IJPAM
Italian Journal of Pure and Applied Mathematics
https://journals.uniurb.it/index.php/ijpam
E-ISSN 2239-0227

DOI: 10.14276/ijpam.5767
Received: 15 April 2026
Accepted: 9 June 2026
Published: 30 June 2026
Peer Review History
Single-blind peer review

Exploring empathy in mathematics feedback: a comparative study of human and AI-generated responses in informal learning contexts

Gennaro Cordasco[a], Umberto Dello Iacono[b], Anna Esposito[c], Antonio Vitale[1d,e], and Carl Vogel[e,f]

[a]Department of Computer Science
University of Salerno, Fisciano, Italy
gcordasco@unisa.it
[b]Department of Mathematics and Physics
University of Campania "Luigi Vanvitelli", Caserta, Italy
umberto.delloiacono@unicampania.it
[c]Department of Psychology
University of Campania "Luigi Vanvitelli", Caserta, Italy
anna.esposito@unicampania.it
[d]Faculty of Education, Cultural Heritage and Tourism Sciences
University of Macerata, Macerata, Italy
a.vitale11@unimc.it
[e]School of Computer Science and Statistics
Trinity College Dublin, Ireland
vogel@tcd.ie
[f]Corvinus Institute for Advanced Studies (CIAS)
Corvinus University of Budapest, Hungary

Abstract: The aim of this study is to analyze feedback perception of empathy generated by humans and Large Language Models (LLMs) in informal mathematics learning contexts. Using the dimensions of Emotion Recognition (ER), Perspective -Taking (PT), and Emotional Contagion (EC), we conducted a comparative evaluation on a dataset of formal logic problems sourced from the Reddit online community. Findings indicate that feedback generated by LLMs, when supported by well - structured prompts, is rated as significantly more empathetic than human feedback, which tends to focus more on procedural accuracy. While ER and EC show the most pronounced gaps in favor of AI, PT emerges as the most complex and least differentiated dimension. Finally, the study suggests that LLMs can effectively integrate effective support into informal mathematics education.

2020 Mathematics Subject Classification: Primary 97C20, 97U50; Secondary 97U70,

[1]Corresponding author.

68T50.
Keywords: empathetic feedback; informal mathematics education; LLMs; Mistral; Gemini; ChatGPT.

1. Introduction

Recently, education has increasingly been considered a process that extends beyond the boundaries of formal educational settings. Learning in a non - formal education environment is mainly characterized by a high level of students' autonomy [1] and is no longer considered a phenomenon restricted to the school environment alone, but rather a continuous process that unfolds across a variety of contexts, interactions, and mediating tools. From this perspective, non - formal learning contexts contribute to building continuous learning trajectories, in which diverse experiences are deeply interconnected [2]. Mathematical learning can also take place through informal practices, that is, outside the school environment, often mediated by digital technologies [3] or everyday social interactions [4]. Due to students' apprehension toward math learning environments involving assessment [5], they are inclined to seek support in various non - formal settings to learn math without the pressure of assessment. Technology - Related informal mathematics Learning Activities (TRLA) are not merely isolated episodes, but rather essential elements for ensuring educational continuity, thereby lending further validity to technology - mediated learning [6].

Within this context, online forums and digital peer - support communities emerge as privileged spaces for non - formal learning, where solving mathematical problems becomes a collaborative and public activity. In such environments, students not only access content but also seek support on specific problems, transforming individual doubts into opportunities for discussion. Learning environments characterized by a supportive atmosphere for seeking help are not only a means of resolving a specific task but also a strategy for active self - regulation. Students, released from the pressures of formal assessment, are able to independently identify their own gaps in knowledge and proactively seek out resources [6].

