instructor select the 16 significant Target Texts (TTs) without reviewing all the tasks? It became clear that random selection is not viable, meaning the instructor would need to correct a significant number of tasks to extract the 16 most suitable formulations for further analysis. It was considered that, as an initial step, the TT could be based on a passage from the course manual that addresses the topics covered in the question being analyzed.

In **test 4**, three TTs were used: (A) The entire paragraph related to the test, (B) A selected portion of the paragraph, and (C) Only the passages specifically related to the question. Unlike previous tests, these TTs did not allow for the classification of texts (in previous trials, TTs were provided for all four levels). Results are provided in Table 9, 10 and 11.

Table 9
Results of Test 4: TT drawn from the textbook

	A		B		C		A+B+C	
	[abs. val.]	*[%]*	*[abs. val.]*	*[%]*	*[abs. val.]*	*[%]*	*[abs. val.]*	*[%]*
Delta 0	115	44	108	41	141	54	135	51
Delta 1	117	44	121	46	102	39	108	41
Delta 2	26	10	29	11	18	7	18	7
Delta 3	5	2	5	2	2	1	2	1

It should be noted that by providing the extracted passage from the manual as the TT, the AA could only rank the texts, not classify them, and the ranking was based on similarity. According to the HA assessments, there were 20 responses in level 1, 38 in level 2, 118 in level 3, and 88 in level 4. Therefore, it was decided to instruct the AA to assign the 20 responses with the lowest similarity scores to level 1, the next 38 responses to level 2, the following 118 responses to level 3, and the 88 responses with the highest similarity scores to level 4. This assignment was possible only because the number of tasks assigned to each level by the HA was already known, but this step was crucial to understand whether there was consistency between the similarity score and the HA's evaluations. In other words, the goal was to determine if the AA's ranking, based on the similarity between the TT (from the manual) and the students' answers, aligned with the HA's evaluations.

Tables 9 and 10 present the results of the queries conducted using each TT, as well as the query performed with a TT created by aggregating the three TTs. Table 9 shows the overall data, while Table 10 breaks down the results by quartile. Table 11 shows the average Delta values obtained in the four quartiles.

Table 10
Results of Test 4: A-B-C TT drawn from the textbook

		A		B		C		A+B+C	
Quartile 1	*delta 0*	18	27%	15	23%	25	**38**%	23	35%
	delta 1	32	48%	33	50%	29	44%	31	47%
	delta 2	13	20%	15	23%	10	15%	10	15%
	delta 3	3	5%	3	5%	2	3%	2	3%
Quartile 2	*delta 0*	32	48%	34	52%	40	61%	37	56%
	delta 1	28	42%	28	42%	24	36%	26	39%
	delta 2	6	9%	4	6%	2	3%	3	5%
	delta 3	0	0%	0	0%	0	0%	0	0%
Quartile 3	*delta 0*	31	48%	32	49%	36	55%	35	54%
	delta 1	32	49%	28	43%	27	42%	29	45%
	delta 2	0	0%	3	5%	2	3%	1	2%
	delta 3	2	3%	2	3%	0	0%	0	0%
Quartile 4	*delta 0*	34	52%	27	42%	40	**62**%	40	62%
	delta 1	25	38%	32	49%	22	34%	22	34%
	delta 2	7	11%	7	11%	4	6%	4	6%
	delta 3	0	0%	0	0%	0	0%	0	0%

The first quartile in Table 10 contains the 66 responses with the lowest similarity scores, and so on, where 66 represents approximately one-quarter of the 263 student responses. From Table 9, the best results were obtained with TT C. However, other valuable insights also emerge.

The analysis of Table 9 and 10 provided interesting insights for the continuation of the research. The results obtained by using all three texts together as TTs are similar to those achieved with Text

C. This suggests that it is worth considering both the results from using separate TTs and from aggregating multiple texts with similar meanings as if they were a single text.

The second insight relates to the data obtained for each quartile. It is immediately noticeable that the average Delta value across all TTs modalities is higher than the Delta obtained within individual quartiles, as shown by the columns in Table 11. When examining Table 11 vertically, it becomes clear that the Delta value decreases significantly from quartile 1 to quartile 4: the Delta across all TTs modalities in quartile 4 is lower, meaning there is a higher agreement between the evaluations of the HA and the AA. This same trend appears in Table 10, which shows that while there is a total agreement between the AA and HA for 38% across all TTs modalities in quartile 1, this agreement rises to 62% in quartile 4. Additionally, in quartile 4, no TTs modality exhibit a Delta of 3, and the percentage of TTs modalities with a Delta of 2 decreases from 15% to 4%. While the values obtained are not yet fully acceptable for evaluation and feedback purposes, they are valuable for supporting the instructor in their interaction with the AA. As already indicated, and as will be further emphasized later, the selection of the TTs is of fundamental importance for the AA's evaluation.

