Learning from Learners: Human-Centered Evaluation of Conversational Agents in Educational Settings

Emily Doherty
emily.doherty@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Michael Buchanan
michael.buchanan@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

E. Margaret Perkoff
margaret.perkoff@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Indrani Dey
idey2@wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

Leanne Hirshfield
leanne.hirshfield@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

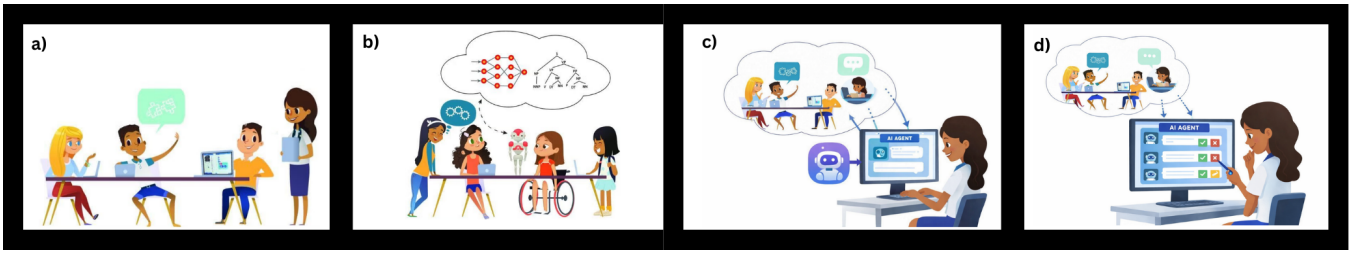

Figure 1: Human–AI workflow for pedagogical agents: (a) Students engage in collaborative learning activities with support provided by human educators. (b) Students collaborate with support from an AI agent, with teacher oversight. (c) In the Wizard of Oz (WoZ) condition, a human expert remotely observes the collaboration and crafts interventions delivered by the agent. (d) In the Human-in-the-Loop (HITL) condition, a human expert reviews LLM-generated interventions in real time, accepting, modifying, or rejecting suggestions before delivery.

Abstract

Conversational pedagogical agents powered by large language models (LLMs) are increasingly deployed in classrooms, yet evaluating their instructional quality remains a persistent challenge. Dominant LLM evaluation methods prioritize semantic similarity, fluency, or surface coherence, but do not account for whether agent interventions are pedagogically aligned with learners' needs. In this paper, we advocate for a human-centered evaluation approach grounded in scaffolding theory. Building on our prior work introducing the Jigsaw Interactive Agent (JIA), we revisit the corpus through two complementary analyses: (1) a post-hoc comparison of LLM-generated and expert Wizard-of-Oz (WoZ) interventions during small group collaboration using theory-grounded scaffolding annotations, and (2) a workflow analysis of Human-in-the-Loop (HITL) deployments examining how real-time oversight shaped intervention delivery. Although LLM responses were often semantically similar to expert responses, they systematically over-scaffolded, providing excessive support regardless of student understanding. The HITL data further demonstrate that human reviewers frequently filtered or replaced agent outputs prior to delivery. Together, these findings expose the limits of similarity-based evaluation and highlight the necessity of structured human oversight. We argue that advancing AI in education requires theory-grounded evaluation and hybrid human–AI workflows that preserve pedagogical discretion in classrooms.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **Empirical studies in HCI**.

Keywords

Large Language Models, evaluation, education, scaffolding

ACM Reference Format:
Emily Doherty, Michael Buchanan, E. Margaret Perkoff, Indrani Dey, and Leanne Hirshfield. 2026. Learning from Learners: Human-Centered Evaluation of Conversational Agents in Educational Settings. In *Proceedings of 3rd Workshop on Human-centered Evaluation and Auditing of Language Models at the 2026 CHI Conference on Human Factors in Computing Systems (HEAL @ CHI'26 Workshop).* ACM, New York, NY, USA, 9 pages.

1 Introduction

Conversational agents are quickly being introduced in K-12 settings at a moment when these systems are more capable and adaptable than ever before. Advances in large language models (LLMs) have enabled these agents to engage in richer, more adaptive dialogue, raising interest in their potential for facilitating collaborative learning experiences, promoting deeper engagement, and bolstering knowledge sharing among students [18, 22, 29, 45, 58]. However,

deploying such agents in classrooms is met with understandable hesitation. Students, educators, and families have raised concerns about privacy, trust, pedagogical alignment, and over-reliance on artificial intelligence (AI) systems [1, 6, 25, 48, 61]. These concerns are not merely about deployment, but also about evaluation. If we lack evaluation frameworks that capture pedagogical alignment, we cannot meaningfully assess whether such systems support or undermine learning. In this sense, this resistance highlights a deeper methodological gap: current evaluation practices are poorly equipped to measure the pedagogical effectiveness of LLM-based agents.

In educational contexts, effectiveness is not determined solely by the linguistic quality or topical relevance of an agent's responses, but by how those responses shape learners' thinking, interaction, and collaboration over time. An intervention may be fluent and factually correct, yet pedagogically misaligned if it provides the wrong type or amount of support at the wrong time. This distinction is particularly important in small-group learning settings, where facilitation requires sensitivity to evolving understanding, participation dynamics, and opportunities for productive struggle [4].

Most existing evaluation approaches for conversational agents are poorly suited to these requirements. Automated evaluation metrics for Natural Language Generation (NLG) primarily measure linguistic fluency [41] and emphasize surface-level properties such as grammar, coherence, or semantic similarity to a reference response [21, 36]. While useful for assessing general language quality, these metrics are largely agnostic to pedagogical intent and learning processes. There have been some attempts to create automatic methods for evaluating the quality of teacher-like responses [46], but they fail to capture whether an agent's intervention actually scaffolds collaboration and learning. Beyond response generation, evaluation of LLM-based agents has focused on task-oriented capabilities such as planning, memory, function calling, or question answering [60], reinforcing evaluation paradigms that overlook judgment criteria central to learning environments.

