CIENTIFIC Publishing Limited

Journal of Qualitative Research in Education
https://ojs.ukscip.com/index.php/jqre

Article

From AI-Generated Student Analytics to Teacher Trust: Evidence from a Longitudinal Classroom Case Study

İbrahim Delen [1,*], Ayten Baykal [1], Bora Şenceylan [2] and Gökhan İnce [2]

[1] Mathematics and Science Education Department, Bogazici University, İstanbul 34470, Türkiye

[2] Faculty of Computer & Informatics Engineering, Istanbul Technical University, İstanbul 34485, Türkiye

[*] Correspondence: ibrahim.delen@bogazici.edu.tr

Received: 26 June 2026; **Revised:** 14 July 2026; **Accepted:** 5 August 2026; **Published:** 11 August 2026

Abstract: This case study examined how teacher trust in student analytics could be built when using Artificial Intelligence (AI) generated assessment. The case study included two data sources: (i) teacher evaluations of AI-generated student reports and (ii) a semi-structured interview with a science teacher who used the system in their classroom for three semesters (first pilot semester and two semesters of implementation). The teacher demonstrated an agreement rate exceeding 80% with Generative Artificial Intelligence (GenAI) generated reports. However, the teacher also stated that she expected the reports to not only describe student characteristics but also provide actionable pedagogical recommendations to support the teaching process. In conclusion, the study provided evidence that the teacher trust in GenAI based student analytics could be developed as the teacher adopts AI based practices in her classroom. When GenAI tools are implemented for a longer period, they have potential to build trust underlined by other studies. Including GenAI represented students' learning characteristics and replicated it consistently across different students. The findings indicate that GenAI-supported learning analytics systems to be developed in the future should be designed not only as tools to describe student behavior, but also as explainable and pedagogical decision support systems that support teachers' professional judgment.

Keywords: AI Generated Assessment; Student Analytics; Teacher Trust

1. Introduction

Artificial intelligence (AI) as a promising field of study in 1960 [1]. However, it took fifty years for educational studies to put strong emphasis on AI. A recent review of AI literature and stated that number of studies in educational disciplines made a leap after 2010, and studies were mainly conducted in United States and China [2]. Delen et al. also added the lack of interdisciplinary studies and lack of connections between authors among countries [2].

The rise of AI in education has influenced different facets of teaching. These include supporting lesson planning [3], managing administration, assignments and assessment [4]. However, educators are reluctant to completely rely on AI for important evaluations because these applications are still in their early stages of development [5]. In this section, we'll examine how Generative Artificial Intelligence (GenAI) supported assessment emerged in different disciplines.

AI could provide individualized assessment in language learning [6]. Offering may personal feedback could offer new venues for music educators. When using AI teachers are not just responding passively to technological changes; instead, they are proactively reshaping their practices [7]. Despite the growing interest and potential of AI, how much it could be integrated into teachers' practices remains uncertain. In this section, we will present how previous studies raised teachers' concerns about using AI for assessment [7].

In a study examining language educators from different countries, teachers were aware of AI's importance in enhancing learning and grading, but teachers are still essential for evaluating evidence. AI can help with objectivity by automating exams, but the evaluation should combine AI analytics with ethical principles and teacher judgment [8]. In another study, researchers evaluated GenAI supported digital multimodal composition language products. The framework developed in this study consisted of analytical rubric that evaluates Generative-AI (GenAI) integration in textual, visual, and auditory dimensions, multimodal checklists that support self and peer assessment, and a formative assessment system to monitor learning progress. The framework extended traditional evaluation criteria with GenAI-specific dimensions such as prompt writing, ethical AI use, multimodal design, transforming AI outputs, and critical AI literacy. Authors suggested that this approach can be used to support learner autonomy, critical AI literacy, and authentic language use in language education [9].

In a study examining GenAI adaptation in a postsecondary setting, authors reported that teachers' adoption of AI differed significantly depending on its intended use, according to a survey of 127 post-secondary educators in Quebec (64 college and 63 university professors; 35 from science, technology, engineering, and mathematics (STEM) and 92 from non-STEM subjects). AI's preferred use consisted of preparing materials, formative assessment, and checking plagiarism. University instructors did not prefer predictive assessment and automatic grading [10]. In another study conducted with university instructors in West Malaysia showed that while resistance to change has a positive but negligible impact, anxiety has a detrimental impact on attitudes and adoption readiness when using AI for assessment [11].

In recent years, research in scientific education has mostly concentrated on the advantages of generative artificial intelligence (GenAI), with little data about its impact in assessment. The consistency of GenAI in evaluating students' socio-scientific issues (SSI) reasoning in relation to human intelligence is empirically investigated in this study. It evaluated responses on SSI subjects taken from the curriculum, including as thermodynamics, mechanics, and electricity, using a case study of 22 first-year Indonesian physics students. Results showed that while GenAI tools/systems, such as ChatGPT, Gemini, and Copilot, scored consistently with human assessors in thermodynamics, there were discrepancies in mechanics and electrical. Although identical keywords were found by both GenAIs and human intelligence, categorization differed across SSI dimensions. The paper addresses the consequences of adopting GenAI as a practical assessment tool for instructors [12].

In another study, different large language models (Claude Sonnet 4, GPT-4o, GPT-5, and Qwen-2.5) have been used in the automated evaluation of student-chatbot interactions. The findings indicate that while AI models can achieve a high level of compatibility with human evaluators when appropriate rubrics and structured prompts are used, there are differences in performance between models. In particular, Claude Sonnet 4 demonstrated the highest inter-rater agreement in both content and dialogue evaluations, while content evaluation was found to be more reliably automated than dialogue and reasoning evaluation. In addition, it has been revealed that multidimensional rubrics that evaluate content and reasoning separately instead of one-dimensional scoring produce more consistent and more meaningful results in terms of instruction [13].

GenAI can be used in the peer review process and its can impact feedback quality. A comparative study showed that approximately some students did not prefer to use GenAI to learn more, while those who did use AI to improve both high-level (content, argumentation) and lower-level (language, spelling) feedback. Students who used AI gave their peers suggestions for further improvement and less superficial and only motivational expressions of praise. The results reveal that generative AI is a useful support tool that can improve the quality of peer review, but it is important to use it consciously and ethically to support students' learning processes [14].

For supporting argumentation scholars implemented various approaches. One of these approaches was argument mapping. Argument mapping supports students' critical reading of science texts generated by ChatGPT. The findings revealed that argument mapping enhances scientific AI literacy by actively involving students in the processes of developing arguments, anticipating counterarguments, and generating rebuttals. In addition, it has been determined that formative assessment practices such as teacher feedback and peer evaluation support this process, and collaborative argument mapping activities enable the production of more qualified arguments compared to individual studies. The research proposes the use of student-centered approaches, collaborative argument mapping, and formative assessment together in teaching practices to improve scientific AI literacy [15].

In another study, researchers used role-playing and group discussions over the course of six weeks. Participants were 28 Korean seventh-grade students. The findings demonstrated that students opposed artificial intelligence systems because they were anxious about them. To support their opinions and resistance, they also described how their worries about the new technology came about. Additionally, authors discovered distinct patterns when students considered other viewpoints when making decisions and encountered disagreements while attempting to reach a consensus as a group [16].

Another approach to support argumentation was using computational grounded theory and an unsupervised clustering technique. Authors examined argumentation patterns of undergrad students, assessed students' written arguments on chemical reactions. Strong agreement between machine and human evaluations was achieved by automating analysis using a comprehensive 20-category rubric. The results demonstrated how deep learning methods can be used to assess the intricacy of student arguments [17].

