Dialoguing with Hattie&TimperlAI: a Research Based AI-Feedback Coach for Math Teacher Training

Silvia Funghi[1], Marta Valentini[2], Agnese Del Zozzo[2(✉)], and Alberto Montresor[2]

[1] Università di Genova, Genova, Italy
silvia.funghi@unige.it
[2] University of Trento, Trento, Italy
{marta.valentini-1,agnese.delzozzo,alberto.montresor}@unitn.it

Abstract. This paper presents Hattie&TimperlAI (H&TAI), a research-based AI agent that acts as a feedback coach for mathematics teacher education grounded in the Hattie and Timperley's model. Building on a previous version (TP-H&TAI), which generates examples of effective feedback for classroom use, we introduce TT-H&TAI, a new configuration designed to help teachers refine their own feedback during training. TT-H&TAI was implemented on three LLM platforms (ChatGPT, Gemini, and Claude) and evaluated through controlled dialogue simulations based on a standardized mathematics task. A qualitative content analysis examined how each implementation supported teachers' reflection on feedback practices. Findings show that all platforms fostered the use of the Hattie and Timperleys framework for analyzing feedback, with Claude additionally supporting the construction and reformulation of feedback. The paper discusses the differences in relation to prompt configuration and proposes a replicable, research-grounded workflow for developing AI tools for teacher education.

Keywords: AI in education · Teacher training · Formative feedback

1 Introduction

Generative AI has rapidly become a key instrument for supporting human work in areas once considered impossible to delegate to machines. As AI permeates daily life, its presence in classrooms and educational contexts is likewise contributing to the ongoing transformation of teaching and learning practices.

From a broad perspective, Holmes and Tuomi [11] offer a comprehensive survey of AI in education, outlining a taxonomy of AI-based solutions targeting students, teachers, and institutions. These systems are primarily designed to relieve teachers of repetitive or routine tasks, such as automated grading or administrative management, thereby freeing up time for human interaction and higher-level pedagogical activities. This category also includes tools that support

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
M. Adamoli et al. (Eds.): HELMeTO 2025, CCIS 2958, pp. 354–371, 2026.
https://doi.org/10.1007/978-3-032-31850-3_22

formative assessment, learning material development, classroom management, and, more broadly, teachers' professional practice.

Nevertheless, the literature reviewed by Holmes and Tuomi supports the view that, rather than treating AI merely as an automation tool, "the transformative effect of AI could be in augmenting human cognition in learning" (p. 543). In this view, AI is not intended to replace human activity, but to enhance and extend human cognitive processes, especially in complex tasks that continue to demand human judgment and oversight.

Even within disciplinary educational research, such as mathematics education, this broader vision of AI as a cognitive partner has begun to take shape. A growing body of work investigates both the opportunities and the risks of AI-supported instruction in this field. For example, a recent special issue [20] includes a scoping survey on ChatGPT's role in mathematics education [19]. These studies show that much of the current work conceptualizes AI primarily as a tool for teachers' professional practice. At the same time, they call for greater attention to AI's role in learning processes themselves, shifting the focus from assistance in professional tasks to support for understanding and reflection. Nevertheless, the literature warns of well-known limitations (such as hallucinations, propagated bias, and potential over-reliance on automated responses), which may undermine both the accuracy of support and learner autonomy.

The shift from using AI merely as a professional aid to exploring how it can mediate learning processes has important implications for teacher education as well. If teachers are to guide students in an AI-rich world, one of the most effective ways to prepare them is to let them experience AI as a learning partner firsthand. Recent research highlights the educational potential of AI to support both pre-service and in-service teacher development [5,15]. Specifically, AI can serve as a scaffolding tool for teachers' professional reflection, helping them articulate reasoning, confront alternative interpretations, and refine their pedagogical decisions. For instance, Noster et al. [17] investigate pre-service teachers' mathematical content knowledge while solving tasks with ChatGPT; Cheng [3] examines AI assistance in effective task formulation; and Schorcht et al. [23] study AI-mediated collaborative task design.

Among the different areas where AI can support teachers' work, formative assessment, and in particular the formulation of formative feedback, stands out as especially significant [5]. Extensive research has shown that formative feedback plays a crucial role in fostering students' learning processes [12]. At the same time, it is a demanding and delicate practice: feedback can easily lose its educational value, or even become counterproductive, when it is vague, overly evaluative, or misaligned with learning objectives [10,13]. A recent study [9] also points to the potential of AI to assist teachers in producing personalized feedback on students' work, while warning of risks such as over-praise (excessive positive evaluation) and over-inference (feedback based on inconsistent interpretations of students' actual performance).

Going beyond the general potential of AI for feedback generation, our previous work proposed a research-based approach to configure *AI agents* that

embody established educational frameworks [7]. In this context, the term *AI agent* refers to a conversational chatbot based on a large language model (LLM) that is explicitly configured through dialogue guidelines and a dedicated knowledge base. This configuration allows the agent to act according to a defined pedagogical identity. Such agents differ from generic chatbots because they are designed to enact theoretical principles within the dialogue itself, turning interaction into a space for reflection rather than mere information retrieval.

Following this approach, we developed Hattie&TimperlAI (H&TAI), an AI feedback coach informed by the model of Hattie and Timperley [10]. The first version of H&TAI (TP-H&TAI, Teacher Practice H&TAI) was designed to provide teachers with examples of effective feedback to support their classroom practice.

Building on this foundation, the present paper introduces a second version of the agent TT-H&TAI (Teacher Training H&TAI), reconceived as a learning companion for teachers. Rather than producing feedback examples, TT-H&TAI engages teachers in a process of reflection and refinement of the feedback they have already produced, thereby promoting metacognitive awareness of the dimensions and levels of formative feedback. This paper outlines the design process that led to this new configuration and its implementation across different LLM platforms, and examines the learning opportunities and dialogic patterns that emerge in user–AI interaction.

