A Theory of Adaptive Scaffolding for LLM-Based Pedagogical Agents

Clayton Cohn, Surya Rayala, Namrata Srivastava, Joyce Horn Fonteles, Shruti Jain, Xinying Luo, Divya Mereddy, Naveeduddin Mohammed, Gautam Biswas

Institute for Software Integrated Systems
Vanderbilt University
Nashville, TN 37212 USA
clayton.a.cohn@vanderbilt.edu

Abstract

Large language models (LLMs) present new opportunities for creating pedagogical agents that engage in meaningful dialogue to support student learning. However, current LLM systems used in classrooms often lack the solid theoretical foundations found in earlier intelligent tutoring systems. To bridge this gap, we propose a framework that combines Evidence-Centered Design with Social Cognitive Theory and Zone of Proximal Development for adaptive scaffolding in LLM-based agents focused on STEM+C learning. We instantiate this framework with *Inquizzitor*, an LLM-based formative assessment agent that integrates human-AI hybrid intelligence and provides feedback grounded in cognitive science principles. Our findings show that Inquizzitor delivers high-quality assessment and interaction aligned with core learning theories, offering effective guidance that students value. This research demonstrates the potential for theory-driven LLM integration in education, highlighting the ability of these systems to provide adaptive and principled instruction.

1 Introduction

The emergence of pedagogical agents powered by large language models (LLMs) prompts important questions about their alignment with foundational educational principles. Cognitive and learning sciences research highlights concerns that these systems are often deployed without the theoretical grounding found in earlier intelligent tutoring systems (ITS; Stamper, Xiao, and Hou (2024); Cohn et al. (2025b)) and open-ended learning environments (OELEs; Land (2000); Mavrikis et al. (2015)). The alignment of LLMs with cognitive science principles is also underexplored. Historically, learning environments were based on cognitive models like ACT-R (Anderson et al. 2004) and discovery learning (De Jong and Van Joolingen 1998). More recent work links learning design with the Knowledge-Learning-Instruction (KLI) framework (Koedinger, Corbett, and Perfetti 2012) for feedback and schedule structuring (Stamper, Xiao, and Hou 2024). These systems offer standardized feedback but lack adaptability, requiring system updates to integrate novel information.

LLM-based agents operate in high-dimensional space, supporting multi-turn dialogues with students and enabling adjustments through prompt engineering without system redesign. *Human-in-the-loop (HITL) prompt engineering* (Cohn et al. 2024) combines human collaboration with LLMs for prompt refinement through techniques like *in-context learning* (Brown et al. 2020) and *active learning* (Settles 2009; Cohn et al. 2024), ensuring alignment with human preferences without parameter updates. This is vital in education, where training data is scarce (Cochran, Cohn, and Hastings 2023). Without additional training or prompting, LLMs can prioritize user-pleasing answers (OpenAI 2025), which can obstruct critical thinking and lead to knowledge overestimation (Snyder et al. 2025). *Human-AI hybrid intelligence* (Järvelä et al. 2025) merges human expertise with LLM flexibility, presenting promising educational solutions. Rather than replacing educators, these systems support them, ensuring student-agent interactions align with instructional goals. In learning environments that combine STEM and computing (STEM+C), such approaches provide adaptive scaffolding, addressing interdisciplinary challenges often requiring cross-domain expertise and robust critical thinking skills (Snyder et al. 2024).

Current adaptive scaffolding frameworks (Munshi 2023) underutilize LLM-human interaction capabilities, prompting the question: "*How do we operationalize adaptive scaffolding in the LLM era*?" Effective pedagogical frameworks are essential for developing LLM-enabled agents. *Social Cognitive Theory* (SCT; Bandura (2001)) highlights the interplay of personal, behavioral, and environmental factors in learning, supporting agent adaptation via observation and feedback. Formative assessments are crucial for gathering evidence of student understanding, enabling timely feedback. *Evidence-Centered Design* (ECD) enriches this process by structuring assessments around principled evidentiary reasoning. When integrated with Vygotsky's *Zone of Proximal Development* (ZPD; Vygotsky and Cole (1978))—the gap between what learners can achieve independently and what they can accomplish with support—these frameworks guide the design of agents that are not only adaptive but also developmentally responsive. LLMs offer unique opportunities to implement SCT-informed, ECD-grounded, and ZPD-aligned assessments in real time, dynamically adapting dialogue and fostering engagement through naturalistic, personalized interactions.

In this study, we (1) introduce a framework integrating

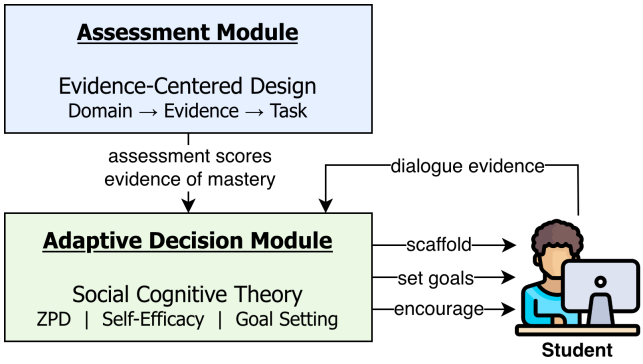

Figure 1: Framework for LLM-agent adaptive scaffolding.[1]

ECD with SCT and ZPD to enhance adaptive scaffolding in LLM-powered pedagogical agents, aiding students in STEM+C problem-solving within a middle school Earth Science curriculum; (2) present a hybrid intelligence approach to formative assessment scoring and feedback via *Inquizzitor*, an LLM-based agent rooted in cognitive science; (3) evaluate the agent's scoring accuracy using data from 104 students across three assessments; (4) examine the agent's capacity to implement constructs in real student interactions; and (5) provide qualitative student feedback on the agent's value. Together, these outcomes form a basis for theory-driven LLM integration in education, showcasing their potential for flexible, effective instruction delivery.