In addition to using various tools, the inclusion of new tools such as chatbots has the potential to improve argumentation abilities. In a study using chatbots, 44 Chinese undergraduate students demonstrated overall structural complexity did not significantly change for arguments. But chatbots enhanced the students' use of assertions, data, and warrants in their arguments. Additionally, compared to standard tasks, students reported higher levels of satisfaction, perceiving their performance as similarly successful despite the increased hurdles. This study demonstrated how incorporating argumentative chatbots into educational environments might improve students' motivation and argumentation abilities [18].

AI generated assessment can reflect on students' decision-making, reasoning and mental processes [19]. Using AI for decision-making necessitates a new strategy that addresses the drawbacks of current evaluation and data collection methods in education while processing student data and improving teaching effectiveness [20]. The implications of generative AI tools on evaluation also present important ethical and pedagogical issues as they become increasingly integrated into education [21]. Researchers suggested that GenAI-connected should include different data sources (e.g., oral tests and project-based assignments) [21]. At its current state, the field is no longer debating about the integration of GenAI for assessment. The discussion focused on how to use different tools [13,14], and how to include different skills [19,21].

While previous studies have looked at the role of GenAI tools and skills to be addressed, studies did not pay close attention to the main user of these tools and skills. This raises the question of teachers' understanding of GenAI and develop trust in these systems. However, teachers are not the ones who directly apply the content produced by GenAI, they are decision-makers who evaluate, verify and reject these contents for pedagogical purposes. For this reason, teachers' understanding of GenAI should be considered as the main factors determining the use of artificial intelligence in the classroom. To explain this gap, three complementary theoretical approaches are brought together in this research. First, the Technology Acceptance Model [22] explains why teachers want to adopt AI based on perceived usefulness and perceived ease of use. However, since GenAI systems are not just tools used but also intelligent systems that generate recommendations and contribute to decision-making processes, technology adoption alone is not enough. At this point, the trust model may come into play [23]. In the context of GenAI, this situation expresses the extent to which the teacher finds the suggestions offered by the system reliable. Teachers' lack of trust in this case may lead to underuse of technology, while over-reliance could lead to misuse without critical evaluation [24]. For teachers to use AI effectively, they need to develop an appropriate level of trust by understanding both the strengths and weaknesses of the system. In the study, it is assumed that teachers' level of understanding of GenAI shapes technology trust in GenAI. To examine teacher trust, our case study investigated how an expert teacher can make use of GenAI provided feedback in a classroom in a longitudinal study. Our research questions were: How does the teacher evaluate AI-generated student analytics? How do AI-generated evaluations build trust?

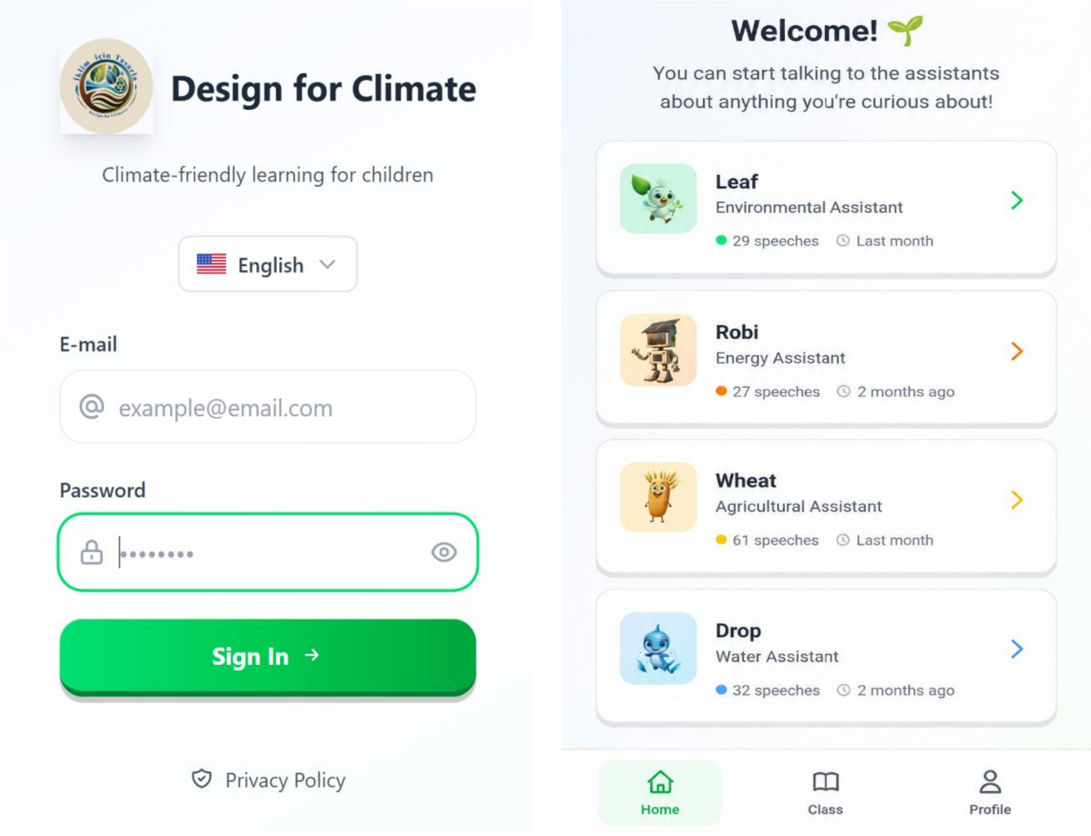
2. Materials and Methods

2.1. Research Design and Participants

This longitudinal case study [25] was carried across three semesters in a public lower secondary school in Türkiye. First semester was the pilot testing, and in the following two semesters the teacher involved a new student group and continued to participate in the study. The teacher was involved in many technology projects, and she was a volunteer to join the pilot. The data sources included (i) teacher evaluations of AI-generated student reports and (ii) a semi-structured interview with a science teacher who used the system in their classroom for approximately one year after pilot test. AI generated assessment was offered by the Design for Climate application. This new application offered an AI-powered educational chatbot platform with four assistants that responded appropriately to off-topic questions and addressed climate-related issues from their own perspective. The system enabled teachers to track activity and message logs. The system also offered metrics for student analytics.

2.2. AI Chatbots and AI Generated Feedback: "Design for Climate" Application

The "Design for Climate" system is an AI-based educational chatbot application developed to increase students' interest and awareness of climate change and sustainability issues. The system has different interfaces for both students and administrators. They can log in using their own email and password information (see **Figure 1a**). The main goal of the system is to provide students with a safe digital learning environment where they can learn at their own pace and increase their curiosity. The system has four chatbot assistants called Drop (Water Literacy Chatbot), Wheat (Agriculture Literacy Chatbot), Robi (Energy Literacy Chatbot), and Leaf (Environment Literacy Chatbot) (see **Figure 1b**).



(a) (b)

Figure 1. (**a**) Design for Climate Application Login Page; (**b**) Design for Climate AI Chatbots.

These chatbots were implemented over a period of time related to the content covered in the curriculum. First a pilot study was carried out on the broader environment domain in which students only interacted with the Leaf chatbot. The pilot phase was successful, and the platform's analytical functions were greatly enhanced by integration of the automated assessment system. In year 2, curriculum was linked to agriculture and Wheat chatbot was primarily used (first semester of the second year). In the second semester of Year 2, the topic in the curriculum was energy. Students primarily used Robi chatbot. **Figure 2** presented a summary of the project's implementation phases. These assistants served as the main characters for student discussions, but students could also use all assistants when they had related questions.

