Too Skeptical or Not Skeptical Enough? Middle School Student and Teacher Perceptions of LLM-Based Project Assessment

XIAOYI TIAN, North Carolina State University, USA
SHAN ZHANG, University of Florida, USA
YUKYEONG SONG, University of Florida, USA
AMOGH MANNEKOTE, University of Florida, USA
JOANNE BARRETT, University of Florida, USA
CHRISTINE FRY WISE, The Findings Group, USA
EMILY DOBAR, The Findings Group, USA
KRISTY ELIZABETH BOYER, University of Florida, USA
MAYA ISRAEL, University of Florida, USA

Large language models (LLMs) have the potential to revolutionize education to improve teaching and learning. LLMs are already being used for automated assessment, but very little is known about teachers' and students' perceptions of AI-based assessment and how we can design systems that support these two key groups of users. This paper addresses the research gap in the context of project-based learning by investigating how students and teachers perceive the trustworthiness, usefulness, and challenges associated with LLM-based assessment of student-created conversational agents. Through focus groups with 30 middle school students (aged 11-12) and interviews with four in-service middle school teachers, we identified differences in trust levels and preferences for LLM-based assessments. We propose design recommendations for effective LLM-based project assessment, including mitigating students' over-trust and misconceptions, addressing teachers' hesitation, and effective experience design for students and teachers to adopt and use.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: large language models, artificial intelligence, assessment, project-based learning

ACM Reference Format:
Xiaoyi Tian, Shan Zhang, Yukyeong Song, Amogh Mannekote, Joanne Barrett, Christine Fry Wise, Emily Dobar, Kristy Elizabeth Boyer, and Maya Israel. 2024. Too Skeptical or Not Skeptical Enough? Middle School Student and Teacher Perceptions of LLM-Based Project Assessment. 1, 1 (September 2024), 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Artificial intelligence (AI) has become an integral part of education and has revolutionized the ways people teach and learn. Among the educational AI applications, generative AI tools and large language models (LLMs) have emerged as particularly influential [40]. These tools support a wide range of educational tasks such as creating course materials [30], providing writing support [55, 82], generating programming codes [9, 42, 43], and automating the evaluation and scoring of student assignments [18, 92]. The use of generative AI-powered applications has brought many benefits to teaching and learning, such as enhancing productivity [46, 55], improving students' language proficiency and confidence [39], leading to an improved learning experience. By leveraging these technologies, educators can reduce the time and energy spent on tedious tasks and thereby improve the efficiency of instructional preparation and the quality of interactions with students.

Project-based learning is a widely adopted instructional approach [49, 78] in computer science (CS) and AI education. This approach engages learners in an open-ended process of designing and developing computational artifacts [26, 28] and

Authors' Contact Information: Xiaoyi Tian, xtian9@ncsu.edu, North Carolina State University, Raleigh, North Carolina, USA; Shan Zhang, zhangshan@ufl.edu, University of Florida, USA; Yukyeong Song, y.song1@ufl.edu, University of Florida, USA; Amogh Mannekote, amogh.mannekote@ufl.edu, University of Florida, USA; Joanne Barrett, jrbarrett@ufl.edu, University of Florida, USA; Christine Fry Wise, christine@thefindingsgroup.org, The Findings Group, USA; Emily Dobar, emily@thefindingsgroup.org, The Findings Group, USA; Kristy Elizabeth Boyer, keboyer@ufl.edu, University of Florida, USA; Maya Israel, misrael@ufl.edu, University of Florida, USA.

Manuscript submitted to ACM

has shown many advantages including increased engagement [49], a deeper understanding of the concepts [27, 47], and enhanced AI literacy [93, 99]. Traditional assessment methods for these projects can be time-consuming and hard to scale, as teachers need to review the artifacts using a rubric. With the recent advancements in generative AI and LLMs, there is a growing potential and early attempts to use these technologies to automate artifact evaluation [19, 92]. This could reduce teachers' workload and provide students with timely and personalized feedback.

However, the adoption of generative AI and LLMs in K-12 classrooms depends on understanding the perspectives of key stakeholders: students and teachers. There are concerns that generative AI might undermine children's creativity, originality in their learning [14, 68] and create an over-reliance on AI-generated content without critical AI literacy [7]. Additionally, AI-generated content could include misinformation or biased and offensive material [94]. Within computer science education, it has shown varied perceptions and levels of trust among students and teachers [1, 43]. Research on adult trust in ChatGPT indicates that they might undertrust it compared to traditional sources like Google Search and Wikipedia [37, 95]. Teachers' trust in these technologies is particularly crucial for the acceptance and use of these applications in classrooms. While current research focuses on improving the accuracy of LLM-based evaluations to match expert grading [12, 57, 100], it is crucial to also consider the perspectives of teachers and students. Understanding their views on the trustworthiness, benefits, and challenges of these technologies is essential for designing effective learning interventions in K-12 classrooms.

In this paper, we explore both students' and teachers' perspectives on the trustworthiness, usefulness, and challenges of LLM-based assessment for student-created conversational agent artifacts in a formal learning context through two studies. In Study 1, we conducted focus group interviews with 30 sixth-grade middle school students (ages 11-12) who participated in a 10-hour science-integrated AI learning module and created conversational agents in their class. After their final projects were assessed using GPT-4, students reviewed the GPT-4-generated scoring and feedback for their projects and shared their thoughts on its accuracy and usefulness. In Study 2, we interviewed four in-service middle school teachers to gather their views on the LLM-based assessment. We compared the perceptions from students and teachers to identify similarities and differences in trust levels and preferences. This study aims to answer the following research questions (RQs):

- RQ1. How do middle school students perceive the trustworthiness and usefulness of LLM-based assessment on their conversational agent artifacts?
- RQ2. How do middle school teachers perceive the trustworthiness and usefulness of the LLM-based assessment for students' conversational agent artifacts?
- RQ3. What are the similarities and differences between students' and teachers' perceptions of the LLM-based assessment for their learning artifacts?

This paper makes the following contributions:

- An investigation of students' and teachers' perceptions on an LLM-based project assessment, from an empirical study in an authentic K-12 classroom.
- A comparison of similarities and differences between middle school students and teachers' perceptions on the trustworthiness, usefulness and needs for LLM-based assessment on computational artifacts.
- A set of design recommendations for effective learning intervention design to support LLM-based project assessment in K-12 classrooms.

2 Related Work

2.1 Project-based Learning and Assessment

Project-based learning (PBL) is a widely adopted instructional approach in STEM and AI education [49, 78, 103]. Rooted in Constructivist principles, PBL emphasizes learning by doing, engaging students in real-world projects that require active problem-solving, decision-making, and collaboration over extended periods [6, 50]. In CS education, PBL particularly aligns with the goals of computational thinking, encouraging students to develop coding, algorithmic thinking, and other practical skills through the creation and refinement of computational artifacts [83]. PBL has been used widely for enhancing students' AI literacy [93, 99]. This teaching method has demonstrated numerous benefits, such as increasing student engagement [49], stimulating students' creativity [22, 32], helping them understand complex concepts [27, 47] and eventually increasing their AI learning [41, 79, 86]. By working on their own projects, students demonstrate deep learning and high-level thinking [3], while teachers can use these projects to assess student progress and provide constructive feedback [97].

One common approach to assessing students' projects and learning is through rubric-based scoring. A rubric offers a description of different levels of performance for a task, outlining what constitutes mastery and the varying degrees of proficiency [53]. Over the past decade, substantial research has focused on the design, construction, rationale, and use of rubrics [29, 34, 51, 76]. In practice, rubrics offer significant value to both instructors and students by providing a clear and concise summary of performance expectations and reasoning for the evaluation. They help define what exemplary work should look like at the highest level, foster constructive learning and encourage students to self-evaluate [35]. Moreover, teachers utilize rubrics to deliver detailed feedback, which further supports student development [80].

Despite the benefits of PBL and the rubric-based evaluation method for PBL, a major challenge is that the manual project-scoring process is highly time-consuming and resource-intensive [27, 63]. This manual approach not only limits the scalability but also poses difficulties in maintaining consistency and objectivity in grading. It can also result in teachers shying away from the student-centered, inquiry-based PBL for more traditional teaching methods such as lectures and quizzes [16]. Furthermore, the delay in feedback due to the manual evaluation process can hinder the immediate reinforcement of learning, which is crucial for students to reflect on their work and make necessary improvements. With that, there is a growing need and interest in automating this process to enhance scalability so that teachers can provide timely, accurate, and formative feedback to support the learning process effectively [66, 85]. Consequently, automated tools that utilize techniques such as machine learning algorithms or rule-based systems, to evaluate student submissions based on predefined criteria have gained attention as a viable solution [67, 75]. These tools offer the potential to streamline grading and evaluation at scale.

2.2 AI and LLMs for Learning Assessment

As AI continues to expand, its role in education, particularly in tutoring and assessment, is coming increasingly significant [25]. There are two typical types of assessments in PBL: summative assessment and formative assessment. Summative assessments evaluate student learning at the end of an instructional period to measure overall achievement, whereas formative assessment involves ongoing feedback during instruction to improve both teaching and student learning [31]. A review of AI applications in student assessment revealed that AI is primarily used for formative assessment across different grade levels (secondary education, university) and subjects (math, English language teaching), mainly by automating the grading process [25]. For example, Kaila et al. [38] utilized active learning methods to develop various types of collaboration-based activities that are automatically assessed. Similarly, Goel and Joyner [23] created an AI-based system in their course for teaching AI, which enables students to automatically receive grades and quickly view their results along with feedback. Choi and McClenen [11] discussed their AI system performs both automatic grading feedback and adjusts subsequent tasks accordingly to enhance learning outcomes. AI-driven systems are also applied in summative assessments, although less commonly used than in formative ones. For instance, Yeung et al. [98] investigated AI-based essay grading tools for automating final essay evaluations by training algorithms on existing grades and testing them for accuracy. GMAT and GRE general test are the most well-known AI-assisted summative assessments that are used frequently [20].

The rapid advancement of LLMs and their application in education indicate even more promising opportunities for project assessments. Unlike traditional methods, LLMs use advanced generative models to simulate a deeper understanding of the content, making them particularly well-suited for providing nuanced feedback on complex and creative student work across various disciplines and tasks, including grading short answers [18, 101], assessing essays [62, 65], scoring the divergent thinking automatically [74], and short textual answers [81].