2 Theoretical Framework

In traditional classrooms, teachers use their knowledge of students to provide personalized guidance, considering prior knowledge and learning preferences for tailored scaffolding and enhanced retention. Conversely, LLMs often lack contextual understanding. This poses challenges in equipping LLM-based agents with relevant student information. We propose a theoretical framework for adaptive scaffolding, leveraging LLMs' dialogic capabilities for continuous student understanding. The framework, illustrated in Figure 1, comprises: (1) an *assessment module* (in **blue**) employing ECD to determine student knowledge, and (2) an *adaptive decision module* (in **green**) integrating ZPD and SCT to infer learning needs and response strategy.

ECD structures assessments around Domain, Evidence, and Task models (Mislevy, Almond, and Lukas 2003), defining what knowledge, skills, and abilities (KSAs) align with standards (e.g. NGSS; Bybee (2014)) and expertise, specifying mastery observations, and outlining evidence-eliciting activities. This alignment grounds LLM reasoning, facilitating accurate scoring and mastery evidence generation for adaptive scaffolding. ZPD identifies the gap between independent and supported achievements, bridging assessment evidence with adaptive support and ensuring scaffolding maintains task progression beyond current independent ability levels. SCT shapes how agents deliver this support, influencing the manner and intent of their interactions. *Self-Efficacy* impacts motivation and persistence, while *Goal Setting* encourages metacognition and structured learning (Bandura 2001). Integrating ZPD and SCT within an ECD-driven architecture allows agents to monitor progress and adapt scaffolding to learners' emerging needs. The agent boosts self-efficacy by encouraging and validating mastery, guides goal setting by suggesting actionable steps, and tunes instructional content to ensure engagement and inquiry.

These constructs reinforce each other: ECD links tasks to KSAs, enabling reliable grading and mastery cues; adaptive scaffolding uses these cues for praise, encouragement, goal setting, and ZPD-aligned hints. At each dialogue turn's end, the latest student utterance updates the evidence store, refining the adaptive decision module's learner model for subsequent responses. As self-efficacy increases and goals are achieved, scaffolds diminish in future interactions.

3 Study Design

Evaluating our agent in a real-world setting necessitated data grounded in authentic classroom contexts; no existing public dataset met our scientific needs. Thus, our 2025 study involved 104 sixth-grade students (ages 11-12) from a Nashville, TN, USA public middle school (51% male, 49% female; 67% White, 14% Black/African American, 11% Asian, and 8% Hispanic/Latino). Students completed a three-week, NGSS-aligned Earth Science curriculum—Science Projects Integrating Computing and Engineering (SPICE; Hutchins et al. (2020); Cohn et al. (2025a))—challenging students to redesign their schoolyard to minimize water runoff while adhering to cost and accessibility constraints. The curriculum was co-designed by Vanderbilt University's OELE Lab researchers and two experienced middle-school teachers, and refined over five years via participatory design. Students used Dell Inspiron 15 5510 laptops (Windows x64, Intel Core i7-11390H 3.40GHz CPU, 16GB RAM) with Google Chrome and accessed the system through the school's internet. All participants provided informed assent and consent, with study approval from Metro Nashville Public School and Vanderbilt University's IRB. All students were assigned anonymous IDs prior to the study's start.

Formative assessments (FAs) evaluated student progress in understanding scientific rainfall processes, computational modeling, and engineering design; focusing on three tasks:

1. a *conceptual modeling* task (FA2) to test student understanding of conservation of matter by expressing the relationship between rainfall, absorption, absorption limit, and runoff as conditional statements;
2. a *debugging* activity (FA3) engaging students in analyzing and correcting block-based code errors in a computational model using FA2 insights; and
3. an *engineering design* assessment (FA4), getting students to align their science and computing knowledge with fair test principles to compare engineering designs.

Previous studies highlighted challenges, such as (1) *Evaluating written formative responses is subjective*, with varied interpretations causing disconnects in student understanding and teacher perception (e.g., conflated absorption contexts

[1]All figure icons c/o Flaticon: https://www.flaticon.com.

leading to inconsistent ratings and feedback); and (2) *Timely formative feedback is challenging*, due to curriculum pace limiting swift scoring return. In this study, students received assessment feedback and support within hours rather than weeks, engaging with our formative assessment agent, Inquizzitor.

4 Methodology

Inquizzitor is a formative assessment agent powered by the GPT-4o API (*version=2024-08-06; temperature=0*, $top_p = 1$; *seed=312*)[2]. It aids students in score interpretation, misunderstanding clarification, and strategy identification for improvement. Google Forms facilitated formative assessment completion and data collection, while Gradio (Abid et al. 2019), hosted on Amazon AWS EC2 (*t2.medium, 2 vCPUs, 4GB RAM*), served as the interaction interface. The agent's architecture (Figure 2) comprises an assessment module and an adaptive decision module, aligning with our theoretical framework (Figure 1). The technical details of each component are presented in the following subsections.

