

Article

A Conversational AI Framework for Teacher Training

Sergio Miranda ^{1,*}  and Antonio Calvani ²

¹ Department of Humanities, Philosophy and Education, University of Salerno, 84084 Fisciano, Italy

² S.Ap.I.E., Society for Learning and Education Informed by Evidence, 10124 Torino, Italy; antonio@calvani.it

* Correspondence: semiranda@unisa.it

Abstract

This paper introduces ETQ-AI (Effective Teaching Questionnaire—Artificial Intelligence), a conversational AI framework designed to support teacher professional development by identifying pedagogical misconceptions and fostering critical reflection. Building on Evidence-Informed Education (EIE) principles, the system integrates a 50-item diagnostic questionnaire with an automated, Gemini-driven Explainable AI (XAI) feedback layer. When teachers complete the diagnostic form, they receive personalized feedback and can engage in dialogic interactions with Google Gemini to explore the underlying meta-analytic evidence (e.g., Hattie's effect sizes, Cognitive Load Theory) or argue personal objections. We evaluated system fidelity against an expert panel benchmark ($N = 64$ items) and conducted a pilot evaluation with 123 educators (70 primary teachers, 40 secondary teachers, 13 school leaders) recruited at an EIE conference. Gemini demonstrated high alignment with expert evaluations (97%), matching or exceeding static baseline feedback in user satisfaction (Mean = 4.1/5 vs. 3.8/5). Qualitative thematic analysis of 34 participant objections revealed key areas of tension, particularly around contextual constraints and pedagogical subjectivity. While the findings validate the technical proof-of-concept, system usability, and user engagement of ETQ-AI as an interactive scaffolding interface, we emphasize that this pilot study measures user perceptions and conversational patterns rather than direct pre/post shifts in critical reflection or classroom instructional practice.

Keywords: generative AI; educational innovation; learning design; evidence-informed education (EIE); teacher training; explainable AI (XAI) in learning; digital learning governance; instructional misconceptions

1. Introduction

Artificial Intelligence (AI) is generating new possibilities across educational contexts, and its integration into school environments is actively promoted by policy frameworks [1]. Its implications for teacher training, however, remain largely exploratory.

To establish conceptual clarity, this study distinguishes between three interrelated cognitive dimensions: awareness, critical thinking, and critical reflection. Awareness refers to the initial, conscious recognition of one's explicit or implicit pedagogical assumptions and potential teaching misconceptions. Critical thinking involves the cognitive, analytical evaluation of teaching strategies against empirical evidence, meta-analytic data, and logical coherence. Critical reflection, following Mezirow's transformative learning theory [2] and Schön's reflective practitioner model [3], goes beyond technical evaluation to examine the underlying epistemological, cultural, and contextual assumptions that guide instructional decisions. ETQ-AI is specifically structured to bridge awareness and critical reflection by



Academic Editors: Rui Chibante and Kleopatra Nikolopoulou

Received: 9 September 2026

Revised: 9 October 2026

Accepted: 13 October 2026

Published: 15 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

using diagnostic disconfirmation to trigger cognitive conflict and dialogic argumentation. With this work, we present an environment that uses AI to identify teaching misconceptions and, more generally, to reflect on the tacitly assumed theoretical assumptions that underlie teachers' ordinary actions.

Over recent decades, the themes of teacher professionalism and reflexivity, including the conceptual frameworks used to understand and apply pedagogical knowledge, have gained prominence in the Italian educational debate [4–9].

In this context, we can also place research on mental frameworks (mind frames), general orientations that should be adopted by school managers and teachers, to which Hattie attributes the greatest relevance for the improvement of learning.

A more specific aspect concerns the models of “good teaching” that teachers implicitly adopt throughout their teaching activities and which underpin their ongoing decisions. While non-cognitive aspects undoubtedly play a role in this regard, it is difficult to deny the influence of tacit assumptions that may nonetheless surface; for example, in a simple lesson study session, when a tutor or classmate, watching a video recording of a lesson, suggests a different action than the one actually performed.

Implicit teaching theories, which are embedded in teachers' decision-making, result from knowledge internalized through personal experience and cultural immersion—an environment not immune to stereotypes, clichés, and biases that may constitute cognitive obstacles to more effective practice.

It is widely acknowledged in school contexts that some teachers consistently achieve better outcomes with students and respond more effectively to challenging situations. And even large-scale research confirms these differences which clearly emerge from a careful observation of a classroom lesson.

Expert teachers, who do not necessarily have more experience or more training credentials, tend to hold qualitatively different pedagogical concepts and action dispositions. They demonstrate, for instance, a clearer vision of lesson organization, learning objectives, pacing, and student interaction, while avoiding certain counterproductive attitudes. These characteristics find strong support in recent scientific research (see Evidence-Based Education, EBE, below).

A core premise of this work is that current research offers sufficiently robust tools to support more targeted teacher training interventions, specifically by transferring expert attitudes and behaviors that must first be understood theoretically and then practiced through laboratory activities and lesson study, which rank among the most effective methods available [10] ($ES = 0.6$). However, it is important to acknowledge that the definition of pedagogical expertise is a subject of ongoing debate within education sciences. While this study adopts a benchmark based on Evidence-Informed Education (EIE) principles, we recognize that educational research involves legitimate epistemological disagreements. The reference model used here is not intended as a definitive truth, but as a highly reliable framework for reflection that remains open to contextual adaptation and critical scrutiny. Along this path, awareness of potentially inadequate concepts, which AI can help identify and raise awareness, can be crucial.

This work situates itself within the growing international discourse on Large Language Models (LLMs) in teacher education, moving beyond traditional static tools toward Conversational Tutoring Systems (CTS) and Explainable AI (XAI). Recent scholarship suggests that AI-mediated professional reflection can bridge the gap between theoretical evidence and classroom practice by providing dialogic, personalized feedback. Recent international research has increasingly highlighted how these tools support continuous professional development by offering real-time insights that allow teachers to refine their instructional methods and address diverse learner needs [11]. Studies have also demonstrated that AI

can foster reflective, collaborative environments, such as AI-enhanced Professional Learning Communities, which help pre-service and in-service teachers transition from initial curiosity to pedagogically informed engagement [12]. Furthermore, the emergence of the ‘hybrid teaching intelligence’ paradigm emphasizes a synergistic collaboration where LLMs assist with task automation and content creation, while the educator maintains agency and intentionality over pedagogical decisions [13]. Despite the accelerating integration of large language models in teacher education, a critical gap remains unaddressed in the current literature. While traditional diagnostic tools, such as static questionnaires, successfully isolate teacher misconceptions, they lack the interactive scalability to provide customized, dialogic explanations or accommodate the situational nuances of actual classrooms without human tutor intervention.

Conversely, general-purpose generative AI systems frequently lack transparency in their reasoning, exhibit risks of ungrounded or unverifiable outputs, and are not strictly aligned with verified evidence-based frameworks, rendering them unreliable for structured teacher training.

By constraining the generative model’s systemic prompts to an Evidence-Informed Education benchmark, the ETQ-AI architecture reduces off-topic domain drift, enforces adherence to prompt constraints, and directly addresses key demands for ethical constraints and algorithmic governance in digital learning processes. Rather than attempting to eliminate algorithmic opacity entirely, the system provides pedagogical rationale explainability by offering teachers auditable, evidence-backed justifications grounded in synthesis literature (e.g., Hattie’s meta-analyses [10,14], Cognitive Load Theory [15]). This structural design transforms user interaction from opaque text generation into a transparent, responsible, and auditable professional dialogue.