In such environments, feedback plays a crucial role. Feedback is expressed through various shapes and forms. The literature defines feedback based on its objectives and ways of interacting. For example, formative feedback is distinguished from summative feedback, where the summative feedback gives an assessment of how well students completed a specific task such as might justify an overall grade for a piece of work, while formative feedback gives specific guidance on how to improve their performance highlighting, for example, what is good or what appears to be missing or mis - understood [7]. Furthermore, there is peer feedback: such feedback has the potential to foster collaborative learning among students, making the entire learning process more meaningful [8]. Lastly, there is affective feedback, which recognizes and supports students' emotional states [9]. Hence, it is essential to recognize the inherent diversity of feedback in order to avoid oversimplified interpretations that view it as a form of communication devoid of nuance. Developments in Artificial Intelligence have led to the introduction of automated methods for providing adaptive, process - oriented, and rapid feedback [10]. However, significant issues arise regarding their ability to address not only students' cognitive needs but also their affective and emotional well - being, especially important in learning processes. A significant turning point was reached with the development of Large Language Models (LLMs) which are capable of identifying the context and relevant sections of students' problems. They provide the technological framework necessary to achieve advanced levels of interaction and to deliver feedback that is as efficient as possible and, importantly, in real time [10]. These capabilities are essential for ensuring the relevance of the feedback. Unfortunately,

this technological process brings with it new challenges in terms of socio - emotional nuances, especially in contexts where emotional support plays a fundamental role in sustaining learning processes. Although these systems are extremely effective at generating content, their empathetic and motivational capabilities are often very limited for effective tutoring. [11] showed that feedback generated by LLMs tends to satisfy objective criteria to a greater extent than emotional ones. This aspect helps us understand how, from a cognitive standpoint, it is possible to compare LLMs and humans, whereas from a purely emotional relational standpoint, there is always a weakness in automated systems. [12] evaluated the performance of LLMs and showed that, from a pedagogical perspective, they are not optimal when assuming the role of educational tutors. For these reasons, it is essential to conduct an in-depth analysis of the empathy in the feedback and examine whether these systems can simulate or modulate the empathy required to address the complex affective needs of students facing formal logic challenges. As part of this study, our research focuses specifically on formative and affective feedback within asynchronous peer feedback dynamics (the providers of feedback are not giving grades but are constructively responding to questions asked by those seeking feedback), exploring how the affective component of feedback can be tailored to students' needs.

2. Theoretical Background

In informal digital environments, interaction plays a pivotal role in supporting learning processes. In these contexts, feedback emerges through interactions among peers or with more experienced members of the community, forming a socially mediated and context - specific process [13, 14]. Feedback is one of the most influential factors supporting learning across various contexts. In mathematics, it not only supports conceptual understanding and problem - solving processes but also impacts the affective dimension of learning, influencing motivation, engagement, and self - perception [9]. Considering feedback as a teaching tool, it can influence students' learning outcomes [15]. However, in order to be effective, feedback must be timely and process - focused to stimulate self - regulation and metacognition. Its function in supporting conceptual understanding and problem - solving, feedback also has a significant affective dimension that influences students' motivation, self - efficacy, and attitudes [9, 16, 17]. From this perspective, the quality of feedback cannot be fully understood without considering its emotional and relational dimensions.

In informal settings, the affective dimension of feedback is particularly significant, due to the self - directed nature of learning, which is strongly linked to intrinsic motivation and self - regulation processes [18, 19]. From this perspective, affective feedback, understood as the set of elements that acknowledge and support students' emotional states, helps shape how students interpret and use the information they receive [9, 17]. In mathematics education, these factors are particularly noteworthy, as emotions such as anxiety can hinder performance and reduce students' willingness to tackle complex tasks [16, 20, 21]. Feedback that integrates both affective and cognitive components can alleviate anxiety and foster perseverance. In informal digital environments, however, asynchronous interactions, heterogeneous participation, and the absence of structured pedagogical guidance make feedback highly variable: alongside constructive responses, there are comments that are unclear or lack sensitivity [22]. In the informal context of online math and logic forums, active help - seeking is a crucial moment in student learning. Whenever they hit a cognitive roadblock, students look not only for a technical solution but also for feedback to validate their efforts. However, the human responses in these forums can be inconsistent both within individual responses and across responses to distinct questions in terms of accuracy, in depth analysis, tone and timeliness. The responses vary from detailed and helpful explanations to brief, ambiguous or even discouraging answers. This inconsistency can affect not only learning outcomes but also student motivation and en-

gagement, making the affective dimension of feedback particularly fragile in informal contexts.