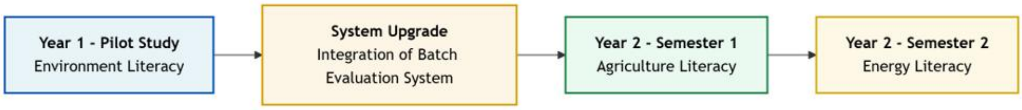


Figure 2. Design for Climate Implementation Phases.

All chatbots can answer questions about any climate topic, not just their specialty. Students can ask any assistant about climate and get detailed answers. **Figure 3** presented a sample student conversation. The flexibility of this system, along with the fact that it could map back pedagogically relevant answers to completely off-topic inputs, ensured an ongoing consistency with the educational goals. Furthermore, the AI chatbots were designed to redirect the conversation with pedagogically appropriate responses, to ensure a continuous alignment with educational goals, in case they were trained to help students get back on related literacy discussion when students ask off-topic questions (see **Figure 4**).

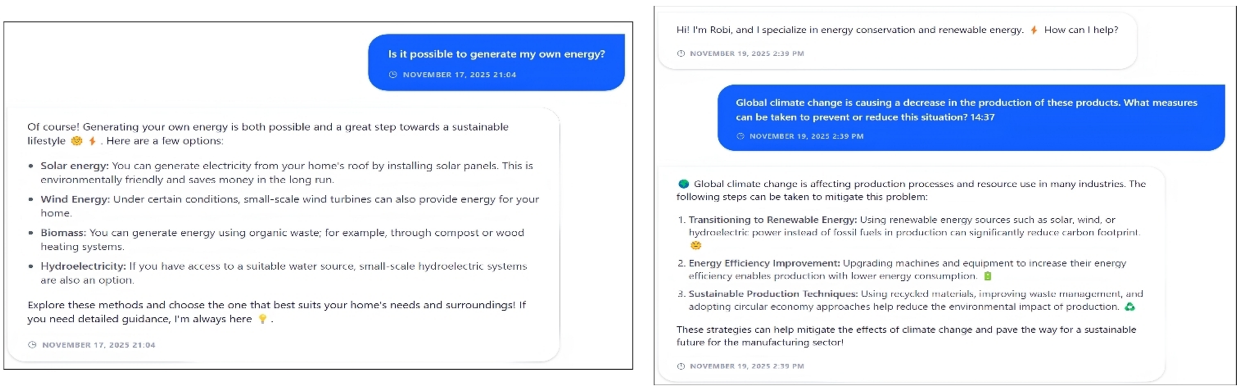


Figure 3. Chatbot Response Examples on to Questions Related to the Literacy Area.

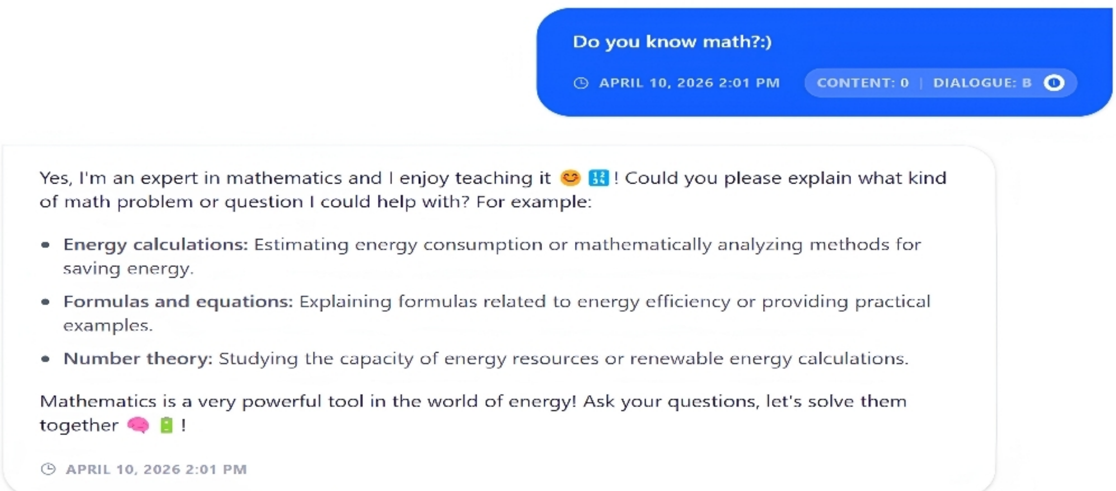


Figure 4. Chatbot Response Example to off Topic Questions.

The system comes with a strong administrative panel that simplifies the data collection process and tracks learning progress. This dashboard enables educators to monitor live engagement data such as trends over time, the platform's total engagement, and distribution of usage across the four AI chatbots (see **Figure 5**).

The platform included an advanced AI-powered Batch Evaluation System. This analytical tool automatically analyzes and assesses students' message logs based on an established set of pedagogical rubrics. It produces individual student diagnostic reports to determine content understanding, dialog complexity and individual cognitive indicators like the presence of scientific reasoning (see **Figure 6**).

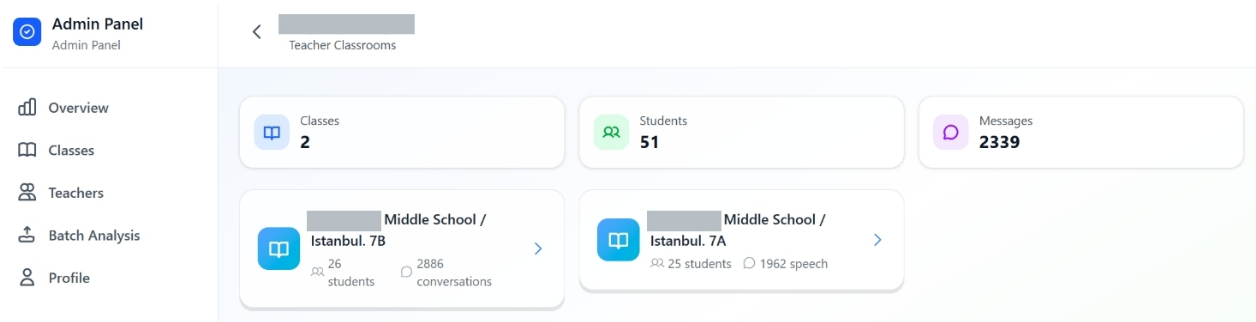


Figure 5. Administrative Panel for Teachers.