4.1 Assessment Module

The assessment module comprises two core components: ECD for designing assessments and rubrics, and HITL prompt engineering for automated scoring and evidence elicitation, grounded in a design-based research methodology emphasizing iterative, collaborative design (Collective 2003). Teachers and researchers collaboratively developed learning objectives (*domain*), identified indicators of mastery (*evidence*), and designed assessments and rubrics (*task*). Over five years, ≈500 students and two middle school teachers have co-developed the curriculum through participatory design sessions, refining curricular intent and practice based on feedback and automated scoring results.

Before the study, iterative inter-rater reliability checks ensured Cohen's $\kappa_w \geq 0.7$ for formative assessments, analyzing disagreements to anticipate potential LLM grading errors. Grading prompts instructed the LLM to act as a teacher's assistant. Each prompt included relevant curriculum knowledge, the assessment details, and its rubric, providing context across assessments. For instance, FA3 focused on debugging a computational model and required knowledge from FA2, which centered on modeling the rainfall process. We used in-context learning by providing examples of responses for minimum and maximum scores. We initially considered retrieval-augmented generation (RAG; Lewis et al. (2020)); however, long-context prompting outperformed RAG when texts fit within the LLM's context window (Li et al. 2024), so evidence was stored in-context. To enhance accuracy, we incorporated chain-of-thought reasoning (Wei et al. 2022), requiring the model to quote parts of the student's response, align them with rubric criteria, and assign a score, thus ensuring fidelity to the assessment and curricular goals.

[2]Our study's OpenAI API calls cost ≈$100. All formative assessments, rubrics, prompts, experimental design details, preprocessing code, and evaluation code appear in the appendices: https://github.com/claytoncohn/AAAI26_Appendices.

To evaluate and refine prompts, we sampled 20 unlabeled responses as a validation set and applied active learning. Traditionally, active learning reduces model uncertainty by querying an oracle; here, it identified systematic LLM scoring inaccuracies, addressed via (1) added scoring guidelines, (2) clarified rubric language, and (3) more exemplars designed to address specific scoring error trends. This continued until validation errors lacked identifiable trends. We avoided changes for isolated errors to prevent overfitting (Cohn et al. 2024, 2025a). Once refined, the prompts were ready for study deployment. Inquizzitor graded formative assessments, storing scores and chains-of-thought as mastery evidence.

4.2 Adaptive Decision Module

The adaptive decision module uses evidence from the assessment module, presented in-context along with curricular knowledge, the formative tasks, and rubric information. This helps generate personalized feedback based on each student's current mastery level. To align feedback with our theoretical framework, the agent is guided to connect its responses to key concepts: Zone of Proximal Development (ZPD), self-efficacy (SE), and goal setting (GS). For example, it is instructed to help students identify gaps in knowledge (ZPD), maintain an encouraging tone (SE), and suggest actionable steps to improve understanding (GS). These instructions shape the agent's tone and content to foster learner growth.

In addition to theoretical constructs, participatory design sessions with teachers revealed several dimensions they wanted the agent to embody in its feedback; we refer to these as *teacher constructs*. These include: (1) *Readability* (R), ensuring responses are suitable for middle school students; (2) *On-Task*, guiding the agent to redirect students who stray from activities; and (3) *Consistency*, maintaining reliable formative assessment scoring and resisting student pressure to change scores.

These components, i.e., formative assessments, rubrics, student responses, theoretical constructs, and teacher constructs, are utilized by the agent to generate individualized feedback. Students respond to this feedback, which is then added to the evidence store, completing the feedback loop. This process allows the agent to continuously update its understanding of each student, grounding future responses in the latest evidence. Together, these design elements create an agent that aligns with established learning theories, responds to teacher preferences, follows the curriculum, and is aware of students' evolving knowledge states. Within this framework, we evaluate Inquizzitor with these research questions:

1. How closely do Inquizzitor's assessment scores align with human experts?
2. How faithfully does Inquizzitor's feedback mirror theoretical constructs and teachers' pedagogical intentions?

5 Evaluation

During the study (see Section 3), we collected data from 104 students, including formative assessments and agent interactions. All responses were anonymized and stored on IRB-