The empathetic aspect of feedback can greatly influence how students interpret and apply it. Empathy is a broader term, encompassing both cognitive and behavioural processes as well as related concepts such as sympathy and compassion [23, 24]. It can be defined in a multidimensional way. On the one hand, there is cognitive empathy, such as the system's ability to accurately identify the specific source of the student's struggle or logical mistake. On the other hand, there is affective empathy, such as the ability to generate helpful, judgment - free responses that validate the student's emotions. [25] list three fundamental components of empathy in artificial systems:

- Emotion recognition: the system is able to identify the user's emotional state (or, more broadly, their experience), deriving it from inputs such as text, speech, or multimodal signals, which are also necessary for successfully performing its assigned task.
- Perspective - taking: the agent interprets the user's thoughts and inferential processes by linking the recognized emotion to their personal context (preferences, personality, goals) and to the communicative situation.
- Emotional contagion: the system simulates an emotion that is consistent with or appropriate to that of the user, generating responses that reflect and integrate emotional cues within their context.

Based on the above components of empathy, empathetic feedback is not limited to correcting mistakes. It takes into account the student's level of competence and the strategies they used, providing gradual support toward the solution. Moreover, it involves recognizing the student's emotional state and responding with understanding and sensitivity and conveys trust, encouragement, and respect, helping to create a positive atmosphere that supports motivation and the willingness to take on new challenges. Adopting the student's perspective, it identifies the exact point at which the student deviated from the correct process, showing that mistakes are part of learning and highlighting what was done correctly. Empathetic feedback thus requires a balance between educational rigor and relational sensitivity. Integrating empathy with cognitive criteria makes instructional feedback more comprehensive and realistic. To be effective, it must be clear, complete, and relevant, while also recognizing emotions, adopting the student's perspective, and responding with emotional contagion, thus supporting learners cognitively, motivationally and relationally.

3. Research aim and question

This study aims to investigate not only the ability of LLMs to generate empathetic feedback but also to recognize and evaluate feedback empathy itself through a cross - evaluation process in non - formal mathematics learning contexts, comparing human feedback with feedback generated by large language models. The aim is to investigate the role and effectiveness of LLMs and human, generated feedback's affective element in the domain of Formal Logic within purely informal educational contexts, such as online forums. Specifically, this paper aims to analyze the socio - emotional dimension of feedback, with a specific focus on the expression and assessment of empathy. The paper focuses on understanding how empathy is expressed in feedback and how it is recognized and assessed within informal mathematical contexts that are characterized by high procedural accuracy. To this end, we adopt an analytical framework that considers the main components of empathy, such as emotional recognition, perspective - taking, and emotional contagion [25]. Specifically, this study is motivated by the following research question: (RQ) How is empathy evaluated in terms of emotional recognition, perspective - taking and emotional contagion across human and LLM - generated feedback in informal mathematics learning contexts?

4. Methodology

In this study, we adopted a comparative research methodology aimed to investigate the affective dimension of feedback in informal mathematics learning contexts. In accordance with the research question, this methodology was designed to analyze how empathy is expressed and evaluated in feedback provided by humans and large language models (LLMs), considering these agents as providers of feedback.

The research process was organized into four sequential phases:

1. Dataset collection and selection from Reddit discussions on formal logic problems;
2. Generation of LLM - based feedback using a standardized emotionally supportive prompt;
3. Affective evaluation of feedback through a cross - evaluation procedure involving both human and AI evaluators;
4. Statistically grounded analysis of the empathy - related dimensions.

These phases are described each in turn in this section; the final phase extends in expanded form with discussion of the results that emerge from our analysis.

To address the research question, we conducted a comparative analysis using a dataset made up of answers to Formal Logic problems collected from an online community site. The answers provided by humans were compared with those generated by three different LLMs (GPT-4.1, Gemini 2.5 Flash, and Mistral non-Large), all specifically trained with a specific prompt to provide emotional support. This study adopted a cross - evaluation methodology, in which both humans and AI models take on the role of evaluators for feedback analysis, based on an analytical framework of empathy that includes *emotional recognition*, *perspective - taking* and *emotional contagion*.