In addition to such applications that directly assist students' learning activities, research explored generative AI applications supporting teacher activities, such as instructional design [10] and assessment question generation [52]. However, their potential for assessing student-created projects in CS and AI learning contexts–where projects often involve both technical and creative aspects–has not yet been fully explored.

2.3 K-12 Students' and Teachers' Perception Toward AI Technologies

As generative AI technologies have increasingly advanced in recent years, more applications of such technologies are introduced and adopted in K-12 classrooms. One of the applications is in the form of pedagogical agents that assist in facilitating learning and enhancing engagement through contingent interaction, automatic grading, and assessments [102]. A recent body of literature suggests the potential positive impacts of generative AI-powered applications for teaching and learning. For example, the assistance of generative AI in students' writing classes has been reported to enhance

students' writing productivity and the quality of their writing outcomes [71, 84]. Similarly, generative AI has been used as a conversational partner for language learners, proving its positive impacts on students' language proficiency and confidence [13]. In addition, generative AI offers the opportunity to provide timely, personalized, and interactive feedback on students' learning and artifacts with highly human-like, naturalistic, and contextualized conversation [89]. Despite the potential benefits of generative AI in education, researchers and practitioners have raised concerns about using generative AI or LLMs in K-12 education settings without a critical review of ethical and pedagogical issues. Generative AI could undermine students' creativity and originality in their learning [14] and make students over-reliant on AI-generated content without critical AI literacy [7]. In addition, AI-generated content could include misinformation or biased and offensive content [94].

While the potential benefits and challenges of using generative AI in education have been widely discussed among researchers, it is paramount to examine the voices of educational stakeholders from authentic K-12 classrooms. Among the stakeholders, the voices of students and teachers are of the most interest as they are the direct users of educational technologies, and their perceptions could largely impact the adoption and acceptance of such technologies [1]. Therefore, there is an increasing interest in teachers' and students' perceptions of AI and generative AI. Kim et al. [44] explored middle school students' naive conceptions of AI in an AI summer camp. The common misconceptions from 14 middle school students were "(1) AI was the same as automation and robotics, (2) AI was a cure-all solution, (3) AI was created to be smart, (4) All data can be used by AI, and (5) AI had nothing to do with ethical considerations" (p. 1). Similarly, Belghith et al. [5] interviewed middle school student groups to elicit their interests, conceptualizations, and approaches to the prevalent generative AI tools. They highlight common conceptions and misconceptions about the definitions, mechanism, and applications of generative AI and suggest design considerations, such as 1) highlighting personally and culturally relevant topics, 2) expanding the scope of AI exploration, 3) leveraging anthropomorphism as an approach to understanding generative AI, 4) leveraging creativity as an approach to understanding generative AI. On the other hand, Antonenko and Abramowitz [2] examined in-service teachers' perceptions of AI in K-12 science education, finding three main themes of teachers' perceived roles of AI in K-12 education: "(1) teachers are overall enthusiastic about the potential of AI for K-12 education, (2) teachers believe it's important for their students to understand the basics of AI, and (3) teachers are overall not concerned or unsure about the ethics of K-12 education" (p. 70).

More recently, researchers have been specifically interested in students' and teachers' perceptions of generative AI technologies and LLMs. Chan and Hu [7] conducted a survey on 399 undergraduate and graduate students and reported their perceptions of generative AI to be positive overall, with a special interest in the potential for personalized learning support, writing and brainstorming assistance, as well as data analysis support. The respondents also reported concerns about generative AI, such as accuracy, privacy, and ethical issues. Another similar study by Obenza et al. [72] reported survey responses from 500 college students, revealing college students' high level of understanding of the strengths and weaknesses of AI and a strong willingness to use generative AI in their college study, along with a moderate level of concerns. In addition, Amoozadeh et al. [1] reported higher education students' trust in generative AI in the context of CS education, highlighting a varied level of trust in generative AI among participants. A closely related work by Kazemitabaar et al. [43] examined student perceptions of LLM-powered programming assistant that offers feedback without revealing the solutions in a semester-long classroom deployment with 700 students, which suggests that students perceived the assistance from the AI agent to be overall correct and helpful, accessible and convenient, and trustable and reliable, while showing concerns, such as skepticism or worrying about being over-reliant. In higher education, Park and Ahn [77] identified the promises and challenges of ChatGPT in terms of usability, user experiences, scalability (promises), algorithmic problems, human and social problems, and usability problems (challenges).

While most studies focus on the use of generative AI in higher education, very limited literature examined the students' and teachers' perceptions of generative AI in K-12 settings. One recent attempt was to explore middle school students' perceptions of a generative AI-powered teachable agent [87]. The study pointed to design considerations for generative AI-based technologies in K-12, including the need for integrating pedagogical theories and promoting effective teacher-AI collaboration. Another study explored elementary school students' and teachers' perceived opportunities and challenges of generative AI-assisted creative mathematical writing [88]. The perceived benefits of generative AI include fostering creativity, enhancing domain understanding, promoting AI literacy, and improving the affective domain. On the other hand, students and teachers found the barriers attributed to their lack of capacities, the inadequate design of the AI technology, and the lack of pedagogical support, of which a similar finding has been presented in Han et al. [30].

Despite the recent body of research that examines students' and teachers' perceptions of the use of generative AI in education, few studies focused on the use of generative AI as an assistant for scoring and providing feedback on students' projects or artifacts. As more applications of LLM-based assessment are created and adopted in K-12 classrooms, it is essential to examine the current perceptions of students and teachers about the trustworthiness, usefulness, potential benefits, and expected challenges of such applications.

Building on the body of literature on project-based learning, automatic assessment, and generative AI technologies in K-12 education, this paper addresses the research gap concerning key stakeholders' perspectives on LLM-based project assessment. Our novelty is having our student participants review and evaluate the LLM assessment of their own chatbot projects created in an authentic K-12 classroom environment. We contrast the perceptions of middle school students and teachers, identify discrepancies between them, and draw design implications to enhance the effectiveness of AI-driven assessment tools.

3 Methods

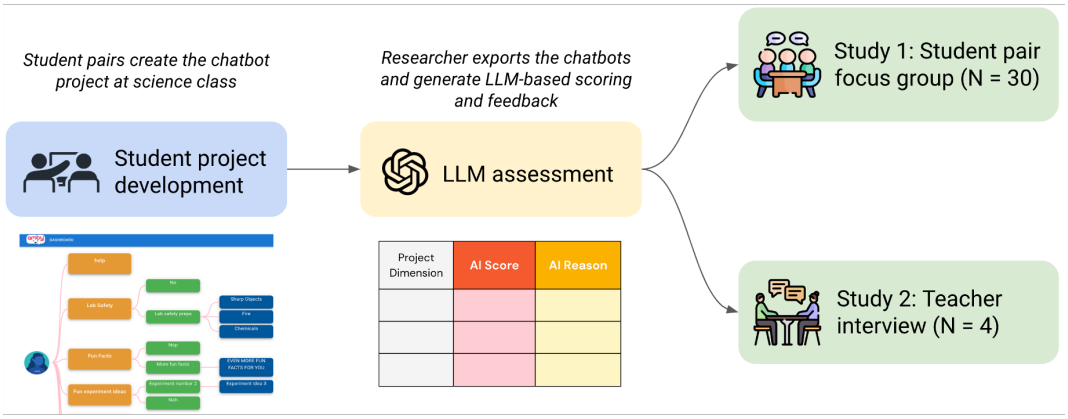

Fig. 1. Overview of Study

In this section, we describe our study methods, as illustrated in Figure 1. First, in Study 1, we implemented an "AI + science" learning module in a middle school science class, where students learned about artificial intelligence and developed AI chatbot projects. Once the projects were completed, their chatbots were evaluated by GPT-4 using a rubric (detailed in Section 3.1). During the classroom study, we conducted focus groups with 15 pairs of students (N=30) to present the LLM-generated scores and feedback for their chatbot projects and gather their perceptions (see Section 3.2.1). Insights from these student focus groups led us to seek a broader perspective by interviewing four science teachers in Study 2 (Section 3.2.2).

3.1 LLM-based Artifact Assessment Implementation

In this section, we describe the implementation of our LLM-based artifact assessment technique. We begin by developing a rubric with high inter-rater reliability to evaluate the chatbot. We then define our LLM prompting strategy and design a prompt template, which we iteratively refined and validated using a dataset of 75 previously collected chatbots. For the evaluation, we preprocess students' chatbot snapshots to extract relevant features such as intents, training phrases, and responses. We utilize GPT-4 [73] and an open-source framework, LLM4Qual [59] to manage the prompt templates and automatically evaluate the artifacts.

Artifact Evaluation Rubric Development. To ensure a systematic evaluation of student-created chatbots, we developed a rubric grounded in our AI curriculum and existing dialogue system evaluation frameworks [96]. The original rubric was collaboratively developed and reviewed by six experts (CS education, AI, software development, middle school classrooms, and educational evaluation) and included ten dimensions using a four-point grading scale. To establish IRR, two human graders independently assessed 40 chatbot artifacts previously collected from middle school summer camps. The resulting Quadratic Weighted Cohen's Kappa was 0.82 across all rubric dimensions, indicating substantial reliability [60].

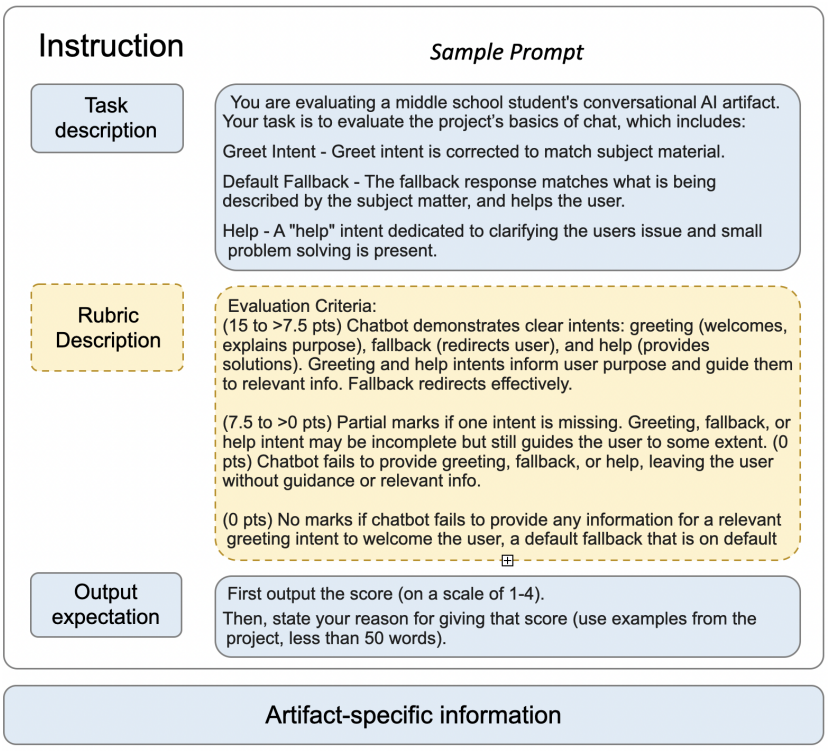

Fig. 2. The prompt template comprises four components: task description, rubric statement, output expectation and artifact-specific information

For this study, the rubric was adapted to better align with the grading practices and student needs in the middle school science classroom. These adaptations included collapsing the ten dimensions into five, transitioning to a 0-15 grading scale, and simplifying the language in the rubric for improved readability among middle school students. A complete description of this adapted rubric is in Appendix A.