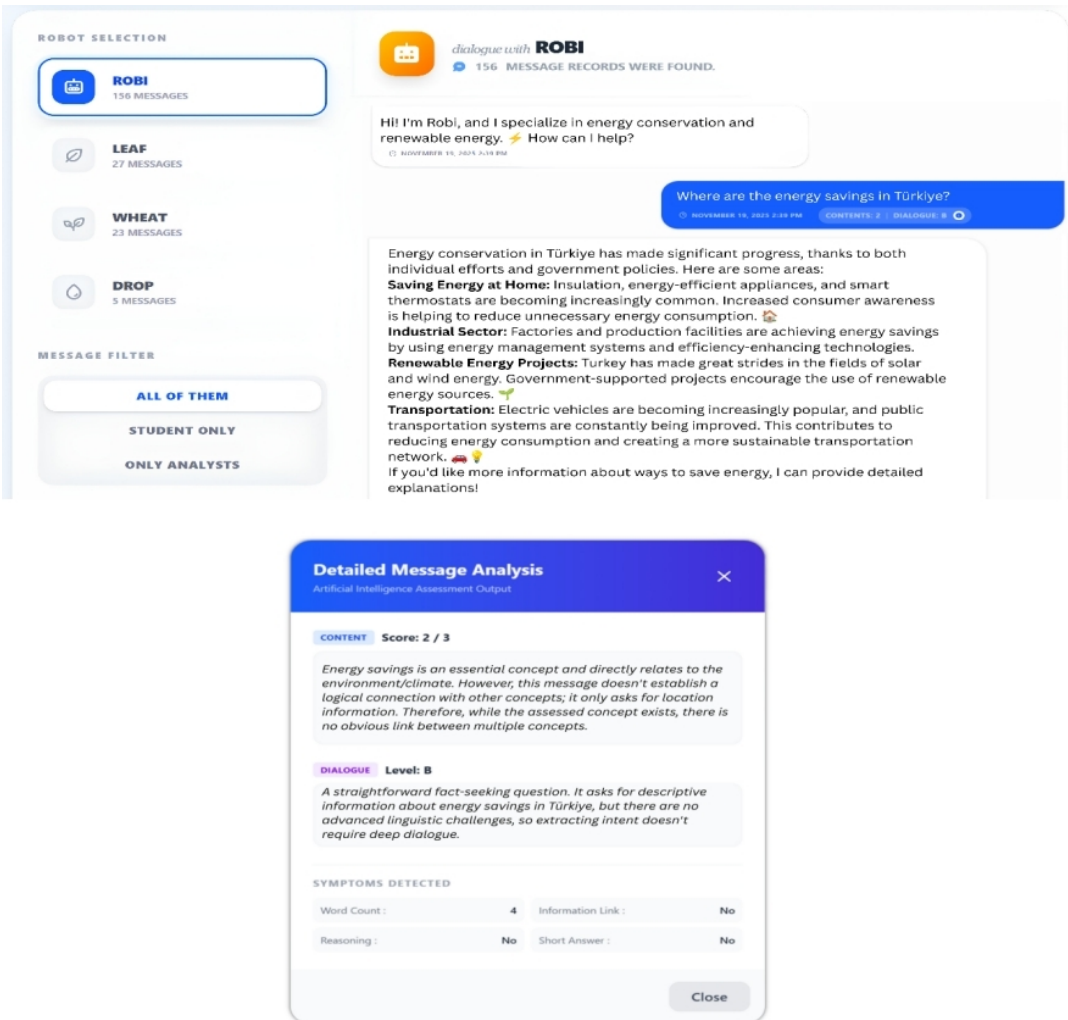


Figure 6. Student Conversation Analysis.

The system combines all these indicators to automatically generate feedback summaries, and represents learning trajectories, allowing for a detailed analysis of student discussion from their interaction data (see **Figure 7**).

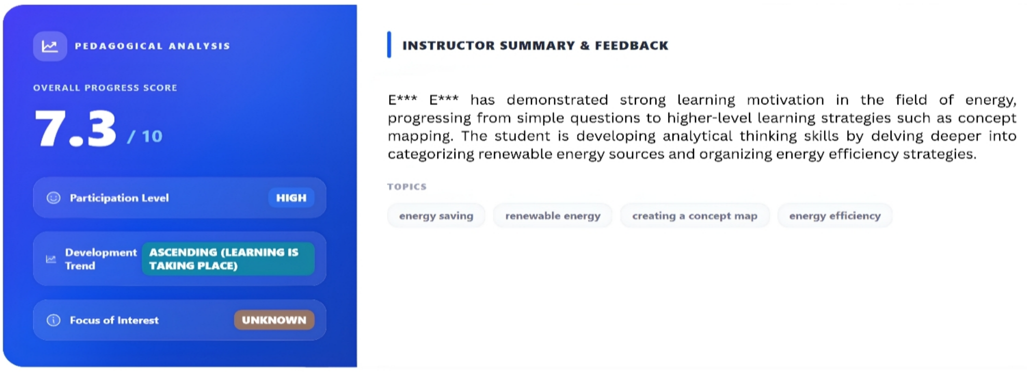


Figure 7. Student Progress Summary.

2.3. Participants

Case study included an expert middle school science teacher and 51 seventh grade students. Students participating in the pilot study were not involved in the main implementation. Conducting the study with a new group of participants after the pilot application eliminated the learning effects that may arise from the pilot process, allowing for a more reliable evaluation of the findings. It also allowed the intervention developed after the pilot to be tested on an independent sample, demonstrating that the results obtained were not exclusive to the pilot group. This approach increased the generalizability of the findings and provided stronger evidence that the intervention can produce similar results in different groups of participants [26].

During the study, the teacher and her 7th grade students used the application in two semesters. 51 students messaged with AI chatbots 2,339 times. The study used the criterion sampling to strengthen both the precision and depth of its qualitative findings [27]. Of the first group of 51 students, those who sent at least 40 messages in both semesters were selected.

The system analyzed all messages, but this study only included students who used the application frequently in both semesters. This threshold was determined by examining pilot data prior to analyses. Pilot analyses show that interactions under 40 messages are mostly short-term, procedural uses; showed that students did not produce enough data reflecting different stages of the learning process. This approach is also in line with research that emphasizes that sustained behavioral engagement [28] should be considered instead of one-time interactions when interpreting students' learning processes in learning analytics studies.

Chat interactions of these 10 students over two semesters were first analyzed by the system and then the teacher analyzed messages. The platform generates personalized learner reports for each learner. Moreover, a blind evaluation approach was adopted to ensure the validity of the study and maintain the objectivity of the expert evaluation. Before reviewing the AI-generated student reports, the teacher independently analyzed each student based on students' interactions with the chatbot.

2.4. Data Collection Tools and Process

Quantitative assessment tools and qualitative interview techniques were used to provide data triangulation. The expert teacher conducted an independent analysis of how students interact with the AI chatbot. Subsequently, custom AI-generated student reports documented for 10 students to present the teacher's review after blind evaluation. The expert teacher filled out a "Student Report Evaluation Form" via Google Forms (see **Appendix A**) that was created to assess the accuracy, consistency, and teaching value of student reports which were generated by AI. The assessment had six items that were presented on a 5-point Likert scale (1 = Strongly disagree; 5 = Strongly agree). Items determined if the student reports (1) accurately reflected the student's characteristics and learning style, (2) identified strengths that were congruent with teacher observations, (3) identified areas of growth that were congruent with teacher observations, (4) the learned tendency classifications (improving, stable or challenged) matched the teacher's perception, (5) was useful in understanding the student better, and (6) guide individual support decisions and instructional planning.

In the second phase, a semi-structured video interview was conducted to examine AI interpretations (see **Appendix A**). The interview was recorded and transcribed. The questions in this step first discussed (1) how evaluations in the report were particularly accurate and useful, (2) which parts were not accurate and useful. In the next part we discussed why teachers gave higher scores for several reports and what aspects were more reliable and more educationally useful for the teacher. Semi structured interview questions were also generated by two faculty members in educational sciences and computer science department similar to the previous data collection tool. Student Report Evaluation Form and semi structured interview questions were generated by two faculty members in educational sciences and computer science departments. Both researchers conducted research on AI and role of using technology in educational settings.

2.5. Data Analysis

To compare AI interpretations of system analysis with the teacher's assessments, the study performed a two-stage analysis. First, we reported teacher's scoring for AI student analytics. For the interview questions, we used an inductive method [29]. We extracted themes from data without predetermined classifications [30]. The first two authors independently coded both teacher evaluations and student reports. Authors had 86% agreement rate and resolved all disagreements through discussion meetings. Both authors coded entire questions since we only had one interview.