4.1. Dataset Collection

The dataset underlying the study consists of textual data extracted from public online discussions regarding problems in formal logic and reasoning, sourced from the Reddit platform (https://www.reddit.com). The selected posts (n = 50) are publicly accessible and were collected without retaining user identifiers. We collected all the interactions analyzed in their original language, which is English. We selected the problems through a systematic search using a predefined Boolean string: (*"symbolic logic" OR "formal logic" OR "natural deduction"*) AND (*"my answer" OR "proof" OR "exercise"*) AND (*"feedback" OR "help" OR "explanation"*). We also filtered the results based on the following inclusion criteria:

a) presence of a clearly formulated logic problem;
b) student - type question;
c) presence of at least one meaningful answer.

After removing duplicates and posts lacking sufficient context, we structured the final set of items. Each item includes the problem text, the student's attempt or question, and one or more pieces of feedback provided by other users.

4.2. LLMs Feedback Generation

For each problem, feedback generated by various large language models (LLMs) was integrated into the dataset to enable a direct comparison with human feedback. To ensure consistency and comparability among the generated feedback, all models were trained using a standardized prompt designed to elicit responses that were not only educationally effective but also emotionally supportive. The prompt used for this process was as follows (Table 1):

*You are an **expert logic tutor** helping a student who is struggling with a formal logic proof. Your goal is to provide a **clear**, **pedagogically effective**, and **emotionally supportive** feedback. Be **concise but helpful**. Do not give the full solution but guide the student toward the next correct step.* ***Problem:** [formal logic problem]* ***Student question:** [student's message]* ***Please provide your feedback.***

Table 1. The standardized prompt for large language models (LLMs), designed to ensure educational effectiveness and emotional support

Notably, the prompt refers to *"emotionally supportive feedback"* instruction as a key element to foster the development of empathetic skills, encouraging recognition of the student's state of mind, attention to their thought process, and the use of an encouraging tone. This approach helped reduce variability related to the wording of the prompt, ensuring consistent conditions in the generation of feedback.

4.3. Evaluation Criteria and Procedure

Following the data collection phase, we developed a feedback corpus drawn from both human and Artificial Intelligence sources, which then enabled us to conduct a rigorous evaluation of the feedback. The primary aim of this evaluation was to quantify the quality and effectiveness of the feedback across three specific dimensions. We carefully selected the socio - emotional dimensions of the feedback (*Emotion Recognition (ER), Perspective - Taking (PT), Emotional Contagion (EC)*), in accordance with current practices in educational research and the evaluation of AI - assisted learning [25]. Each of these dimensions was then evaluated using a 5 - point Likert Scale. The evaluation was conducted using a cross - tabulation approach, in which each item was analyzed comparatively across the entire set of feedback (both human and LLMs - generated) regarding the same Logical problem. This allowed us to ensure consistency in evaluation criteria, minimize inter - rater variability, and facilitate a systematic and direct cross - comparison between the different types of feedback. In the context of human evaluation of the feedback, this was performed by the first author.

The prompt used to perform the affective process evaluation was as follows (Table 2):

You are evaluating the empathy of a "XLLMs or Human" feedback message given to a student about a math/logic problem.

Input will be structured as follows:

Problem: [formal logic problem]

Student question: [student's message]

Feedback to evaluate: [XLLMs or Human feedback]

Evaluate the feedback according to the three dimensions below. For each dimension, assign a score from 1 to 5 using the operational descriptions.

Dimension 1 - Emotion Recognition

Does the feedback show awareness of the student's emotional state (e.g., frustration, confusion)?
1 = No emotional recognition.
2 = Minimal or unclear recognition.
3 = Generic acknowledgment (e.g., "I understand").
4 = Clear and context - based recognition.
5 = Specific, accurate, and sensitive recognition.