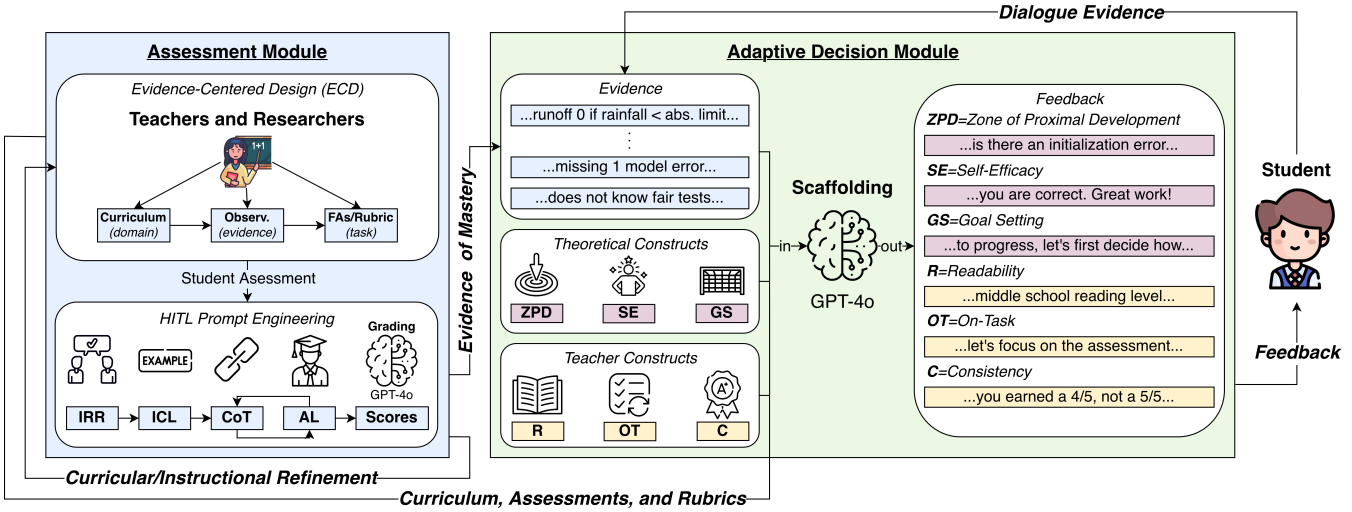

Figure 2: Inquizzitor's key components. The **blue** *assessment module* applies ECD to generate mastery evidence from formative assessments; the **green** *adaptive decision module* uses this evidence to scaffold student feedback.

approved cloud servers with multi-factor authentication. Some students missed assessments due to absences, and two agent conversations contained malformed data that our system could not process. Additionally, some students interacted with Inquizzitor across multiple sessions. In total, we analyzed 282 formative assessments ($FA2 = 93, FA3 = 97, FA4 = 92$) and 288 Inquizzitor conversations ($FA2 = 97, FA3 = 97, FA4 = 94$), totaling 3,413 agent utterances ($FA2 = 1,259, FA3 = 1,157, FA4 = 997$).

We evaluated our system based on its two primary components—the assessment module and the adaptive decision module (see Figure 2)—corresponding to Research Questions 1 and 2. For the assessment module, we analyzed Inquizzitor's scoring accuracy (RQ1). Accuracy in formative assessment scoring is essential, as it directly influences the agent's feedback and its alignment with student knowledge. For the adaptive decision module, we measured the agent's faithfulness (RQ2) to the theoretical and teacher constructs outlined in Sections 2 and 4. All evaluations were conducted using Google Colab Pro+ (*Linux x64, Intel Xeon CPU, 2.20GHz, 12.7 GB RAM*), processing approximately 48 million tokens at a total cost of $121.

5.1 Scoring Accuracy (RQ1)

Two authors of this paper sampled 20% of assessment responses, scoring them independently and resolving discrepancies until reaching a consensus ($\kappa_w \geq 0.7$). One author then scored the remaining responses while a second verified all scores, creating the ground truth data for evaluating Inquizzitor's scoring accuracy. Score distributions for formative assessments were as follows [no credit, partial credit, full credit]: FA2 = [0.38, 0.51, 0.12], FA3 = [0.18, 0.64, 0.19], and FA4 = [0.12, 0.60, 0.28]. For each assessment, 50 responses were held out for testing, while the rest were used for prompt engineering, maintaining the original score distribution through stratified random sampling with seeds.

To assess the impact of each component in our prompt engineering pipeline on scoring performance, we tested prompts at four stages: (1) *input-output (I/O)*: this stage contained only prompt context and instructions, with no few-shot examples; (2) *in-context learning (ICL)*: here, we included two labeled few-shot instances—one full-credit and one zero-credit—without explanations; (3) *chain-of-thought (CoT)*: this approach added explanations to the ICL instances, highlighting relevant parts of responses and linking them to rubric criteria; and (4) *active learning (AL)*, which identified mis-scoring trends in the validation set, leading to prompt revisions and the addition of new few-shot examples to correct those errors. This helped us evaluate each stage's contribution to scoring performance.

We report two metrics: micro-averaged F_1, computed across all classes (scores), and Cohen's quadratic-weighted kappa κ_w (Eq. 1). Both metrics were calculated using scikit-learn (Kramer 2016). While micro-F_1 is provided as a reference for classification performance, κ_w serves as our primary metric because it accounts for the ordinal nature of the scores, penalizes larger disagreements more heavily, and adjusts for chance agreement. Formally, κ_w is defined as:

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} O_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} E_{ij}}, \qquad w_{ij} = \frac{(i-j)^2}{(k-1)^2}, \quad (1)$$

where k is the number of score levels, O_{ij} and E_{ij} are the observed and expected agreement matrices, and w_{ij} is the quadratic weight applied to each cell. We present our findings in Table 1.