The analysis included three themes. **Figure 8** presented how different codes emerged under different analysis sections regarding to these themes. The theme "AI-Generated Learning Patterns" includes cognitive patterns, behavioral indicators, and learning trends. "Teacher Validation of AI Profiles" emerged from teacher evaluations and demonstrated agreement between system assessments and the teacher's perceptions of student profiles and instructional usefulness. Final theme "Calibrated Trust" emphasized the connection system assessments, highlighting the necessity for specificity and useful insights shaped by her three semesters of experience with the system.

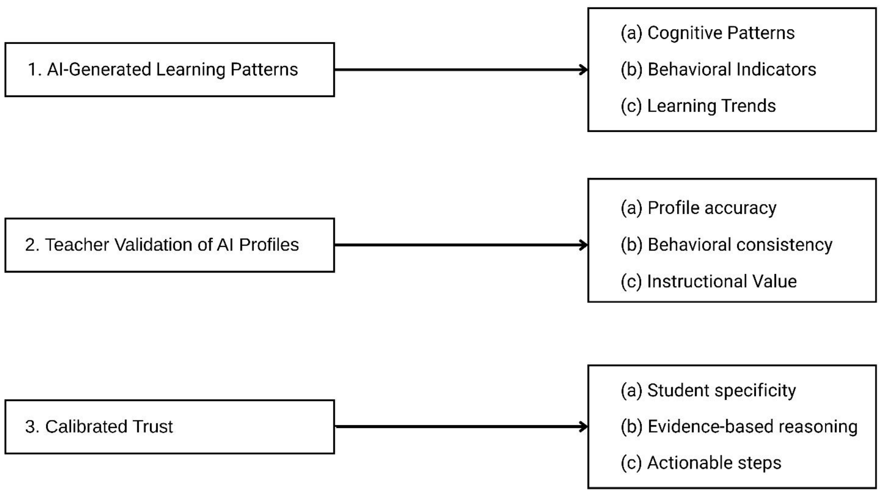


Figure 8. Themes and Codes.

In this research, calibrated trust is defined as the teacher's neither unconditional acceptance nor complete rejection of AI outputs; instead, they evaluate the system together with their pedagogical expertise, considering its strengths and limitations. This concept is based on the concept of credibility [23] and trust [24]. Therefore, in this study, trust refers to the balanced relationship that the teacher establishes between the artificial intelligence outputs and his/her own professional judgment rather than adopting technology.

3. Results

This case study focused on the evaluation of learning analytics reports generated for students who interacted with the AI system for at least 40 messages in both implementation periods. This selection aimed to examine situations where students interact adequately with the system and AI can generate data that can create meaningful patterns about students' learning processes. In addition, student reports were evaluated by an experienced teacher who participated in the pilot implementation process of the system and used the AI-supported learning environment for three semesters. Therefore, the findings reveal not only the student analytics generated by AI, but also how these analytics are interpreted by an experienced teacher.

3.1. AI-Generated Learning Patterns and Teachers' Interpretations

The teacher first evaluated 10 students in six steps. When teachers' opinions on student evaluation reports generated by artificial intelligence are examined, it is seen that positive evaluations stand out in all items. While the average scores in the first four dimensions (accuracy of the student profile, strengths, areas open to development and accurate reflection of learning tendency) were 4.0, the average scores regarding the contribution of the report to getting to know the student better and its usability in instructional planning and support decisions for the student were determined as 4.2. In all items, 100% of the teachers answered "Agree" or "Strongly agree". While the mean was 4.00 in the descriptive accuracy dimensions, it increased to 4.20 in the instructional usefulness dimension. This shows that teachers see reports not only as accurate, but also as tools to support their teaching decisions. Overall, the teacher demonstrated an agreement rate exceeding 80% with GenAI generated reports (see **Figure 9**). These results indicate that teachers perceive AI-generated assessment reports not only as outputs that accurately reflect the student's characteristics, but also as a practical and reliable support mechanism that can be used in instructional decision-making processes.

Figure 9. Teacher Agreement with System Generated Reports.

When teachers' written evaluations were examined, the first important finding revealed in the reports is that artificial intelligence can make long-term patterns of students' learning processes visible. However, high engagement does not always translate to deep learning. Although all students are in the same learning environment, their learning tendencies differ from each other. While some students initially asked superficial questions, they started to ask more complex and research-oriented questions over time, while some students showed high interest at the beginning but lost focus in the future. This shows that students' learning journeys are not linear.

AI-generated reports revealed not only what subjects that students were interested in, but also how those interests changed over time. Learning tendency indicators have enabled the monitoring of students' development processes. For some students, "learning is taking place" was evaluated, while for some students, "loss of focus or strain" patterns were determined.

Another noteworthy point is that artificial intelligence can capture not only cognitive processes but also behavioral patterns. Behaviors such as constantly asking for short answers, changing subjects, sending meaningless messages or persistently working on certain topics provided important indicators of students' learning approaches. This finding suggests that AI-driven learning analytics not only generate performance indicators but can also provide process-based insights into students' learning processes.

3.2. Teacher Validation of AI Profiles

When the teacher's evaluations are examined, it is generally thought that the reports generated by artificial intelligence represent the students correctly. However, the teacher's evaluations are not shaped only by accuracy. It has become more important to what extent it can define the student and how visible it can make student-specific patterns.

It is seen that the characteristic learning behaviors of the students are defined especially in the reports with high scores. The teacher stated that for some students, characteristics such as the tendency to structure the specified information, the effort to make connections between concepts, and the constant use of certain learning tools accurately reflected the students. Similarly, she stated that behaviors such as distraction, constantly asking for summaries, or going off topic for some students were compatible with her own classroom observations. For example, during the interview, when discussing a high-scoring report for a specific student, she stated: "The inferences are very accurate. Although he is an academically very bright and quick-learning child, his attention gets distracted quickly, he gets bored easily, and he often misses learning opportunities in science subjects. So, there is a highly consistent observation here regarding his general learning behavior."

Likewise, in assessing another student, the teacher highlighted that AI analysis was valuable for instruction: "He is generally a child who needs extra time to make sense of and structure subjects, and he successfully does this when he focuses. I can safely say that this perfectly matches his in-class learning behavior. Seeing that these inferences were consistent across multiple students naturally increased my trust."

The remarkable point here is that the teacher's ideas do not come from estimates of academic achievement, but from the accurate representation of the student's learning process. For the teacher, inferences about how the student learns seem to be more meaningful than the evaluations about how successful the student is. This was clearly seen in one of the assessments: the reason she gave a high score to AI generated summary was not because of the student's academic achievement, but because it accurately identified behaviors such as distractibility and disengagement that she had personally observed in the classroom.

When open-ended responses are examined, it is seen that the teacher finds the sections that describe the students' interests, participation styles, questioning habits and learning behaviors useful. This shows that artificial intelligence is not considered by teachers as a scoring system, but as an analytical tool that makes students' learning processes visible.

She also stated that the reports helped her to see the students in a more holistic way. However, this contribution was realized in the form of systematizing the observations she had about the students rather than providing completely new information. For this reason, artificial intelligence is not positioned as an expert replacing the teacher, but as a tool that supports her professional evaluations.

3.3. Calibrated Trust

The most important finding of this case study is that the teacher's trust in artificial intelligence is not unconditional. The teacher, who uses the system for three semesters starting from the pilot application, does not automatically accept the evaluations produced by artificial intelligence, but constantly compares them with their own observations.