LLM Prompting Strategy and Template. To develop the LLM-based assessment system, we employed a zero-shot prompting strategy with a description of the rubric within the prompt. Zero-shot LLM prompting is a popular strategy when there are no in-context examples available (in our case, graded student examples using the new classroom rubric have not yet been collected) [48]. Research suggests that rubric-integrated prompts can enhance accuracy and even produces results comparable to those of few-shot prompting strategies [92].

We adapted the zero-shot prompt template which comprises four components: 1) Task Description, which gives high-level background information about the artifact evaluation task; 2) Rubric Statement, which breaks down the range of the grading scale into parts depending on the artifact quality; 3) Output Expectation, which defines the expected format of the evaluation output being the score for the dimension and accompanying rationale; and 4) Artifact-Specific Information, which consists of specific components extracted from each chatbot snapshot under evaluation. An example of the prompt template is shown in Fig. 2. In this example prompt, which is used to evaluate the chatbot's "basics of chat" dimension, the artifact-specific information includes the textual content of the chatbot's greet intent, default fallback intent, and help intent, all processed from the original student-project chatbot.

Our evaluation of this prompt template using a dataset of 75 chatbot artifacts collected in previous summer camps showed moderate correlations between the LLM-predicted and human-rated scores (Spearman correlation coefficient ranging from 0.127 to 0.782 across five dimensions). While it is possible to improve the prediction accuracy, this paper does not rely on an accurate scoring because we sought to investigate human perceptions of the current, readily available LLM-based artifact assessments. The variable accuracy in scoring can provide us deeper insights into how students and teachers perceive the effectiveness and limitations of LLM-based evaluations.

3.2 Study Procedure

3.2.1 *Study 1: Middle School Classroom Study and Focus Group.*

Classroom Study Context. This study was conducted in a public middle school in the southeastern United States in Spring 2024. A total of 128 sixth-grade students participated in a 10-hour AI-integrated learning module as part of their regular science curriculum. The instructional hours were distributed over 10 days across four weeks, with one hour of class time per study day. The study has obtained ethical approval from the university's institutional review board (IRB). Parental consent and children assent were collected before the beginning of the study.

AI Learning Module. The learning module covers fundamentals concepts of AI and conversational agents in a science context. Over the initial four days, students learned key concepts and engaged in hands-on practice creating conversational agents using AMBY, a graphics-based conversational AI development environment [91] (Figure 3). For the subsequent four days, students worked in randomly assigned pairs to design and develop their own science-themed conversational agents. They also participated in a peer review session to provide feedback to other students and refine their projects. Students were provided with the rubric at the beginning of their project development to guide their work.

Student Project Assessment. After the students completed their projects, we exported their final chatbot snapshots. We then used GPT-4 to generate scores and feedback for each project using the prompt template and rubric detailed in Section 3.1. During the focus groups, we presented students with a printed copies of their chatbot's LLM-generated scoring and feedback (an example paper prototype of student chatbot assessment is in Appendix B.1).

Data Collection. On the final day, we conducted artifact-based focus groups with 15 student pairs (Student IDs are coded as S1 or S2 from G1-G15) to gather in-depth feedback. Given that the students collaborated in pairs on developing the artifact, it was methodologically appropriate for them to collectively review the LLM evaluation for their group project. To ensure the students were comfortable talking to the researchers during the interview, selection priority was given to those who had more personal interactions with the researchers (e.g., through asking questions during previous lessons, requesting for help with debugging). During these focus groups, we first introduced students the concept of LLM and how their projects were assessed by the LLM, then we showed them the LLM-generated assessment results, and asked them to share their perceptions of the accuracy, trustworthiness, and usefulness of the generated result.

Participants. Out of the 30 focus group participants, 29 students provided their demographic information: 16 identified as boys and 13 as girls. Racial/ethnic distribution was as follows: 11 Asian, 9 White, 6 Black or African American, 3 Hispanic or Latinx, and 3 self-described[1]. The average age of participants was 11.7 years (SD = 0.48).

3.2.2 *Study 2: Teacher Interview.*

The student responses from the study 1 (detailed in Section 4.1) inspired us to further investigate by interviewing teachers, as they are the key stakeholders in the successful implementation of AI-based assessments in the classroom.

Participant Recruitment. Ideally, the classroom teacher from Study 1, who possessed contextual knowledge of the student projects, would have been the primary participant for this study. However, due to the unavailability of this teacher, we identified a new group of teachers who were prepared with the relevant backgrounds and expressed interest in participating.

We recruited four middle school science teachers from a professional development workshop about AI education happening in the southeastern United States in Summer 2024. Invited by the workshop organizer, we gave a 30-minute guest lecture introducing our conversational AI middle school study and the features of the development environment, AMBY. During the workshop, we announced our teacher interview study and initially recruited five teachers (one dropped later due to schedule conflicts).

Data Collection. About a week after the workshop, we conducted remote one-on-one interviews with the four teacher participants. Each interview lasted 30-60 minutes and was video- and audio-recorded. The recordings were automatically transcribed by Zoom, and corrected manually by researchers. Table 1 presents the demographic and background information of the interviewed teachers.

[1]Participants may report multiple race/ethnic groups.

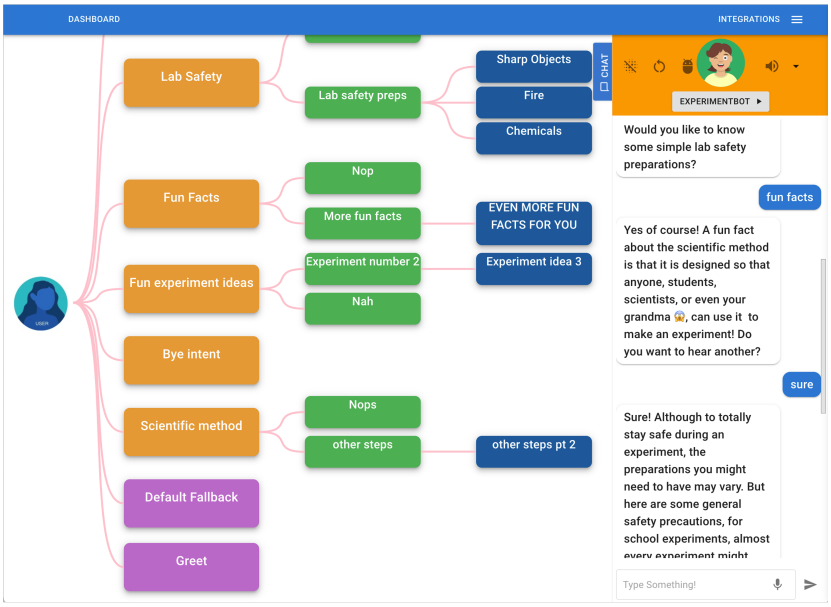

Fig. 3. AMBY chatbot development environment and an example chatbot artifact "ExperimentBot" students created during the class

During the interviews, we asked about the teachers' backgrounds, their experiences with project-based learning, and their familiarity with LLMs. Similar to the student focus groups, we introduced how our LLM-based assessment was generated, then presented two to three examples of LLM-generated assessments of student-created chatbots (same as we showed to students in study 1), along with descriptions of the relevant rubric dimensions (see Appendix B.2 for an example). After reviewing the assessment examples, teachers shared their perspectives on the trustworthiness, usefulness, and challenges of using AI-generated grading and feedback in their classrooms. At the end of the interview, participants completed a brief demographic questionnaire.

Participant Demographics. Our teacher participants all identified as women, and two identified their race/ethnicity as White, one as Asian, and one preferred not to disclose. Detailed information about their backgrounds and experiences is in Table 1.

Table 1. Teacher Demographics and LLM Experience

Teacher ID	Grade and Subject	Teaching Experience	Student Population	Experience with LLMs
T1	7th grade, science	25 years	Majority White; many English-as-second-language learners; 45% from low-income families, large autistic unit in school	Learned about LLMs in training; uses LLMs to modify reading levels; occasional use (3 times a year).
T2	7th/8th grade, biology/robotics	22 years	55% White, 33% Black, 31% Hispanic; over 40% students from low-income families	Recently learned about LLMs; Used once for lab instructions; excited to use more (e.g., quizzes).
T3	6th/7th grade, science	12 years	Majority White, some Black/Hispanic; many English-as-second-language learners	Never used LLMs; concerned about privacy.
T4	7th grade, life science	8 years	68% students from low-income families	Used ChatGPT for bell work; occasional use.

3.3 Qualitative Data Analysis

In this study, we report the qualitative findings from the student focus group responses and teacher interviews. Following the detailed guidance from Naeem et al. [64] and Lester et al. [54], we employed thematic analysis using an inductive approach (as opposed to deductive) for both data sources [8] because of the exploratory nature of the study.

For the analysis of student perceptions gathered from study 1, we utilized a qualitative coding approach as outlined by Naeem et al. [64]. One primary coder coded the responses across all questions, and a second coder reviewed these codes and applied secondary coding as necessary. The two coders then met to discuss their coding decisions and resolve any disagreements. In cases where disagreements cannot be resolved after discussion, both codes were retained.

Regarding teacher perceptions of the LLM-based assessment from study 2, we employed affinity diagramming methodology using Miro software[2]. Affinity diagramming is a common HCI research technique for bottom-up analysis of interview and observational data [24, 33]. To construct the affinity diagrams, the interview transcripts were first broken down into individual "notes" (short quotes) related to specific topics. These notes were then grouped and organized based on thematic similarities and themes derived from study 1. This thematic analysis process was conducted collaboratively with another researcher to ensure consensus on the interpretation and synthesized themes.