Dimension 2 - Perspective -Taking

Does the feedback consider how the student reasoned and where the mistake originated?
1 = Generic feedback that could apply to anyone.
2 = Minimal reference to the student's process.
3 = Partial attempt to connect to the student's reasoning.
4 = Clearly identifies the origin of the mistake and uses it constructively.
5 = Reconstructs the student's mental path and guides them step - by - step.

Dimension 3 - Emotional Contagion (Appropriate Answers)

Is the tone supportive and encouraging, without being judgmental or patronizing?
1 = Judgmental, cold, or harsh tone.
2 = Neutral tone without encouragement.
3 = Respectful tone but not particularly motivating.
4 = Warm and supportive tone.
5 = Strongly encouraging tone that conveys confidence in the student.

Return the evaluation in the following format (do not add anything else):

Emotion Recognition (ER): [score]

Perspective - Taking (PT): [score]

Emotional Contagion (EC): [score]

Table 2. The prompt used to perform the affective process evaluation

We systematically examined the differences in the scores assigned to the affective dimensions of empathy (*Emotion Recognition, Perspective - Taking, Emotional Contagion*) across the feedback generated by the various evaluators (humans and LLMs), using a comparative quantitative approach.

4.4. Statistical Analysis

Analysis was conducted using R Studio Software. The final Dataset included a total of 693 observations. Feedback distribution by provider involved Human responses ($n = 243$) and responses for each of the three LLMs ($n = 150$). These were evaluated by a group of Evaluators (Gemini, ChatGPT-4, Mistral, with $n = 181$ while Humans, with $n = 150$).

Given the nonparametric nature of the data, we conducted the comparisons using nonparametric statistical tests between groups (*Wilcoxon test*), distinguishing between comparisons of independent groups (feedback providers) and paired evaluations of the same items (evaluators). Furthermore, to control for the Type I error rate resulting from performing multiple group pairwise tests, we applied the Bonferroni correction to all resulting p-values ($p < 0.05$).

5. Results

Table 3 shows a perceived quality generated by each source, as assessed by the group of evaluators. Combining the scores by feedback provider, we can identify which source (Human or LLMs generated feedback) demonstrates a greater ability to express empathy in an informal mathematical learning context.

EMPATHY DIMENSION	GEMINI	GPT-4	HUMAN	MISTRAL
Perspective Taking (PT)	4.05	4.29	3.77	3.89
Emotion Recognition (ER)	3.83	3.75	2.22	3.28
Emotional Contagion (EC)	4.15	4.15	2.95	3.86

Table 3. Average of Empathy Dimensions scores

Data highlights a clear ranking of performance and shows significant differences in how emotional indicators are used by humans in comparison to large language models. Specifically, in the domain of *Perspective - Taking*, we can see that AI systems emerge as the primary providers of feedback. Regarding the ability to adapt to the student's point of view and understand their cognitive state, both GPT-4 (4.29) and GEMINI (4.05) demonstrate an advantage over Mistral (3.89) and the Human (3.77). The low score for human feedback suggests that, in this context, it is more task - focused rather than centered on the empathetic perspective necessary for understanding the student's subjective experience.

Meanwhile, regarding *Emotion Recognition*, there is a significant difference in the ability of the systems to identify and recognize the student's emotional state. Once again, AI systems outperform humans. Specifically, GEMINI (3.83) and GPT-4 (3.75) are capable of understanding emotional nuances and reflecting them in the feedback they generate. The feedback generated by MISTRAL (3.28) receives a lower score than the other two LLMs, indicating a more moderate but less precise ability to recognize emotions. Significantly, the score for human feedback (2.22) is markedly lower than that of the LLMs: this suggests that human feedback appears less focused on addressing students' emotional states.

Finally, the analysis of *Emotional Contagion* shows that both GEMINI and GPT-4 (4.15) have the highest scores. This indicates that their feedback is highly aligned with students' needs, showing a high level of emotional contagion. MISTRAL (3.86) has also achieved a high

score, demonstrating that LLMs exhibit a high level of emotional accuracy. Again, human feedback scores lower (2.95) in this domain: this indicates that the feedback provided may appear less emotionally aligned and fails to offer students the emotional resonance they need.