It is worth clarifying that the notion of *agent* can refer to two different operational instantiations. In our prior study, the agent was deployed as a preconfigured system, with the configuration embedded in the runtime setup and not exposed to end users (e.g., custom GPTs or Gemini Gems). In the present paper, by contrast, we make the configuration explicit and portable by providing a configuration prompt that users can manually paste at the beginning of a chat to instantiate the same pedagogical behavior within a standard LLM interface.

2 Theoretical Framework

LLM-based chatbots are essentially dialogic entities that simulate human language conversations. Human languages, as used by AI, typically require a grammatical *subject*. As a result, simulating human dialogue almost inevitably takes the form of utterances that involve an "I" corresponding to the speaker. It follows that the chatbot is, in a sense, compelled to adopt some form of identity, however loosely defined it may be. In particular, Qian et al. [21] state: "*the personality of a chatbot refers to the character that the bot plays or performs during conversational interactions. In this scenario, personality can be viewed as a composite of the identity (the background and profile) that a chatbot is endowed with, linguistic style that an agent exhibits during interactions, and many more explicit and implicit cues that may portray character*" (p. 1). In this regard, the example provided by Cheng et al. [4] is especially relevant: they propose a chatbot whose identity emerges progressively through interaction with the user.

To use the terms of Qian et al. [21], the character we intended for H&TAI to embody was that of a feedback expert. As a scientific reference for shaping

H&TAI's identity, we chose the seminal work of Hattie and Timperley [10]. In their paper, feedback is defined as "*information by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding, given in order to fill the gap between what is understood and what is aimed to be understood*" (pp. 81–82). Through an analysis of existing literature on feedback, the authors observe that it can be a highly influential tool in students' learning process, but that its impact (positive or negative) depends greatly on how and when feedback is delivered. They note that praise, rewards, and punishments, which are typically not accompanied by information for improvement, are generally minimally effective. Conversely, feedback that highlights an error but also offers pathways for improvement can have a significant positive impact on student progress.

To better frame what constitutes effective feedback, Hattie and Timperley proposed a model in which feedback addresses three major questions:

- *Where am I going?* (feed-up dimension - FU), concerning the expected learning objectives;
- *How am I going?* (feed-back dimension - FB), concerning information about the student's progress towards the learning objectives;
- *Where to next?* (feed-forward dimension - FF), concerning guidance for the student on further steps to "close the gap".

Each question can address one of the following four levels (descriptions taken from [10], p. 87):

- *Task-level Feedback* (TF): focused on how well the task is understood or performed;
- *Process-level Feedback* (PF): focused on the main process needed to understand how to do the task;
- *Self-regulation-level Feedback* (SRF): focused on the student's self-monitoring, directing, and regulating of actions;
- *Self-level Feedback* (SF): focused on personal evaluations and affect about the learner.

Nevertheless, Hattie and Timperley highlight that feedback is most powerful when it promotes deeper understanding and self-regulation; in other words, PF and SRF are the feedback types most likely to have the greatest impact on student learning. TF can be useful when it does not merely point out an incorrect product but also invites reflection on the thought process behind that product. Finally, SF is considered the least effective type of feedback.

The other theoretical foundation that informed the H&TAI design concerns the role of the teacher's interpretation of errors in structuring an instructional intervention, i.e. a feedback response. In mathematics education research, it is widely acknowledged that handling students' errors is a fundamental teaching competence [18]. This has also been conceptualized as a specific component of teacher knowledge related to the interpretation of students' reasoning behind their productions, often referred to as interpretative knowledge [22]. However,

providing students with feedback that facilitates their understanding of their own errors and fosters reflection on the underlying thought processes is far from straightforward. Zan [24], in particular, highlights that the effectiveness of a teacher's intervention (in this case, feedback) depends on its coherence with the actual difficulty that led the student to make a specific error. Zan emphasizes that teachers' actions always stem from their interpretation of the causes underlying a student's mistake–a notion consistent with Guo et al. [9] regarding AI's tendency to over-infer when generating feedback–and that teachers should have a "repertoire of interpretations" of students' behaviors.

A first version of H&TAI, hereafter TP-H&TAI (Teacher-Practice H&TAI), was developed to provide example feedback to teachers and to support them in their teaching practice. In [7], we compared the feedback produced by this agent with that generated by basic chatbots. We found that interactions with TP-H&TAI (both in its ChatGPT and Gemini implementations) were consistent with the Hattie and Timperley's framework, whereas basic chatbots tended to produce short lectures rather than coaching dialogues.

Building on these findings, we collected perspectives from in-service teachers who interacted with TP-H&TAI. Their insights guided the design of a new version, focused more specifically on fostering teachers' learning about formative feedback exchange in the context of teacher training. This paper presents this new version of H&TAI, hereafter TT-H&TAI (Teacher-Training H&TAI), obtained by modifying the previous configuration prompt to support a process of refinement of user-generated feedback rather than providing AI-generated feedback proposals.

Once the configuration prompt for TT-H&TAI was finalized, the model was implemented on ChatGPT, Gemini, and Claude. We then investigated the following research questions:

1. To what extent does the dialogue with TT-H&TAI offer teachers the opportunity to learn about formative feedback exchange?
2. Considering the implementation of TT-H&TAI on the ChatGPT, Gemini, and Claude platforms, what differences are there in the way they provide support for prompting teacher reflections?

3 Methodology

To provide a complete overview, this section describes the design of both TP-H&TAI and TT-H&TAI, with greater emphasis on the latter.