4 Results

In this section, we show the findings of the two studies and highlight the similarity and divergence of the two groups' results.

4.1 Student Perceptions

4.1.1 ***Overall Positive Reactions on LLM-generated Score and Reasoning.*** Most students generally agreed with the project score provided by the LLM. All students appreciated the comprehensive feedback and how relevant this explanation is about their chatbot.

> *"I'm actually really surprised on how good it was. I thought it would just come up with some random rubric and then grade it, but this actually gave a good detailed explanation on why it gave us that score."* (from S1, G7)

Students even went on to analyze the reasoning and explain why they get points off.

> "*I would think maybe 'topic' might be a little small. We did write our responses what's about biosphere was in the middle of being too much and too little. So I feel like it would've marked us down a bit.*" (from S1, G2).

Seven out of nine groups of students found the scores aligned with their expectations and considered them fair. In addition, they found the reasoning helpful as the detailed explanations from the LLM helped them identify areas for improvement. For example, students quoted:

> "*I like the reasoning it gives because it doesn't just put out a grade. I'd imagine if you got a five out of 10 on the last one, it could explain what you did wrong, how to fix it.*" (from S2, G10)
> "*the reasoning also helps us know why it was bad.*" (from S2, G12)

Some students felt the score was higher than they expected:

> *"Oh, we got a hundred. I was not expecting that."* (from S1, Group 11)
> *"Perfect. Wait, is it 15 out of 15? If 15 is the most you can get, oh, that's good."*(from S1, G14).

4.1.2 ***Student Mixed Perspectives and Trust on AI Grading.*** In terms of student trust over AI grading and feedback, we observed mixed perspectives. Despite 20 out of 30 students indicates they would trust AI for grading their projects, 12 students state that they would trust their teacher's grading more than the AI.

Reasons for Trusting AI. Several factors contributed to student confidence in AI grading:

(1) Detailed Explanations and Reduced Errors. The AI's ability to provide specific, reasoned feedback and its perceived objectivity resonated with students. They appreciated the clarity of explanations and the potential for comprehensive assessment.

[2]miro.com

"I like the reasoning it gives because it doesn't just put out a grade." (from S2, G10)

"*It says all of our follow-up intent and stuff and it says what we talked about and how our chatbot is specific with the answers and stuff.*" (from S2, G2).

Interestingly, students believe that LLMs such as ChatGPT are particularly knowledgeable in grading their AI chatbots because they share the same conversational AI nature:

"For this I think I would trust it because it's grading on another chatbot. So I think it would [better] understand the chatbot and understand all the intents and stuff and it will give a more accurate grade." (from S1, G1).

However, they also expressed misconceptions about AI of being smarter than them or less error-prone than teachers:

"*I think this AI is probably smarter than an average sixth grader. So I can't say anything with this.*" (from S2, G7)

"*It [AI] doesn't make as many mistakes on grading. Teachers can make mistakes. AI, very rarely.*" (from S2, G5).

(2) Rubric-Based Assessment and contextual data. Students noted the use of a pre-defined rubric and training data based on previously graded, student-created chatbots by the LLM increased their perceived fairness and reliability.

"*I would trust it because since it's a rubric based on other chatbots, it would compare human work with human work and based on that, we could be graded if our chatbot is good. So I would trust it. If the AI just create a random rubric, then obviously no one would trust it. But since it's based on other chatbots, then it would be trustworthy.* " (from S2, G3).

Reasons for Skepticism. Despite acknowledging the AI's strengths, 12 out of 30 students clearly expressed a preference for teacher grading, driven by several concerns:

(1) Accuracy. There is an interesting tension as for comparing students' concerns about AI accuracy with their previous statements praising AI's intelligence and reduced error-rate. While some students viewed AI as more objective and less error-prone, other questioned its general accuracy and ability to effectively extract information:

"Maybe since it (AI) scans really fast, they could have missed something important." (from S1, G7).

"I put my trust in it, but not that much because it can still get sometimes things wrong. It can overlook something sometimes." (from S2, G10).

Students also expressed uncertainty about the AI's broader knowledge base and whether it consistently applied the rubric in the context of their specific learning experiences.

"[Teacher name] he has the rubric and maybe the chatbot AI didn't use the rubric itself. It just gave [a grade] based on its own algorithm." (from S2, G6).

"[Teacher name] is better? I think so, only because it is a human. The AI can [be] wrong... the student will not [get] the true credit." (from S2, G9).

This contrast reflects student diverse perceptions regarding how AI works. Some believe that AI is more accurate than humans, while others know it can make mistakes, especially when processing complex information in their chatbot. This tension in their perception impacts their trust in it grading fairly.

(2) Lack of classroom contextual awareness. A major concern is the lack of awareness regarding specific concepts covered in class and individual student progress comparing to their teachers. As one student noted:

"I wouldn't really trust it because it doesn't know what we have learned in class. I trust the teacher more because he knows what concepts we went over and [will] grade by that." (from S2, G10).

4.1.3 ***Perceived Helpfulness of AI Feedback.*** The majority of students welcomed the idea of AI providing feedback during their project development, as long as it does not directly impact their grade:

"*If it didn't count towards our final grade, then I think it would be extremely useful [...] I think [feedback] would be good because then we basically guaranteed to get a hundred because it would help you with what do you need to improve on and [suggest] 'Oh, you missing this, you need to put this in, it'll help you.'* " (from S2, G1).

They view the AI-generated feedback similarly to a peer-review process where they get advice from a classmate:

"*It's kind of the same as having a peer review. They notice things that you don't quite notice and AI definitely has that information beforehand.*" (from S1, G1).

However, some students are concerned about potential threats to student agency and autonomy. One group of students worried that the AI might be overly directive and limiting their creative freedom and ownership over their project. They liked the idea of AI suggestions but also need to preserve their autonomy in the creative process.

> *"I think it might take away some of the freedom that you have in creating the chatbot. It might make you change some things when it doesn't need to be changed because the AI thinks it's bad. It might be helpful, but as long as it gives suggestions and not changes the chatbot."* (from S2, G13).

4.2 Teacher Perceptions

4.2.1 ***Mixed Reactions of the Artifact Assessment Example.*** During the focus group, we asked the teachers to review two chatbot grading examples randomly picked from our previous student artifacts. Through their review, teachers identified both pros and cons. They acknowledged the chatbot's capacity for providing detailed feedback and clear examples linked to specific rubric criteria, which could be valuable for student understanding. One teacher points out the possibility that AI scoring and feedback could facilitate better instruction. By automating the grading process for certain assignments, teachers could have more time and flexibility to address the individual needs of their students and foster deeper engagement in the classroom. However, they also raised concerns regarding some AI-generated content.

A common concern is the AI's focus on aspects such as the language style of the chatbot. For example, one student's chatbot was deducted points by the LLM because of "language appears to be too playful and informal." This may not align with a science teacher's priorities. Teachers point out that understanding and communicating scientific concepts is more important than stylistic fluency in a science classroom. T4 said:

> *"Perhaps with a language AI, It's very, very important to have the correct tone. But that's not really what I care about as a science teacher. I care about 'Did they understand the science?' So I'm not taking points away for necessarily being the the best chat."* (from T4)

There were also questions about consistency in evaluating different dimensions of the chatbot. Teachers also point out that the chatbot may not account for individual student effort, creativity, or important learning differences. Teachers acknowledge that grading a project is inherently subjective because their normal grading practices often take account for individual student needs such as creativity, effort in project development, language competencies, and individual students' background. Student might demonstrate some degree of understanding through the wrong answers. They believe AI cannot factor these in the generated scoring and assessment.

> *"Sometimes students might choose the wrong answer. But maybe they have a really good understanding of why they chose that answer. So in the past I give credit or partial credit when they're able to explain [...] their explanation might get them the credit. I think that is really valid, and I don't know if AI would be able to decipher that nuance."* (from T3)

Finally, simplifying the feedback language and providing a more detailed rubric were identified as key steps for enhancing clarity and transparency for students.

4.2.2 ***Teachers Do Not Trust the AI-generated Scores.*** Although all our interviewed teachers noted grading is one of their most time-consuming and tedious tasks, they expressed significant reservations about relying on AI for direct student assessment. All four teachers we interviewed pointed out that teachers take responsibility for students' grades and the teachers want control over their students' project grading. To not mislead the students with potential discrepancies in the AI-generated scores, they would need to personally verify the AI's output or even completely re-grade assignments, which might generate more work rather than alleviating their burden of grading. T3 said:

> *"Because I'm ultimately responsible for their education, but also their grade. I could see a parent calling me and wondering why their kid only got 5 points out of 15 points, and then I would say, 'Sorry. Let me go back and find that AI reason.' "* (from T3)

Teachers also note that the grading opportunity offers valuable insights on student misconceptions on the class material and their progresses. They were concerned about"*how do I know what my students are missing?*" which inform how they would adjust their teaching subsequently.

4.2.3 ***Teachers' Concerns.*** The teachers raised several potential challenges and concerns for AI grading. First is the need for AI to offer personalized scoring, considering aspects like creativity, effort, student prior knowledge, and individual learning backgrounds (e.g., English learner, students with disabilities). Teachers emphasize the importance of transparency and understanding how AI arrives at its evaluations, as they remain ultimately responsible for student grades and addressing

any discrepancies. **The consensus is that unless the AI demonstrates exceptional accuracy (e.g., 90%), the time spent deciphering and correcting its assessments outweighs any potential time-saving benefits.** Identifying student misconceptions and ability from the incorrect aspects are crucial for building teacher trust and ensuring fair assessment. Concerns also arise around the practicalities of implementation, including the effort required for data input and algorithm training, Learning Management System (LMS) integration, and ensuring student data privacy.