The significance analysis, performed using the Wilcoxon test for pairedwise comparisons of groups, also confirms this trend (see Table 4, Table 5, Table 6):

	GEMINI	**GPT-4**	**HUMAN**
GPT-4	0.0786	-	-
HUMAN	0.1879	1.7e-0.5	-
MISTRAL	1.0000	0.0067	1.0000

Table 4. Perspective Taking variable p_value matrix

Table 4 shows that GPT-4 is statistically more effective than HUMANS (p = 1.7e-05) and MISTRAL (p = 0.0067). There was no significant difference between GPT-4 and GEMINI (p = 0.0786), neither between GEMINI, MISTRAL, and HUMANS. These data suggest that GPT-4 is clearly more effective in the domain of PT, while the other systems perform within a statistically similar range.

	GEMINI	**GPT-4**	**HUMAN**
GPT-4	1.0000	-	-
HUMAN	$<$ 2e-16	$<$ 2e-16	-
MISTRAL	5e-05	0.0016	$<$ 2e-16

Table 5. Emotion Recognition variable p_value matrix

Table 5 shows a highly divergent pattern. Notably, the difference between HUMANS and LLMs is extremely significant ($p <$ 2e-16). Furthermore, while GEMINI and GPT-4 show no significant differences between themselves (p = 1000), they significantly outperform MISTRAL (with p = 5e-05 and p= 0.0016, respectively). This confirms that GEMINI and GPT-4 show significantly higher performance in emotion recognition.

	GEMINI	**GPT-4**	**HUMAN**
GPT-4	1.0000	-	-
HUMAN	$<$ 2e-16	$<$ 2e-16	-
MISTRAL	0.0019	0.0037	$<$ 2e-16

Table 6. Emotional Contagion variable p_value matrix

Table 6 shows that the difference between HUMAN feedback and all LLMs is highly significant ($p <$ 2e-16). Again, while GEMINI and GPT-4 show no significant differences between themselves (p = 1.000), they significantly outperform MISTRAL (with p = 0.0019 and p = 0.0037, respectively). All of this confirms that the LLMs (GEMINI and GPT-4) are perceived as significantly more effective on an emotional level than both HUMAN and MISTRAL.

6. Discussion and Conclusion

This study explored how empathy is evaluated by humans and LLMs feedback in informal mathematics contexts. Specifically, RQ was addressed by investigating the dimensions of *Emotion Recognition (ER), Perspective - Taking (PT), and Emotional Contagion (EC)*, as reported by [25], are perceived and evaluated by comparing human - generated feedback with that generated by various language models. While the data explored here was solely anchored in English,

we expect that similar findings would emerge for other languages - this is an exploration that we leave for the future.

Findings highlight statistically significant differences in how empathetic quality is rated for different feedback providers. Across all dimensions GPT-4 and GEMINI achieved the highest scores while HUMAN showed the lowest scores, specifically in emotion recognition and emotional contagion.

Feedback generated by LLMs is rated as more empathetic than human feedback across all three dimensions considered. This is particularly noteworthy in light of previous studies in Literature. Specifically, the studies by [11] and [12] highlight how LLMs are very effective from a cognitive perspective but less so in managing the emotive dimension. On the other hand, our study suggests that, if LLMs generate feedback using well - structured prompts [26], they are perceived as capable of producing feedback that adequately addresses students' emotional needs.