3.1 TP-H&TAI Design

In order to design H&TAI, we[1] began in April 2025 with a very first version of the AI Agent structured around Hattie and Timperley's paper. Many cycles of

[1] In all the paper, "we" refers exclusively to the author team: AI was used only to refine the formulation of prompts for H&TAI and not to support or perform any reasoning about its development.

test and refinement were set in motion by the critical analysis of early versions' behavior in dialogue simulations in the subsequent phases.

Moreover, two documents were inserted as a knowledge base for the chatbot: i) the foundational article on feedback exchange by Hattie and Timperley [10]; ii) an institutional document describing Italy's national learning goals for grades 1–8 school mathematics [16].

Phase 0 – H&TAI's *identity definition.* We created a prompt embedding the key traits of effective feedback, drawing on the framework by Hattie and Timperley [10]. The agent's role was strictly limited to providing feedback based on these theoretical principles, and it was not specified whether it was addressed to students or teachers (potentially both).

Phase 1 – Simulation of dialogue with a student. To evaluate whether H&TAI internalized its theoretical identity, we simulated a dialogue between the agent and a "student" prompted by a middle school math task. The student was played by us to control responses and elicit realistic misconceptions. This controlled setting allowed us to fine-tune the agent's behavior and assess whether its feedback aligned with Hattie and Timperley's four levels. A first revision of the configuration prompt was done after this phase, to require the AI to give different interpretations for the student's error [24].

Phase 2 – Teacher–Coach–Student Simulation. We simulated a realistic classroom scenario inspired by Cusi et al. [6], involving three roles: a teacher (played by ourselves), a student (simulated by a chatbot) and H&TAI. In this setting, the "teacher" interacted with H&TAI describing the task and the student's mistake, asking for advice. H&TAI's answer was used by the teacher to give feedback to the student, whose response was turned over again by the teacher to H&TAI to ask for new advice, and so on. This controlled setting enabled us to fine-tune the agent's behavior and to observe the agent's support in a dynamic, practice-oriented context. The prompt was structurally optimized, also with the aid of AI itself, to guide AI's internal reasoning process consistently and to emphasize references to the different levels of feedback.

Phase 3 – Field Evaluation. We involved in-service teachers, who interacted with H&TAI using examples drawn from their own classroom practice. This was done to inform the refinement of the agent's configuration and provide insight into its perceived usefulness and usability in authentic teaching contexts.

The progressive refinement process of TP-H&TAI's configuration prompt ultimately led to the last version of TP-H&TAI[2]. The design process involved several key iterations, each step building on the previous one to sharpen the agent's pedagogical and structural capabilities. Crucially, insights from teachers in the final phase led to a revision of H&TAI's objective: instead of suggesting feedback, it now offers proposals for reflection on or refinement of a feedback statement already formulated by the teacher, resulting in a new type of agent, TT-H&TAI.

[2] https://tinyurl.com/ICTMT17.

3.2 TT-H&TAI Design

Building on the final prompt revised during Phase 3 of the TP-H&TAI design process, a new round of revisions was conducted to fine-tune the agent's behavior and better support teachers' reflection. In the TP-H&TAI configuration prompt, the key rules of conduct for the agent include:

1. ask the teacher what are the learning objectives (feed-up);
2. not provide complete solutions for students;
3. ask questions one at a time (to temper the intrinsic verbosity and proactivity of generative AI);
4. propose a list of possible interpretations of the error;
5. provide PF–SRF based on the interpretation identified by the teacher;
6. give priority to PF and SRF with respect to TF, avoid SF;
7. invite the teacher to reflect on how they will use feedback with the student (feed-forward).

In configuring TT-H&TAI, this list of key rules was revised and expanded asking also to:

8. accompany the list of possible error interpretations with reflective questions, inviting the teacher to consider whether further interpretations are possible;
9. ask the teacher for spontaneously offered feedback;
10. provide a synthesis outlining the three dimensions (FU, FB, FF) and the four levels (TF, PF, SRF, SF), and ask the teacher to classify their feedback according to these descriptions;
11. provide a justified classification of the teacher's feedback in terms of levels and dimensions, and engage in dialogue with them to propose reformulations that incorporate any missing dimensions, operating at the PF/SRF levels;
12. adopt a Socratic approach and formulate a series of questions to the teacher to stimulate meaningful reflection on the construction and phrasing of feedback, and when needed facilitate a deeper comprehension of the Hattie and Timperley model.

To test the resulting configuration prompt[3], we simulated a scenario involving two roles: a teacher (played by ourselves) and TT-H&TAI. In this case as well, the "teacher" interacted with TT-H&TAI by describing the task and the student's mistake, and by providing indications related to learning objectives, chosen interpretation, and corresponding feedback.

3.3 A TT-H&TAI-Teacher Simulation

Simulations with TT-H&TAI included tests with ChatGPT (GPT-5), Gemini (Gemini 2.5 Flash), and Claude (Claude Sonnet 4.5)

[3] https://tinyurl.com/CONF-PROMPT-V7.

This paper analyzes a simulated teacher-agent interaction with TT-H&TAI, where we (the researchers) assumed the teacher role and consulted the agent to critically reflect on feedback related to an incorrect task solution. The task is derived from an item of the Italian standardized test (INVALSI) designed for Grade 8 students, which features an image of a racetrack (Fig. 1). The problem is the following: "*The figure represents the layout of a track composed of the following elements: two circular arcs, each with a radius of 50 cm; two straight segments, each 100 cm long, which are perpendicular to each other at their midpoint. What is the total length of the track? Show the calculations used to find the answer and finally state the result. Result: approximately ___ cm.*" (our translation).