Table 11

Results of Test 4: mean values of Delta for each quartile obtained using the four different TT

	A	**B**	**C**	**A+B+C**
1st quartile	1,380	1,414	1,228	1,251
2nd quartile	0,848	0,864	0,652	0,727
3rd quartile	0,636	0,591	0,409	0,439
4th quartile	0,600	0,615	0,492	0,492

In the initial trials, the TTs were derived from an initial manual evaluation by the instructor, which cannot be random. To select, for example, 16 significant TTs, the instructor must analyze far more than 16 students' answers. Test 4 suggests a way to assist the instructor in identifying the texts to use for constructing the TTs. Of the 15 students' answers to which the AA assigned the highest similarity scores, 14 of them were placed in the highest level by the HA. In other words, this preliminary automatic analysis by the AA gives the instructor an indication of where to find students' answers that are most likely to be assigned to the highest level. Shifting the focus to the students' answers with the lowest similarity scores, among the 15 which were assigned the lowest similarity by the AA, 9 of them were placed in level 1 by the HA.

In summary, test 4 provides some interesting insights. It suggests that a text identified by the instructor and independent of the students' tasks can be used as a TT. This text must be related to the questions, and good results can also be achieved by aggregating multiple texts. The results obtained with such texts are relatively reliable but not yet suitable for definitive evaluations, as the total consistency between the AA and HA is, at best, 54%.

This initial procedure can be used by the instructor to identify tasks belonging to levels 1 and 4 from which to extract the TTs. By reviewing only 20 out of 263 tasks, the instructor can focus on tasks that are highly likely to belong either to level 4 (the ten with the highest similarity) or to level 1 (the ten with the lowest similarity). In this way, the first step of the interaction between HA and AA has been identified, making the entire process more sustainable.

4.3. Answer decomposition

The next step arose from the observation that the AA identifies level-4 tasks more easily. This result is understandable: while level-4 tasks show greater consistency, as they are evaluated for their similarity to a clear model, the lower-level tasks lack coherence with a specific model and instead display consistency with a range of diverse, and not always predictable, possibilities. This led to further reflection on the assignment itself. The prompt required students to address four distinct concepts: 1. what appropriation is, 2. what in-situation is, 3. what in-corporation is, and 4. what in-culturation is. If level-4 texts were expected to contain all four concepts, the tasks in levels 1, 2, and

3 could still be coherent in one or more of the four themes; however, inconsistency could manifest in various ways. Therefore, it was decided to provide distinct TTs for each of the individual concepts. A set of texts was constructed, each relating to one of the four required themes (appropriation, in-situation, in-corporation, in-culturation), with a fifth category added for examples. For each theme, two sets of TTs were prepared, ranging from 5 to 15 short phrases: the first set consisted of level-4 TTs, and the second set of level-1 TTs (see Table 12).

Table 12
TT examples for the "in-situazione" concept

in-situazione: implica un cambiamento dell'ambiente e di alcuni suoi elementi che, da non pertinenti che erano, divengono costitutivi della situazione dell'attore;
in-situazione: l'oggetto di apprendimento entra a far parte dell'ambiente del soggetto, nel senso che il soggetto si rende conto che l'oggetto esiste per lui (poteva essere presente anche prima nell'ambiente ma non nel SUO ambiente)
l'appropriazione in-situazione implica un cambiamento dell'ambiente e di alcuni suoi elementi che, da non pertinenti che erano, divengono costitutivi della situazione dell'attore
rispetto alla situazione (IN – SITUAZIONE): si acquisiscono elementi nuovi che prima non erano pertinenti ma che diventano significativi nella nuova situazione di apprendimento
rispetto alla situazione (in-situazione): che prevede il cambiamento dell'ambiente e di alcuni elementi che prima erano insignificanti ma ora diventano parti costitutive dell'attività del soggetto
la prima è in-situazione la quale evidenzia come alcuni elementi dell'ambiente che inizialmente erano irrilevanti acquisiscono importanza

In this case, the procedure carried out by the AA was no longer a single process. Instead, for each concept, the entire text of each student was compared against a set of TTs. Additionally, the student's text was compared with each individual TT, and the system reported three similarity values—those obtained from the three TTs with the highest similarity scores. Moreover, the students' texts were compared with the TT created by combining all 10 TTs related to the same concept.