4.3 Student vs Teacher: Similarity and Differences

Table 2. Comparison of Student and Teacher Perspectives on LLM-based artifact assessments

Aspect	**Similarity**	
Detailed reasoning and feedback	Both groups recognized the value of AI's ability to provide detailed feedback linked to specific rubric criteria and deliver potentially unbiased assessments.	
Concerns about accuracy for direct scoring	Both acknowledged the importance of accuracy and the need for AI to understand the specific learning context, including classroom content and individual student progress.	
Role of AI grading in classrooms	Both believed AI cannot take over teacher's grading. Teachers emphasized their responsibility of student grading and needs control over it. Students stated they trust teacher's grading more even if the score is lower. AI also needs to be transparent about the assessment process in order to be trustworthy.	
Importance of human factors	In project-based learning, both students and teachers care about creativity and individual differences, this part AI cannot supplement teacher's role.	
Value in Formative Feedback	Both believed AI for offering feedback during project development would be extremely beneficial, both for teachers to save time and for students to receive feedback early to improve their projects.	
	Differences	
Aspect	**Students Perspective**	**Teachers Perspective**
Reactions on the chatbot assessment example	Mostly positive about the scores and reasoning	More critical about the generated content
Trust in AI-Generated Grades	Mixed trust: 20/30 trusted AI's judgment for grading, 12/30 preferred teacher grading. All agreed AI assessment is more trustworthy than assessment from a random adult outside of classroom.	Lower trust: All interviewed teachers expressed significant reservations about relying on AI for directly assigning grades.
Perceptions of AI Accuracy	Some viewed AI as more accurate and less error-prone than human, particularly in evaluating chatbots due to their shared AI nature.	More concerned about the accuracy, noting the importance of human judgment in assessing subjective aspects like creativity and student effort.
Practical Concerns	Concerned about potential threats to their own agency and autonomy, such as AI being overly directive in feedback.	Focused on potentially increased workload to safeguard the AI-generated content. Also noting logistical challenges related to implementation, data privacy and cross-platform integration.

We compare the student and teacher perspectives on LLM-based artifact assessment and summarize the differences and similarity in Table 2. Both students and teachers recognized the ability for LLMs to provide detailed, rubric-specific feedback for the student work. They shared concerns about the accuracy of LLMs to offer direct scoring, which could negatively impact student grades and their learning. Both acknowledged the need to factor in classroom and individual contexts and agreed that AI should not replace teacher-led grading but rather serve as a supportive tool. Especially in project-based learning, both parties value some degree of subjectivity, such as creativity and individual differences–they believe AI cannot fully replace. Both groups prefer using LLM-based assessment for formative feedback so that students can receive early and constructive feedback while alleviating teacher workload.

Despite these shared perspectives, they differed in several aspects. When student groups examined the LLM assessment of their chatbots, they reacted mostly positively, with a majority agreeing with the AI-generated scores and a few groups finding them higher than expected. Teachers, however, were more critical of the AI-generated content. They point to the inconsistencies in the criteria application across different chatbots and the tendency for the LLM to focus on language quality instead of the subject knowledge (e.g., science). This might lead to their varied trust levels in AI-generated grades. There is a mixed response from students–some reported trusting AI more than humans, while others preferred their teacher grading. Conversely, all interviewed teachers expressed strong reservations about relying solely on AI for grading, especially in evaluating subjective aspects such as creativity and effort. In terms of practical concerns, students worried about AI potentially undermining their autonomy, and teachers were concerned about increased workload to verify the AI-generated content, as well as broader important implications related to a school's policy regarding data privacy and learning management system integration.

5 Design Implications: Where and How AI-Based Assessment Can Be Useful in the Classroom

As we explore the potential of AI-based project assessments in K-12 classrooms, it is important to address the challenges and opportunities identified through our research. The findings inform our design implications and recommendations related to student over-trust, teacher under-trust, interface barriers, alignment with educational objectives, and the effective implementation of LLM-based assessments in the K-12 classroom environment.

5.1 Mitigating Students' Over-trust in AI feedback

From the student response findings, it appears that some students may over-trust AI-generated feedback due to misconceptions about AI. Some viewed AI as "*rarely making mistakes*" and considered it to be smarter or even more intelligent than humans. A study by Belghith et al. [5] highlights similar misconceptions about AI capabilities and operations, which echos the importance to enhance students' AI literacy regarding generative AI [58]. Addressing these misconceptions also involves making the algorithm and AI's decision-making process transparent. In the context of project assessments, this means showing students how the project data was processed and trained, and how the AI's assessment aligns with the teacher's rubric.

In addition to increasing AI literacy, specific interaction strategies may help mitigate users' over-trust [4, 61]. For instance, Geiskkovitch and Young [21] suggests that intentionally including errors in accuracy and responsiveness, along with certain error recovery strategies, may reduce trust in children and encourage more critical thinking about the system. In human-LLM interactions, Metzger et al. [61] found that the LLM's communication style (e.g., highly vs. lowly authoritative) can significantly influence user trust. For children, it is crucial to design effective communication styles and error-handling strategies to calibrate trust to an appropriate level.

5.2 Addressing Teachers' Hesitance Toward LLM Project Assessment

Assessment is a significant burden in the teaching profession, and while teachers generally welcome tools that assist with their tasks, grading remains a high-stakes and critical activity. Our findings indicate that although teachers appreciate the detailed reasoning and feedback provided by LLMs, they have reservations about trusting AI for grading student projects. Concerns about LLMs' accuracy, their ability to understand classroom context, and their sensitivity to individual student differences undermine their trust. Inaccurate AI outputs, particularly in critical tasks like grading, can discourage teachers from using the technology.

Drawing from the technology acceptance model (TAM) [15], teachers' beliefs about a technology's usefulness and ease of use significantly impact their intention to use it in the future. Teachers are likely to be more comfortable adopting LLM technologies if they perceive their previous experiences with LLMs as useful. Providing clear, reliable examples and demonstrating the effectiveness of LLMs in these less critical areas such as lesson planning or slide creation can help foster trust and facilitate their adoption in grading and other teaching activities. This gradual approach allows teachers to become familiar with the tools and develop trust before applying them to more critical tasks.

As teachers start exploring LLMs and new tools, it is important to acknowledge that they, like much of society, may have limited prior experience with LLMs. This lack of experience can hinder their willingness to adopt new technology. According to [90], teachers encounter significant information gaps and misconceptions when using ChatGPT for personalized learning

tasks and aligning it with diverse learner needs. This amplifies the issue of undertrust–where teachers are hesitant to fully embrace the technology–can be a significant barrier to effective LLM use in the classrooms.

5.3 Designing User-Friendly and Low-Barrier Experiences for Teachers

Another key consideration for LLM adoption is the user experience with these tools. Given that teachers often lack both time and professional development on tool usage, any tool considered must be flexible and user-friendly to teachers [69]. This includes straightforward navigation, from access and setup to clear output. For a project-based assessment tool, rubric implementation is critical. The tool should facilitate the customization of rubrics and allow for reusing the teacher-created rubrics [56]. Additionally, the system design must be adaptable and simple. During our interviews, the teacher expressed concerns about the effort required to "feed data into it" and "get the model ready to evaluate." Thus, an interface should support teachers to easily import their own rubrics and examples for training or fine-tuning the LLM to suit their classroom needs.

When using automatic grading tools, it is essential to keep teachers informed about student progress. Our teachers was concerned about potentially losing insights about their student learning to improve their subsequent teaching. Classroom visualization tools that offer dashboards on participant activities [17] or provide feedback to help teachers reflect on and improve their practices [36, 69] can mitigate this concern.

From our investigation, integrating automated scoring tools across learning platforms can significantly lower the barrier for teachers adopting these tools. Moving all grades to existing learning management (LMS) systems, such as Canvas, can be burdensome and may not necessarily save time compared to grading manually. Therefore, when designing LLM-based automated assessments, it is crucial to explore ways to integrate these tools with existing learning management systems. This integration can streamline workflows and make the transition smoother for educators.

5.4 Fitting AI-based assessments to Learning Objectives and Classroom Needs

In our chatbot scoring examples, some feedback generated by the LLM penalized students for their language style in projects be "too playful" or "informal." While language style may be relevant to AI learning objectives (our classroom learning context), it is not a learning goal for science education. This reflect that LLMs may penalize the scoring due to the literal application of the rubric. This issue also highlights a broader challenge: in many projects, the domain-specific learning objectives exist alongside the modality of the project. For effective LLM-based grading, it is crucial to ensure that the AI aligns with the learning objectives specific to the subject matter and the classroom needs, rather than incorporating irrelevant criteria. Clear guidelines and parameters must be established to prevent LLMs from applying inappropriate standards so that assessments accurately reflect the intended educational goals.

5.5 LLM as Formative Feedback, Not a Diagnostic Tool

One notable finding from our research is that both students and teachers highly value the formative feedback generated by LLMs. This suggests that LLMs may be more effective when used for formative feedback rather than for scoring in a classroom project-based learning. In classrooms with a high student-to-teacher ratio, timely feedback from instructors can be challenging to provide, which can lead to student disengagement. Implementing LLMs for feedback generation can provide timely assistance, enhance the classroom experience, and make it easier for teachers to integrate PBL into their practices. Offering early feedback is also beneficial to guide student learning, support revisions, encourage self-reflection, and foster student ownership of their learning before formal grading occurs.

To be effective, AI feedback needs to be tailored to a student audience: clear, concise, and presented in accessible language. Effective feedback should highlight specific areas for improvement, reference relevant standards or concepts, and provide supplemental resources for students to learn more. Balancing sufficient scaffolding with learner autonomy is crucial. Feedback should provide enough support to guide students while allowing them the space to apply their own creativity and critical thinking [70].

In terms of interaction style, our participants preferred having a button to request overall feedback from the chatbot. However, prior research indicates that students often prefer interactive, conversational AI feedback systems rather than receiving a "single-shot hint or bit of advice" [45]. Future research should explore different interaction styles for AI feedback in project-based learning to determine which methods best support student engagement and learning outcomes.

6 Limitations

One limitation for study 1 is that the student projects have not been personally graded by their teacher at the time we conduct the focus group, so we cannot directly compare our LLM-generated scoring and the teacher's scoring. Students' perceptions toward our LLM-generated score might be shifted after seeing the actual score given by their teacher.

Our limitation for study 2 is that the interviewed teachers were not the classroom teachers of the students from Study 1 and they had not implemented the conversational AI learning module in their classrooms or reviewed student-created conversational agents before. However, due to their familiarity with our curriculum and the learning environment through the guest lecture and AI literacy gained from the workshop, their perspectives would still be valuable for our investigation.