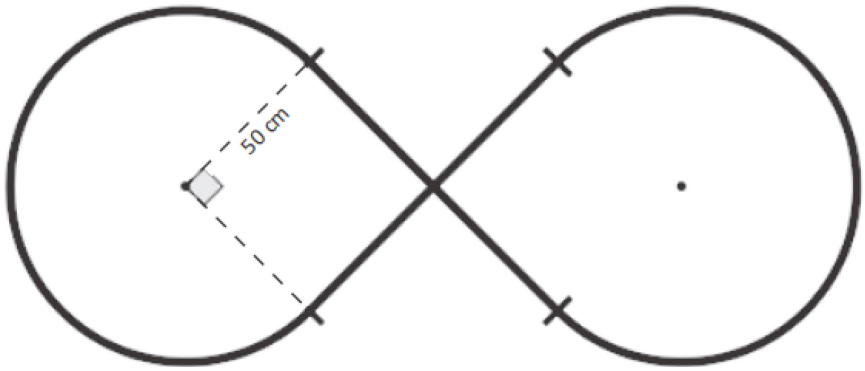


Fig. 1. The figure associated with the test problem

The specific erroneous student solution is adapted from Asenova et al. [1], who documented a student protocol consisting of the following steps: (a) $100 + 100 = 200$; (b) $\pi \times radius^2 = 7850$; (c) $7850 \div 2 = 3925$; (d) $7850 + 3925 = 11775$; with the student's final answer being approximately 11975 cm. The authors propose a possible interpretation of the reasoning underlying this solution, suggesting that step (a) corresponds to the sum of the straight segments' lengths, and step (b) to the calculation of the full circumference (although the student appears to have used the formula πr^2 instead of $2\pi r$). Steps (c) and (d) allow for the calculation of the lengths of the circular arcs, which together comprise two parts, each equivalent to $3/4$ of the full circumference. Since $3/4 + 3/4 = 3/2 = 1 + 1/2$, we can infer that the student may have computed half of the full circumference in step (c) and then added it to the full circumference in step (d). The final answer would thus correspond to the sum of the values obtained in steps (a) and (d).

Considering the INVALSI classification, which identifies problem-solving as the primary competence addressed by this task, one learning objective in proposing this problem can be defined as follows: "to solve accessible problems across all content domains, maintaining control over both the resolution process and the final results. The student can describe the procedure followed and recognize solution strategies that differ from their own" ([16], p. 61, our translation).

This item was selected because it allows for feedback formulation at all levels. A TF would address just the incorrectness of result or formula used in (b) (e.g.,

"the formula you used in the second step computes the area, not the circumference"). A PF would address the more general strategy selection process (e.g., "explain how you determined the length of the curvilinear parts"). A SRF could focus on strategies to check solution acceptability (e.g., "try to consider how you can check if the result is plausible within the context"). Furthermore, a teacher's management of the FF dimension will depend on their specific interpretation of the student's mistake in the formula in step (b). If the teacher considers it a minor oversight, then, in relation to the aforementioned learning objective, the emphasis should predominantly be on control processes, particularly the verification of results. Conversely, if the teacher suspects the student is conceptually confusing the perimeter and area of a figure, then they may place greater emphasis on the definitions of the two concepts and on differences between them.

Once the task was selected, we implemented TT-H&TAI across Gemini, Claude, and ChatGPT and initiated the conversation with each of them. In all cases, the first input was the configuration prompt (see Footnote 3).

4 Data Analysis

All the chats were collected and organized as transcriptions of real conversations. To facilitate comparison, we created a single table containing all three conversations together[4]. We then applied an approach inspired by qualitative content analysis [14] to these three conversations, examining the interventions of the AI acting as a coach and focusing on comparing the support modalities implemented by TT-H&TAI across the three platforms.

In general, among the three dialogues with TT-H&TAI, the one with Claude appears to be the most concise in its verbal output, whereas the other two are more verbose and analytical. The overall structure of the three conversations is very similar, apart from some supplementary "digressions" encountered with TT-H&TAI implemented on Claude before returning to the main conversational flow—which adheres to the structure of the configuration prompt. The conversational flow unfolds as follows (referring to the lines of the table):

1. Agent presentation and request for the instructional objective (lines 3–4).
2. Request for the student's production (lines 5–6).
3. Transcription of the task and the student's solution, and request for confirmation (lines 7–10).
4. Error analysis and proposal of interpretations, asking the teacher for their point of view (lines 11–12).
5. Request for feedback from the teacher (lines 13–14).
6. Synthesis of the three dimensions and four levels of Hattie and Timperley and request for the teacher to classify the feedback (lines 15–16).
7. Analysis of feedback from TT-H&TAI, accompanied by questions and, potentially, proposals for reformulation (lines 17–21).

[4] https://tinyurl.com/SIMULATION-V7.

Given general chatbots' interaction modalities and the specific structure of the configuration prompt, all three TT-H&TAIS behave consistently with the request to activate the teacher and all conclude their turns with a question.

Considering the teacher's turns, several reflective processes can be activated: reflection on learning objectives (line 4); monitoring and, if necessary, correction of the AI's response (line 8); reflection on the evidence collected to support their own interpretation, critical examination of alternative interpretations, and explicit articulation of their position (line 12); formulation of feedback (line 14); application of Hattie and Timperley's dimensions and levels as tools for feedback analysis (line 16); and further reflection on feedback (line 18 and following).

The learning opportunities related to feedback exchange therefore concern two main aspects: the dependence of the provided feedback on the chosen interpretation (a point not yet explicitly addressed in the three dialogues with TT-H&TAI), and the use of Hattie and Timperley's framework both to analyze feedback and to improve its quality. In the dialogue with TT-H&TAI in Claude, these opportunities are further enriched by the possibility of reflecting on a level not present in the initial response and of applying Hattie and Timperley's framework for feedback construction and reformulation.