This is especially seen in the scores given to the reports. She evaluated some reports which described specific learning behaviors of students with full points (5 out of 5). However, she gave lower points to some reports that seemed to rely just on the students' interaction with the system, rather than students' overall learning behavior, attention span or pedagogical tendencies. The main reason for this difference is the level of student-specificity rather than the accuracy of the evaluations. The teacher found the reports reflecting the behavioral patterns that distinguish the student from others more reliable. On the other hand, general expressions that can be used for many students were evaluated lower.

In assessing pupils' motivation, focus, and attitudes toward learning, a teacher stressed the value of evidence and insisted that the reasoning behind findings be made clear. She pointed out that although AI technologies are good at identifying learning problems, they frequently fall short of offering practical solutions. She stated that she wanted reports to provide specific recommendations for development, such improving focus or controlling distractions, in addition to correctly describing student behavior: "In summary, it would have been fantastic if it had included leading recommendations for growth and improvement after identifying the pupils' strong and weak points. How might the learner improve their focus, for instance? How can they carry on the discussion without using wrong terminology or losing focus?"

These results imply that long-term use of GenAI helped her gain knowledge of the system's advantages and disadvantages over time, as well as how to draw conclusions and when to seek further proof. In other words, the situation that emerged at the end of the three-semester experience is not a complete surrender to artificial intelligence, but a critical and calibrated trust relationship.

Overall, this case study shows that AI-powered learning analytics make more sense when teachers are positioned as analytical partners that support them rather than replace their professional judgment. The teacher's evaluations show that student analytics produced by GenAI are not only based on accuracy. Student analytics also depend on factors such as student specificity, explainability and pedagogical usefulness.

4. Discussion

This study examined a teacher's involvement in a project that focused on developing and implementing Design for Climate application. In the pilot year, the teacher tested the application and in the second year, she continued with another student group. Conducting the study with a new group of participants following the pilot application increased the strength and generalizability of the findings, showing that the effects of the intervention were not applicable to the pilot sample [31]. In addition, a teacher's continued use of the system in their classroom for two semesters after the pilot was an important indicator supporting the sustainability of the teacher trust for AI generated assessment. In this study, teacher trust was evaluated with student analytics generated by an AI chatbot and teacher evaluations. The first evaluation source consisted of the evaluation scores given by the teacher to the student reports generated by AI, and the second data source consists of a semi-structured interview with the science teacher. This approach differed from existing studies in that it reveals not only which reports the teacher found more successful, but also the pedagogical reasons behind this evaluation. Previous studies underlined that AI based assessment can be used for assessment [8,32]. In this section, we will present how AI based assessment gained teacher's trust.

4.1. AI-Generated Learning Patterns and Teachers' Interpretations

The first theme revealed how the student analyzes generated by artificial intelligence are interpreted by the teacher. When the teacher evaluations and interview findings are examined together, it is seen that the common feature of the highly rated reports is that they do not only summarize the student's chat history. The reports that the teacher finds most valuable are the reports that produce pedagogical interpretations of the learning processes of the students. For example, one of the reports stated that the student needed more time to make sense of and structure the information, and that the ability to establish relationships between concepts improved; The teacher, on the other hand, stated that these evaluations completely coincided with the student's learning behaviors in the classroom. Similarly, in the Ibrahim Edhem report, comments about the student's short attention span, impatience with detailed explanations and therefore missing learning opportunities were found to be consistent with the information the teacher gained by observing the student for a long time.

This finding suggests that the teacher didn't expect AI to merely depict student behavior. What is valuable for the teacher is not what the student does in the system, but what these behaviors say about the student's learning style. In other words, teachers value pedagogical interpretations that explain the student's learning characteristics rather than summarizing chat records. This situation reveals that AI-supported learning analytics should go beyond just producing descriptive analyses and offer pedagogical inferences that make sense of the student's learning process [21,33].

The teacher had an agreement rate exceeding 80% with student analytics. This indicated that the teacher finds GenAI student reports both accurate and instructionally beneficial. Open-ended responses reveal the reasons behind this assessment. The teacher stated that the reports largely coincided with their own classroom observations of individual characteristics such as students' difficulty in focusing, interest in the subject, research tendency, academic potential, difficulties in the learning process, and even language barrier. When these results are evaluated in terms of the Technology Acceptance Model [22], it showed that teachers perceive GenAI not only as an easy-to-use tool but also as a useful technology that makes a concrete contribution to teaching processes.

The teacher stated that these reports could provide deep enough inferences about the student's general learning behaviors, attention characteristics, or learning tendencies. Therefore, the value of AI-generated analyses seems to depend not only on the amount of information it contains, but also on how meaningful and explanatory this information is for the teacher. Our results confirm that AI supported analytics can support analysis [34], and including student input could be useful to improve the integrity of this analysis [35].

4.2. Teacher Validation of AI Profiles

One of the most significant contributions of the research is the comparison of AI-generated student profiles with long-term classroom observations. The teacher used the system in her own classroom for about a year from the pilot implementation process and had the opportunity to observe the same students both in their learning processes in the classroom and in their interactions with artificial intelligence. This allowed the evaluation of the validity of AI-generated student profiles in the context of the actual classroom. The interview findings show that the teacher largely validates the AI-generated student profiles. In particular, the inferences regarding students' attention patterns, learning tendencies, conceptual development and learning behaviors overlap significantly with the teacher's long-term observations. However, the important finding of this study is that trust is not formed through only two successful examples. The teacher especially emphasized that inferences with similar accuracy were repeated in other student reports.

These findings show that teachers do not evaluate artificial intelligence systems based on single examples [10]. The reliability of AI-generated student models depends on their ability to work consistently across different student profiles. In other words, for the teacher, trust is not fed by a single correct analysis, but by repeated patterns of truth in different students. In this respect, the study reveals that consistency as well as accuracy is a critical element for the teacher [36]. In addition, these findings show that AI is found more meaningful by teachers when it can model students based on the student's general learning characteristics, not just in a specific chat session. This indicates that learning analytics should evolve from systems that only report interaction data to systems that explain the student's learning profile [9,37].

4.3. Calibrated Trust

The third important conclusion of the research is that teacher trust is a process. Throughout the interview, the teacher stated that she did not directly accept any inferences generated by artificial intelligence; she stated that she constantly evaluates them by comparing them with her own classroom observations and professional experience. Thus, the trust revealed in this study is not an absolute trust in AI, but a calibrated trust that is constantly validated and reproduced by the teacher's professional judgment [8].

The teacher's evaluations also show that their expectations from artificial intelligence go beyond the current reports. Although the teacher found it valuable to identify the strengths and areas that needed improvement among the students, they considered the lack of suggestions on how to use these findings in the teaching process as a significant limitation. In particular, she stated that suggestions such as how to maintain the student's attention, how to increase learning motivation, how to support classroom learning behaviors or how to make the conversation process more efficient will significantly increase the instructional value of the reports [38].

The findings also coincide with the trust model [23]. To establish trust, users must first believe that the system is sufficient and reliable. The fact that the teacher finds the student profiles in the reports highly compatible with their own observations shows that artificial intelligence is considered sufficient to analyze students' learning characteristics. This finding supports the teacher trust [24]. Teacher neither unconditionally trusts GenAI nor rejects the system. Instead, she acknowledges the strengths of the reports while critically highlighting the aspects that need improvement. In particular, the expectation of more personalized recommendations and intervention strategies suggests that the teacher views GenAI as a tool to support their decisions, rather than as an authority that replaces their pedagogical expertise. Teacher's understanding of artificial intelligence allows them to evaluate the accuracy of system outputs. Effective use of GenAI in education could pave the way for trust [24] and perceived usefulness [22].