This evolving landscape provides a broader context for our proposed environment, which focuses on integrating such reflective AI capabilities with Evidence-Informed Education principles.

Rather than treating technology in isolation, this study explores the dynamic interplay between generative AI capabilities and the pedagogical structures of learning design [16]. By utilizing the Gemini Large Language Model as an evaluative and ideational partner, the proposed framework supports educators during the critical phases of lesson formulation and curriculum design. The ETQ–Gemini environment serves as a predictive sandbox where teachers can test instructional hypotheses against Evidence-Informed Education (EIE) tenets, transforming generative AI into an interactive tool for professional inspiration and conceptual correction.

Related Work and Comparative Positioning

While the application of Large Language Models (LLMs) in teacher education has expanded rapidly, existing approaches generally fall into three distinct categories: (a) Static Diagnostic Questionnaires (e.g., OECD TALIS framework), (b) General-Purpose Conversational LLMs without theoretical prompt boundaries, and (c) Rule-Based Automated Intelligent Tutoring Systems (ITS, e.g., AutoTutor [17]) that offer structured dialogic turns but lack generative adaptability. As summarized in Table 1, ETQ-AI combines evidence-informed diagnostic constraints with a dynamic conversational layer.

Table 1. Comparative matrix of AI frameworks in teacher education (comparative criteria (High/Medium/Low) represent heuristic classifications grounded in comparative analyses of AI and conversational agents in teacher education literature [11,13]. Diagnostic Rigor is the presence of validated item scales; Dialogic Adaptability is the capacity for unscripted, natural language response; Theoretical Scaffolding is structural constraint to verified evidence frameworks; Pedagogical Traceability is the ability to trace system rationale to published meta-analytic data).

Framework Type	Diagnostic Rigor	Dialogic Adaptability	Theoretical Scaffolding	Pedagogical Traceability
Static Surveys (e.g., OECD TALIS framework; traditional ETQ, paper instruments)	High	None	High	Low
Unconstrained LLMs (e.g., general GPT-4/Gemini queries without system prompt boundaries)	Low	High	Low	Very Low
Automated Intelligent Tutors (ITS) (e.g., AutoTutor [17], rule-based pedagogical agents)	Medium	Low/rule-based	Medium	Medium
ETQ-AI (Proposed Framework)	High	High	High (EIE-Constrained)	High (XAI Evidence Grounding)

2. Materials and Methods

2.1. EBE/EIE

This work is based on the observation that research advances now allow us to identify a set of findings relating to teaching practices that we can consider consolidated (or, more precisely, highly reliable). These findings should serve as the foundation for any critical reflection, rejecting the unproductive tendency to abandon accumulated evidence in favor of untested novelties driven by claims of innovation or by the self-referential dynamics that frequently characterize educational discourse.

It is essential to formally distinguish between Evidence-Based Education (EBE) and Evidence-Informed Education (EIE). While EBE historically draws inspiration from clinical paradigms, emphasizing strict protocol adherence based on empirical findings, EIE represents a more nuanced and flexible orientation. EIE recognizes empirical research (such as meta-analyses and systematic reviews) not as rigid prescriptions, but as highly reliable sources of evidence designed to inform professional decision-making. Under EIE, scientific evidence serves as a critical benchmark, while educators retain full professional autonomy and accountability to adapt strategies to specific classroom contexts, affective factors, and pedagogical goals.

It is not necessary here to present this orientation and the complex of issues that are the subject of long-standing discussion and which are also linked to it in Italy and for which we refer to previous works [18,19].

In brief, the proliferation of experimental and quasi-experimental designs, the increasing availability of systematic reviews and meta-analyses, and the development of methodologies for assessing study quality and effect size robustness have made it possible to achieve large-scale verification of certain theoretical orientations, confirming what expert teacher observation had already suggested in some cases, and refuting it in others.

In short, large-scale evidence has confirmed the effectiveness of theories such as those on the general principles of instruction (based on references to instructional design,

direct instruction, and explicit instruction [20–24] (Education Endowment Foundation (EEF), <https://educationendowmentfoundation.org.uk/>, accessed on 15 February 2026), on models from cognitive sciences such as those on Cognitive Load Theory [15], on the recognition of the role of feedback and formative assessment [24].

In the field of effective instruction, central constructs include the clarity of learning objectives, activation of prior knowledge, identification of misconceptions, and deliberate strategies to address them. These include, in particular, guided modeling, anticipation, cognitive load management, consistent use of feedback, and metacognition. Other aspects that appear “traditional” include activities such as teaching students to understand texts by thinking aloud (thinking aloud), teaching them to summarize and actively return to the same topics over time (spaced learning and retrieval practice).

This approach does not ignore the complex nature of teaching but rejects its translation into forms of decision-making paralysis or into excuses for choices driven by more or less hidden self-referential criteria. Within a broad range of possibilities, we can highlight what appears most reliable based on experimental data. With the ever-legitimate warnings regarding the risk of reductive simplifications and the neglect of contextual specificities [19], it is therefore of fundamental importance that a mentality oriented in this direction be given strong prominence in teacher training.

2.2. Teaching Misconceptions and ETQ

As we advance the acquisition of principles and programs that have proven to work best, the world of misconceptions also emerges. This is an area we have been studying for several years, collecting them and presenting them to teachers in the most practical and effective way possible. This activity has led to the development and use of an interactive questionnaire available online, in which participants submit statements or teaching decisions for evaluation, then compare their responses with those provided by a panel of expert EBE teachers and researchers (ETQ, Effective Teaching Questionnaire) [25–27]. The construction of this expert model involved a selection process of five researchers and three master teachers with extensive experience in EIE. While this panel provides a stable reference point for the ETQ, the development of the ETQ-AI environment specifically aims to move beyond a rigid ‘correct/incorrect’ binary. By integrating AI, the system allows teachers to problematize the expert model itself, facilitating a dialogue where the inherent variability of pedagogical situations can be discussed rather than standardized.

The questionnaire has a dual purpose: to provide an immediate “X-ray” of the teacher’s teaching theories, immediately highlighting areas that indicate a greater departure from the expected model; and to provide feedback explaining to the teacher the reason for the response that would be preferable based on the expert benchmark, citing the evidence supporting it.

The original item pool of the Effective Teaching Questionnaire (ETQ) comprised 64 items used in preliminary benchmark alignment tests [26,28]. For the interactive ETQ-AI platform, a curated 50-item version was deployed to optimize completion time and minimize respondent fatigue while preserving comprehensive domain representation. The selection of these 50 binary items was conducted through an expert consensus procedure involving five educational researchers and three master teachers specializing in EIE (achieving high initial inter-rater agreement, Fleiss’ $\kappa = 0.88$). Items were selected based on three practical criteria: (a) direct mapping onto established meta-analytic effect sizes ($d \geq 0.40$); (b) operational capacity to distinguish common teacher misconceptions from evidence-aligned practice; and (c) balanced representation across core pedagogical domains. While previous extended versions of the ETQ demonstrated sound internal consistency (Cronbach’s $\alpha = 0.87$) [26,27], readers should treat this curated 50-item scale strictly as a tailored

diagnostic scaffold rather than an independently re-validated psychometric inventory. The final scale is distributed across five dimensions: Planning (12 items), Cognitive (12 items), Evaluative (10 items), Management (8 items), and Special Education/Inclusion (8 items).

It was also believed that the questionnaire could serve as a basis for a more in-depth reflection on the foundations of one's teaching assumptions, an aspect that, however, points to the possibility of developing a personal relationship between the teacher and an expert tutor.