We can look at this apparent discrepancy in light of how empathy is assessed. As [17] and [9] have pointed out, the affective aspect of feedback significantly influences how students perceive and interpret it. Therefore, it is likely that LLMs, through the use of systematic language that provides emotional support and validation, may be perceived as empathetic. Specifically, our data analysis showed that the most significant differences emerge in the dimensions of Emotion Recognition and Emotional Contagion, where human feedback scores are significantly lower than LLMs. One interpretation of these findings is that, in mathematical contexts, human feedback is more likely to focus on procedural correctness and task resolution. On the other hand, LLMs, designed to generate complete and appropriate answers, are able to incorporate affective elements, making them more aligned with the components of *emotion recognition* and *emotional contagion* [25]. With regard to the dimension of *perspective - taking*, findings reveal a more nuanced scenario. Although GPT-4 outperforms humans significantly, the differences between it and other LLMs are less notable. Such findings suggest that the ability to adopt the student's perspective is the most complex dimension of empathy, which aligns with the study by [23], who distinguish between cognitive and affective empathy, noting that the former requires more complex and profound inferential processes and is therefore less easily simulated than affective empathy. Overall, these results help redefine the role of LLMs in informal educational contexts. While the literature tends to highlight the limitations of such systems in terms of empathy, the data from this study suggest that they can be highly effective in producing feedback perceived as emotionally appropriate. Nevertheless, we must highlight a significant aspect: within the educational context, the role of the expert differs from that provided by LLMs. Specifically, although in this study LLMs demonstrate a strong capacity to simulate immediate and supportive empathy, human support remains the only source of comprehensive pedagogical guidance. However, the scores achieved by human answers do not indicate a shortcoming but rather a focus on the educational process that takes into account the student's entire long - term progress, unlike LLMs, which are artificially programmed to be immediately effective.

Results suggest that LLMs have the potential not only to be cognitive tutors but also to provide effective support in the context of informal mathematics education. LLMs' added value lies in their ability to incorporate an emotional and relational dimension alongside their technical capabilities. In informal mathematics contexts, where anxiety toward the subject often acts as a learning obstacle, the use of LLMs can foster a learning environment where feedback is not only corrective but also emotionally effective and validating. This is not inherent in LLMs but

is mediated by the quality of the prompt. According to [26], the structuring of the prompt is important but, above all, effective. Therefore, it is essential that teachers and developers collaborate to "train" LLMs not to limit themselves to mere procedural correctness but to also be able to incorporate elements of emotional and affective support. As a result, the use of these technologies in educational settings should not be viewed as an alternative but rather as a perfectly complementary tool, with humans serving as guides and supporters.

Another aspect to consider regarding the analysis of these findings is the nature of LLMs. Namely, these systems should not be viewed as entirely neutral. Their high scores on affective dimensions do not necessarily imply that they are truly capable of understanding the student's emotional status. On the other hand, these results show that LLMs primarily operate by following a set prompt. Consequently, the empathy demonstrated by LLMs risks being merely "ideal" empathy. Hence, it is essential that future studies not only measure the effectiveness of feedback but also, and above all, analyze the cultural and expressive biases hidden within these systems.

Concluding, future research could explore which specific components of feedback influence the perception of empathy by conducting a qualitative analysis aimed at identifying the most influential text segments. Specifically, it might be interesting to classify how certain linguistic elements (i.e. the use of modal auxiliaries, expressions acknowledging mistakes, or the explicit expression of emotional support) actually affect the dimensions of *emotion recognition* and *emotional contagion*.

Acknowledgements

This research received funding by the EU-H2020 program, grant No. 101182965 (CRYSTAL)

Conflict of interest

The authors declare that there are no conflicts of interest.

References

[1] S. Romi, M. Schmida, Non-formal education: A major educational force in the postmodern era, *Cambridge Journal of Education*, **39**(2) (2009), 257–273

[2] S. Akkerman, A. Bakker, Boundary crossing and boundary objects. *Review of Educational Research*, **81** (2011) 132–169.

[3] T. He, S. Li, A comparative study of digital informal learning: The effects of digital competence and technology expectancy. *British Journal of Educational Technology*, **50**(4) (2019) 1744–1758, doi: https://doi.org/10.1111/bjet.12778

[4] G. Benigno, The everyday mathematical experiences and understandings of three, 4-year-old, African-American children from working-class backgrounds (Dissertation) University of Maryland, College Park, MD, (2012), doi: http://drum.lib.umd.edu/handle/1903/12562

[5] Z. Yuan, T. Tan, R. Ye, A cross-national study of mathematics anxiety, *The Asia- Pacific Education Research*, **32** (2023), 295–306, doi. https:// doi: org/ 10. 1007/ s40299- 022-00652–7