5. Conclusions

This study examined teacher trust in student analytics generated by an AI chatbot by integrating two different data sources. While teacher evaluations revealed which reports were found to be more successful, a semi-structured interview with a teacher with about a year of usage experience explained the pedagogical reasons behind this evaluation. Thus, it has been possible to understand the trust in AI-supported student analytics not only with the scores, but also with the teacher's professional evaluation process.

Research findings show that teacher trust is built on three basic elements. The first is that AI-generated student analytics accurately reflect students' long-term learning behaviors and learning characteristics. The second is that this accuracy is consistently observed not only in specific students but also in different student profiles. The third is that the AI-generated analyses provide meaningful and actionable recommendations that can support the teacher's pedagogical decisions.

In this respect, the study makes a significant contribution to GenAI-supported learning analytics research. The findings show that teachers' trust in artificial intelligence is based on how accurately artificial intelligence recognizes students, how much it makes pedagogical sense of this information, and to what extent it can contribute to the teaching process, rather than algorithmic accuracy. In other words, teachers expect GenAI to understand the student pedagogically, not just identify them.

As a result, the success of future GenAI chatbots should be judged not only by their ability to create more accurate student models, but also by their ability to transform these models into explainable, student-specific and actionable pedagogical recommendations that support teachers' professional decision-making. The role of artificial intelligence in education is not to replace the teacher; to be a reliable pedagogical partner that strengthens the teacher's many years of professional experience, helps them make sense of the classroom reality and supports their teaching decisions.

The findings of this study are limited to data obtained from a science teacher who voluntarily participated in the use of technology and was open to AI-assisted teaching practices. Therefore, the results cannot be generalized to all teachers. It should be considered that trust and the use of GenAI may develop in different ways, especially in teachers with different characteristics in terms of attitude towards GenAI, teachers' competence and teaching experience. It is important that future studies test the findings with multi-case studies involving teachers from different subjects, different levels of experience, and different school contexts.

It is also important to note that the use of a 40-message threshold may have led to a certain selection bias. This criterion has caused students who use artificial intelligence in a limited way or stopped using it at an early stage to be excluded from the analysis, leading to qualitative findings reflecting the experiences of students with high levels of interaction. Therefore, the results obtained should be interpreted for students who use GenAI in a longer span rather than for all participants. In future studies, comparing low, moderate, and high-interaction students may reveal the effect of this possible bias in more detail. The learning analytics literature also draws attention to the fact that direct comparison of users with high engagement and low engagement users can create self-selection and participation bias [39].

The system evaluated in this study is a field-oriented artificial intelligence platform developed specifically for the climate change curriculum and "Design for Climate" activities. Therefore, the system is not only based on general language model capabilities; It also accesses contextual information based on students' activity history, learning objectives, teaching process and interaction records. In contrast, general-purpose generative GenAI systems such as ChatGPT or Gemini can analyze the same chat recordings, but by default they do not know the student's previous learning progress, curriculum objectives, and classroom context. For this reason, it can be expected that teachers' trust in these systems will differ. Domain-oriented educational GenAI has the potential to provide more explainable and contextual assessments that support teacher confidence because they work in integration with the curriculum, pedagogical objectives, and learning analytics. Future studies can make significant contributions to analyzing the same student data with general-purpose and education-oriented generative GenAI systems and comparing their effects on teacher confidence.

Author Contributions

İ.D. was the principal investigator for the project and led the entire process. A.B. led data collection and analysis. B.Ş. and G.İ. led development of the application and created AI generated reports. All authors have read and agreed to the published version of the manuscript.

Funding

This study was conducted within the scope of the project titled “Design for Climate: Learning and Retention in Design-Based Learning.” The project is supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 1001 Scientific and Technological Research Projects Funding Program (Project No: 223K222).

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by Boğaziçi University Social and Human Sciences Ethics Committee (SBİNAREK) (protocol code 2025/08 October 13, 2025).

Informed Consent Statement

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

Data Availability Statement

Data will be provided based on reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this manuscript, the authors used Grammarly solely for language refinement. All outputs were critically reviewed and edited by the authors.

Appendix A. Student Report Evaluation Form and Interview Questions

Student Report Evaluation Form

Please select the student you will evaluate: ________________
Please rate the following statements on a scale from
1 (Strongly Disagree) to 5 (Strongly Agree).

Statement	Strongly Disagree (1)	Disagree (2)	Undecided (3)	Agree (4)	Strongly Agree (5)
1. The student profile presented in the report (areas of interest, participation style, learning behavior) accurately reflects the student.	○	○	○	○	○
2. The strengths of the students identified in the report are consistent with my own observations.	○	○	○	○	○
3. The areas for improvement identified for the student in the report are consistent with my own observations.	○	○	○	○	○
4. The learning tendency assessment (improving, stable, struggling, etc.) accurately reflects the student’s current situation.	○	○	○	○	○
5. This report helps me get to know the students better.	○	○	○	○	○
6. The information in this report can be used in instructional planning.	○	○	○	○	○

Interview Questions

General Questions

1. Which of the assessments included in the report did you find particularly accurate and useful?

2. What are the aspects of the report that are missing, inaccurate, or in need of improvement?

Student-Specific Questions

1. S1 *(you assigned a score of 5 for this student)*
In S1’s report, artificial intelligence commented that the student “makes an effort to structure the knowledge he has learned” and “has developed the ability to make sense of topics and establish connections between them.” Which of these comments align with your own observations in the classroom? Which characteristics of the student that you already knew did the artificial intelligence confirm here, and which new perspectives did it offer? If similarly, student-specific inferences had been present in the other student reports at a comparable level, would your confidence in them have increased?

2. S5 *(you also assigned a score of 5 for this student)*
In S5’s report, artificial intelligence made behavioral inferences such as the student “repeatedly giving the command to shorten,” “not taking the learning process seriously,” “having a short attention span,” and “showing impatience with detailed explanations.” Which of these comments align with your own observations in the classroom? Which points do you think artificial intelligence captured accurately about the student and which points do you think it interpreted incompletely or incorrectly?

3. While you rated the reports for S1 and S5 at the highest level, you gave lower scores to the reports for the other students. In these reports, what were the points where you felt the artificial intelligence failed to sufficiently understand the student? What additional evidence or inferences regarding the student’s learning behaviors, motivation, or in-class characteristics would have increased your confidence in these reports?