From a learning sciences perspective, these judgments are captured by the construct of *scaffolding*: adaptive, temporary support that is contingent on learners' demonstrated understanding and gradually withdrawn as learners become more independent [37, 49]. Effective scaffolding in collaborative settings requires deciding not only how to intervene, but when to intervene—and when to refrain. Excessive or mistimed support can undermine learner agency and disrupt collaborative knowledge building [40], even when the content of an intervention is correct.

In this paper, we advocate for a human-centered approach to evaluating LLM-based pedagogical agents that prioritizes pedagogical theory and expert judgment. This paper builds on our prior CHI publication introducing the Jigsaw Interactive Agent (JIA) [11], a LLM-based conversational agent designed to support youth working in small groups on an open-ended programming and brainstorming activity. In [11], we conducted an empirical study with youth aged 12-17 to evaluate JIA. Student groups were placed into one of two treatment groups: (1) The first group worked with the LLM-based JIA agent while an expert in the learning sciences moderated LLM-generated interventions in real time, acting as a human-in-the-loop (HITL), with the ability to approve, modify, or reject each intervention in real-time. (2) The second treatment group worked with a Wizard of Oz (WoZ) version of JIA, where a human expert crafted interventions delivered by JIA in real time. In our prior work we demonstrated the effectiveness of JIA by making comparisons at the group level between the groups exposed to the HITL vs. the WoZ treatments, respectively.

Our prior work did not evaluate the HITL or WoZ interactions at the intervention level, which is the focus of this workshop paper. Thus, we present two complementary analyses: (1) a theory-grounded post-hoc audit comparing JIA's LLM-generated interventions and the human expert WoZ interventions from 16 student groups, using annotations of scaffolding level and student understanding alongside automated metrics, and (2) a workflow analysis of the HITL sessions examining how the real-time human oversight shaped which JIA agent interventions were delivered. Together, these analyses illustrate both the limitations of similarity-based evaluation and the role of structured human oversight in governing AI in classrooms. Figure 1 illustrates the broader human–AI workflow examined in this paper, spanning real-time intervention, human oversight, and post hoc pedagogical auditing.

Our results reveal a systematic misalignment between similarity-based automated metrics and pedagogical quality. Although LLM-generated responses are often semantically similar to expert interventions, they tend to provide higher levels of scaffolding regardless of students' demonstrated understanding—directly opposing the principle of contingent scaffolding [49]. In contrast, expert WoZ interventions decrease support as understanding increases. These findings illustrate a critical challenge for evaluating pedagogical agents: automated metrics (e.g., surface-level similarity) do not capture important dimensions of pedagogical effectiveness. The real-time HITL data further reinforce this concern. In deployment, human reviewers frequently filtered, modified, or replaced the LLM-generated interventions before delivery, and over one-third of messages ultimately sent to students were entirely written by the human expert. Together, these findings suggest that even linguistically appropriate responses require human calibration to maintain pedagogical alignment.

The goal of this paper is therefore to demonstrate how pedagogical expertise can be incorporated into both evaluation and deployment workflows to audit agent behavior in context. By treating scaffolding as a human-centered evaluation lens, we surface dimensions of instructional support that automated metrics routinely overlook. Our findings demonstrate that sustained human oversight is not merely complementary, but essential for meaningfully assessing and governing educational AI in classroom settings.

2 Background

2.1 Scaffolding to Support Learning

Scaffolding refers to dynamic, temporary support, adapted to the learner's needs [59]. The scaffolding construct is rooted in sociocultural learning theories, particularly Vygotsky's Zone of Proximal Development (ZPD), which is the gap between what learners can do independently and what they can achieve with support from more knowledgeable others [53]. As the learner becomes more independent, the scaffolding can be gradually removed or faded. Van de Pol and colleagues [49, 50] emphasized the importance of *contingent scaffolding*, where guidance is adjusted according to student capabilities—reducing support when students demonstrate

greater understanding and independence and increasing it when they face difficulties. Although traditionally focused on individual learning and conceptual understanding, scaffolding can be extended to collaborative settings where groups of learners may need more structured support to manage group dynamics and interactions [19, 56]. However, asking teachers to monitor multiple groups and provide continuous adaptive scaffolding simultaneously in real time is not feasible [9] and researchers have been investigating tools [13, 31] and pedagogies [52] that can potentially provide more structure and/or support for collaboration. A further challenge is that many of these tools offer uniform support rather than adaptive support, in the true nature of scaffolding [37]. AI-driven technology that leverage LLMs and Natural Language Processing (NLP) offer a promising solution by analyzing student conversations as they work together in groups, identifying conceptual or collaborative challenges, and providing more targeted, adaptive support. One such application is a conversational pedagogical agent.

2.2 Conversational Pedagogical Agents

Pedagogical agents have facilitated student learning for nearly thirty years now, providing support through various digital mediums [44]. Specifically, we focus on *conversational* pedagogical agents (CPAs), which allow for dialog-based human-agent interaction [2, 16, 28]. Recent surveys have categorized educational LLM agents into two main types: pedagogical agents that support teachers (e.g., through classroom simulation or feedback generation) and those that provide direct support to students in real time [7]. Our work falls into the latter category, we are evaluating an LLM-based agent designed to provide adaptive collaborative support during peer learning. Such agents can play different pedagogical roles including tutors [10], teachers [5], or peers [47]. CPAs in a teacher-like role help learners perform tasks by imitating the gold standard of educators and presenting instructions, providing examples, and asking questions [32, 55, 57]. There are many examples of these CPAs, including Wambsganss et al.'s *ArgueTutor* [54], which supported student learning with adaptive argumentation feedback, and Ruan's *EnglishBot* [39], that provided adaptive support during language learning. Winkler et al. [57] also developed *Sara*, a web-based CPA that provides voice- and text-based scaffolds to learners during online video lectures. However, many of these agents are primarily evaluated based on student performance, which may introduce potential biases—for example, students might perform better simply due to the novelty or engagement of interacting with an AI system, rather than the instructional quality of the responses [3].