The teaching situations were selected in alignment with the EIE learning outcomes described above. Across all questionnaire versions, the items address: the importance of well-structured and purposeful instructional design; the cognitive overload and distraction risks associated with technology use; and the centrality of formative assessment and timely feedback. We have included aspects that imply the ability to recognize different types of cognitive processes, such as those that distinguish notional knowledge from deeper knowledge, or insights derived from classroom management (proactive management, lack of stigma, etc.), or the widespread idea in Italy of a naive conception of inclusion, identified with the idea that all students should always be kept together in the same class, including those with severe intellectual disabilities, and with an almost "salvage" role attributed to group work.

Other aspects considered are related to the lack of updating on research advances in specific fields. For example, in the case of autism, it is well known that myths and psychotherapeutic methods based on affective deficits have declined in favor of the consolidated efficacy of behavioral methods (such as ABA and TEACCH). In teaching reading and writing, comprehensive methods have declined in favor of phono-syllabic ones. Learning styles have also proven to be a sterile reference, incapable of leading to effective interventions.

With this content, which certainly does not exhaust the scope, the items for the ETQ were developed. Notably, across all versions of the questionnaire—administered to both pre-service and in-service teachers—significant gaps recur systematically around specific critical points, reflecting a form of widespread common sense in which persistent misconceptions and naïve teaching beliefs are deeply embedded.

This internalized common sense precedes entry into school, is maintained over time, and persists despite university education and teaching practice.

Figure 1 provides an overall idea of the significant size of the gap between teachers' responses and those of the reference model (scored at 100) through items of different types (for ease of summary, they have been grouped into four categories: planning, cognitive, evaluative, and managerial). As can be seen, the values are around 50% for three of the four areas, with a slight improvement in the management side, where practical experience can evidently play in teachers' favour.

Figure 2 highlights the indicators of misconceptions in detail; note the critical points where the lowest histograms are obtained.

2.3. Artificial Intelligence (AI)

While ETQ applications with pre-arranged feedback have demonstrated that this tool allows teachers to quickly highlight the areas that most diverge from the expert-validated framework, the question of enabling more in-depth and personalized reflection has arisen. The world of Evidence-Informed Education (EIE) evolves, and it is legitimate to critically examine teaching items, which in their concise form cannot be entirely free from contextual ambiguity. This prompted us to explore whether AI could make in-depth, on-demand learning accessible, acting as a dialogic partner when human expert tutors are not readily available.

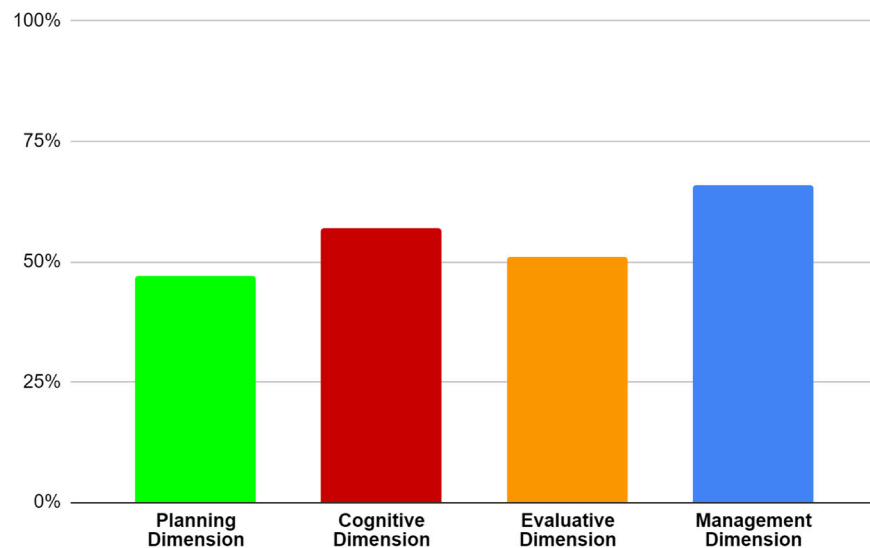


Figure 1. Average of teachers' responses in the ETQ (ETQ version 7 with 64 items) dimensions compared to the reference model (100% threshold).

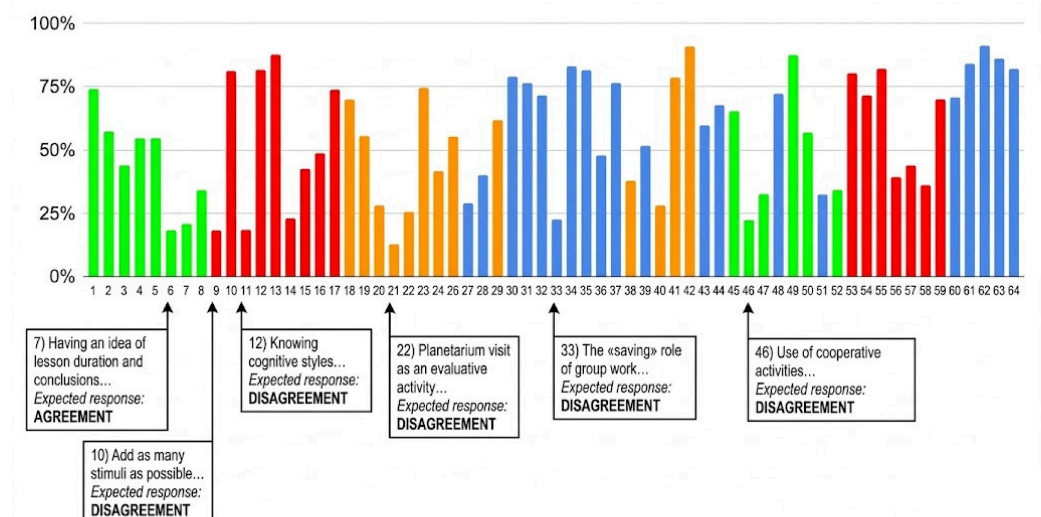


Figure 2. Average of teachers' responses to individual ETQ items across the original 64-item benchmark pool prior to 50-item scale curation.

By moving from static feedback to a dialogic interface, the system transitions from a traditional questionnaire toward a Conversational Tutoring System (CTS). This approach aligns with international research on Explainable AI (XAI) in education, where the goal is to offer transparent justifications of pedagogical logic grounded in empirical evidence, bridging theoretical principles with situated practice.

Building upon our prior work [28], which established the initial feasibility of using static ChatGPT prompts to emulate EBE expert reasoning, the present study introduces three critical advancements: (a) a comparative fidelity evaluation incorporating Google Gemini alongside ChatGPT, (b) a custom-built, web-integrated ETQ-AI platform with dynamic API callback loops, and (c) an empirical mixed-methods evaluation analyzing live teacher-AI argumentative interactions and user perceptions.

To evaluate baseline model fidelity prior to deployment, a standardized benchmark evaluation was conducted across 64 validation items. We evaluated Google Gemini (specifically the gemini-1.5-flash endpoint, snapshot November 2024) against OpenAI ChatGPT (gpt-4o). To account for LLM stochasticity, each item was evaluated across 5 independent

runs per model under fixed hyperparameters (Temperature $T = 0.2$, $Top_P = 0.95$, Max Tokens = 450). An item was classified as “Aligned” if a minimum of 4 out of 5 runs (80% majority agreement) produced evaluations conforming to the expert benchmark. Intra-model consistency was operationally defined as the percentage of benchmark items where the model generated identical binary alignment classifications (Aligned vs. Misaligned) across all 5 runs ($N = 64 \text{ items} \times 5 \text{ runs} = 320 \text{ evaluations per model}$). Under this operational metric, intra-model consistency reached 98.1% for Gemini 1.5 Flash and 95.3% for ChatGPT (gpt-4o). Two independent educational researchers evaluated model outputs against the human expert benchmark using a binary alignment coding protocol (1 = Aligned, 0 = Misaligned/Drifted). Inter-rater reliability was high (Cohen’s $\kappa = 0.89$), with minor discrepancies resolved through joint consensus review.