The results obtained from Test 5 (Table 13), while not showing significant progress compared to previous tests, suggest some promising potential: knowing the correspondence between the student's text and each individual concept allows for providing the student with precise feedback on each specific part of the task.

Table 13
Results of Test 5

	Total		**1st quartile**		**2nd quartile**		**3rd quartile**		**4th quartile**	
Delta 0	116	44%	17	26%	33	52%	32	49%	34	53%
Delta 1	121	46%	31	47%	28	44%	32	49%	30	47%
Delta 2	22	8%	15	23%	5	8%	2	3%	0	0%
Delta 3	3	1%	3	5%	0	0%	0	0%	0	0%

4.4. Refining the procedure: Sorting and Classification

The subsequent steps took three directions: refining the writing of the TTs; improving the procedure for ranking the students' texts; understanding how to extract feedback from the obtained data to provide to students.

Test 6 aimed to understand whether small modifications, even graphical ones, to the TT would significantly affect the results. For example, the impact of including or omitting capitalized words, numbered or bulleted lists, or presenting the same list in a narrative format was explored. It was observed that a level-4 student text, written in all capital letters, was classified as level 1. Test 2 was repeated after normalizing the students' texts.

The data showed a slight improvement, though minimal. Moreover, this test prompted further investigation into the AI model's indicators to fine-tune it in relation to the desired objectives. In the analysis, the percentage of Delta 0 exceeded 50%, and the average Delta dropped below 0.6. While these results are still not fully satisfactory, they are an improvement over previous test (Table 14).

Table 14

Results of Test 6

Mean Delta	Delta 0		Delta 1		Delta 2		Delta 3	
0,574	121	51%	95	40%	17	7%	2	0,8%

In **Test 7**, the focus was more specifically on the difference between ranking and classification (Table 15). In the initial tests, the AA was asked to classify the students' texts, clustering them into the predefined levels based on the TTs, which were also provided for each level. This method did not yield satisfactory results. Subsequently, the decision was made to work on ranking: based on the TTs, the students' texts were ranked, with similarity as the key indicator. In other words, texts with higher similarity scores were more similar to the TT.

Table 15

Results of Test 7

TT- Level 1

Actual Score	No. ST	1st Q	2nd Q	3rd Q	4th Q	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]
Level 1	**16**	5	2	4	5	31	13	25	31	9	3	7	8
Level 2	**33**	11	10	9	3	33	30	27	9	19	17	15	5
Level 3	**113**	25	26	25	37	22	23	22	33	43	44	42	63
Level 4	**73**	17	21	21	14	23	29	29	19	29	36	36	24
total	**235**	58	59	59	59					100	100	100	100

TT- Level 2

Actual Score	No. ST	1st Q	2nd Q	3rd Q	4th Q	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]
Level 1	**16**	10	4	2	0	63	25	13	0	17	7	3	0
Level 2	**33**	16	8	4	5	48	24	12	15	28	14	7	8
Level 3	**113**	25	31	28	29	22	27	25	26	43	53	47	49
Level 4	**73**	7	16	25	25	10	22	34	34	12	27	42	42
total	**235**	58	59	59	59					100	100	100	100

TT- Level 3

Actual Score	No. ST	1st Q	2nd Q	3rd Q	4th Q	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]
Level 1	**16**	13	1	2	0	81	6	13	0	22	2	3	0
Level 2	**33**	19	7	4	3	58	21	12	9	33	12	7	5
Level 3	**113**	23	37	26	27	20	33	23	24	40	63	44	46
Level 4	**73**	3	14	27	29	4	19	37	40	5	24	46	49
total	**235**	58	59	59	59					100	100	100	100

TT- Level 4

Actual Score	No. ST	1st Q	2nd Q	3rd Q	4th Q	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]	1st Q [%]	2nd Q [%]	3rd Q [%]	4th Q [%]
Level 1	**16**	13	2	0	1	81	13	0	6	22	3	0	2
Level 2	**33**	21	7	5	0	64	21	15	0	36	12	8	0
Level 3	**113**	21	39	33	20	19	35	29	18	36	66	56	34
Level 4	**73**	3	11	21	38	4	15	29	52	5	19	36	64
total	**235**	58	59	59	59					100	100	100	100