The qualitative analysis of the conversations led to the identification of several categories describing differences in the dialogic behavior of the three TT-H&TAI implementations, referred to as H&TAI-GEM (Gemini), H&TAI-CLD (Claude), and H&TAI-GPT (ChatGPT).

We considered each interaction with TT-H&TAI as a unit of analysis. In the following, we refer to the lines of the table mentioned above (linked in Footnote 4), each of which corresponds to a single conversational turn between our research team and the chatbot.

The categories were constructed in two steps. First, we read all conversations and conducted an initial coding phase to describe the chatbots' actions in their turns and to identify differences between them (e.g., line 1 was coded as: "Presentation of the agent and request for indication of the learning objective. "Anticipated" request by H&TAI-CLD for the student's solution. Anticipation by H&TAI-GPT of the Hattie and Timperley model."). After this first round of coding, we observed that chatbot responses did not fully align with the intended coach behavior. In some cases, however, these issues were due to general technical limitations, such as H&TAI-CLD's difficulty in recognizing that the curvilinear parts of the track corresponded to 3/4. A second round of coding allowed us to construct four categories through a process of grouping the previous open codes: consistency with the teacher's initial stated learning objective, consistency with the interpretation provided by the teacher, consistency with the Hattie and Timperley model, and consistency with interpretation relative to the configuration prompt requests. We will comment on each category in detail below.

Consistency with the Teacher's Initial Stated Learning Objective. The stated learning objective of the chosen task was "solving problems, maintaining control over both the resolution process and the final results" (line 4), in agreement with the INVALSI test guidelines. When synthesizing the three dimensions and

four levels (line 15) and analyzing feedback (line 17), H&TAI-GEM and H&TAI-GPT shift focus to calculation of the track's length (e.g., *"Remember, the goal of the problem is to compute the total length of the track."*, line 17, H&TAI-GEM; *"it does not clearly state the overall objective ("finding the total length of the track")"*, line 17, H&TAI-GPT), whereas H&TAI-CLD maintains correspondence with the teacher's initially declared objective–even in the analysis of the reformulated feedback (*"How could you remind the student of the complete goal of the activity? (not just "find the answer" but also "maintain control over the resolution process")"*, line 19). It is noteworthy that despite this main trend, H&TAI-GEM recalls from time to time the initial stated objective, but not in the FU dimension (see the mention to *"key objectives (maintaining control over the process)"* when justifying why the formulated feedback lacks SRF, line 17).

Consistency with the Interpretation Provided by the Teacher. The teacher's interpretation referred to a confusion between the formula for the area of a circle and that for the circumference length (*"I believe the student incorrectly retrieved the formula for the circumference instead of the area formula from memory"*, line 12, conversation with H&TAI-GEM). In line 17, H&TAI-GPT generates a PF that is inconsistent with this interpretation (*"To improve, start again from the drawing: identify which fractions of the circumference are present and use the formula* $L = fraction \times 2\pi r$*"*): if the student correctly identified the curvilinear parts but simply recalled the wrong formula, asking them to start again from the drawing is unlikely to be useful. Moreover, some of the questions generated by H&TAI-GPT and H&TAI-CLD focus on the conceptual difference between area and length, rather than aligning with the teacher's initial interpretation. For instance, the self-regulation question *"Am I looking for an area or a length?"* (line 17, conversation with H&TAI-GPT) is minimally useful if the problem is a lapse in memory regarding the formula. Also process questions from H&TAI-CLD (line 19) appear slightly out of focus relative to the interpretation provided (*"How could you help the student conceptually distinguish between 'measuring a boundary' (circumference) and 'measuring a surface' (area)? What question could you ask them to make them reflect on the difference between the two measures?"*).

Consistency with Hattie and Timperley Model. In the synthesis of Hattie and Timperley's framework (line 15), H&TAI-GEM, reports also SF, while H&TAI-CLD and H&TAI-GPT do not, even if it could be helpful of SF to be explicitly stated. Despite this, the rationale for H&TAI-GPT discarding SF is unclear (*"This level should be avoided, as it is too generic (e.g., "You are intelligent", "Well done")."*). Both H&TAI-GEM and H&TAI-GPT accompany the definitions with example questions, which somewhat risks directing the teacher's feedback revision: the "quick examples" from H&TAI-GPT actually constitute the construction of feedback accordingly with Hattie and Timperley's framework (e.g., *"SRF (example): "Always check: (a) units of measurement, (b) whether you are using area or perimeter, (c) a quick estimate to check plausibility"."*, line 15).
The feedback provided by the teacher was formulated as a TF, which we deemed likely to be among those most commonly used by a teacher: *"Attention! The rea-*

soning is correct but you used the area formula instead of the circumference formula". The feedback analysis (line 17) across the three TT-H&TAIS is consistent: all of them evaluate the provided feedback as lacking FU and FF, and classify it as TF. The arguments concerning the different levels are generally coherent, except for the rationale behind FU absence (see the category *Consistency with the teacher's initial stated objective*).