Overall, Gemini achieved an alignment rate of 96.8% (62 out of 64 items aligned across runs), whereas ChatGPT achieved 92.2% (59 out of 64 items aligned). Misalignments primarily occurred in complex items related to special education and inclusion, where models produced overly permissive responses.

We emphasize that this fidelity benchmark measures prompt-adherence and framework compliance, specifically the model’s capacity to emulate a restricted evidence-based framework without domain drift. High benchmark alignment demonstrates structural fidelity; it does not constitute an automated guarantee of factual accuracy or a substitute for independent citation auditing. Detailed operational prompt parameters and middleware controls governing this pipeline are specified in Section 2.4.3, while raw interaction summaries are compiled in Appendix A.

To evaluate the integrated environment, this study addresses three core Research Questions (RQs):

- RQ1 (Fidelity): To what extent can generative AI models accurately emulate the Evidence-Informed Education (EIE) expert framework when prompted with specific teaching misconceptions?
- RQ2 (Interaction): What qualitative patterns emerge from the personal objections and argumentative interactions between teachers and the AI?
- RQ3 (Utility): How do teachers perceive the technical functionality and pedagogical relevance of receiving AI-mediated feedback compared to the original questionnaire model?

2.4. The Developed System

To operationalize the ETQ-AI framework, a custom web-based platform was developed. The system integrates Google Forms, Google Apps Script, a PHP middleware server, an encrypted MySQL relational database, and the Google Gemini API. The overall technical architecture and data execution loop are illustrated in Figure 3.

2.4.1. System Architecture and Data Flow

As illustrated in the updated architectural diagram (Figure 3), all incoming payloads pass through local SHA-256 hashing before database persistent logging, while the API middleware actively monitors rate limits via a rate-limiting manager and triggers pre-cached expert fallback responses in the event of endpoint timeouts or HTTP 429 status codes. The platform operates through an automated three-tier architecture designed for security, GDPR compliance, and real-time processing:

- Client Tier: The educator interacts initially with an embedded 50-item ETQ Google Form interface. Upon form completion, a native Google Apps Script trigger executes an onFormSubmit() event handler.

- **Privacy & Pseudonymization Module:** To comply with European data protection standards (GDPR), the user's raw email address is processed locally by the script worker to generate a pseudonymized SHA-256 hash token prior to database logging. Furthermore, free-text entries submitted during interactive dialogue are automatically screened and scrubbed for personally identifiable information (PII) at the middleware layer before payload transmission to the external LLM API endpoint, ensuring that no identifying user data is exposed to third-party providers.
- **Middleware & Database Tier:** The extracted response vector is transmitted via an HTTPS POST request to a secure Apache/PHP middleware server. The PHP backend logs the user's responses, score vectors, and interaction timestamps into an encrypted MySQL relational database. Concurrently, an automated script generates a personalized feedback report dispatched to the user's email.

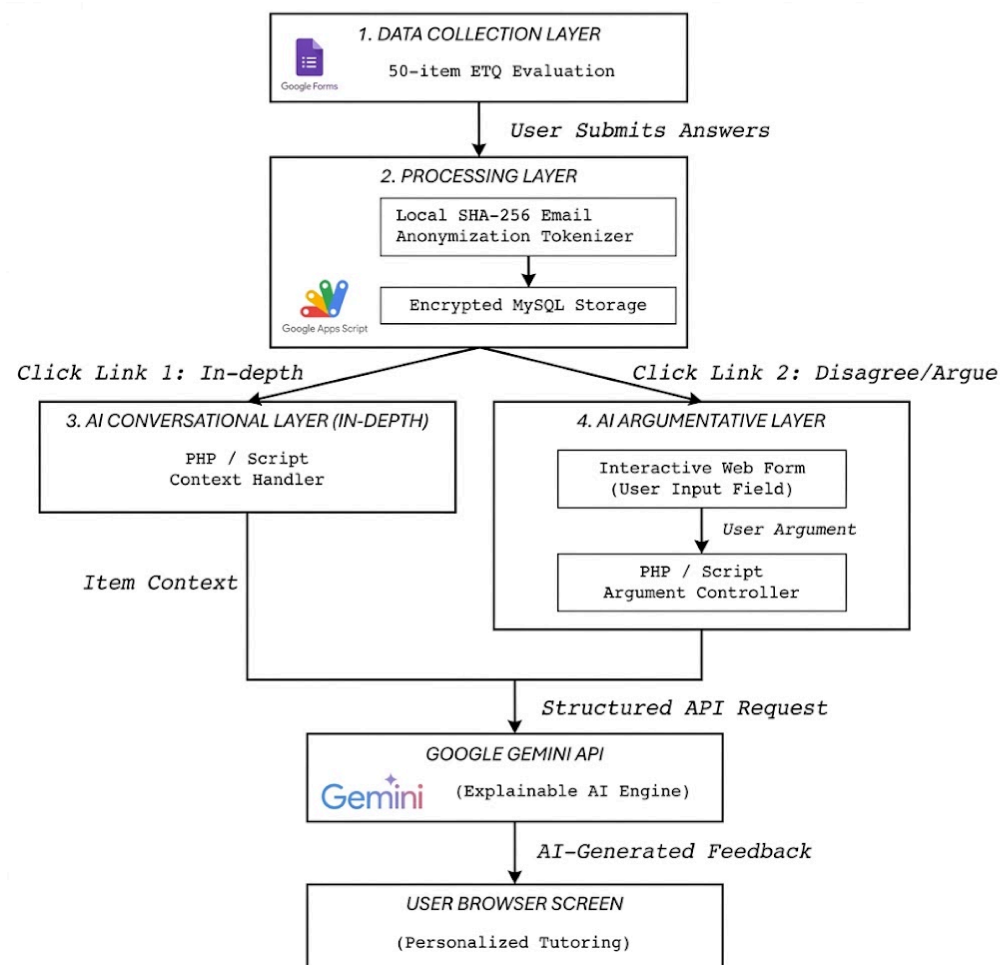


Figure 3. System architecture of ETQ-AI, illustrating data ingestion, SHA-256 pseudonymization, middleware rate-limiting/fallback execution, and external Gemini API routing.

2.4.2. User Interface and Interactive Workflow

The user workflow consists of three distinct pedagogical phases, designed to move the educator from static evaluation to dialogic reflection:

1. **Diagnostic Phase (Google Form):** The user completes the 50 ETQ items, selecting Agree or Disagree for each statement.
2. **Individualized Email Feedback Phase:** Within minutes of submission, the user receives an automated HTML email report (Figure 4). For each item where the user's answer diverged from the EIE benchmark, the email provides: (a) the original statement,

- (b) the baseline expert response, (c) a concise commentary written by the authors, and (d) two dynamic action links: “In-depth EIE Analysis” and “Argue/Discuss with AI”.

Dear participant,

Thank you for participating in this process of interaction and discussion.

Hoping that these suggestions can assist you in your reflections on effective teaching practices, we indicate the degree of AGREEMENT/DISAGREEMENT between your responses and those of the reference model adopted by the authors, for the four dimensions (planning, cognitive, evaluative, managerial), to which is added separately the level of AGREEMENT regarding inclusion and special education.