While in most previous tests, the focus was on classification—where TTs from four levels were provided and each student text (TS) had to be placed into one level— the goal of test 7 was simply to order the TSs according to similarity, using TTs from a single level. Additionally, a separate analysis was conducted for the four main concepts (plus one additional) required by the task. In previous tests, most TTs included all four concepts. Now, separate queries were performed with TTs representing a single concept. Four queries were conducted, each using seven TTs for one specific concept and level. The overall evaluation was then obtained by aggregating the results for each TS from the separate queries. Logically, level-1 student texts should show a higher relative similarity when compared with level-1 TTs and lower similarity when compared with level-4 TTs. The results are based on 235 student texts, as 28 TTs were excluded from the total of 263. The analysis (Table 15) partially confirms the hypothesis, as the data obtained with the seven level-4 TTs produced a significant ranking. However, the results from the queries using TTs from the other three levels provided much less consistent rankings compared to the HA's evaluations.

In **Test 8**, TTs extracted from the manual, similar to those used in Test 4, were reintroduced, but they were modified to better align with the task prompt, removing any unnecessary or accessory information. The results were almost identical to the previous ones, but the small difference allowed for the identification of potential improvements (Table 16). When working with a few TTs and shorter texts, the presence of adverbs, adjectives, lists, examples, and additional details tends to confuse the AA.

Table 16

Results of Test 8 compared with test 4

	n°	Mean Delta	Mean Similarity	delta 0	delta 1	delta 2	delta 3	delta 0	delta 1
Test 4	263	0,548	0,950	141	102	18	2	54%	39%
Test 8	263	0,515	0,947	142	107	14	0	54%	41%

In **Test 9**, 28 sentences were provided as TTs, all extracted from the students' texts. Each sentence was numbered and analyzed separately. Since the AA provides the three highest similarity scores and the corresponding three TTs out of the 28 provided, this test allowed for the identification of which TTs were more likely to have higher similarity scores. It was observed that the highest similarity was concentrated on 4 TTs. However, it was challenging to understand why these particular TTs were so frequently matched.

4.5. Procedure for test 10 and 11

At this point, the task was to engineer the procedure based on the rules extracted from the tests. From the tests, it became clear that short texts—those with a character count below the mean minus the standard deviation (see Test 3)—were difficult to analyze. In our case, these short texts made up less than 10% of the total. In the majority of cases, these texts were classified by the HA at level 1. Therefore, the first rule was to assign all short texts to level 1. For the analysis of the remaining texts, the following considerations were made:

1. Use ranking logic instead of classification. We can graphically represent the texts as points on a plane, similar to the dispersion of points around a center. The texts in level 4 are closer to the center, while texts in levels 1, 2, and 3 are not only farther from the center but also dispersed in different directions. The similarity measure calculates the distance from the center but not the direction. It was found that the AA's ability to group the texts was not particularly useful, while ranking them by similarity proved effective.
2. Divide the results into quartiles and work within each quartile. The most reliable quartiles were the first and fourth.

3. Perform a query for each concept required by the task, using specific TTs for each concept. In the studied case, the four required concepts were appropriation, in-situation, in-corporation, and in-culturation. Additionally, many student texts included examples. Therefore, four queries should be made, one for each concept.
4. Construct TTs without redundant or conflicting information. These could be sentences with similar meanings but using different language, terms, or graphical structures. A set of about ten sentences per concept should suffice.
5. Prepare the TTs in two phases. In the first phase, student texts are queried using TTs extracted from the manual or the instructor's notes. This identifies the level 4 texts. In the second phase, TTs are extracted from student texts to evaluate the remaining submissions.
6. Use TTs both individually and aggregated. For each concept, a set of around ten TTs is provided. Queries can be run with each individual TT or by aggregating all ten TTs as if they were a single sentence. This approach yields different results—one for each TT and one for the aggregated text—that can either reinforce or weaken the highest similarity result.

Based on these considerations, the following procedure was developed and validated in Tests 10 and 11. The final evaluation is obtained by summing the results from the queries performed for each concept. In this experiment, the same weight was assigned to each concept; however, in other cases, it might be appropriate to assign different weights to each concept. The value for each individual concept is determined as follows:

a) Texts in the fourth quartile for each concept are assigned 1 point. Similarly, texts in the first quartile are assigned 0 points.
b) Texts in the sixth octile are assigned 1 point if their other similarity values (second, third, and/or aggregated) are consistent. Similarly, texts in the third octile are assigned 0 points if the other evaluations are consistent.