7 Conclusion and Future Work

In this paper, we investigated the perceptions of students and teachers regarding LLM-based assessments for student-created conversational agents in a classroom. Through thematic analysis of interviews with 30 sixth-grade students and four middle school teachers, we uncovered key insights into the effectiveness, trustworthiness, and challenges associated with LLM-based assessments. Our findings suggest that there is a gap between students' perceptions and trust and teachers' perceptions. LLMs might be more effective when used for formative feedback rather than final grading, as they can provide timely, actionable insights that enhance learning and support teaching practices. To improve adoption and effectiveness, it is crucial to address students' misconceptions about AI, support teachers with user-friendly tools, integrate AI assessments smoothly into existing systems, and ensure it is in aligned with learning objectives. Overall, AI-based assessments have the potential to significantly enhance the educational experience by providing meaningful feedback and supporting both student and teacher needs.

As the first step to investigate the potential to implement LLM technologies in K-12 classrooms at scale, this research highlights several directions for future work. First, investigating how LLM-based feedback influences learning outcomes in specific subject areas or skill sets will help determine its impact on different types of learning. Additionally, research should examine the effects of various feedback modalities, such as textual, auditory, and interactive, on student engagement and retention to identify which formats are most effective. Understanding learners' perceptions of the credibility and usefulness of LLM-based feedback is also crucial, as these perceptions can significantly impact their learning behaviors and actions. Finally, identifying the training and support educators need to effectively integrate LLM-based feedback into their teaching practices will be essential for maximizing the potential of these tools and ensuring their successful adoption in educational settings.

Acknowledgments

We thank the members of Project DIALOGS, including Tom McKlin and Carly Solomon for their feedback on the rubric development, as well as Oluwatomisin Obajemu for his careful copy-editing to the manuscript. We also thank all the student and teacher participants for elaborating their thoughts and allowing us to collect data. This research is supported by the United States National Science Foundation under grant DRL-2048480. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Matin Amoozadeh, David Daniels, Daye Nam, Aayush Kumar, Stella Chen, Michael Hilton, Sruti Srinivasa Ragavan, and Mohammad Amin Alipour. 2024. Trust in Generative AI among students: An exploratory study. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 67–73.

[2] Pavlo Antonenko and Brian Abramowitz. 2023. In-service teachers'(mis) conceptions of artificial intelligence in K-12 science education. *Journal of Research on Technology in Education* 55, 1 (2023), 64–78.

[3] Patricia Armstrong. 2010. Bloom's taxonomy. *Vanderbilt University Center for Teaching* (2010), 1–3.

[4] Alexander M Aroyo, Jan De Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, et al. 2021. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 423–436.

[5] Yasmine Belghith, Atefeh Mahdavi Goloujeh, Brian Magerko, Duri Long, Tom Mcklin, and Jessica Roberts. 2024. Testing, Socializing, Exploring: Characterizing Middle Schoolers' Approaches to and Conceptions of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[6] Stephanie Bell. 2010. Project-based learning for the 21st century: Skills for the future. *The clearing house* 83, 2 (2010), 39–43.

[7] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 43.

[8] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* sage.

[9] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[10] Gi Woong Choi, Soo Hyeon Kim, Daeyeoul Lee, and Jewoong Moon. 2024. Utilizing Generative AI for Instructional Design: Exploring Strengths, Weaknesses, Opportunities, and Threats. *TechTrends* (2024), 1–13.

[11] Younyoung Choi and Cayce McClenen. 2020. Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences* 10, 22 (2020), 8196.

[12] Keith Cochran, Clayton Cohn, Jean Francois Rouet, and Peter Hastings. 2023. Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation. In *International Conference on Artificial Intelligence in Education*. Springer, 217–228.

[13] Edwin Creely. 2024. Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges. *International Journal of Changes in Education* (2024).

[14] Yun Dai, Ang Liu, and Cher Ping Lim. 2023. Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP* 119 (2023), 84–90.

[15] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[16] Vicki Dobbs. 2008. *Comparing student achievement in the problem-based learning classroom and traditional teaching methods classroom.* Ph. D. Dissertation. Walden University.

[17] Gloria Fernandez-Nieto, Pengcheng An, Jian Zhao, Simon Buckingham Shum, and Roberto Martinez-Maldonado. 2022. Classroom dandelions: Visualising participant position, trajectory and body orientation augments teachers' sensemaking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[18] Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto, and Kentaro Inui. 2023. Reducing the Cost: Cross-Prompt Pre-finetuning for Short Answer Scoring. In *International Conference on Artificial Intelligence in Education*. Springer, 78–89.

[19] Rujun Gao, Hillary E Merzdorf, Saira Anwar, M Cynthia Hipwell, and Arun Srinivasa. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence* (2024), 100206.

[20] John Gardner, Michael O'Leary, and Li Yuan. 2021. Artificial intelligence in educational assessment:'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning* 37, 5 (2021), 1207–1216.

[21] Denise Y Geiskkovitch and James E Young. 2023. Trust Calibration Through Intentional Errors: Designing Robot Errors to Decrease Children's Trust Towards Robots. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1402–1406.

[22] Murat Genc. 2015. The project-based learning approach in environmental education. *International Research in Geographical and Environmental Education* 24, 2 (2015), 105–117.

[23] Ashok K Goel and David A Joyner. 2017. Using AI to teach AI: Lessons from an online AI class. *Ai Magazine* 38, 2 (2017), 48–59.

[24] Ariel Goldman, Cindy Espinosa, Shivani Patel, Francesca Cavuoti, Jade Chen, Alexandra Cheng, Sabrina Meng, Aditi Patil, Lydia B Chilton, and Sarah Morrison-Smith. 2022. Quad: Deep-learning assisted qualitative data analysis with affinity diagrams. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[25] Víctor González-Calatayud, Paz Prendes-Espinosa, and Rosabel Roig-Vila. 2021. Artificial intelligence for student assessment: A systematic review. *Applied sciences* 11, 12 (2021), 5467.

[26] Michael M Grant and Robert Maribe Branch. 2005. Project-based learning in a middle school: Tracing abilities through the artifacts of learning. *Journal of Research on technology in Education* 38, 1 (2005), 65–98.

[27] Pengyue Guo, Nadira Saab, Lysanne S Post, and Wilfried Admiraal. 2020. A review of project-based learning in higher education: Student outcomes and measures. *International journal of educational research* 102 (2020), 101586.

[28] Zeliha Zuhal Guven. 2014. Project based learning: A constructive way toward learner autonomy. *International Journal of Languages' Education and Teaching* 2, 3 (2014), 182–193.

[29] John Hafner and Patti Hafner. 2003. Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. *Int. J. Sci. Educ.* 25, 12 (2003), 1509–1528.

[30] Ariel Han, Xiaofei Zhou, Zhenyao Cai, Shenshen Han, Richard Ko, Seth Corrigan, and Kylie A Peppler. 2024. Teachers, Parents, and Students' perspectives on Integrating Generative AI into Elementary Literacy Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[31] Wynne Harlen and Mary James. 1997. Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in education: Principles, policy & practice* 4, 3 (1997), 365–379.

[32] Pedro Hernández-Ramos and Susan De La Paz. 2009. Learning history in middle school by designing multimedia in a project-based learning experience. *Journal of Research on Technology in Education* 42, 2 (2009), 151–173.

[33] Karen Holtzblatt and Hugh Beyer. 2016. *Contextual design: Design for life.* Morgan Kaufmann.

[34] Hiroshi Ito. 2015. Is a rubric worth the time and effort? Conditions for its success. *International Journal of Learning, Teaching and Educational Research* 10, 2 (2015).

[35] Gerriet Janssen, Valerie Meier, and Jonathan Trace. 2015. Building a better rubric: Mixed methods rubric revision. *Assessing writing* 26 (2015), 51–66.

[36] Emily Jensen, Meghan Dale, Patrick J Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K D'Mello. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[37] Yongnam Jung, Cheng Chen, Eunchae Jang, and S Shyam Sundar. 2024. Do We Trust ChatGPT as much as Google Search and Wikipedia?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.

[38] Erkki Kaila, Einari Kurvinen, Erno Lokkila, and Mikko-Jussi Laakso. 2016. Redesigning an object-oriented programming course. *ACM Transactions on Computing Education (TOCE)* 16, 4 (2016), 1–21.

[39] Sukran Karaosmanoglu, Elisabeth L Fittschen, Hande Eyicalis, David Kraus, Henrik Nickelmann, Anna Tomko, and Frank Steinicke. 2024. Language of Zelda: Facilitating Language Learning Practices Using ChatGPT. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–5.

[40] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[41] Gloria Ashiya Katuka, Yvonika Auguste, Yukyeong Song, Xiaoyi Tian, Amit Kumar, Mehmet Celepkolu, Kristy Elizabeth Boyer, Joanne Barrett, Maya Israel, and Tom McKlin. 2023. A Summer Camp Experience to Engage Middle School Learners in AI through Conversational App Development. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 813–819.

[42] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI code generators on supporting novice learners in introductory programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.

[43] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[44] Keunjae Kim, Kyungbin Kwon, Anne Ottenbreit-Leftwich, Haesol Bae, and Krista Glazewski. 2023. Exploring middle school students' common naive conceptions of Artificial Intelligence concepts, and the evolution of these ideas. *Education and Information Technologies* 28, 8 (2023), 9827–9854.

[45] Bailey Kimmel, Austin Lee Geisert, Lily Yaro, Brendan Gipson, Ronald Taylor Hotchkiss, Sidney Kwame Osae-Asante, Hunter Vaught, Grant Wininger, and Chase Yamaguchi. 2024. Enhancing Programming Error Messages in Real Time with Generative AI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[46] Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024. " If the Machine Is As Good As Me, Then What Use Am I?"–How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[47] Joyce Hwee Ling Koh, Susan C Herring, and Khe Foon Hew. 2010. Project-based learning and student knowledge construction during asynchronous online discussion. *The Internet and Higher Education* 13, 4 (2010), 284–291.

[48] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[49] Dimitra Kokotsaki, Victoria Menzies, and Andy Wiggins. 2016. Project-based learning: A review of the literature. *Improving schools* 19, 3 (2016), 267–277.

[50] Joseph S Krajcik and Phyllis C Blumenfeld. 2006. *Project-based learning*. na.

[51] T. Larkin. 2015. A Rubric to Enrich Student Writing and Understanding. *International Journal of Engineering Pedagogy* 5, 2 (2015), 12–19. https://www.learntechlib.org/p/207461/ Retrieved August 26, 2024.

[52] Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. 2023. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies* (2023), 1–33.

[53] Angela Leonhardt. 2005. Using rubrics as an assessment tool in your classroom. *General Music Today* 19, 1 (2005), 10–16.