This general information will be followed by an analytical comparison, item by item.

The feedback we provide should be considered merely as an invitation to further study, without claiming to be exhaustive. It is also possible, and entirely legitimate, to have disagreements. For each item, we have added a prompt requesting further information that can be automatically sent to Google Gemini's Artificial Intelligence, to also consider the perspective that may arise from this perspective.

Given the exploratory nature of this device, we would appreciate it if you could send us feedback, including your assessment and any changes we might need to make.

[Click here](#) to complete the evaluation questionnaire.

Dimension	Evaluation
PLANNING	56 % (5 / 9)
COGNITIVE	67 % (6 / 9)
EVALUATION	67 % (4 / 6)
MANAGEMENT	20 % (1 / 5)
OVERALL EVALUATION	55 % (16 / 29)

As you may have noticed, questions 33 and later concern aspects related to inclusion.

An additional percentage of affinity with the authors' position was calculated for this purpose:

Context	Evaluation
INCLUSION	71 % (15 / 21)

In the form below, you can compare your choices with the reference answers.

If there are any discrepancies, you can consult the feedback item by item and, if necessary, explore the topic further with the help of Artificial Intelligence.

Two modes are available: an in-depth one and an interactive one where you can argue your position.

Kind regards,

The authors of the System.

Antonio Calvani & Sergio Miranda

	Dimension	Given answer	Expected answer	Feedback
1	PLANNING n.1	AGREEMENT	DISAGREEMENT	You should respond by AGREEING or DISAGREEING with the following statement: <i>In most teaching activities, it is preferable for the student themselves to suggest the topic, the activities, and the time to devote to them.</i> The expected response is DISAGREE This statement is unacceptable. Evidence shows that whenever too much room is left for students' intentions for a prolonged period, teaching activity becomes disjointed and learning deteriorates. Your response does not align with the opinion of the questionnaire authors. If you would like to explore this topic further with additional EIE knowledge, click here. If you disagree and would like to discuss your arguments with the AI, click here.
...	...	...	...	...
4	PLANNING n.4	DISAGREEMENT	DISAGREEMENT	Please answer whether you AGREE or DISAGREE with the following statement: <i>When entering the classroom, it is especially important to rely on your creativity and ability to improvise.</i> OK! Your answer is in line with the opinion of the questionnaire authors (DISAGREE) Creativity and the ability to improvise are important teacher skills, but a good lesson requires adhering to a predefined structure with clear goals to achieve. A wealth of evidence highlights the benefits of well-organized lessons. If you would like to delve deeper into this topic with additional EIE knowledge, click here. If you would like to interact with the AI by adding a comment, click here.
...	...	...	...	...

Figure 4. Feedback email in response to completing the ETQ-AI questionnaire with the first feedback predefined by the authors from which to launch the prompts for further investigation with the AI.

3. Dialogic Argumentation Phase (Web Interface): Clicking either link directs the user to a dedicated web environment hosted on the PHP platform (Figures 5–7). If the user selects “In-depth EIE Analysis”, the system triggers an API call to present a detailed, evidence-grounded synthesis. If the user selects “Argue/Discuss with AI”, an interactive modal opens, allowing the teacher to input specific objections or contextual constraints (e.g., “This strategy is unfeasible in large classes”). The system then initiates an interactive dialogue powered by Google Gemini.

ETQ-AI: In-depth analysis of item #1

I summarize the interaction...

To the question:

You must answer whether you AGREE or DISAGREE with the following statement: In most teaching activities, it is preferable for the student themselves to suggest the topic to be covered, the activities, and the time to devote to them.

You answered: AGREE.

The expected response was: DISAGREE, with the following justification:

This statement is unacceptable. Evidence shows that whenever too much room is left for students' intentions for a prolonged period, the teaching activity becomes dispersed and learning deteriorates.

To the clarification you requested, I can answer:

The EIE confirms that teacher guidance is crucial. According to Cognitive Load Theory (Sweller), "novice" students lack the mental schemas to independently structure learning: delegating time and content to them overloads their working memory. Meta-analyses (Hattie) indicate that "Direct Instruction" and clear objectives (Teacher Clarity) have a very high effect size ($d > 0.60$), while unguided discovery learning is poorly effective, as it lacks the sequentiality necessary for knowledge consolidation.

Figure 5. The prompt received by Gemini for more information on the topic related to question #1, where the first link is activated.

ETQ-AI: In-depth analysis of item #1

I summarize the interaction...

To the question:

You must answer whether you AGREE or DISAGREE with the following statement:

In most teaching activities, it is preferable for the student themselves to suggest the topic to be covered, the activities, and the time to devote to them.

You answered: AGREE.

The expected response was: DISAGREE, with the following justification:

This statement is unacceptable. Evidence shows that whenever too much room is left for students' intentions for a prolonged period, teaching activity becomes disjointed and learning deteriorates.

My reasons:

However, this approach seems to me to risk failing to appreciate students' needs and curiosities.

Figure 6. The prompt received by Gemini when the interlocutor also wants to offer personal observations or objections.

Considering your request, given the conflicting data between your position and the expected response from the questionnaire authors, I can answer:

The EIE positively values Student Voice ($d=0.47$) for engagement, but full autonomy contradicts the evidence. Meta-analyses highlight the effectiveness of highly structured strategies, such as Explicit Teaching ($d>0.60$) and Teacher Clarity ($d=0.75$). Requiring students to determine most of the topic, activity, and time risks generating a high Cognitive Load, disrupting learning. Scaffolding and sequencing must be guided by the teacher to maximize outcomes.

Figure 7. The answer provided by Gemini summarizing the interaction and the ongoing controversy.

2.4.3. Prompt Engineering, API Management, and Fallback Handling

The integration of the Google Gemini API (gemini-1.5-flash endpoint) is governed by strict system prompts and technical constraints designed to enforce prompt adherence, preserve framework compliance, and minimize off-topic domain drift. These structural

constraints aim to anchor generative outputs to the Evidence-Informed Education (EIE) benchmark, serving as operational heuristics to reduce domain drift rather than automated guarantees against factual or citation inaccuracies.

To manage API execution stability during high-concurrency periods, the middleware implements a rate-limiting manager and an automated fallback engine. If the primary API endpoint returns an HTTP 429 status code or fails to respond within a 5-s timeout window, the pipeline bypasses dynamic generation and delivers a pre-cached, expert-validated response matching the specific ETQ item diagnostic parameters. In particular:

- **System Prompt Construction:** Every API payload dynamically incorporates: (a) Item ID and text, (b) user choice vs. expert benchmark, (c) author baseline commentary, (d) user text entry/objection, and (e) system Prompt instructions constraining Gemini to cite specific meta-analytic metrics (e.g., Hattie's effect sizes, Cognitive Load Theory).
- **Hyperparameter Locking:** To balance natural linguistic fluency with rigorous adherence to evidence, API parameters are fixed across all interactive calls: Temperature $T = 0.3$ (slightly increased from $T = 0.2$ benchmark testing to foster conversational engagement while maintaining deterministic boundaries), $Top_P = 0.95$, and Maximum Output Limit $Max_Tokens = 450$.
- **Rate Limiting & Error Fallback:** To handle network latency or API rate limits (HTTP status 429), the PHP middleware implements an exponential backoff retry logic (up to 3 attempts). If the API endpoint remains unresponsive, the system seamlessly displays a pre-cached expert response from the database, ensuring zero disruption to the user experience. All user-AI interaction logs are stored in the MySQL database for post-hoc qualitative analysis.