Because CPAs can adopt diverse instructional roles (e.g., tutor, teacher, peer), their success cannot be evaluated with a single outcome metric; instead, evaluation must account for pedagogical intent and interactional context. Some CPA applications have measurable evaluation metrics, e.g., improved learning [10] or engagement [39]. Other applications, however, are challenging to evaluate or measure, including facilitating collaboration [47] and promoting critical thinking skills [20, 35]. Many existing methods do not capture content related to pedagogical goals, demonstrating a need for improved evaluation criteria for CPAs.

2.3 Current Evaluation Methods

In this section, we review existing evaluation methods to illustrate how many assess linguistic form and task completion over pedagogical outcomes, creating systematic blind spots for CPAs. Current evaluation methods consist of: automated and pre-trained metrics, and human evaluation methods. Automated metrics such as BLEU [36] and ROUGE [30] assess lexical overlap between generated and reference text. While effective for tasks like machine translation and summarization, these metrics capture surface-level similarity rather than contextual or instructional quality. Despite known limitations outside their original domains [38], they remain widely used in conversational agent research [41]. While text overlap can indicate general semantic alignment, it fails to capture specific pedagogical-related measures, such as appropriate level of scaffolding.

Pre-trained metrics such as BERTScore [62] and BLEURT [42] leverage contextual embeddings to measure semantic similarity and correlate with human judgments of fluency and coherence. Although more robust than n-gram metrics, they remain similarity-based and implicitly assume that semantic alignment corresponds to response quality. In educational settings, however, the same idea may be delivered in pedagogically appropriate or inappropriate ways depending on the level of control, contingency, and timing.

Educational research has developed theory-grounded frameworks for analyzing teacher discourse, operationalizing instructional moves such as question types, goal specificity, cognitive demand, and conversational uptake [8, 12, 15, 26, 33, 34, 43]. For example, Jensen et al. [23] showed that computational models can approximate structured pedagogical dimensions, though higher-order judgments involving cognitive complexity remain difficult to model reliably. Together, this work demonstrates that pedagogically meaningful features of teacher talk can be operationalized and measured—standards that conversational pedagogical agents should likewise meet in classroom settings.

Human evaluation is frequently used for assessing conversational agents, typically through Likert-scale ratings or pairwise comparisons [14, 51]. However, annotation standards are inconsistent, and evaluator expertise is often unspecified [14]. In educational contexts, domain knowledge is critical: pedagogical appropriateness depends on instructional intent, learner understanding, and interactional context. Structured comparison frameworks such as the AI Teacher Test [46] and domain-specific annotation schemes [17, 27, 57] highlight the value of theory-grounded evaluation, yet such approaches are rarely integrated systematically into LLM evaluation workflows of pedagogical agents.

The prior research in this space demonstrates that instructional discourse can be meaningfully characterized using theory-grounded frameworks, but reveals a gap in applying these standards to LLM-based classroom agents [41]. In this paper, we address this gap by grounding evaluation in scaffolding theory and incorporating expert human annotation alongside pre-trained metrics.

3 Human-Centered Evaluation & Auditing of JIA

In order to investigate human-centered evaluation of LLMs in small group K-12 settings, we leverage the JIA LLM-based interactive

agent, which was constructed to support small group collaboration, as presented in our prior work [11].

JIA and Prior Empirical Study. The JIA LLM-based agent was informed by a rule-based dialogue policy that monitors discourse in real-time and detects when students are in a certain state (*i.e., Parallel Interaction, Contributing to the Shared Problem Space, and Unproductive Perseverance)*, so that JIA can intervene accordingly. The states detected via student discourse and suggested intervention types are included in a prompt to *Mistral-7B-Instruct-v0.3 model* [24], which generates JIA's responses. JIA was evaluated in an empirical study where dyads and triads (total n = 145) aged 12–17 collaborated together on a group programming activity. Participant groups were placed into one of three conditions: (1) A control condition, where youth worked on the programming activity without any support, (2) a Wizard-of-Oz subject matter expert (WoZ-SME) condition where groups were supported in real time by a human expert, and (3) a JIA-HITL condition, where groups were supported by an LLM-agent, whose interventions were overseen and moderated by a human-in-the-loop expert, while completing the activity. Specifically, the HITL could accept, modify, or reject each intervention, preserving instructional discretion in real time. This design reflected concerns about deploying fully autonomous AI systems in classrooms and allowed us to examine how human oversight shapes intervention quality. For more detail on the experimental design, creation of the JIA's underlying dialogue policy, and prompt construction, please see [11].