[54] Jessica Nina Lester, Yonjoo Cho, and Chad R Lochmiller. 2020. Learning to do qualitative data analysis: A starting point. *Human resource development review* 19, 1 (2020), 94–106.

[55] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.

[56] Ally Limke, Marnie Hill, Veronica Cateté, and Tiffany Barnes. 2024. A Survey of K-12 Teacher Needs for an Online Programming Learning System. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[57] Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. 2022. Open-ended knowledge tracing for computer science education. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

[58] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[59] Amogh Mannekote. 2024. LLM4Qual. https://github.com/msamogh/llm4qual

[60] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[61] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-) Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[62] Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* 2, 2 (2023), 100050.

[63] Chrystalla Mouza, Alison Marzocchi, Yi-Cheng Pan, and Lori Pollock. 2016. Development, implementation, and outcomes of an equitable computer science after-school program: Findings from middle-school students. *Journal of Research on Technology in Education* 48, 2 (2016), 84–104.

[64] Muhammad Naeem, Wilson Ozuem, Kerry Howell, and Silvia Ranfagni. 2023. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods* 22 (2023), 16094069231205789.

[65] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 394–403.

[66] Nilupulee Nathawitharana, Qing Huang, Kok-Leong Ong, Peter Vitartas, Madhura Jayaratne, Damminda Alahakoon, Sarah Midford, Aleks Michalewicz, Gillian Sullivan Mort, Tanvir Ahmed, et al. 2017. Towards next generation rubrics: An automated assignment feedback system. *Australasian Journal of Information Systems* 21 (2017).

[67] Sidhidatri Nayak, Reshu Agarwal, and Sunil Kumar Khatri. 2022. Automated Assessment Tools for grading of programming Assignments: A review. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–4.

[68] Michele Newman, Kaiwen Sun, Ilena B Dalla Gasperina, Grace Y Shin, Matthew Kyle Pedraja, Ritesh Kanchi, Maia B Song, Rannie Li, Jin Ha Lee, and Jason Yip. 2024. " I want it to talk like Darth Vader": Helping Children Construct Creative Self-Efficacy with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[69] Tricia J Ngoon, S Sushil, Angela EB Stewart, Ung-Sang Lee, Saranya Venkatraman, Neil Thawani, Prasenjit Mitra, Sherice Clarke, John Zimmerman, and Amy Ogan. 2024. ClassInSight: Designing Conversation Support Tools to Visualize Classroom Discussion for Personalized Teacher Professional Development. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[70] Le Thi Cam Nguyen and Yongqi Gu. 2013. Strategy-based instruction: A learner-focused approach to developing learner autonomy. *Language Teaching Research* 17, 1 (2013), 9–30.

[71] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192.

[72] Brandon Obenza, Alexa Salvahan, Alexandra Nicole Rios, Althea Solo, Rea Ashlee Alburo, and Rey Jose Gabila. 2024. University Students' Perception and Use of ChatGPT: Generative Artificial Intelligence (AI) in Higher Education. *International Journal of Human Computing Studies* 5, 12 (2024), 5–18.

[73] OpenAI. 2024. OpenAI. https://www.openai.com Accessed: 2024-09-11.

[74] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49 (2023), 101356.

[75] José Carlos Paiva, José Paulo Leal, and Álvaro Figueira. 2022. Automated assessment in computer science education: A state-of-the-art review. *ACM Transactions on Computing Education (TOCE)* 22, 3 (2022), 1–40.

[76] Ernesto Panadero and Margarida Romero. 2014. To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice* 21, 2 (2014), 133–148.

[77] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[78] Robert Pucher and Martin Lehner. 2011. Project based learning in computer science–a review of more than 500 projects. *Procedia-Social and Behavioral Sciences* 29 (2011), 1561–1566.

[79] Ingrid Russell, Susan Coleman, and Zdravko Markov. 2011. A contextualized project-based approach for improving student engagement and learning in AI courses. In *Annual Conference on Innovation and Technology in Computer Science Education*. https://api.semanticscholar.org/CorpusID:27862535

[80] María Consuelo Sáiz-Manzanares, Isidoro Iván Cuesta Segura, Jesús Manuel Alegre Calderón, and Lorena Peñacoba Antona. 2017. Effects of different types of rubric-based feedback on learning outcomes. In *Frontiers in Education*, Vol. 2. Frontiers Media SA, 34.

[81] Johannes Schneider, Bernd Schenk, Christina Niklaus, and Michaelis Vlachos. 2023. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508* (2023).

[82] Antonette Shibani, Simon Knight, Kirsty Kitto, Ajanie Karunanayake, and Simon Buckingham Shum. 2024. Untangling Critical Interaction with AI in Students' Written Assessment. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[83] Namsoo Shin, Jonathan Bowers, Joseph Krajcik, and Daniel Damelin. 2021. Promoting computational thinking through project-based learning. *Disciplinary and Interdisciplinary Science Education Research* 3 (2021), 1–15.

[84] Jovan Shopovski. 2024. Generative Artificial Intelligence, AI for Scientific Writing: A Literature Review. (2024).

[85] Tamara Sladoljev-Agejev and Jan Šnajder. 2017. Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in l2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 181–186.

[86] Yukyeong Song, Gloria Ashiya Katuka, Joanne Barrett, Xiaoyi Tian, Amit Kumar, Tom McKlin, Mehmet Celepkolu, Kristy Elizabeth Boyer, and Maya Israel. 2023. AI Made By Youth: A Conversational AI Curriculum for Middle School Summer Camps. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Innovative Applications of Artificial Intelligence Conference and Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence*.

[87] Yukyeong Song, Jinhee Kim, Zifeng Liu, Chenglu Li, and Wanli Xing. 2024. Students' Perceived Roles, Opportunities, and Challenges of a Generative AI-powered Teachable Agent: A Case of Middle School Math Class. *arXiv preprint arXiv:2409.06721* (2024). https://arxiv.org/abs/2409.06721

[88] Yukyeong Song, Jinhee Kim, Wanli Xing, Zifeng Liu, Chenglu Li, and Hyunju Oh. 2024. Elementary School Students' and Teachers' Perceptions Towards Creative Mathematical Writing with Generative AI. *arXiv preprint arXiv:2409.06723* (2024). https://arxiv.org/abs/2409.06723

[89] Yuan Sun, Eunchae Jang, Fenglong Ma, and Ting Wang. 2024. Generative AI in the Wild: Prospects, Challenges, and Strategies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[90] Mei Tan and Hari Subramonyam. 2024. More than model documentation: uncovering teachers' bespoke information needs for informed classroom integration of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[91] Xiaoyi Tian, Amit Kumar, Carly E Solomon, Kaceja D Calder, Gloria Ashiya Katuka, Yukyeong Song, Mehmet Celepkolu, Lydia Pezzullo, Joanne Barrett, Kristy Elizabeth Boyer, and Israel Maya. 2023. AMBY: A Development Environment for Youth to Create Conversational Agents. *International Journal of Child-Computer Interaction* 38 (2023), 100618. https://doi.org/10.1016/j.ijcci.2023.100618

[92] Xiaoyi Tian, Amogh Mannekote, Carly E. Solomon, Yukyeong Song, Christine Fry Wise, Tom Mcklin, Joanne Barrett, Kristy Elizabeth Boyer, and Maya Israel. 2024. Examining LLM Prompting Strategies for Automatic Evaluation of Learner-Created Computational Artifacts. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*. 1–4. In press.

[93] Manuel Vargas, Tabita Nunez, Miguel Alfaro, Guillermo Fuertes, Sebastian Gutierrez, Rodrigo Ternero, Jorge Sabattin, Leonardo Banguera, Claudia Duran, and Maria Alejandra Peralta. 2020. A project based learning approach for teaching artificial intelligence to undergraduate students. *Int. J. Eng. Educ* 36, 6 (2020), 1773–1782.

[94] Stéphan Vincent-Lancrin and Reyer van der Vlies. 2020. Trustworthy Artificial Intelligence (AI) in Education: Promises and Challenges. OECD Education Working Papers, No. 218. *OECD Publishing* (2020).

[95] Francesco Walker, Matteo Favetta, Linde Hasker, and Richard Walker. 2024. They Prefer Humans! Experimental Measurement of Student Trust in ChatGPT. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[96] Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering* 6, 3-4 (2000), 363–377.

[97] Michelle Hoda Wilkerson-Jerde. 2014. Construction, categorization, and consensus: Student generated computational artifacts as a context for disciplinary reflection. *Educational Technology Research and Development* 62 (2014), 99–121.

[98] WE Yeung, Cong Qi, JL Xiao, and FR Wong. 2023. Evaluating the effectiveness of AI-based essay grading tools in the summative assessment of higher education. In *ICERI2023 Proceedings*. IATED, 8069–8073.

[99] Iris Heung Yue Yim and Jiahong Su. 2024. Artificial intelligence (AI) learning tools in K-12 education: A scoping review. *Journal of Computers in Education* (2024), 1–39.

[100] Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic Short Math Answer Grading via In-context Meta-learning. In *Proceedings of the 15th International Conference on Educational Data Mining*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 122–132. https://doi.org/10.5281/zenodo.6853032

[101] Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic short math answer grading via in-context meta-learning. *arXiv preprint arXiv:2205.15219* (2022).

[102] Shan Zhang, Chris Davis Jaldi, Noah L Schroeder, and Jessica R Gladstone. 2024. Pedagogical agents in K-12 education: a scoping review. *Journal of Research on Technology in Education* (2024), 1–28.