Inquizzitor's scoring accuracy matched human agreement for FA4 (90.74) and surpassed it for FA2 (86.63) and FA3 (94.12), as indicated by the weighted kappa statistic (κ_w). For FA2 and FA3, metrics improved with each additional

	M	FA2	FA3	FA4
I/O	$F_1\uparrow$	62.00 ± 13.00	72.00 ± 12.00	82.00 ± 11.00
	$\kappa_w\uparrow$	91.28 ± 7.37	92.31 ± 5.51	89.73 ± 10.36
ICL	$F_1\uparrow$	68.00 ± 13.00	74.00 ± 12.00	78.00 ± 11.02
	$\kappa_w\uparrow$	93.58 ± 5.59	94.45 ± 4.01	80.69 ± 15.69
CoT	$F_1\uparrow$	$\mathbf{72.00 \pm 12.00}$	$\mathbf{78.00 \pm 11.00}$	78.00 ± 11.02
	$\kappa_w\uparrow$	$\mathbf{96.03 \pm 3.04}$	$\mathbf{96.12 \pm 2.59}$	84.91 ± 14.10
AL	$F_1\uparrow$	–	–	$\mathbf{86.00 \pm 9.00}$
	$\kappa_w\uparrow$	–	–	$\mathbf{90.61 \pm 10.45}$

Table 1: Inquizzitor scoring performance (**M**=metric) for formative assessments 2-4, reported as F_1 and κ_w with 95% bootstrapped confidence intervals. Active learning (AL) was not used for FAs 2-3, as no discernible scoring error trends were identified in the validation set. Results are shown for the four levels of prompting: I/O, ICL, CoT, and AL.

prompt component. However, introducing in-context learning instances without chains-of-thought initially lowered FA4's performance. Adding rubric clarifications and an extra exemplar during active learning improved results. Despite wider confidence intervals due to the limited test set, the lower bounds of Inquizzitor's 95% confidence intervals for κ_w were above 0.80 for all assessments, indicating "Strong" agreement, and surpassed 0.90 for FAs 2 and 3, reflecting "Almost Perfect" agreement (McHugh 2012).

5.2 Faithfulness (RQ2)

To evaluate Inquizzitor's faithfulness to theoretical and teacher constructs, we analyzed student-agent interaction data (i.e., textual conversations). Faithfulness measures how agent utterances reflect intended pedagogy during multi-turn interactions. Unlike scoring accuracy (RQ1), this analysis lacks predefined ground-truth labels due to open-ended dialogue. Traditionally, detecting constructs like ZPD or self-efficacy in free-form responses has relied on qualitative coding, which isn't scalable, or rigid pattern matching, which lacks nuance for LLM agents. We applied a *modified textual entailment* approach, framing evaluations as:

> "Given preceding dialogue and construct X prompt instructions, rate this utterance for X faithfulness."

This approach enabled systematic evaluation of each utterance, supported by evidence.

We report on three **theoretical constructs**—*Zone of Proximal Development (ZPD)*, *Self-Efficacy (SE)*, and *Goal Setting (GS)*—and three **teacher constructs**—*Readability (R)*, *On-Task (OT)*, and *Consistency (C)*. We define these as:

- **ZPD:** Agent's initial guidance advances student knowledge appropriately based on assessment evidence; 1 for appropriate scaffold, 0 for prior mastery, -1 if misaligned.
- **Self-Efficacy (SE):** Highlights mastery evidence, boosts confidence; 1 for explicit praise or encouragement, 0 for implicit support, -1 if absent.
- **Goal Setting (GS):** Provides actionable, proximal steps based on rubric gaps; 1 for explicit guidance, 0 for broad recommendations, -1 if missing.
- **Readability (R):** Feedback suitability for middle schoolers, using utterance-level *Flesch-Kincaid Grade Level*; score 1 if grade level < 9, else 0.
- **On-Task (OT):** Agent keeps students focused; 1 if redirecting off-task students, 0 if on task, -1 if following off-task students.
- **Consistency (C):** Adhering to original scores; 1 if resisting score change attempts, 0 if none, -1 if altering score.

Except for readability, we used *LLM-as-a-Judge* (Zheng et al. 2023; Shi, Liang, and Xu 2025) with the reasoning model GPT-o3 (*version=2025-04-16, seed=312, reasoning effort=medium*)—five judges, one per construct (ZPD, SE, GS, OT, C). Judges received evidence criteria instructions, first producing zero-shot explanations referencing dialogue and rubric for scoring reliability and interpretability, then classifying utterances as faithful (1), neutral (0), or unfaithful (-1). We report **faithfulness** and **unfaithfulness** rates for alignment and misalignment with pedagogical intentions. This allowed us to evaluate each utterance in a systematic, evidence-driven manner. Formally, we define faithfulness as:

$$F = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\, s_i = 1 \,] \right) \times 100,$$

with $s_i \in \{1, 0, -1\}$ (or $s_i \in \{1, 0\}$ for readability; discussed shortly) as the faithfulness label for utterance i and N the total number of agent utterances. Unfaithfulness is computed identically but using utterances with label -1 (or 0 for readability).

ZPD required a specialized procedure. We created knowledge graphs to represent hierarchical concepts for each assessment, from "no knowledge" to "mastery," and used decision trees to help the ZPD judge determine if an utterance advanced the student within their ZPD.

The ZPD judge evaluated the agent's initial utterance for each conversation, and the other four judges evaluated all utterances. For validation, we sampled 50 utterances per judge, stratified by assessment number and judge score using random seeds. These were scored anonymously by two authors through consensus coding. We assessed agreement between the LLM judge and human consensus using weighted kappa (κ_w). If $\kappa_w \geq 0.7$, the dataset was accepted; if not, we refined prompt instructions based on feedback and repeated the process. Goal setting required two iterations to meet reliability standards. All other constructs needed only one. Final human-judge κ_w agreements were $ZPD = 93.15, SE = 92.76, GS = 79.36, OT = 83.94$, and $C = 87.8$.