3.1 Comparing JIA-LLM & WoZ-SME Interventions

For this set of analyses, we revisit the student dialogue and interventions during the JIA-HITL and WoZ-SME data collections. We also generated a new set of *post hoc* JIA responses based on the student dialogue from the WoZ-SME data collections, by prompting the agent to respond to the same student dialogues addressed by WoZ-SMEs, enabling direct comparison between human- and AI-generated interventions. We compared 67 SME-written interventions from 16 WoZ-SME sessions to 67 post hoc JIA LLM-generated responses to the same student dialogue. All conversation histories and interventions were double-coded by four human annotators, who are graduate students in the U.S. with backgrounds in learning sciences or human–computer interaction. Raters annotated the preceding student dialogue for *Level of Understanding* and evaluated both WoZ-SME and JIA LLM-based interventions for *Appropriateness* and *Level of Scaffolding*. To assess appropriateness, raters were shown the 30 seconds of dialogue preceding each intervention in order to judge how well the response aligned with students' evolving needs. Appropriateness was rated on a scale from 0 to 2. Level of Understanding was annotated using a rubric adapted from Van de Pol et al. [49], classifying group comprehension as Too Off-Topic to Tell, Poor (0), Partial (1), or Good (2) (see Table 4). Our scaffolding metric was adapted from Van de Pol's (2019) five levels of teacher control, ranging from minimal support with open-ended questions (Level 1) to more structured guidance (Level 5) [49]; according to contingent scaffolding, the lower the student understanding, the greater the amount of control the teacher should use to guide the conversation.

Table 1: Mean (SD) Appropriateness and Scaffolding ratings across 67 matched WoZ-SME and JIA intervention contexts.

	Appropriateness	**Scaffolding**
WoZ-SME	1.22 (0.53)	0.57 (0.47)
JIA	1.23 (0.45)	0.81 (0.49)
t-test	$t(66) = -0.18$, $p = .86$	$t(66) = -2.76$, $p = .008$

While these levels primarily address conceptual understanding, students may also need support for collaboration. For example, while monitoring groups, teachers may intervene to encourage more equal or active participation from group members, help negotiate group norms, or support the group's knowledge-building by asking questions [56]. Therefore, we extended Van de Pol's framework to include low, medium, and high levels of collaborative support in our Levels of Scaffolding (see Table 3).

Two independent annotator pairs coded separate subsets of the data (n = 55; n = 12). Inter-rater reliability was assessed via Cohen's κ (quadratic-weighted for ordinal variables). For the larger subset, agreement was substantial to near-perfect for Level of Understanding ($\kappa = .82$) and Scaffolding (WoZ-SME $\kappa = .86$; JIA $\kappa = .81$), and moderate for Appropriateness (WoZ-SME $\kappa = .43$; JIA $\kappa = .57$). For the smaller subset, reliability estimates were comparable for Level of Understanding ($\kappa = .80$) and WoZ-SME Level of Scaffolding ($\kappa = .80$), while agreement was lower for JIA Level of Scaffolding ($\kappa = .31$) and Appropriateness (WoZ-SME $\kappa = -.06$; JIA $\kappa = .22$), likely due to the much smaller sample size.

Additionally, we computed BERTScores to compare JIA LLM-based responses to WoZ-SME interventions, using the latter as the reference text. While BERTScore only evaluates for semantic similarity and does not account for pedagogical effectiveness, it can serve as a complementary metric to human annotations by providing a quantitative measure of textual overlap. This helps identify whether LLM-generated responses align conceptually with human interventions, while human annotations capture deeper aspects such as scaffolding, appropriateness, and contextual relevance that pre-trained metrics overlook.

Results. We report the means and standard deviations of Appropriateness and Scaffolding across the 67 matched intervention contexts (Table 1). Paired-samples t-tests were conducted to compare WoZ-SME and JIA LLM-based responses.

JIA's LLM-generated responses were slightly more appropriate than the WoZ-SME responses, yet there was no statistical significant difference. However, JIA's responses did provide overall higher scaffolding than WoZ-SME responses ($t(66) = -2.76$, $p = .008$). While this suggests that JIA offered more structured support, it does not necessarily reflect greater instructional quality. In fact, providing high levels of scaffolding regardless of student need may indicate a lack of pedagogical sensitivity, whereas WoZ-SME responses more often adapted support to the students' demonstrated understanding.

We then break down ratings of Scaffolding by the Level of Understanding to see whether students with lower understanding received a high level of scaffolding, as expected (Table 2). These descriptive statistics reflect annotator-level ratings across all 134 coded responses (67 responses, double-coded). We retain coder-level

Table 2: Mean (SD) Scaffolding ratings by Level of Understanding across all 134 annotator-level coded responses.

Level of Understanding	WoZ-SME	JIA
Too off-topic (n = 26)	0.44 (0.51)	0.625 (0.58)
Poor (0) (n = 26)	0.8077 (0.63)	0.4231 (0.50)
Partial (1) (n = 52)	0.5882 (0.70)	0.9608 (0.63)
Good (2) (n = 30)	0.4667 (0.68)	1.0714 (0.72)

observations for this breakdown to preserve variation in annotators' judgments of student understanding. Collapsing to intervention-level averages would require resolving disagreements into a single Level of Understanding category per intervention, thereby constraining the distribution of cases across categories and obscuring how scaffolding ratings co-varied with individual annotator perceptions.

When broken down by student understanding, WoZ-SME responses followed expected patterns of contingent scaffolding (i.e., increased Scaffolding with lower Levels of Understanding), while JIA responses did not. In contrast, JIA responses exhibit an opposite trend, with scaffolding increasing as students demonstrate greater understanding. This suggests that WoZ-SME responses were more reactive to struggling students, whereas JIA may have applied more structured scaffolding, even when students showed higher comprehension. Qualitative inspection of the intervention pairs helps explain this divergence. When students demonstrated higher understanding, often by discussing concepts or giving examples, WoZ-SME responses frequently validated or extended students' ideas, resulting in lower scaffolding ratings. Table **??** illustrates this pattern: following a rich student discussion connecting multiple sensors to real-world examples, the WoZ-SME response affirmed the connection without adding additional structure, whereas the JIA response redirected the group toward task alignment through more explicit guidance.