3. Results

This section presents the empirical findings derived from the pilot evaluation of the ETQ-AI platform. Section 3.1 covers the quantitative system evaluation, subgroup analysis, and identification of critical pedagogical friction points. Section 3.2 details the qualitative thematic analysis of teacher-AI argumentative interactions.

3.1. Quantitative System Evaluation and Subgroup Analysis

A total of $N = 123$ educators participated in the pilot evaluation during a national Evidence-Informed Education (EIE) conference (response rate: 82%). The sample comprised primary school teachers ($n = 70$, 56.9%), secondary school teachers ($n = 40$, 32.5%), and school leaders/headteachers ($n = 13$, 10.6%), with a mean teaching experience of 7.2 years ($SD = 4.1$).

Across the questionnaire administration, participants generated 218 automated API queries for in-depth EIE explanations and submitted 34 structured argumentative objections. Analysis of user response patterns revealed that disagreement with the EIE expert benchmark was strongly clustered around specific pedagogical items:

1. Item #12 (Cognitive Styles): 78% initial disagreement (teachers defending visual/auditory learning style matching).
2. Item #33 and #46 (Cooperative Group Work & Inclusion): 68% disagreement (teachers arguing for social-affective inclusion over cognitive effect sizes).
3. Item #10 (EdTech & Information Overload): 54% disagreement.

All 123 participants evaluated both author feedback and Gemini feedback. During the interactive workshop, every participant received the automated email report containing static author baseline comments and completed at least one interactive EIE deep-dive query (218 total queries generated across the 123 participants). Subsequently, participants evaluated both feedback modes on identical 5-point Likert subscales (1 = Very Poor/Strongly

Disagree, 5 = Excellent/Strongly Agree) assessing “Clarity and Actionable Depth”. As shown in Table 2, ratings were high across all dimensions.

Table 2. Participant Evaluation of the ETQ-AI System across Professional Roles ($N = 123$).

Evaluation Dimension	Overall Mean (SD)	Median	IQR	Primary ($n = 70$)	Secondary ($n = 40$)	Leaders ($n = 13$)	p -Value
Technical Functionality	4.41 (0.62)	5.0	1.0	4.43	4.38	4.46	0.842
Pedagogical Tool Relevance	4.23 (0.71)	4.0	1.0	4.20	4.25	4.31	0.780
Clarity and Actionable Depth of Static Author Feedback	3.82 (0.89)	4.0	1.0	3.75	3.88	4.00	0.412
Clarity and Actionable Depth of Gemini AI Feedback	4.12 (0.75)	4.0	1.0	4.08	4.15	4.23	0.651

A paired-samples t -test revealed that participants rated the dynamic, Gemini-generated explanations significantly higher in pedagogical clarity and actionable depth ($M = 4.12$, $SD = 0.75$) than the static baseline author comments ($M = 3.82$, $SD = 0.89$; $t(122) = 3.41$, $p < 0.001$, Cohen’s $d = 0.36$). A one-way ANOVA was conducted to compare evaluation ratings across professional roles (Primary Teachers $n = 70$, Secondary Teachers $n = 40$, School Leaders $n = 13$). No statistically significant differences were detected across groups for technical functionality ($F(2, 120) = 0.17$, $p = 0.842$, $\eta_p^2 = 0.003$), pedagogical relevance ($F(2, 120) = 0.25$, $p = 0.780$, $\eta_p^2 = 0.004$), or Gemini consistency ($F(2, 120) = 0.43$, $p = 0.651$, $\eta_p^2 = 0.007$). While these results indicate that no subgroup-specific satisfaction barriers were detected in this sample, the absence of statistically significant differences does not establish formal equivalence across professional profiles, particularly given the smaller sample size of school leaders ($n = 13$).

3.2. Qualitative Analysis of Teacher-AI Argumentative Interactions

To examine how teachers argue against automated feedback, the corpus of 34 text objections was analyzed using a Codebook approach to Thematic Analysis [29,30]. Unlike fully reflexive thematic analysis, a codebook approach is epistemologically aligned with structured coding reliability and inter-coder agreement metrics. Two independent researchers developed an initial coding frame based on a 20% sample, iteratively refined code definitions, and independently coded the full corpus, achieving strong inter-rater reliability (Cohen’s $\kappa = 0.84$). Disagreements were resolved through consensus review. We note that these 34 objections represent a self-selected sub-sample of participants who actively chose to contest feedback; results describe the argumentative patterns within these specific interactions rather than the entire participant cohort. Three primary themes were identified, summarized in Table 3, with additional participant qualitative feedback detailed in Appendix A (Tables A1 and A2).

3.2.1. Pattern Analysis of Gemini’s Pedagogical Responses (“Acknowledge-And-Refine”)

Analysis of the interactive logs revealed that Gemini consistently adopted a three-phase “Acknowledge-and-Refine” conversational strategy:

- Phase 1 (Affective Validation): Gemini begins by validating the teacher’s practical experience (e.g., “You are completely correct that student rapport and emotional safety are fundamental prerequisites...”).

- Phase 2 (Evidence Reframing): The model re-introduces the EIE evidence without dismissing the teacher's context (e.g., "However, Hattie's meta-analyses ($d = 0.75$) show that clear teacher structure actually reduces student anxiety by lowering cognitive load...").
- Phase 3 (Contextual Adaptation): The model proposes a pragmatic compromise tailored to the user's constraint (e.g., "In a class of 28, explicit instruction can be chunked into 10-min micro-lessons followed by guided pair work...").

Table 3. Qualitative Coding Matrix of Teacher Objections ($N = 34$).

Theme Category	Freq. n (%)	Operational Coding Definition	Representative Participant Verbatim Quotes
Contextual Constraints & Resource Deficits	14 (41.2%)	Objections asserting EIE rules are unfeasible due to class size, time, or curriculum bounds.	Direct instruction works in theory, but with 28 pupils including 4 with special needs, pacing breaks down. (P24, Primary) "We must cover state syllabi; spending time on retrieval practice limits coverage." (P81, Sec.)
Pedagogical Subjectivity & Affective Primacy	12 (35.3%)	Objections prioritizing intuition, emotional rapport, or unique pupil needs over meta-analytic data.	Teaching is an art based on relationships, not effect sizes. Every child is unique. (P11, Prim.) "Focusing solely on cognitive metrics neglects emotional well-being." (P105, Secondary)
Item Ambiguity & Decontextualization	8 (23.5%)	Objections arising from binary wording rather than rejecting the underlying evidence.	The statement forces a Yes/No on group work, which depends on whether it is exploratory or consolidative. (P43, Leader)

3.2.2. Outlier/Negative Case Analysis

In 2 out of 34 interactions (5.9%), participants completely rejected Gemini's reframing, characterizing the system as "an algorithmic dogma machine that ignores real classroom reality." These outlier cases demonstrate that when AI feedback is perceived as overly normative, it can trigger professional defensiveness rather than reflective learning.

4. Discussion

The empirical results from this pilot evaluation demonstrate the technical viability and high user acceptance of ETQ-AI as an interactive scaffolding tool for teacher professional development. By combining standardized EIE diagnostic items with a Gemini-powered dialogic layer, the system successfully transitions diagnostic surveys from static measurements into active reflective spaces. However, interpreting these findings requires a critical examination of explainability paradigms, architectural dependencies, sample constraints, and ethical risks.