It is interesting to observe that H&TAI-GEM and H&TAI-GPT explain not only why certain dimensions/levels are present in the feedback, but they also justify why they say others are *not* present (e.g., *"Absent or underexpressed dimensions: Feed-Up (FU): it does not clearly state the overall objective ("finding the total length of the track"). Feed-Forward (FF): it does not suggest concrete actions for improvement ("what to do to correct"). PF / SRF: it also does not yet address strategies or personal reflection on the error."*, line 17, conversation with H&TAI-GPT). H&TAI-CLD instead explains why the feedback lacks FU and FF dimensions, but does not motivate why the feedback is not PF or SRF: this is noteworthy, since H&TAI-GEM emphasizes the presence of a PF "nuance" (*"The feedback is predominantly situated at the Task Level (TF), but with a slight touch of the Process Level (PF)"*, line 17). In its feedback proposal (line 17), H&TAI-GEM suggests a structure consistent with the Hattie and Timperley model but classifies its feedback as PF, even though it offers the student no stimulus for personal reflection and merely instructs them to *"find the correct formula for length and recalculate"*. In the PF proposed by H&TAI-GPT (*"To improve, start again from the drawing: identify which fractions of the circumference are present and use the formula (L = *fraction* $\times 2\pi r$)"*, line 17), the formula presented (L = *fraction* $\times 2\pi r$) combines verbal and symbolic language and risks leading the student to memorize a new "rule" rather than reason about the process.

Consistency and Interpretation Relative to the Configuration Prompt Requests. We have observed that the general conversational flow follows the prompt's instructions. Regarding point 8 in Sect. 3.2, all three TT-H&TAIS accompany the interpretations with follow-up questions directed at the teacher (line 11). However, while H&TAI-CLD and H&TAI-GPT limit themselves to questions seemingly aimed at gathering evidence to support one or more interpretations (e.g., *"Have you noticed if the student systematically shows confusion between area and perimeter formulas?"*, line 11, conversation with H&TAI-CLD), H&TAI-GEM mixes these types of questions with those that anticipate reflection on the teacher's feedback (e.g., *"How can I help them build the habit of always writing down what each calculation step represents?"*, line 11).

Regarding point 11 in Sect. 3.2, a different interpretation of the meaning of "engage in dialogue [...] to propose reformulations of the feedback" emerges from H&TAI-CLD compared to the other two chatbots. H&TAI-GEM and H&TAI-GPT automatically generate reformulation proposals:

Reformulation Proposal (integrates TF, PF and SRF; FU-FB-FF). "The objective was to calculate the length of the track (FU). Your reasoning about combining the arcs and straight segments is correct (FB), but you applied the area formula

instead of the circumference formula, so the result does not measure a length (TF). To improve, start again from the drawing: identify which fractions of the circumference are present and use the formula ($L = fraction \times 2\pi r$) (PF). Before concluding a similar problem, always ask yourself: 'Am I looking for an area or a length?'—this will help you self-regulate in future exercises (SRF, FF)." (line 17, conversation with H&TAI-GPT).

H&TAI-CLD, instead, asks a question to help the interlocutor delve deeper into SRF (*"What could help the student distinguish between'when to use the area' and 'when to use the circumference' autonomously, without having to rely solely on memory?"*, line 17). In subsequent turns, it poses further questions to help the teacher add the missing feedback components (e.g., *"How could you phrase a concrete invitation to verify the reasonableness of the result?"*, line 19). Once new feedback is received from the teacher, H&TAI-CLD re-proposes an analysis against the three dimensions and four levels of Hattie and Timperley model and concludes with further probing questions (e.g., *"On the Missing Part: The student failed to explicitly state their reasoning in writing (as required by the assignment). In your feedback, you ask them to "explain in writing what each operation is for." Do you think this is sufficient guidance, or might the student need an example/model of how to articulate a step?"*, line 21).

Regarding point 12 in Sect. 3.2, once the user's feedback has been analyzed and commented on by the agent, the questions intended to stimulate further reflections on feedback are not always "Socratic":

- H&TAI-GEM proposes a generic question (*"What are your thoughts, Teacher? Do you believe this reformulation is more effective?"*, line 17) after providing its own formulation of the feedback, which, rather than triggering users' reflection, risks ending the conversation, since the goal of building complete feedback according to the Hattie and Timperley model has been achieved.
- H&TAI-GPT provides a question explicitly labeling it as 'Socratic' (*"Socratic Questions for Reflection (Step 5f). In your opinion, how might this type of reformulation help the student understand their error, instead of merely correcting it?"*, line 17), but since, also in this case, the question is asked after the agent has provided its own formulation of the feedback, it is unlikely to give rise to new reflections.
- The behavior of H&TAI-CLD, on the other hand, is more in line with the initial prompt requests: it explicitly invites the user to work on feedback formulation, without any suggestions (*"Let's Discuss Possible Reformulations. Now I would like to work with you to enrich this feedback, also including the FU and FF dimensions, and moving towards the PF and SRF levels."*, line 17). Then, H&TAI-CLD poses a question that effectively redirects the user to reflection on the initially identified problem, namely, the difficulty of remembering the correct formula. (*"Let's start with a Socratic question: What could help the student distinguish between 'when to use the area' and 'when to use the circumference' autonomously, without having to rely solely on memory?"*, line 17). In our simulation, the question actually presented us with an interesting issue that we had not yet considered, and allowed us to highlight both

"practical" strategies (checking the textbook) and metacognitive strategies (checking the plausibility of the result obtained with the formula remembered in relation to the units of measurement), which had not been made explicit until then (line 18). In the following, H&TAI-CLD is even more aligned with a Socratic approach, as in subsequent exchanges (lines 19 and 21) it continues to accompany its response with interesting questions for reflection (in line 19 on each of the FU, FB, and FF dimensions; in line 21, after the user's feedback has been classified as complete with respect to the Hattie and Timperley's framework, on other possible aspects of the feedback).