> **Context (30s Student Utterances)** 'I know.' 'Oh my gosh.' 'Shut up.' "I can't wait, hear me out." 'Let me try.' 'So we have the moisture, we have the sound.' 'And the temperature, whenever we walk outside, the moisture attributes a lot.' "To the temperature outside it's always crazy, and inside they keep it super cold so that all the fish and the butterflies…" 'Have you ever been to there? Yeah, of course, they have fish in the water.' **Level of Understanding: Good**
> **WoZ-SME:** "I love butterfly pavilions! Great connection to something you've experienced :)" **Level of Scaffolding: Low**
> **JIA:** "Let's focus on our assignment and share our sensor expertise to progress together." **Level of Scaffolding: High**

For pre-trained metrics, we selected BERTScore and used the WoZ-SME responses as the reference text. The average F1 score calculated using BERTScore was 0.865, indicating that the JIA and WoZ-SME responses shared overall semantic similarity. However, when we analyzed BERTScore values by human-annotated categories, we found minimal variation: for Appropriateness levels 0, 1, and 2, the average BERTScores were 0.853, 0.857, and 0.863 respectively; for Scaffolding levels 0, 1, and 2, they were 0.853, 0.860, and 0.852. These negligible differences suggest that BERTScore does not meaningfully differentiate between varying levels of educational quality, reinforcing the need for human annotation to capture critical pedagogical dimensions.

3.2 Human-in-the-Loop Workflow Outcomes

In the HITL condition, JIA provided LLM-generated interventions to a human expert throughout the session. The human expert then chose to modify, reject, or send the generated intervention as is. Our analysis focuses on two stages of this workflow: (1) review decisions for interventions that the HITL actively engaged with, and (2) the composition of messages ultimately delivered to participants.

Results. At the review stage (N = 588 reviewed interventions across 22 HITL sessions), SMEs marked 23.0% as Accept, 5.1% as Modify, 45.1% as Reject, and 26.9% as Ignore. Importantly, these counts reflect only interventions that the SME interacted with, not all interventions generated by the system. Instead, they reflect the necessity of human calibration when deploying generative systems in sensitive instructional contexts.

Examining the delivered messages provides a clearer view of the hybrid human–AI collaboration. Of the 253 total messages sent to students, 53.4% were direct AI accepts, 10.3% were modified AI responses, and 36.4% were authored entirely by the human expert. In other words, humans chose to accept the LLM-generated response over half of the time, while the remainder reflected included some form of human oversight or authorship. Most modifications involved making the agent's response more succinct and to more directly answer the student's question, especially when they specifically requested help.

4 Discussion

Evaluating CPAs remains a major challenge, as existing evaluation methods are poorly suited to assess the effectiveness of agent support. Most LLM evaluation methods focus on fluency and coherence, but these do not capture educational-related goals, including whether a response scaffolds learning or supports collaboration. Some pre-trained metrics (e.g., semantic similarity) can be complementary to human annotations, but only provide a surface-level analysis of responses. As Chu et al. (2025) note, most current evaluation frameworks focus on task automation and model utility, overlooking alignment with pedagogical goals and learner-centered outcomes [7]. This highlights the importance of human annotation in capturing the nuances of student-agent interactions.

Our case study demonstrates the disconnect between surface-level pre-trained evaluation metrics and finer-grained expert human annotations, highlighting the need for human-centered practices grounded in learning sciences. While scaffolding of the JIA LLM-based responses increased with student understanding, this was the opposite for WoZ-SME responses, which generally decreased with increasing understanding. This suggests that the WoZ-SME responses offered more support when students struggled, aligning with the theory of contingent scaffolding [50], while JIA's LLM-generated responses may have offered more support when it was not warranted. One likely explanation for the LLM's tendency to over-scaffold is that the model lacks an explicit representation of student understanding or pedagogical intent. While the prompt includes dialogue state and suggested intervention types, it does not

encode principles of contingent scaffolding (e.g., reducing support as understanding increases). As a result, the model may default to providing more directed or structured responses, which are often statistically associated with "helpfulness" in general-purpose training data, but are pedagogically misaligned in collaborative learning contexts. This limitation further explains the role of HITL oversight observed in our deployment: human experts frequently filtered or revised over-scaffolded responses, effectively reintroducing contingent pedagogical judgment that the model itself lacks.

Further, the pre-trained metrics fail to capture the differences between the scaffolding of the responses. The high average BERTScore proves that WoZ-SME and JIA discuss similar ideas in their responses, but it does not show how the WoZ-SME provided less scaffolding. In fact, when the BERTScore results were averaged across both of our human annotated values, there was minimal difference between each of the levels and the corresponding automated score value. One driving factor in the discrepancies between the human evaluation and pre-trained metrics is that the human annotators had access to the conversation history and therefore understood the context in which the advice is being given. Since pre-trained metrics can only compare a reference response to a prediction, they fail to use the context of the lesson and conversation history in evaluating responses. A promising direction is the development of agentic systems that incorporate richer context, including full conversation history and inferred student understanding, enabling more context-sensitive evaluation and intervention generation.

Furthermore, these metrics are trained to assess general semantic similarity scores, which may not adequately capture the nuances between responses that convey similar concepts at different levels of control. Together, these findings suggest that evaluating CPAs requires attention to learner context and instructional function, rather than relying solely on automated similarity metrics. Incorporating expert-informed, context-sensitive evaluation approaches may therefore be essential for accurately assessing pedagogical quality and alignment.

Human-in-the-Loop as a Safeguard: Beyond post hoc evaluation, our HITL analysis highlights the importance of integrating human oversight into the intervention workflow. In the original JIA deployment [11], LLM-generated interventions were surfaced to an expert for review prior to delivery. Our analysis examined two stages of this process: (1) SME review decisions over interventions that entered the review interface, and (2) the composition of messages ultimately delivered to students.