[6] H. Jiang, R. Chugh, D. Turnbull, X. Wang, S. Chen, Exploring the effects of technology-related informal mathematics learning activities: A structural

equation modeling analysis., *CQUniversity. Journal contribution.*, **32** (2025), doi: https://hdl.handle.net/10779/cqu.28331093.v1

[7] P. Black, D. Wiliam, *Assessment and Classroom Learning*, Assessment in Education: Principles, Policy & Practice, London, 1998, ISSN: 1465–329X

[8] M. Alqassab, J. W. Strijbos, S. Ufer, The impact of peer solution quality on peer feedback provision on geometry proofs: Evidence from eye movement analysis, *The Journal of the European Association for Research on Learning and Instruction (EARLI)*, **58** (2018), 182–192

[9] J. Hattie, H. Timperley, The power of feedback. *Review of Educational Research*, **77**(1), (2007) 81–112, doi https://doi: org/10.3102/003465430298487

[10] H. Y. Durak, A. Onan, A systematic review of Al-based feedback in educational settings, *Journal of Computational Social Science*, **8**(96), (2025), doi: https://doi.org/10.1007/s42001-025-00428-1

[11] E. Rudolph, H. Seer, C. Mothes, J. Albrecht, Automated feedback generation in an intelligent tutoring system for counselor education, *19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, (2024), 501–512, doi: 10.15439/2024F1649

[12] K. K. Maurya, K. V. A Srivatsa, K. Petukhova, E. Kochmar, Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors, *arXiv*, (2025), Available online: https://arxiv.org/abs/2412.09416

[13] S. Hrastinski, Asynchronous & Synchronous E-Learning, *Educause Quarterly*, (2008), 51–55.

[14] B. Rogoff, Developing Understanding of the Idea of Communities of Learners, *Mind, Culture, and Activity*, **1**, (1994), 209-229

[15] E. Faulconer, J. Griffith, A. Gruss, The impact of positive feedback on student outcomes and perceptions, *Assessment & Evaluation in Higher Education*, **47**, (2021), 1–10, doi: 10.1080/02602938.2021.1910140

[16] P. Di Martino, R. Zan, "Me and Maths": Toward a Definition of Attitude Ground on Students' Narratives. *Journal of Mathematics Teacher Education*, **13**, (2010), 27–48, doi: https://doi.org/10.1007/s10857-009-9134-z

[17] A. A. Lipnevich, J. K. Smith, Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, **15**(4), (2009), 319–333, doi: https://doi.org/10.1037/a0017841

[18] R. M. Ryan, E. L. Deci, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being, *American Psychologist*, **55**(1), (2000), 68–78, doi: https://doi.org/10.1037/0003-066X.55.1.68

[19] B. J. Zimmerman, Becoming a Self-Regulated Learner: An Overview, *Theory Into Practice*, **41**(2), (2002), 64–70

[20] M. H. Ashcraft, Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, **11**(5), (2002), 181–185, doi: https://doi.org/10.1111/1467-8721.00196

[21] A. Dowker, A. Sarkar, C. Y. Looi, Mathematics Anxiety: What Have We Learned in 60 Years?, *Frontiers in Psychology*, **7**(508), (2016), doi: 10.3389/fpsyg.2016.00508

[22] P. Ferguson, Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education*, **36**(1), (2011), 51–62, doi: https://doi.org/10.1080/02602930903197883

[23] B. M. Cuff, S. J. Brown, L. Taylor, D. J. Howat, Empathy: A review of the concept. *Emotion review*, **8**(2), (2016), 144–153

[24] S. D. Preston, F. B. De Waal, Empathy: Its ultimate and proximate bases, *Behavioral and brain sciences*, **25**(1), (2002), 1–20

[25] A. Debnath, O. Conlan, A Critical Analysis of Empathetic Dialogues as a Corpus for Empathetic Engagement. In *EmpathiCH workshop (EMPATHICH '23)*, (2023) doi: https://doi.org/10.1145/3588967.3588973

[26] L. J. Jacobsen, K. E. Weber, The Promises and Pitfalls of Large Language Models as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback, *AI*, **6**(35), (2025), doi: https://doi.org/10.3390/ai6020035