4.1. Re-Framing Explainable AI (XAI): Pedagogical Rationale vs. White-Box Transparency

It is vital to distinguish internal algorithmic transparency (i.e., 'white-box' visibility into token probabilities or attention weights) from pedagogical rationale explainability. ETQ-AI does not expose the internal mathematical mechanics of Gemini. Rather, it enforces domain-constrained system prompting to ensure that output justifications are grounded in explicit synthesis literature (e.g., Hattie's effect sizes, Cognitive Load Theory). While this structural boundary mitigates off-topic algorithmic drift, it does not guarantee complete immunity from hallucinatory references or subtle textual inaccuracies. Consequently, out-

put justifications should be understood as contextually constrained pedagogical arguments rather than infallible absolute truths.

4.2. Architectural Dependency, API Governance, and Open-Source Alternatives (DeepSeek)

A critical structural issue identified in this study concerns its reliance on proprietary commercial LLM APIs (Google Gemini). Commercial API endpoints are subject to unannounced model updates, safety guardrail modifications, and policy shifts by platform providers, introducing long-term risks of pedagogical drift. Future developments should explore open-weights models (e.g., Llama-3, Mistral) and emerging reasoning-focused architectures (e.g., DeepSeek-R1) hosted on local institutional infrastructure. From a theoretical standpoint, local deployment ensures complete data sovereignty and eliminates commercial API dependency, provided that system prompt constraints are rigorously maintained to preserve epistemological alignment with verified educational evidence.

4.3. Methodological Limitations and Sample Selection Bias

The generalizability of our empirical findings is constrained by the pilot study's sampling design. Participants were recruited via convenience sampling among attendees of a national Evidence-Informed Education (EIE) conference ($N = 123$). Consequently, the sample represents a self-selected cohort already predisposed toward evidence-based methodologies and educational research. While this cohort provided an ideal baseline environment for testing technical mechanics and dialogic interaction, the high satisfaction ratings ($M = 4.23/5$) cannot be directly generalized to the broader teacher population, where resistance to quantitative metrics and standardized pedagogical frameworks may be substantially higher.

4.4. Ethical Considerations: Guarding Against Pedagogical Echo Chambers

Finally, system developers and educational leaders must address the ethical risks of pedagogical homogenization. Enforcing a single expert benchmark risks creating a normative “echo chamber” that marginalizes context-sensitive, intuitive, or culturally responsive teaching practices. ETQ-AI must not be deployed as an authoritative, infallible arbiter of absolute pedagogical truth. Rather, it functions as a “boundary object”—a reflective scaffold engineered to induce productive cognitive conflict, challenge tacit assumptions, and invite teachers into professional argumentation. The ultimate objective is not strict compliance with an algorithmic benchmark, but the cultivation of critical, evidence-informed professional judgment.

5. Conclusions

This study evaluated the design, implementation, and empirical validation of the ETQ-AI platform—an interactive environment designed to integrate generative AI models into evidence-informed teacher professional development. By transitioning from traditional, static diagnostic scaffolds toward dynamic, conversational feedback, this investigation addressed three fundamental Research Questions regarding model fidelity, qualitative interaction dynamics, and user-perceived utility.

Regarding RQ1 (Fidelity), the empirical findings confirm that modern generative AI architectures, specifically Gemini 1.5 Flash and GPT-4o, can emulate the Evidence-Informed Education (EIE) expert framework with high fidelity and reliability. Across five independent runs evaluating the benchmark item pool, the models achieved strong inter-rater alignment ($\kappa = 0.89$) under controlled decoding parameters ($T = 0.2$ during testing and $T = 0.3$ during deployment). Crucially, the system successfully operationalized EIE principles by providing nuanced, context-sensitive counterarguments to common teaching

misconceptions without reverting to the rigid, prescriptive dogmatism often associated with traditional Evidence-Based Education (EBE).

Concerning RQ2 (Interaction), thematic analysis of the conversational exchanges ($\kappa = 0.84$) revealed clear qualitative patterns in how educators articulate personal objections and engage with AI-mediated pedagogical feedback. The primary argumentative trajectory followed a pattern, wherein the model validated the teacher's lived classroom experience before introducing research-informed nuances. Furthermore, the systematic identification of a 5.9% resistance rate through negative case analysis highlighted the boundary conditions of automated dialogue. Rather than indicating system failure, these interactions demonstrated that deep-seated pedagogical beliefs require careful negotiation, reinforcing the principle that AI feedback must serve as a reflective heuristic rather than an infallible authority.

In terms of RQ3 (Utility), user perception data demonstrated that teachers found the AI-mediated feedback environment significantly superior in both technical functionality and pedagogical relevance compared to the original static questionnaire model. Educators reported that the dynamic dialogue provided a valuable scaffold for professional engagement, creating structured opportunities to connect initial Awareness and Critical Thinking with reflective self-evaluation [2,3]. Rather than demonstrating direct pre/post gains in critical reflection, which were not directly measured in this pilot study, these findings indicate that the platform offers an accessible, dialogic interface for challenging implicit pedagogical assumptions. By offering pedagogical rationale explainability rather than opaque numerical scores, the platform supported teachers in re-examining their instructional choices while preserving their professional autonomy.

Contributions, Limitations, and Future Work

Beyond addressing the primary research questions, this work offers key theoretical and methodological contributions to the AI in Education (AIED) literature. Epistemologically, it operationalizes a crucial distinction between prescriptive EBE and flexible EIE within automated dialogue systems. Methodologically, it establishes a transparent protocol for benchmarking model fidelity and evaluating human-AI qualitative interactions.

Nevertheless, several limitations define the scope of these findings. First, while the study measured teacher perceptions and reflective patterns within the digital scaffold, further empirical work is required to investigate the longitudinal transfer of these cognitive shifts into actual classroom teaching practice and student learning outcomes. Second, reliance on commercial proprietary APIs introduces operational and privacy constraints. Future research should examine the deployment of fine-tuned, open-weights reasoning architectures (such as Llama-3 or DeepSeek-R1) hosted on local infrastructure. Finally, expanding the item pool and validating the framework across broader, culturally diverse educational systems will be essential to ensure that AI-driven professional scaffolds remain inclusive, ethically sound, and fundamentally supportive of teacher agency.

Author Contributions: Conceptualization, A.C. and S.M.; methodology, A.C. and S.M.; software, S.M.; validation, A.C. and S.M.; data curation, S.M.; writing—original draft preparation, A.C. and S.M.; writing—review and editing, S.M.; supervision, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and institutional guidelines for educational research. Formal ethical review was waived by the Ethics Committee of the Department of Human Sciences, Philosophy and Education at the University of Salerno, as the study protocol involved non-invasive educational diagnostic surveys

and collection of pseudonymized operational data. In accordance with institutional research policies, non-invasive surveys utilizing pseudonymized data do not require a formal protocol tracking number. All participant email addresses were pseudonymized locally via SHA-256 hashing prior to database storage, and free-text inputs were scrubbed of personally identifiable information (PII) before external API processing. All research data will be stored on encrypted infrastructure for a maximum duration of three years post-study completion, after which all database logs will be permanently purged.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: All authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

- EBE Evidence-Based Education
- EIE Evidence-Informed Education
- ETQ Effective Teaching Questionnaire
- AI Artificial Intelligence

Appendix A

Interactions with the AI of Google Gemini

Table A1. Summary of participants and interactions.

Detail	Count
Participants	123
AI calls for further information	218
AI calls for interaction	34
Personal assessments collected	18

Table A2. Participant Comments on the ETQ-AI System.