At the review stage, SMEs frequently rejected or ignored interventions. However, these decisions reflect filtering within a forced-generation design (interventions were proposed at fixed intervals), rather than a global measure of system failure. More informative is the composition of delivered messages. Of all instructional messages sent to students, over half were sent without modification and one third required full human authorship. Thus, while the agent contributed substantially to classroom discourse, many of delivered messages required full human authorship or modification. Notably, the tendency of JIA's LLM-generated responses to over-scaffold in the post-hoc analysis helps explain why real-time filtering was frequently necessary. The HITL data therefore complement the evaluation findings: even semantically appropriate responses may require human calibration to maintain pedagogical alignment.

Toward a Hybrid Human–AI Evaluation Workflow: Taken together, the misalignment observed in post-hoc evaluation and the filtering patterns observed in HITL deployment point toward a hybrid evaluation and governance model. First, humans remain in the loop during live interventions, reviewing and curating agent interventions before they reach students. Second, after the collaboration, videos and/or conversation logs can be audited by human annotators using pedagogically grounded criteria (e.g., scaffolding level, understanding). These annotations can then inform iterative updates to dialogue policies and model prompting strategies. This two-stage workflow: 1) real-time human filtering followed by 2) post hoc pedagogical auditing, reframes evaluation as an ongoing governance process rather than a one-time benchmarking exercise. Rather than optimizing solely for semantic similarity or fluency, evaluation becomes an ongoing process of aligning agent behavior with instructional goals and contextual classroom realities. In this way, resistance to fully autonomous AI becomes a productive design constraint, motivating evaluation practices that embed human judgment as a structural feature rather than a temporary safeguard.

Overall, our findings suggest that JIA functioned most effectively as a intervention generator within a human-governed system, rather than as autonomous support. Particularly in youth classrooms, where pedagogical decisions are sensitive and developmentally consequential, sustained human oversight is not merely a temporary safeguard but a core design requirement. Real-time review enables educators to filter over-scaffolded or mistimed interventions, preserving instructional discretion and learner agency. At the same time, whether teachers can feasibly monitor AI interventions in real time while also circulating, supporting, and monitoring students warrants careful consideration, as classroom realities may constrain the extent of continuous oversight.

To improve CPA evaluation, future work should focus on developing context-aware metrics that incorporate pedagogical features, conversation history, and student understanding levels. One promising direction is training evaluation models on expert-annotated datasets, ensuring that response quality is measured against instructional effectiveness rather than linguistic similarity alone. Additionally, integrating human-centered approaches into evaluation can help create adaptive, pedagogically informed agents that respond dynamically to student needs. Without such improvements, evaluation methods will continue to misrepresent the effectiveness of CPAs, limiting their potential for supporting meaningful learning experiences.

5 Conclusion

This study highlights the misalignment between existing evaluation metrics and pedagogically meaningful CPA assessment. We analyzed the ability of LLMs to provide appropriate levels of support in a collaborative learning context. LLM-generated responses exhibited higher scaffolding levels than WoZ-written interventions, including at times when providing more scaffolding could take away from learning opportunities. The analysis of BERTScores further demonstrates that pre-trained metrics fail to account for instructional effectiveness, as they lack access to essential context.

These findings underscore the need for education-specific evaluation frameworks that move beyond surface-level textual comparisons. Beyond metric misalignment, our findings demonstrate that even when responses appear semantically aligned with expert interventions, real-time human oversight remains critical in youth classrooms. Evaluation and governance are therefore intertwined: meaningful CPA assessment requires both theory-grounded annotation and structured human–AI workflows. We advocate for future work in developing CPA evaluation benchmarks based on expert-annotated datasets and incorporating human-centered approaches. By bridging the gap between current evaluation and pedagogical expertise, we can create more effective conversational agents that truly enhance student learning and collaboration.

Acknowledgments

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grants DRL 2019805 and DRL 2454151. The opinions expressed are those of the authors and do not represent views of the NSF.

References

[1] Zishan Ahmed, Shakib Sadat Shanto, Most Humayra Khanom Rime, Md Kishor Morol, Nafiz Fahad, Md Jakir Hossen, and Md Abdullah-Al-Jubair. 2024. The generative AI landscape in education: Mapping the terrain of opportunities, challenges, and student perception. *IEEE access* 12 (2024), 147023–147050.

[2] L. Bradeško and D. Mladenić. 2012. A survey of chatbot systems through a Loebner Prize competition. In *Proc. of Slovenian Language Technologies Society Eighth Conference of Language Technologies*. 34–37.

[3] Yu Chen, Scott Jensen, Leslie J Albert, Sambhav Gupta, and Terri Lee. 2023. Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers* 25, 1 (2023), 161–182.

[4] Ying-Chih Chen. 2022. Epistemic uncertainty and the support of productive struggle during scientific modeling for knowledge co-development. *Journal of Research in Science Teaching* 59, 3 (2022), 383–422.

[5] Nalin Chhibber and Edith Law. 2019. Using conversational agents to support learning by teaching. *arXiv preprint arXiv:1909.13443* (2019).

[6] Seongyune Choi, Yeonju Jang, and Hyeoncheol Kim. 2023. Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human–Computer Interaction* 39, 4 (2023), 910–922.

[7] Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733* (2025).

[8] Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). ACL, Online, 1638–1653. doi:10.18653/v1/2021.acl-long.130

[9] P. Dillenbourg, L. P. Prieto, and J. K. Olsen. 2018. Classroom orchestration. In *International handbook of the learning sciences*. Routledge, 180–190.

[10] Sidney D'mello and Art Graesser. 2013. AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 4 (2013), 1–39.

[11] Emily Doherty, E Margaret Perkoff, Sean von Bayern, Rui Zhang, Indrani Dey, Michal Bodzianowski, Sadhana Puntambekar, and Leanne Hirshfield. 2025. Piecing Together Teamwork: A Responsible Approach to an LLM-based Educational Jigsaw Agent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[12] Patrick J Donnelly, Nathaniel Blanchard, Andrew M Olney, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proc. of the 7th International Learning Analytics & Knowledge Conference*. ACM, New York, NY, USA.