The values assigned using these rules have approximately 90% accuracy.

Using this procedure, each student text is assigned up to four values, one for each concept. Summing these values yields the overall score. Additionally, the instructor receives information on how the student handled each individual concept.

Some texts, after this initial scoring, have blank cells. In the studied case, applying the two rules resulted in a score strongly consistent with the HA's evaluation for 60% of the sample: 87% had a Delta 0, and 13% had a Delta 1. It should be noted that the Delta 0 between HA2 (non-triangulated evaluation) and HA (triangulated evaluation) was 56%. In other words, in this case, the AA was more reliable than a human evaluation that did not include triangulation.

While the results were satisfactory, having useful outcomes for only 60% of the entire sample was still considered low. Therefore, the results of different queries were used. From these analyses, the following rules emerged, which increased the number of reliable evaluations:

c) Texts in the third quartile in the queries with the single TT showing the highest similarity and in the third or fourth quartile in the queries with other individual TTs (that did not show the highest similarity) or with aggregated TTs were assigned 1 point. Similarly, texts in the second quartile in the queries with individual TTs and in the first or second quartile with aggregated texts were assigned 0 points.

Applying these additional rules yielded evaluations for 71% of the sample, although reliability decreased: 79% had a Delta 0, and 14% had a Delta 1.

The next step involved reviewing the texts that had 2 or 3 evaluations instead of 4 after the previous analyses. In these cases, the missing texts fell into the second or third quartile, so an intermediate value between 0 and 1 was assigned. This deduction was only partially correct, as the data analysis, as previously mentioned, indicated that intermediate zones contained less reliable data.

In the sample analyzed, 67 texts had 2 or 3 evaluations out of 4. For these 67 texts, the missing cells were filled in with values of 0.25, 0.50, or 0.75 points based on the quartiles of the various similarity measures.

By intervening in this way, evaluations were obtained with approximately 70% accuracy (Delta 0) and 25% Delta 1. This allowed for the correction of 217 texts out of 263, or 85% of the sample, with a Delta 0 of 74% and a Delta 1 of 20%—a performance superior to that of human agents who did not use triangulation.

5. Conclusions

The adopted methods allowed for reliable data on approximately 80% of the student submissions. Additionally, the system highlights which texts still require further evaluation and indicates the uncertainty of different blocks, identifying those with more reliable evaluations. It should be noted that even in triangulated evaluation, there were "gray areas" where the three evaluators found it difficult to reach a consensus. It is important to emphasize that in the AA's evaluation, over 70% of the texts had a Delta of 0, while the remaining had a Delta of 1.

Although the level of agreement between the AA and HA is higher than that obtained in triangulated evaluation between three evaluators (HA) and the evaluation performed by a single teacher, from a psychological standpoint, especially for the student, such uncertainty can be difficult to accept. Therefore, the evaluation system includes an interactive phase between teacher and student: the system sends the evaluation to the student, who can then respond with comments. This process serves a dual purpose: on the one hand, it allows for the sharing and negotiation of the evaluation; on the other hand, it equips the student with the tools to understand the reasoning behind it. The student is informed about which parts of their work are effective and where there are deficiencies. This process reinforces the recursive feedback system introduced earlier in this paper.

The established procedure is a first step toward automated evaluation and has also highlighted certain limitations. One key limitation, linked to the decision to focus on ranking rather than classification, automatically divides the submissions into four quartiles, each containing the same number of tasks. It is up to the teacher to set the threshold to define the level of acceptability, which may not necessarily align with the average value. In this experiment, threshold values were set to define the four levels: from the automatic calculation, each submission received a score based on the sum of the values obtained for each concept, and the three thresholds (between levels 1 and 2, 2 and 3, and 3 and 4) allowed for the identification of tasks in each level. The teacher independently determines these three thresholds after the AA has ranked the 263 submissions.

Furthermore, continued experimentation is needed to determine how this procedure may be influenced by different types of tasks and academic disciplines. In this context, after the first experiment described here, the procedure was applied to another pedagogical task in a real-world educational setting, and although the reliability obtained was not identical, it was comparable to the initial results. These trials were part of a preliminary investigation into the feasibility of the system, laying the groundwork for future pilot studies in collaboration with schools and other educational institutions. Future research will involve larger-scale pilot studies across various educational institutions to validate these findings and ensure the system's scalability and effectiveness in different academic contexts.