Comment
I found the questions on the topic of inclusion interesting; I would have added a few more.
I found every topic interesting.
I don't have any at the moment.
I found this meeting very useful, especially the section dedicated to planning for improvement, because it offers concrete and practically applicable insights.
I really appreciated this method of interaction and reflection.
For me, the instructions were absolutely clear and comprehensive.
Specify the items more clearly and add items for middle and high school.
Frankly, a questionnaire like this is useless. There's not much to add. Certainly, being able to agree or disagree with statements taken out of context is extremely limiting.
It seems like an excellent reflection tool, useful for sharing in working groups to subsequently share methods, strategies, and approaches useful for everyday practice.
An interesting tool.
I struggled to understand how some questions (1 or 2) differed from each other, and personally, I don't find them mutually exclusive.

Table A2. Cont.

Comment
I feel inadequate in answering this question, having never used such a tool.
The feedback was very good, as were the insights and comparisons.
The AI clarification, which provides precise references, was interesting.
I believe that the tool, in its simplicity (Y/N), could be useful for evaluating the teacher's work with the class, but some questions needed more specificity, and others were unclear (related to primary school) for a teacher who teaches secondary school. Sometimes the answers were not given a real choice between YES and NO.
The feedback provided by Gemini is closely tied to EBE, and while accurate and fairly acceptable, it is predictable and does not address the specific nature of the profession, which must relate to the sheer diversity of students' individualities and the related range of solutions to be adopted.
Furthermore, the area where I personally "lacked" most concerns inclusion, where many answers could have been acceptable based on my 25 years of experience, where each case is unique (especially regarding questions about autism), and where each teacher seeks solutions on a case-by-case basis, seeking those that are most effective for the individual.
This tool has considerable potential; it would be interesting to further interrogate the AI to obtain "some" clarifications that would help us gain a more consistent picture of the topic being discussed. Accessing a questionnaire to, for example, evaluate one's learning in a course of study and having a discussion with AI in the same context to clarify any doubts and improve one's experience is undoubtedly very interesting. It's like having a personal tutor. Thank you.

References

- Ministero dell'Istruzione e del Merito (MIM). Linee Guida per l'Intelligenza Artificiale Nelle Istituzioni Scolastiche. 2025. Available online: https://www.mim.gov.it/documents/20182/0/MIM_Linee+guida+IA+nella+Scuola_09_08_2025-signed.pdf/b70fdc45-4b75-1f7e-73bf-eab12989b928?t=1756468797694 (accessed on 12 October 2026).
- Mezirow, J. *Transformative Dimensions of Adult Learning*; Jossey-Bass: San Francisco, CA, USA, 1991.
- Schön, D.A. *The Reflective Practitioner: How Professionals Think in Action*; Basic Books: New York, NY, USA, 1983.
- Mortari, L. *Ricerche e Riflettere. La Formazione del Docente Professionista*; Carocci: Roma, Italy, 2009.
- Vannini, I. Introduzione. Fare ricerca educativa per promuovere la professionalità docente. Il "qui ed ora" del Centro CRESPI. In *La Ricerca-Formazione. Temi, Esperienze, Prospettive*; Asquini, G., Ed.; FrancoAngeli: Milano, Italy, 2018; pp. 13–24.
- Bertagna, G. La formazione iniziale dei docenti: Una proposta epistemologica, lo scenario nazionale e i suoi problemi. *Nuova Second.* **2019**, XXXV, 16–33.
- Perla, L.; Martini, B. (Eds.) *Professione Insegnante Idee e Modelli di Formazione*; FrancoAngeli: Milano, Italy, 2019.
- Trevisan, O.; Phillips, M.; De Rossi, M. Unpacking teacher decision-making: Connecting complex elements. *Ital. J. Educ. Res.* **2021**, 27, 13–26. [\[CrossRef\]](#)
- Albanese, M.; Fiorello, E. The concept of Pedagogical Content Knowledge in light of a neuroeducational perspective in initial teacher education. *Ital. J. Educ. Res.* **2025**, 19–29. [\[CrossRef\]](#)
- Hattie, J. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*; Routledge: London, UK, 2009.
- Lakhe Shrestha, B.L.; Dahal, N.; Hasan, M.K.; Paudel, S.; Kapar, H. Generative AI on professional development: A narrative inquiry using TPack framework. *Front. Educ.* **2025**, 10, 1550773. [\[CrossRef\]](#)
- Arefian, M.H. AI-enhanced professional learning communities: A new era of personalized teacher education. *Cogent Educ.* **2026**, 13, 2626629. [\[CrossRef\]](#)
- Giannakos, M.; Cukurova, M.; Sharma, K.; Papavlasopoulou, S. Toward hybrid teaching intelligence: Investigating the potential of teacher–AI collaboration using large language models. *Behav. Inf. Technol.* **2025**, 1–19. [\[CrossRef\]](#)
- Hattie, J. *Visible Learning for Teachers: Maximizing Impact on Learning*; Routledge/Taylor & Francis Group: Abingdon, UK, 2012.
- Sweller, J. Cognitive load theory and educational technology. *Educ. Technol. Res. Dev.* **2019**, 68, 1–16. [\[CrossRef\]](#)
- Rizzo, A.; Legrenzi, P. *Pensare Con L'Intelligenza Artificiale. Un'alleata Possibile*; Il Mulino—Ebook: Bologna, Italy, 2025.
- Graesser, A.C.; Lu, S.; Jackson, G.T.; Mitchell, H.H.; Ventura, M.; Olney, A.; Louwerse, M.M. AutoTutor: A tutor with dialogue in natural language. *Behav. Res. Methods Instrum. Comput.* **2004**, 36, 180–192. [\[CrossRef\]](#)
- Vivianet, G. *Che Cos'è L'Evidence Based Education*; Carocci: Roma, Italy, 2014.
- Calvani, A. Evidence based education. Perché è necessario orientare l'istruzione verso una cultura dell'evidenza. *Nuova Second.* **2025**, 1, 45–58.
- Gagné, R.M.; Briggs, L.J. *Principles of Instructional Design*; Holt, Rinehart and Winston: New York, NY, USA, 1974.

21. Merrill, M.D. First principles of instruction. *Educ. Technol. Res. Dev.* **2002**, *50*, 43–59. [[CrossRef](#)]
22. Rosenshine, B. Principles of instruction: Research-based strategies that all teachers should know. *Am. Educ.* **2012**, *36*, 12–39.
23. Bell, M. *The Fundamentals of Teaching: A Five-Step Model to Put Research Evidence into Practice*; Routledge: New York, NY, USA, 2020.
24. Hattie, J.; Timperley, H. The power of feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [[CrossRef](#)]
25. Menichetti, L.; Pellegrini, M.; Gola, G. Cornici mentali e stereotipie didattiche nella formazione degli insegnanti. *Form. Insegn.* **2019**, *17*, 351–374. [[CrossRef](#)]
26. Calvani, A.; Marzano, A.; Morganti, A. *La Didattica in Classe Casi, Problemi e Soluzioni*; Carocci: Roma, Italy, 2021.
27. Miranda, S. Orienting the attitudes of future teachers towards effective interventions: Restructuring misconceptions and naïve didactic points of view. *J. Educ. Cult. Psychol. Stud. (ECPS J.)* **2022**, *25*, 141–160. [[CrossRef](#)]
28. Calvani, A.; Miranda, S. Can AI help teachers become teaching experts? *Ital. J. Educ. Res.* **2025**, *35*, 67–85. [[CrossRef](#)]
29. Braun, V.; Clarke, V. Reflecting on reflexive thematic analysis. *Qual. Res. Sport Exerc. Health* **2019**, *11*, 589–597. [[CrossRef](#)]
30. Boyatzis, R.E. *Transforming Qualitative Information: Thematic Analysis and Code Development*; SAGE Publications: Thousand Oaks, CA, USA, 1998.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.