[13] Cheng-Yu Fan and Gwo-Dong Chen. 2021. A scaffolding tool to assist learners in argumentative writing. *Computer Assisted Language Learning* 34, 1-2 (2021), 159–183.

[14] S. Gehrmann, E. Clark, and T. Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *The Journal of Artificial Intelligence Research* 77 (2023), 103–166.

[15] A Graesser, Y Ozuru, and J Sullins. 2009. What is a good question? In *Threads of coherence in research on the development of reading ability*, M G Mckeown And Kucan (Ed.). Guilford Press, 112–141.

[16] A. Gulz, M. Haake, A. Silvervarg, B. Sjödén, and G. Veletsianos. 2011. Building a social conversational pedagogical agent: Design challenges and methodological approaches. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, 128–155.

[17] Kunal Handa, Margarett Clapper, Jessica Boyle, Rose Wang, Diyi Yang, David Yeager, and Dorottya Demszky. 2023. "mistakes help us grow": Facilitating and evaluating growth mindset supportive language in classrooms. In *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing*. ACL, Stroudsburg, PA, USA, 8877–8897.

[18] Yugo Hayashi. 2020. Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: an experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning* 15, 4 (2020), 469–498.

[19] N. Hoang, J. B. Bush, I. Dey, E. Watts, C. Clevenger, and W. R. Penuel. 2024. MOSAIC Protocol: Analyzing Small Group Work to Gain Insights into Collaboration Support for Middle School STEM Classrooms. In *Proc. of the 17th International Conference on Computer-Supported Collaborative Learning-CSCL 2024*. International Society of the Learning Sciences, 297–300.

[20] Y Huang. 2025. Developing Customized AI-Based Conversational Tools to Support Personalized Learning and Promote Critical Thinking in Higher Education Institutions. (2025).

[21] HuggingFace. 2025. Perplexity of fixed-length models. https://huggingface.co/docs/transformers/en/perplexity

[22] Patrícia Jaques, Adja Andrade, João Jung, Rafael Bordini, and Rosa Vicari. 2023. Using pedagogical agents to support collaborative distance learning. In *Computer support for collaborative learning*. Routledge, 546–547.

[23] 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 *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.

[24] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[25] Mohd Kashif, Mohammad Ammar, Abdellatif Sellami, Thomas KF Chiu, Saddam Akber Abbasi, and Zubair Ahmad. 2025. Teachers' perspectives on AI integration in K-12 education: challenges, opportunities, and preliminary assessment model–a systematic review. *Computers in the Schools* (2025), 1–27.

[26] Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educ. Res.* 47, 7 (Oct. 2018), 451–464.

[27] Bijan Khosrawi-Rad, Paul Keller, Linda Grogorick, and Susanne Robra-Bissantz. 2022. Introducing Vicky: A Pedagogical Conversational Agent for the Classification of Learning Styles. *International Conference on Design Science Research in Information Systems and Technology (DESRIST)* (June 2022).

[28] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Educ. Inf. Technol.* 28, 1 (Jan. 2023), 973–1018.

[29] R Kumar and Carolyn P Rosé. 2011. Architecture for building conversational agents that support collaborative learning. *IEEE Trans. Learn. Technol.* 4, 1 (Jan. 2011), 21–34.

[30] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[31] P. C. Lin, H. T. Hou, and K. E. Chang. 2022. The development of a collaborative problem solving environment that integrates a scaffolding mind tool and simulation-based learning: an analysis of learners' performance and their cognitive process in discussion. *Interactive Learning Environments* 30, 7 (2022), 1273–1290.

[32] Michael Frederick McTear, Zoraida Callejas, and David Griol. 2016. *The conversational interface*. Vol. 6. Springer.

[33] National Research Council, Division of Behavioral and Social Sciences and Education, and Board on Behavioral, Cognitive, and Sensory Sciences. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press, Washington, D.C., DC.

[34] Newmann, M Fred, and Ed. 1992. *Student engagement and achievement in American secondary schools*. Teachers' College Press, New York, NY.

[35] A A Olawale. 2025. Conversational AI Review of Literature on the Role of Artificial Intelligence as a Tool for Bolstering Critical Thinking Skills in Mobile and Adaptive Systems. *AI and Curriculum Development for the Future* (2025).