5 Results and Discussion

The analysis of the three conversations highlighted both how TT-H&TAI supports teachers' learning and how the three AI implementations differ in their dialogic support for reflection. All three versions of TT-H&TAI offer the learning opportunity to use Hattie and Timperley's framework for analyzing feedback. They could also support reflection on the consistency between feedback and error interpretation, but none of the three explicitly prompted this reflection. Only the interaction with H&TAI-CLD enabled teachers to use Hattie and Timperley's framework for constructing or reformulating feedback.

Regarding differences in the interactions with the three TT-H&TAIS, we identified several coherence issues, the most relevant of which relate to how specific parts of the configuration prompt were interpreted. In all three cases, the overall structure of the conversation proved to be quite faithful to the sequence of key rules listed within the configuration prompt (so much so that H&TAI-GEM and H&TAI-GPT explicitly refer to the points in the list of key rules in several turns). However, the synthesis of Hattie and Timperley's framework shows interesting differences: H&TAI-GEM and H&TAI-GPT generate accompanying examples related to the various dimensions and levels, which could preemptively suggest possible feedback components to the teacher, thereby diminishing the reflective potential of the interaction.

Furthermore, when analyzing the teacher's feedback, H&TAI-GEM and H&TAI-GPT not only provide their analysis but also generate a new formulation of a possible feedback; the subsequent proposal of follow-up questions might lead the interlocutor to believe that the work on the feedback is concluded, when in fact the proposed feedback is only one of the possibilities. H&TAI-CLD, on the other hand, refrains from providing reference examples throughout the conversation and prompts the teacher to reformulate the feedback, offering a reflection question consistent with the instructional objective indicated initially as support.

6 Conclusions

In this paper, we proposed a research-based approach to design AI personalized dialogical partners configurations for teacher training, exemplified by the

development of H&TAI, a prototype AI feedback coach grounded in the model by Hattie and Timperley [10]. Because LLM technology evolves rapidly, our goal was not to define a *definitive* prompt for designing a scientifically grounded AI agent, but rather to develop a stable and coherent process that supports comparison and adaptation over time.

In this paper, we have sought to outline such a process, consisting of the following stages: (i) prompt definition, (ii) internal testing, (iii) internal simulations, (iv) context-based trials, and (v) public release of the configuration prompt. This process is the main contribution we wish to emphasize, as it extends beyond the specific objective for which H&TAI was designed.

This study targets the internal testing and simulation phases of the TT-H&TAI design process, with researchers acting as teachers for the simulated interactions. Here, our aim was not to present TT-H&TAI as a finished "product", but to highlight its potential and reflect on possible prompt refinements that emerged during the process. We consider this important as a second contribution, since it allowed us to present a concrete example of the refinement process within our broader AI agent development workflow.

We distinguished between TP-H&TAI, developed to assist teachers by providing AI-generated feedback proposals, and TT-H&TAI, designed to support the refinement of user-generated feedback. The design of TT-H&TAI builds on the final version of the TP-H&TAI prompt configuration, which was developed through multiple rounds of simulation and refinement. We first described how TP-H&TAI's identity was defined and what changes were introduced to shape TT-H&TAI's identity. We then conducted an internal validation of the chatbot's behavior through a simulated coaching session involving a teacher and TT-H&TAI implemented on Gemini, Claude, and ChatGPT. The coaching dialogue focused on an INVALSI item and on an interpretation of the student's solution that allowed for feedback at different levels. Finally, we transcribed and analyzed the conversations using a qualitative content analysis approach.

Regarding the first research question, namely the opportunities that TT-H&TAI offers for teachers' learning about feedback exchange, the results suggest that TT-H&TAI can serve as a metacognitive tool to support teachers' reflection, particularly with respect to the three dimensions and four levels of feedback defined by Hattie and Timperley. These constructs proved effective as structured references for feedback analysis. Only the Claude implementation of TT-H&TAI allowed us to use this structure in producing feedback. This aspect should be improved in future versions of the configuration prompt. The chatbot's tendency to generate examples or guiding questions based on Hattie and Timperley's framework should be limited, to keep teachers actively engaged in creating feedback that covers all three dimensions and higher levels. Future updates of TT-H&TAI should also include prompts that ask users to reflect on how their feedback aligns with the identified error interpretation, and mechanisms that help the AI maintain consistency throughout the dialogue, especially in relation to the learning objective, the error interpretation, and the elements of Hattie and Timperley's framework. Another improvement to the configuration prompt

would be revising the key rules to avoid ambiguity or unwanted behavior (e.g., clarifying that teachers should reformulate feedback first, not AI).

Regarding the second research question, the interactions with the three TT-H&TAI implementations all showed some issues. However, the version implemented on Claude was, in this case, the closest to the desired behavior. The overall conversation flow, except for the final segment, was similar across the three implementations. This suggests that the key regulatory structure defined in the current prompt provides a certain degree of control over the trajectory of the dialogue. We hope that further refinements of the prompt will lead to interactions with stable characteristics, regardless of the implementation platform.

Future work will involve further exploration of how to improve the definition of rules of conduct, and in-field studies in two main directions. The first one is the teacher training dimension - with a field evaluation involving pre-service and in-service teachers to test TT-H&TAI in real contexts of teacher education or teacher professional development. This work is still in its preliminary stages: while initial tests have been conducted with teachers, the findings remain to be evaluated through an extensive trial. Nevertheless, it is possible to propose some initial hypotheses. Given its functional design, TT-H&TAI could prove useful in enabling teachers to familiarize themselves with the Hattie and Timperley's framework—not through a purely theoretical study of the model, but through "guided reformulation exercises" of feedback.