[103] Chengbo Zheng, Kangyu Yuan, Bingcan Guo, Reza Hadi Mogavi, Zhenhui Peng, Shuai Ma, and Xiaojuan Ma. 2024. Charting the Future of AI in Project-Based Learning: A Co-Design Exploration with Students. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

A Full Description of the rubric

Criteria	Explanation	Full Marks (15 to >7.5 pts)	Partial Marks (7.5 to >0 pts)	No Marks (0 pts)
Project Ideation Demonstrates Purpose	The students show clear expectations of what the bot will be able to answer, navigate issues, and help the user to express knowledge about concept.	(15 to >7.5 pts) Full Marks: The Chatbot is able to demonstrate: State the information the Chatbot can provide. Directs the user to relevant content buzz words. Helps to shift the user back to original information in the event of unknown content.	**(7.5 to >0 pts) Partial Marks: The Chatbot cannot demonstrate the following: State the information the chatbot can provide. Directs the user to relevant content buzz words. Helps to shift the user back to original information in the event of unknown content**	No Marks (0 pts) The ChatBot is completely unable to discuss information with the user, demonstrate understanding of the intent, or share what it can do.
Main Topic Intents	Overall Intents - how material is set up in the AI Main Intents - Subject matter is split appropriately into the different intent sections.	(15 to >7.5 pts) Full Marks: Information in the data for the chatbots is clear, organized into distinct parts, and can be followed. The topic is noted in the overall intents, with the main topics split into clear areas of interest.	(7.5 to >0 pts) Partial Marks: Information in the data for the chatbots fails to complete one of the following. Organized into distinct parts, and can be followed. The topic is noted in the overall intents, with the main topics split into clear areas of interest.	No Marks (0 pts) Information in the data for the chatbots falls into a single pattern, information is not displayed easily, and other programmers would be unable to locate specific information.
Basics of Chat	The ChatBot will be able to full demonstrate the following: Greet Intent - Greet intent is corrected to match subject material. Default Fallback - The fallback response matches what is being described by the subject matter, and helps the user. Help - A "help" intent dedicated to clarifying the users issue and small problem solving is present.	(15 to >7.5 pts) Full Marks: Chatbot will be able to demonstrate the following clearly when presented with general use questions. A greet intent that welcomes the user, explains the purpose of the chat bot, and relevant information. The default fallback response helps to redirect the user towards information the bot is able to define. A help intent is included in the event the user struggles or has issues with the chatbot, allowing the bot to provide simple suggestions to correct issues. Greet intent expresses the purpose of the chatbot, and information it can share. Default feedback appears and helps the user to redirect to purpose / information available. Help - there is information about how the Chatbot works or quick solutions to redirect the user to the relevant user.	(7.5 to >0 pts) Partial Marks: Chatbot is missing one of following intents. A greet intent that welcomes the user, explains the purpose of the chat bot, and relevant information. The default fallback response helps to redirect the user towards information the bot is able to define. A help intent is included in the event the user struggles or has issues with the chatbot, allowing the bot to provide simple suggestions to correct issues. Great intent expresses the purpose of the chatbot, and information it can share. Default feedback appears and helps the user to redirect to purpose / information available. Help - there is information about how the Chatbot works or quick solutions to redirect the user to the relevant user.	No Marks (0 pts) Chatbot fails to provide any information for a relevant greeting intent to welcome the user, a default fallback that is on default mode, and no Presence of any help / suggestions available.
Follow Up Intents	All intents have at least one of the following subjects clearly added to a main attempt. Follow up that helps to explain, redirect, or continues the flow of the subject matter appropriately.	(5 to >0 pts) Full Marks: The main intents have at least two of the following to receive full content: The intent demonstrates knowledge of the content. The intent redirects the user to other information in a conversational way. There is an intent that allows you to dive deeper into the topic.	No Marks (0 pts) Follow up intents fail to include 2 of the following: The intent demonstrates knowledge of the content. The intent redirects the user to other information in a conversational way. There is an intent that allows you to dive deeper into the topic.	
Topic [Science]	The chatbot demonstrates sufficient knowledge about the scientific topic being presented. Overall, 70% of the chatbots information and function demonstrates adequate knowledge about the scientific topic and helps the user to learn more about it.	10 to >0 pts Full Marks: The chatbot is focused on a scientific topic, ideally one from the list of examples OR was pre-approved by the teacher. Depending on the amount of knowledge provided, the score will reflect that. A minimum of 70% is required for full marks.	No Marks (0 pts) The Chatbot fails to or in no way is focused on a scientific topic. Other reasons for 0 points are that the bot talks about an unapproved topic or is conversationally unhelpful in learning about the topic.	

B LLM-generated chatbot assessment example

B.1 Chatbot assessment example showed to students during study 1

	Score	AI reason
Basics of Chat (0-15 pts)	12	The chatbot gets 12 out of 15 points. The greet intent is present and it introduces the chatbot's purpose, but the language is too playful and might confuse users. The default fallback is also present and it redirects users back to the topic of the spheres of the earth. The help intent is included and it explains what the chatbot can do. However, the language and tone used throughout the chatbot's responses are overly playful and could be confusing, hence the deduction of 3 points.
Main topic intents (0-15 pts)	12	"The chatbot's main topic is about the spheres of the Earth, which is clearly divided into distinct parts: Geosphere, Hydrosphere, Atmosphere, Biosphere, and Cryosphere. Each sphere has its own intent with training phrases and responses. The chatbot also includes other intents like Greeting, Help, and Thanks. However, the language used in the responses is a bit informal and might be confusing for some users. The Moon intent seems out of place and doesn't provide any useful information. Score: 12/15"
Topic [Science] (0-10 pts)	10	"The chatbot focuses on the scientific topic of the ""Spheres of the Earth"", which is one of the approved topics. It provides information about the Geosphere, Hydrosphere, Atmosphere, Biosphere, and Cryosphere. The chatbot uses a playful and engaging tone, which could be appealing to some users. However, the language and tone might be confusing for others. The chatbot covers the topic well and provides detailed information about each sphere. Score: 10 points."

B.2 Chatbot assessment example showed to teachers during study 2

Criteria	Explanation	Score	AI reason
Basics of Chat (0-15 pts)	The ChatBot will be able to full demonstrate the following: Greet Intent - Greet intent is corrected to match subject material. Default Fallback - The fallback response matches what is being described by the subject matter, and helps the user. Help - A "help" intent dedicated to clarifying the users issue and small problem solving is present.	10	The chatbot beta01_BingBong has a clear greet intent that welcomes the user and explains its purpose. The default fallback response is also well designed, redirecting the user towards the bot's capabilities. However, there is no dedicated help intent present to assist users if they encounter issues. Therefore, I would give this chatbot 10 out of 15 points.
Main topic intents (0-15 pts)	Overall Intents - how material is set up in the AI Main Intents - Subject matter is split appropriately into the different intent sections.	15	The chatbot is well-organized and easy to follow. The main topic is about living and non-living organisms, which is clearly stated in the overall intents. The main topics are divided into distinct parts such as ""Greet"", ""facts about living organisms"", ""facts about nonliving things"", and ""Default Fallback"". Each part has its own training phrases and responses, making it easy for other programmers to locate specific information. The student has done a good job of splitting the subject matter appropriately into different intent sections. Score: 15/15.
Topic [Science] (0-10 pts)	The Chatbot demonstrates sufficient knowledge about the scientific topic being presented. Overall, 70% of the chatbots information and function demonstrates adequate knowledge about the scientific topic and helps the user to learn more about it.	8	The chatbot focuses on the scientific topic of living and non-living organisms, which is related to the suggested topic ""Characteristics of living things"". It provides detailed information about the characteristics of living organisms using the acronym ""Mrs. Gren"" and also explains about non-living things. However, the examples of non-living ""animals"" like plastic, gold, etc. are incorrect as they are not animals. The chatbot is mostly helpful in learning about the topic. Score: 8 out of 10.

C Interview Questions for Study 1 Student Focus Group

Now, I am going to ask some questions about scoring your chatbot project.

(1) Have you heard about Large Language models like ChatGPT?
(2) How familiar are you with them? Prompt: used it regularly? Occasionally? Maybe used once or twice? Or never used?

These AI applications can do a lot for us. For example, when you submit your homework or project, usually your teacher will have a rubric and they will grade your homework based on the rubric, and your teacher will give you a score, right? Nowadays, AI applications like ChatGPT can actually create this kind of rubric after seeing many example homeworks from others, and it can grade your homework almost instantly. However, we don't know whether the grade provided by AI is accurate or not. So we need you to help us understand whether these AI language models can do a good job evaluating students' projects. Are you interested in seeing how the AI grades the project you created on AMBY?

[If yes, show the evaluation result]

[Give students a few minutes to read]

(1) What do you think about the score that the AI gives for your project?
(2) Do you agree with the AI score? Why or why not?
(3) How does it compared with the score your teacher gives or what you expected to have? Which part is not true? - ask about specific dimensions (ideation, main topic, follow ups etc)
(4) What do you think about the AI's reasoning? Do they actually reflect what you did for your project?
(5) What would you do to improve your project after seeing these scores?
(6) How much would you trust these scores? Why?
(7) Because the AI can give the score and suggestions almost instantly after seeing your project, what do you think about having an AI that gives you feedback while you're working in AMBY?
(8) How helpful would it be if AMBY can grade your projects automatically and give you feedback while you're working?
(9) How would you like to receive that feedback from AMBY?

D Interview Questions for Study 2 Teacher Interview

LLM Perception

(1) Have you heard about Large Language Models like ChatGPT? How familiar are you with them? Do you use it regularly? Occasionally? Maybe used once or twice? Or never used?
(2) What do you use it for?
(3) Do you use it for your own teaching? How? For activity design? Assessment? Concept learning?
(4) How has your experience been with LLMs?

LLM Feedback

In AMBY, we actually developed an early prototype to automatically grade student chatbots and offer specific feedback for their chatbots. It uses GPT-4, which is the same model that ChatGPT uses, and we trained the model using the evaluation rubric on specific dimensions, along with some example chatbots that students created in the past. The goal is to reduce the teacher's burden of manually evaluating every student project and offering feedback along the way. However, we don't know whether the grade that the AI provided can be trustworthy or helpful for teachers. So we need your input to help us understand how we can support your needs. Is it okay if I show you our AI-generated grade and feedback?

[If yes, show the evaluation result: share screen]

[Explain if needed: To give you some background, this rubric was used by the teacher to grade student's chatbot. It has multiple dimensions, each dimension covering one aspect of the project. The first dimension, basics of chat, is looking at some basic elements like greet, default fallback and help; the second dimension, main topics, is looking at how well the students organized different intents to talk about their topics; The third dimension is about topic relevance, that is whether the topic they chose is appropriate and whether it can teach their user about the topic]

[Give the teacher a few minutes to read]

(1) What do you think about the grade that the AI gives for this student project?
(2) Imagine, how does it compare with the score that you would give for your students? Which part needs improvement?
(3) What do you think about the AI's reasoning for its scoring? Would you give something similar to your students?
(4) How much would you trust these scores and reasoning? Why?
(5) How much do you expect your students to trust these scores and the reasoning?
(6) In what ways do you think these scores and reasoning can be useful? Do you think it would reduce your burden of grading and offering feedback? Why or why not?
(7) What would be some challenges for using such AI-based assessments in your classroom?