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.
[37] S. Puntambekar and R. Hubscher. 2005. Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist* 40, 1 (2005), 1–12.
[38] Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44, 3 (2018), 393–401.
[39] S. Ruan, L. Jiang, Q. Xu, Z. Liu, G. M. Davis, E. Brunskill, and J. A. Landay. 2021. EnglishBot: An AI-powered conversational system for second language learning. In *26th International Conference on Intelligent User Interfaces*. 434–444.
[40] Nikol Rummel, Erin Walker, and Vincent Aleven. 2016. Different Futures of Adaptive Collaborative Learning Support. *International Journal of Artificial Intelligence in Education* 26, 2 (1 June 2016), 784–795.
[41] P. Schmidtová, S. Mahamood, S. Balloccu, O. Dušek, A. Gatt, D. Gkatzia, D. M. Howcroft, O. Plátek, and A. Sivaprasad. 2024. Automatic Metrics in Natural Language Generation: A Survey of Current Evaluation Practices. In *Proc. of the 17th International Natural Language Generation Conference*. 557–583.
[42] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).
[43] David J Shernoff, Sean Kelly, Stephen M Tonks, Brett Anderson, Robert F Cavanagh, Suparna Sinha, and Beheshteh Abdi. 2016. Student engagement as a function of environmental complexity in high school classrooms. *Learn. Instr.* 43 (June 2016), 52–60.
[44] Robert F Siegle, Noah L Schroeder, H Chad Lane, and Scotty D Craig. 2023. Twenty-five years of learning with pedagogical agents: History, barriers, and opportunities. *TechTrends* 67, 5 (Sept. 2023), 851–864.
[45] Pieta Sikström, Chiara Valentini, Anu Sivunen, and Tommi Kärkkäinen. 2022. How pedagogical agents communicate with students: A two-phase systematic review. *Computers & Education* 188 (Oct. 2022), 104564. doi:10.1016/j.compedu.2022.104564
[46] Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. *arXiv preprint arXiv:2205.07540* (2022).
[47] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2015. Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Comput. Educ.* 87 (1 Sept. 2015), 309–325.
[48] Chad C Tossell, Nathan L Tenhundfeld, Ali Momen, Katrina Cooley, and Ewart J De Visser. 2024. Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies* 17 (2024), 1069–1081.
[49] Janneke Van de Pol, Neil Mercer, and Monique Volman. 2019. Scaffolding student understanding in small-group work: Students' uptake of teacher support in subsequent small-group interaction. *Journal of the Learning Sciences* 28, 2 (2019), 206–239.
[50] J. Van de Pol, M. Volman, and J. Beishuizen. 2010. Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review* 22 (2010), 271–296. doi:10.1080/10508406.2018.1522258
[51] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proc. of the 12th International Conference on Natural Language Generation*, Kees van Deemter, Chenghua Lin, and Hiroya Takamura (Eds.). ACL, Tokyo, Japan, 355–368.
[52] F. Vogel, C. Wecker, I. Kollar, and F. Fischer. 2017. Socio-cognitive scaffolding with computer-supported collaboration scripts: A meta-analysis. *Educational Psychology Review* 29 (2017), 477–511.
[53] L. S. Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press.
[54] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proc. of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13.
[55] T. Wambsganss, R. Winkler, M. Söllner, and J. M. Leimeister. 2020. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.
[56] Noreen M Webb. 2013. Collaboration in the classroom. In *International guide to student achievement*. Routledge, 215–217.
[57] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.
[58] Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet? - A systematic literature review on chatbots in education. *Front. Artif. Intell.* 4 (July 2021), 654924.
[59] D. Wood, J. S. Bruner, and G. Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* 17, 2 (1976), 89–100.
[60] Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on Evaluation of LLM-based Agents. arXiv:2503.16416 [cs.AI] https://arxiv.org/abs/2503.16416
[61] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart learning environments* 11, 1 (2024), 28.
[62] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

A Appendix

Received 19 February 2026; Accepted 2 March 2026

Table 3: Levels of Scaffolding, Based on the Amount of Control over the Conversation

Scaffolding Level and Description	Example
Misc: Mostly procedural or task support (e.g., giving students instructions to look something up), details about the task (e.g., time remaining), or generic responses.	*"Got it! Thank you for explaining that!"* *"You still have 15 minutes. Can you think of any other problems?"*
Low: Does not provide new content. May ask students to elaborate on their response, but not provide an explanation. Asks students to share or attempts to give everyone a chance to talk. Generic praises or validates work.	*"Great work, team!"* *"What do you know about each other's sensors?"* *"Tell me more about how the sensors would solve that problem."*
Medium: Does not provide new content. Asks why questions, prompting students to explain or elaborate. Asks a more detailed but still open-ended question. Validates student idea that prompts more discussion of that idea. Asks students to share their knowledge with each other and work together.	*"How could you use your sensors to help your community?"* *"That's an interesting idea about the zoo."* *"Let's share our sensor expertise and work together to complete the assignment.* *Can someone explain their sensor's function and how it contributes to the project?"*
High: May provide new content. May not be based on or build off students' existing ideas. Provides a hint or suggestive question. Provides an explanation.	*"What if you used the sensor for detecting plant activity?"* *"To measure humidity, you can use a hygrometer...* *Let's share our findings with each other."*

Table 4: Levels of Understanding and Appropriateness

Level and Description	Example
0 – No Understanding / Active Confusion: The response shows little to no comprehension of the topic. The student may be confused, provide incorrect information, or show a lack of engagement with the concept.	*"I don't get what this is asking."* *"Is this about animals or something else?"*
1 – Partial Understanding: The response shows some understanding but lacks clarity, is incomplete, or contains minor inaccuracies. The student may recognize key ideas but struggle to fully explain them.	*"I think it means the sensor measures heat, but I'm not sure how."* *"It's kind of like tracking movement, I guess."*
2 – Good Understanding: The response demonstrates a strong grasp of the concept. The student can explain key ideas clearly, provide relevant examples, and make logical connections.	*"The humidity sensor measures moisture in the air, which helps us track plant health."* *"This works because the sensor detects changes in temperature and converts them into data."*
Too Off-Topic to Tell: The response is unrelated, disorganized, or lacks enough context to determine comprehension.	*"My favorite animal is a turtle."* *"Did you know it's almost lunchtime?"*
0 – Not Appropriate (Intervention): The intervention is irrelevant, ineffective, or does not support learning. It may be confusing, dismissive, or too vague to be helpful.	*"Just try harder."* *"That's wrong. Move on."*
1 – Somewhat Appropriate (Intervention): The intervention is somewhat helpful but may not fully align with the student's needs. It may be too generic, lack depth, or miss opportunities to clarify misconceptions.	*"Maybe think about what the sensor does."* *"Try explaining it again."*
2 – Highly Appropriate (Intervention): The intervention directly addresses the student's needs and fosters deeper understanding. It is well-aligned with the conversation and encourages critical thinking.	*"Your idea is close — the sensor measures humidity, which helps detect moisture changes. How might that help your design?"* *"Let's break down what the sensor detects and how that connects to your project goal."*