The fact that TT-H&TAI (the H&TAI-CLD version in particular) enabled users to reflect on how their feedback integrates across dimensions and levels is noteworthy, as it may help teachers view feedback not merely as error correction, but as a flexible tool tailored to specific instructional aims. For instance, reflecting on the possibility of providing process-oriented or self-regulatory guidance shifts the focus from the error itself to the student's stance toward a mathematical task. Furthermore, reflecting on the explicit articulation of learning objectives implicitly means to share the goal of fostering greater metacognitive awareness in students. Both of these aspects partially transcend performance on a specific task and relate to broader conceptions of the teaching-learning process. Consequently, the use of TT-H&TAI could indirectly encourage reflection on the role of the teacher. Teachers are no longer merely "judges" of task correctness, but rather, they are experts who guide each student toward reflective awareness of their own learning. The extent to which such AI-mediated reflection can reshape professional identity remains a compelling avenue for future longitudinal research.

As a second direction, we intend to embrace more recent lines of research that shift the attention from feedback formulation (which is the actual focus of both versions of H&TAI) to the feedback exchange as a process [8]. This shift requires studies that focus not only on how teachers formulate feedback, but also on how students use feedback, considering also long-term perspectives [2]. These studies would provide a clearer understanding of possible pedagogical implications of using an AI feedback coach like TT-H&TAI, not only in teacher training but also in classroom practice.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Asenova, M., Del Zozzo, A., Santi, G.: Unfolding teachers' interpretative knowledge into semiotic interpretative knowledge to understand and improve mathematical learning in an inclusive perspective. Educat. Sci. **13**(1), 65 (2023)
2. Carless, D.: Feedback loops and the longer-term: towards feedback spirals. Assessment Evaluat. Higher Educat. **44**(5), 705–714 (2019)
3. Cheng, E.C.: Leveraging generative AI in science lesson study: transforming density concept instruction through ChatGPT integration. Int. J. Lesson Learn. Stud. (2025)
4. Cheng, Y., et al.: Evolving to be your soulmate: personalized dialogue agents with dynamically adapted personas. arXiv e-prints pp. arXiv–2406 (2024)
5. Chu, Z., et al.: LLM agents for education: Advances and applications (2025). https://arxiv.org/abs/2503.11733, arXiv preprint
6. Cusi, A., Morselli, F., Sabena, C.: Promuovere strategie di valutazione formativa in matematica con le nuove tecnologie: l'esperienza del progetto FaSMEd. Annali online della Didattica e della Formazione Docente **9**(14), 91–107 (2017)
7. Del Zozzo, A., Funghi, S., Montresor, A., Valentini, M.: H&TAI: Improving formative feedback through a research-driven teacher coach. In: Proceedings of the 17th International Conference on Technology in Mathematics Teaching (ICTMT17). London, UK (2026). https://discovery.ucl.ac.uk/id/eprint/10223410/
8. Grion, V., Serbati, A.: Valutazione sostenibile e feedback nei contesti universitari. Pensa Multimedia (2019)
9. Guo, S., Latif, E., Zhou, Y., Huang, X., Zhai, X.: Using generative AI and multi-agents to provide automatic feedback. arXiv preprint (2024). https://arxiv.org/abs/2411.07407v1
10. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007)
11. Holmes, W., Tuomi, I.: State of the art and practice in AI in education. Eur. J. Educ. **57**(4), 542–570 (2022)
12. Hughes, G., Smith, H., Creese, B.: Not seeing the wood for the trees: developing a feedback analysis tool to explore feed forward in modularised programmes. Assessment Evaluat. High. Educat. **40**(8), 1079–1094 (2015)
13. Kluger, A.N., DeNisi, A.: The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol. Bull. **119**(2), 254–284 (1996)
14. Kuckartz, U.: Qualitative text analysis: a systematic approach. In: Compendium for early career researchers in mathematics education, pp. 181–197. Springer International Publishing Cham (2019)
15. Lee, U., et al.: Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers. In: NeurIPS'23 Workshop on Generative AI for Education (GAIED). Advances in Neural Information Processing Systems, Curran Associates, Inc. (2023)
16. MIUR - Ministero dell'Istruzione, dell'Università e della Ricerca: Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo d'istruzione. Annali della pubblica Istruzione (2012)

17. Noster, N., Gerber, S., Siller, H.S.: Pre-service teachers' approaches in solving mathematics tasks with ChatGPT: a qualitative analysis. Digital Experiences Math. Educat. **10**(4), 543–567 (2024)
18. Novotná, J., Moraová, H., Tatto, M.T.: Mathematics teacher education organization, curriculum, and outcomes. In: Encyclopedia of mathematics education, pp. 1–7. Springer (2018)
19. Pepin, B., Buchholtz, N., Salinas-Hernández, U.: A scoping survey of ChatGPT in mathematics education. Digital Exper. Math. Educat. **11**, 9–41 (2025)
20. Pepin, B., Buchholtz, N., Salinas-Hernández, U.: "Mathematics education in the era of ChatGPT: investigating its meaning and use for school and university education"—editorial to special issue. Digit. Exper. Mathematics Education **11**, 1–8 (2025)
21. Qian, Q., Huang, M., Zhao, H., Xu, J., Zhu, X.: Assigning personality/identity to a chatting machine for coherent conversation generation. arXiv preprint arXiv:1706.02861 (2017). https://arxiv.org/abs/1706.02861v3
22. Ribeiro, M., Mellone, M., Jakobsen, A.: Interpreting students'non-standard reasoning: insights for mathematics teacher education. Learn. Math. **36**(2), 8–13 (2016)
23. Schorcht, S., Peters, F., Kriegel, J.: Communicative AI agents in mathematical task design: a qualitative study of GPT network acting as a multi-professional team. Digit. Exper. Math. Educ. **11**(1), 77–113 (2024)
24. Zan, R.: Difficoltà in matematica: osservare, interpretare, intervenire. Springer (2007)