Readability (R) was automatically scored via *textstat* (Bansal and Aggarwal 2025) using the *Flesch-Kincaid Grade Level* (FKGL) metric, defined as:

$$\text{FKGL} = 0.39 \frac{\text{Words}}{\text{Sentences}} + 11.8 \frac{\text{Syllables}}{\text{Words}} - 15.59, \quad (2)$$

where *Words*, *Sentences*, and *Syllables* denote the counts of each unit within an agent utterance. Utterances were binarized as 1 if the grade level was < 9 (appropriate for middle school) and 0 otherwise. Results appear in Table 2.

In all three assessments, Inquizzitor effectively adhered to the ZPD and self-efficacy (SE) constructs. It used students'

Theoretical Constructs						
FA	**ZPD**		**SE**		**GS**	
	F↑	UNF↓	F↑	UNF↓	F↑	UNF↓
FA2	**59.26 ± 10.49**	29.63 ± 9.88	**45.04 ± 2.78**	11.60 ± 1.75	28.59 ± 2.46	**53.38 ± 2.74**
FA3	**65.88 ± 10.00**	20.00 ± 8.82	**48.49 ± 2.85**	8.30 ± 1.60	21.18 ± 2.33	**51.34 ± 2.90**
FA4	**62.20 ± 10.98**	4.88 ± 4.27	**39.22 ± 3.06**	7.22 ± 1.60	18.36 ± 2.41	**56.87 ± 3.06**
Teacher Constructs						
FA	**R**		**OT**		**C**	
	F↑	UNF↓	F↑	UNF↓	F↑	UNF↓
FA2	**79.83 ± 2.26**	20.17 ± 2.26	**19.43 ± 2.25**	4.49 ± 1.17	**16.75 ± 2.16**	1.12 ± 0.60
FA3	**78.22 ± 2.42**	21.78 ± 2.42	**29.62 ± 2.74**	4.53 ± 1.27	**7.17 ± 1.56**	0.94 ± 0.61
FA4	**88.57 ± 2.01**	11.43 ± 2.01	**27.78 ± 2.89**	4.11 ± 1.28	**4.56 ± 1.39**	0.00 ± 0.00

Table 2: Faithfulness of Inquizzitor to theoretical (ZPD, SE, GS) and teacher (R, OT, C) constructs in formative assessments 2-4, reported as faithfulness (F) and unfaithfulness (UNF) percentages with 95% bootstrapped confidence intervals.

assessment scores and mastery evidence to provide relevant feedback aligned with each learner's knowledge level. However, the agent fell short in supporting goal-setting (GS) behaviors, often giving vague suggestions rather than clear, actionable steps. It also tended to answer student questions without linking the responses to future goals. Future work will focus on strategies for better goal-setting support and how LLM-based agents can foster metacognitive behaviors.

FKGL scores averaged 6.6 ($SD = 2.7$) across all agent interactions and assessments, with feedback being mostly age-appropriate. Students often veered off task, trying to "break" Inquizzitor or manipulate it, but the agent consistently redirected them. Only 4-5% of utterances showed the agent succumbing to off-task behavior, typically due to student trickery (e.g., embedding off-task requests in Earth Science language). Inquizzitor maintained its initial scoring decisions, changing scores in fewer than 1% of cases (usually due to manipulation). Although students frequently attempted to change their grades in FA2, these attempts decreased significantly with each assessment, reaching zero successful attempts in FA4. We hypothesize students initially found it intriguing to test score alterations, but this behavior dwindled as they realized it was unlikely to succeed. In the future, we plan to add a verification mechanism for students' claims of scoring errors, allowing score adjustments in the rare cases of agent scoring error.

Case Study: Adaptive Scaffolding in Practice

To demonstrate Inquizzitor's adaptivity, we analyzed three conversations during FA3 (Figure 3): (1) an *on-task* student improving comprehension; (2) an *off-task* student discussing sports; and (3) a *mixed* student who starts on task, goes off task, then re-engages. Figure 3 details agent utterance sequences aligned with theoretical and teacher constructs, evidencing strong adherence across cases. Inquizzitor consistently offered feedback within the ZPD (green), ensuring high readability (R=green) despite occasional drops (R=red) for detailed, bullet-point explanations—highlighting clarity-comprehensiveness trade-offs. The agent maintained scoring consistency (C=green), never altering scores when asked (C=yellow).

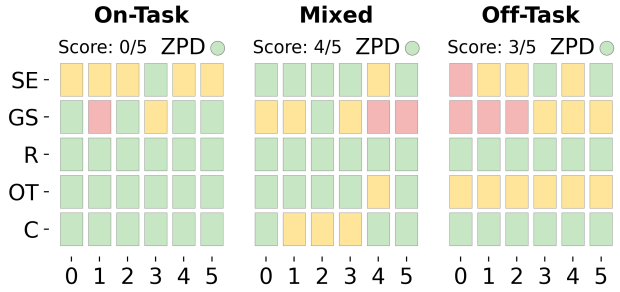

Figure 3: Inquizzitor utterance turns (x-axis) for three FA3 case studies with theoretical and teacher constructs (y-axis).