References

1. Shubik, M. Bibliography on Simulation, Gaming, Artificial Intelligence and Allied Topics. *J. Am. Stat. Assoc.* **1960**, *55*, 736–751. [CrossRef]

2. Delen, I.; Sen, N.; Ozudogru, F.; et al. Understanding the Growth of Artificial Intelligence in Educational Research through Bibliometric Analysis. *Sustainability* **2024**, *16*, 6724. [CrossRef]
3. Moundridou, M.; Matzakos, N.; Doukakis, S. Generative AI Tools as Educators' Assistants: Designing and Implementing Inquiry-Based Lesson Plans. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100277. [CrossRef]
4. Ndjama, J.D.N. The Use of Artificial Intelligence in Lesson Delivery and Evaluation in Large-Scale Teaching Environments. *Interdiscip. J. Educ. Res.* **2025**, *7*, 1. [CrossRef]
5. García Laborda, J.; Magal Royo, T.; Fernández Alvarez, M. An Approach to the Advance of Tools Based on Generative Artificial Intelligence for Language Assessment: A SWOT Analysis. *Rev. Iberoam. Estud. Educ.* **2024**, *19*, e19060. (in Spanish)
6. Shaw, B.P. Artificial Intelligence and Assessment: Three Implications for Music Educators. *Music Educ. J.* **2024**, *111*, 19–25.
7. Breen, P. Towards a Pedagogy of Negotiation: EAP Teachers Rethinking University Teaching and Assessment in the Digital Age. *Lond. Rev. Educ.* **2026**, *24*, 9.
8. Huertas-Abril, C.A.; Palacios-Hidalgo, F.J. New Possibilities of Artificial Intelligence-Assisted Language Learning (AIALL): Comparing Visions from the East and the West. *Educ. Sci.* **2023**, *13*, 1234.
9. Söğüt, S.; Çorbacı, E.C.; Hegelheimer, V. Designing Feedback and Assessment for GenAI-Assisted Digital Multimodal Composition in L2. *TESOL Q.* **2026**. [CrossRef]
10. Lepage, A.; Roy, N. Studying Artificial Intelligence Adoption by Postsecondary Teachers in Quebec through Five Use Cases. *Int. J. Technol. High. Educ.* **2025**, *22*, 13. (in French)
11. Shahid, M.K.; Zia, T.; Bangfan, L.; et al. Exploring the Relationship of Psychological Factors and Adoption Readiness in Determining University Teachers' Attitude on AI-Based Assessment Systems. *Int. J. Manag. Educ.* **2024**, *22*, 100967.
12. Gumilar, S.; Ong, Y.S.; Bhakti, D.D.; et al. Examining Generative Artificial Intelligence (GenAI) in Assessing Students' Responses to Socio-Scientific Issues in Physics. *Int. J. Sci. Educ.* **2025**, 1–24.
13. Çeltik, E.; Şenceylan, B.; Delen, İ.; et al. Automated Evaluation of Student-Chatbot Interactions in Climate Education Using Large Language Models. In *Artificial Intelligence in Education*; Blanchard, E.G., Chen, G., Chi, M., et al., Eds.; Springer: Cham, Switzerland, 2026.
14. Noroozi, O.; Haddadian, G.; Gao, X.; et al. The Value of GenAI for Peer Feedback Provision: Student Perceptions and Impacts. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 61.
15. Archila, P.A.; Ortiz, B.T.; Truscott de Mejía, A.M.; et al. Critically Reading Science-Related Texts Produced by ChatGPT. *Sci. Educ.* **2025**, *34*, 4627–4662.
16. Mun, J.; Kim, M.; Kim, S.W. How Seventh-Grade Students Experience the Complexity of Socioscientific Issues through Decision Making on the Autonomous Vehicle Issue. *Asia-Pac. Sci. Educ.* **2022**, *8*, 43–71.
17. Martin, P.P.; Kranz, D.; Wulff, P.; et al. Exploring New Depths: Applying Machine Learning for the Analysis of Student Argumentation in Chemistry. *J. Res. Sci. Teach.* **2024**, *61*, 1757–1792.
18. Guo, K.; Zhong, Y.; Li, D.; et al. Effects of Chatbot-Assisted In-Class Debates on Students' Argumentation Skills and Task Motivation. *Comput. Educ.* **2023**, *203*, 104862.
19. Corbin, T.; Bearman, M.; Boud, D.; et al. Assessment after Artificial Intelligence: The Research We Should Be Doing. *J. Univ. Teach. Learn. Pract.* **2025**, *22*, 1–12.
20. Yin, H. Development and Testing of a Teaching Quality Assessment and Examination Data Collection System Based on Artificial Intelligence. *Int. J. Contin. Eng. Educ. Life Long Learn.* **2026**, *36*, 40–56.
21. Elsawah, W. Rethinking Student Assessment in the Age of AI: Insights and Framework Proposals from a Gulf-Based Educational Program. In *Integrity, Assessment, and Collaboration in the AI Classroom*; IGI Global Scientific Publishing: Hershey, PA, USA, 2026; pp. 211–240.
22. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **1989**, *13*, 319–340.
23. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. An Integrative Model of Organizational Trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734.
24. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80.
25. Street, C.T.; Ward, K.W. Improving Validity and Reliability in Longitudinal Case Study Timelines. *Eur. J. Inf. Syst.* **2012**, *21*, 160–175.
26. Ismail, N.; Kinchin, G.; Edwards, J.A. Pilot Study, Does It Really Matter? Learning Lessons from Conducting a Pilot Study for a Qualitative PhD Thesis. *Int. J. Soc. Sci. Res.* **2018**, *6*, 1–17.
27. Palinkas, L.A.; Horwitz, S.M.; Green, C.A.; et al. Purposive Sampling for Qualitative Data Collection and Analysis in Mixed Methods Application Research. *Admin. Policy Ment. Health Ment. Health Serv. Res.* **2013**, *42*, 533–544.

28. Pan, J.; Liang, J.; Hu, Y. Gamified Behavioral Engagement as an Indirect Association Linking Ideal L2 Self and Vocabulary Achievement in Mobile-Assisted EFL Vocabulary Learning. *Front. Psychol.* **2026**, *17*, 1836665.
29. Glaser, B.G.; Strauss, A.L. *Grounded Theory: Strategien Qualitativer Forschung*; Huber: Bern, Switzerland, 1998.
30. Vears, D.F.; Gillam, L. Inductive Content Analysis: A Guide for Beginning Qualitative Researchers. *Focus Health Prof. Educ.* **2022**, *23*, 111–127.
31. Beets, M.W.; Weaver, R.G.; Ioannidis, J.P.; et al. Identification and Evaluation of Risk of Generalizability Biases in Pilot versus Efficacy/Effectiveness Trials: A Systematic Review and Meta-Analysis. *Int. J. Behav. Nutr. Phys. Act.* **2020**, *17*, 19.
32. Zhai, X.; Nehm, R.H. AI and Formative Assessment: The Train Has Left the Station. *J. Res. Sci. Teach.* **2023**, *60*, 1390–1398.
33. Kim, J. Leading Teachers' Perspective on Teacher-AI Collaboration in Education. *Educ. Inf. Technol.* **2024**, *29*, 8693–8724.
34. Li, Y.; Shan, Z.; Raković, M.; et al. When AI Explains in Natural Language: Unveiling the Impact of Generative AI Explanations on Educators' Grading and Feedback Practices. *Educ. Inf. Technol.* **2025**, *30*, 24931–24964.
35. Lee, S.; Song, K.S. Teachers' and Students' Perceptions of AI-Generated Concept Explanations: Implications for Integrating Generative AI in Computer Science Education. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100283.
36. Ishaq, S.M.; Bangulzai, A.R.; Asif, M.; et al. Artificial Intelligence as a Co-Teacher: Enhancing Educator Efficiency through Intelligent Systems. *Crit. Rev. Soc. Sci. Stud.* **2025**, *3*, 1324–1342.
37. Sajja, R.; Sermet, Y.; Cwiertny, D.; et al. Integrating AI and Learning Analytics for Data-Driven Pedagogical Decisions and Personalized Interventions in Education. *Technol. Knowl. Learn.* **2025**, 1–31.
38. Hu, J.; Huang, Z.; Li, J.; et al. Real-Time Classroom Behavior Analysis for Enhanced Engineering Education: An AI-Assisted Approach. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 167.
39. Cabral, L.; Pinto, R.; Gonçalves, G. AI-Powered Learning Analytics Dashboards: A Systematic Review of Applications, Techniques, and Research Gaps. *Discover Educ.* **2025**, *4*, 525.