Future iterations of this system could incorporate topic extraction and clusterization techniques, such as BERTopic [14] and the exploration of other Large Language Models (LLMs) for assessing open-ended responses.

The development of this research will be closely linked to the evolution of AI models. Every day, new and more advanced projects are being introduced by various companies. By the time this contribution is published, the previous model could already be outdated. Looking ahead, generative AI models could assist teachers in the early stages of the process, including the preparation of the task itself. Although there is still a long way to go and many developments are possible, the results obtained so far are promising for the adoption of models based on recursive interaction between AA and HA to make the widespread use of feedback in daily university practice more sustainable.

Acknowledgements

This research activity is part of the PRIN 2022 project no. 2022ZMYYTH, titled "AI&F – Artificial Intelligence & Feedback for Effective Learning" – CUP D53D23013110006, funded under the PRIN 2022 Call (Funded by the European Union – Next Generation EU – PNRR Mission 4, Component 2, Investment 1.1).

References

[1] F. Gratani, L. Screpanti, L. Giannandrea, D. Scaradozzi, L.M. Capolla, Personalized Feedback in University Contexts: Exploring the Potential of AI-Based Techniques, in: G. Casalino, R. Di Fuccio, G. Fulantelli, P. Raviolo, P. C. Rivoltella, D. Taibi, G. A. Toto (Eds.), 5th International Conference on Higher Education Learning Methodologies and Technologies Online, HELMeTO 2023, volume 2076 of Communications in Computer and Information Science, Springer, Cham, 2024, pp. 440-454. doi: 10.1007/978-3-031-67351-1_30.

[2] A. A. Lipnevich, E. Panadero, A review of feedback models and theories: Descriptions, definitions, and conclusions, in Frontiers in Education, volume 6, pp. 720195, Frontiers, 2021. doi: 10.3389/feduc.2021.720195

[3] S. Pozdniakov, J. Brazil, S. Abdi, A. Bakharia, S. Sadiq, D., Gašević, D., P. Danny, and H. Khosravi, Large language models meet user interfaces: The case of provisioning feedback, Computers and Education: Artificial Intelligence 7 (2024) 100289. doi: 10.1016/j.caeai.2024.100289

[4] S. S. Lee and R. L. Moore, Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review, Online Learning 28, 3 (2024) 82 - 106. doi: 10.24059/olj.v28i3.4593

[5] P. X. Lam, P. Q. H. Mai, Q. H. Nguyen, T. Pham, T. H. H. Nguyen, and T. H. Nguyen, Enhancing educational evaluation through predictive student assessment modeling, Computers and Education: Artificial Intelligence 6 (2024) 100244. doi: 10.1016/j.caeai.2024.100244

[6] A. Bewersdorff, K. Seßler, A. Baur, E. Kasneci, and C. Nerdel, Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters, Computers and Education: Artificial Intelligence 5 (2023) 100177. doi: 10.1016/j.caeai.2023.100177

[7] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, A. Srinivasa, "Automatic assessment of text-based responses in post-secondary education: A systematic review." Computers and Education: Artificial Intelligence 6, (2024): 100206. doi:10.1016/j.caeai.2024.100206.

[8] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, J. De Weerdt, "A review of automated feedback systems for learners: Classification framework, challenges and opportunities." Computers & Education 162.3 (2021): 104094. doi: 10.1016/j.compedu.2020.104094.

[9] S. Gombert, A. Fink, T. Giorgashvili, I. Jivet, D. Di Mitri, J. Yau, A. Frey, H. Drachsler, "From the Automated Assessment of Student Essay Content to highly informative feedback: A case study." International Journal of Artificial Intelligence in Education, (2024): 1-39. doi:10.1007/s40593-023-00387-6.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (2018). doi: 10.48550/arXiv.1810.04805.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv:1706.03762 (2017). doi: 10.48550/arXiv.1706.03762.

[12] A. Petukhova, A., J.P. Matos-Carvalho, N. Fachada, Text clustering with LLM embeddings, arXiv:2403.15112 (2024). doi: 10.48550/arXiv.2403.15112

[13] J. Wang, Y. Dong, "Measurement of text similarity: a survey." Information 11.9 (2020): 421. doi: 10.3390/info11090421

[14] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv preprint arXiv:2203.05794 (2022). Doi: 10.48550/arXiv.2203.05794