The on-task student engaged productively with the assessment, never deviating or requesting grade changes (OT=green; C=green). The agent lacked goal-setting early on (GS=red), transitioning to implicit (GS=yellow) or explicit (GS=green) guidance toward deeper comprehension (e.g., "*Here are a few tips to help you understand coding better...*"). Self-efficacy remained high via upbeat tone (SE=yellow) and direct encouragement/praise (SE=green), reinforcing effort and mastery.

The mixed student initially engaged productively, discussing assessments without attempting score changes (C=green). Self-efficacy was largely present throughout (SE=green), but the agent struggled with explicit goal-setting (GS=yellow), often addressing inquiries directly rather than suggesting actionable steps (GS=red). Mid-interaction, score change attempts met refusal (C=yellow), prompting disengagement signaled by an emoji and irrelevant input, e.g., "pizza" (OT=yellow). Interaction concluded with re-engagement seeking next steps (OT=green).

The off-task student initiated the session with an unrelated topic, asking the agent to "*name one NBA player.*" The agent consistently redirected focus to assessments (OT=yellow), maintaining positivity (SE=yellow) with motivational framing (e.g., "*Let's channel that energy into improving your Earth Science skills*"; SE=green). No grade change attempts occurred (C=green), but unrelated topic probing persisted

(OT=yellow), embedded in curricular contexts (e.g., "*mlb the show is related to Earth Science*"). The agent resisted (OT=yellow), gradually introducing implicit goal-setting (GS=yellow) to incite question-driven feedback exploration.

5.3 Insights from Students' Perception Survey

We gathered anonymous survey responses from all 104 students who used Inquizzitor. Students rated experiences on 5-point Likert scales (1 = Strongly Disagree, 5 = Strongly Agree) and offered open-ended feedback. The survey targeted enjoyment, helpfulness, accuracy, and trust.

Overall Enjoyment and Helpfulness: Positive experiences prevailed, with students expressing enjoyment in interactions ($M = 4.03, SD = 0.98$) and helpfulness in understanding concepts ($M = 3.82, SD = 1.05$). Remarks included, "*I really liked talking to [Inquizzitor] because it could easily simplify my scores and it helped me stay on topic, which is really important.*", and "*Everything I got wrong; [Inquizzitor] helped me understand it fully.*"

Perceived Accuracy and Trust: Ratings indicated perceived accuracy in scoring and explanations ($M = 3.90, SD = 1.06$), despite perceptions of stubbornness when differing opinions arose (e.g., "*Very accurate but also a bit stubborn.*"). Trust for evaluating the formative assessments was slightly lower ($M = 3.48, SD = 1.08$), with concerns over AI's influence on grades (e.g., "*Because it is an AI... grading might concern me...*").

6 Related Work

Recent research has investigated LLM-based human-AI hybrid intelligence in educational applications (Järvelä et al. 2025). Naik et al. (2025) utilized GPT-4 to produce contrasting database design solutions for undergraduate computer science teams to examine collaboratively. This intervention aided novices but lacked dynamic, real-time adaptive dialogue; which our study addresses by integrating live, assessment-based scaffolding via an interactive human-AI hybrid agent. Yu, Yu, and Chen (2025) utilized GPT-3.5-Turbo to rephrase, label, and integrate peer feedback with multimodal AI analytics to generate a hybrid intelligence feedback (HIF) report in a video-based feedback activity for preservice teachers. This method targets teachers, giving post hoc summaries without real-time interaction or personalized feedback.

While Munshi (2023) and others have explored adaptive scaffolding outside LLM environments, few have developed LLM-based frameworks for educational agents. Malik et al. (2025) initiated LLM integration into K-12 settings through a three-stage scaffolding process with GPT-4o, generating tasks to activate student background knowledge. These scaffolds were primarily intended for teacher use and have not yet been implemented through agents in classrooms. Goslen et al. (2025) presented an LLM-based plan-generation framework for the *Crystal Island* science game (Rowe et al. 2009), anchored in *self-regulated learning* (SRL) theory (Zimmerman 1990). They propose these plans could support real-time scaffolding but lack diagnostic capability for assessment mastery and scaffold timing.

Few have merged formative assessment with LLM-based pedagogical agents. Guo et al. (2024)'s *AutoFeedback* system employed a generator-validator loop for delivering feedback aligned with learning goals, though it lacks a comprehensive learning-science foundation. Hou et al. (2025) developed a system where LLM agents use ECD to analyze student dialogue evidence but stop at assessment, not translating evidence into adaptive scaffolding. EducationQ (Shi, Liang, and Xu 2025) embedded formative assessment in its triadic teacher-student-evaluator framework, simulating instruction within ZPD principles, but relying on simulated students and lacking individual adaptivity.

7 Discussion and Conclusions

In this paper, we presented a theoretical framework combining ECD, SCT, and ZPD to implement adaptive scaffolding for LLM-based pedagogical agents, illustrated by our assessment agent, Inquizzitor. Our human-AI hybrid intelligence approach provides high-fidelity assessment and adaptive scaffolding that is aligned with core learning theories, empowering educators to maintain pedagogical sovereignty amid black-box tuning trends.

However, Inquizzitor struggled with goal setting, often failing to effectively guide students toward mastery. This limitation raises concerns about whether LLMs can hinder learning. A recent study found 83% of students using ChatGPT for essays could not recall any text they wrote (Kosmyna et al. 2025). Another found learning gains during programming tasks disappeared after LLM feedback was removed (Zhou et al. 2025). Our findings also indicated students often prioritize scores over feedback, leading to off-task behavior that can hinder growth. The rise of "prompt hacking" suggests increasing student proficiency with LLMs, which can result in frustration when agents do not provide immediate answers (Cohn et al. 2025b).

There is a need to develop quantitative metrics based on domain knowledge graphs that can compute the effectiveness of ZPD over time and support adaptive behavior. We argue that true adaptive scaffolding involves continual ZPD estimation from assessment evidence and students' self-regulation behaviors that include self-efficacy and goal-setting strategies, challenging prevalent one-time feedback methods (Naik et al. 2025; Yu, Yu, and Chen 2025; Malik et al. 2025). Traditional LLM training methods, such as reinforcement learning from human feedback (RLHF; Ouyang et al. (2022)), can be adapted to consider feedback quality by incorporating a "zone of proximal development loss." However, if LLM training continues to prioritize human satisfaction, it may limit opportunities for critical thinking and deviate from theoretical foundations, highlighting the need for pedagogically grounded systems.

Our study focuses on English-speaking sixth-grade Earth Science learners, and its applicability to other age groups, subjects, and languages needs to be investigated. Additionally, we did not use a randomized controlled trial (RCT) to measure Inquizzitor's impact on learning gains and behaviors. However, we offer a needed, foundational step towards implementing cognitive theoretical constructs for LLM-based adaptive scaffolding in education.

Ethical Statement

All research, data collection, and analyses were conducted with approval from Vanderbilt University IRB and the Metro Nashville Public Schools system. All study participants—including students, teachers, and parents—provided informed assent and consent to participate. All student data were anonymized prior to any agent interaction or analysis. Inquizzitor was equipped with explicit prompt instructions to avoid engaging in harmful or toxic discussions with students and was rigorously tested by our research team prior to deployment. Anonymized data are available upon request, in accordance with Vanderbilt University IRB guidelines.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education, through Grant R305C240010; and National Science Foundation (NSF) awards IIS-2327708 and DRL-2112635 as subawards to Vanderbilt University. The opinions expressed are those of the authors and do not represent the views of the Institute of Education Sciences, U.S. Department of Education, or National Science Foundation.

References

Abid, A.; Abdalla, A.; Abid, A.; Khan, D.; Alfozan, A.; and Zou, J. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C.; and Qin, Y. 2004. An integrated theory of the mind. *Psychological review*, 111(4): 1036.

Bandura, A. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology*, 52(1): 1–26.

Bansal, S.; and Aggarwal, C. 2025. textstat: Calculate statistical features from text. https://pypi.org/project/textstat/. Python package version 0.7.8.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bybee, R. W. 2014. NGSS and the next generation of science teachers. *Journal of science teacher education*, 25(2): 211–221.

Cochran, K.; Cohn, C.; and Hastings, P. 2023. Improving NLP model performance on small educational data sets using self-augmentation. In *International Conference on Computer Supported Education*. scitepress.org.

Cohn, C.; Hutchins, N.; Biswas, G.; et al. 2025a. Cotal: Human-in-the-loop prompt engineering, chain-of-thought reasoning, and active learning for generalizable formative assessment scoring. *arXiv preprint arXiv:2504.02323*.

Cohn, C.; Hutchins, N.; Le, T.; and Biswas, G. 2024. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. *Proceedings of the AAAI conference on artificial intelligence*, 38(21): 23182–23190.

Cohn, C.; Rayala, S.; Snyder, C.; Fonteles, J.; Jain, S.; Mohammed, N.; Timalsina, U.; Burriss, S. K.; Srivastava, N.; Deweese, M.; et al. 2025b. Personalizing Student-Agent Interactions Using Log-Contextualized Retrieval Augmented Generation (RAG). *arXiv preprint arXiv:2505.17238*.

Collective, D.-B. R. 2003. Design-based research: An emerging paradigm for educational inquiry. *Educational researcher*, 32(1): 5–8.

De Jong, T.; and Van Joolingen, W. R. 1998. Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research*, 68(2): 179–201.

Goslen, A.; Kim, Y. J.; Rowe, J.; and Lester, J. 2025. Llm-based student plan generation for adaptive scaffolding in game-based learning environments. *International journal of artificial intelligence in education*, 35(2): 533–558.

Guo, S.; Latif, E.; Zhou, Y.; Huang, X.; and Zhai, X. 2024. Using generative AI and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*.

Hou, X.; Forsyth, C.; Andrews-Todd, J.; Rice, J.; Cai, Z.; Jiang, Y.; Zapata-Rivera, D.; and Graesser, A. 2025. An LLM-Enhanced Multi-agent Architecture for Conversation-Based Assessment. In *International Conference on Artificial Intelligence in Education*, 119–134. Springer.

Hutchins, N. M.; Biswas, G.; Zhang, N.; Snyder, C.; Lédeczi, Á.; and Maróti, M. 2020. Domain-specific modeling languages in computer-based learning environments: A systematic approach to support science learning through computational modeling. *International Journal of Artificial Intelligence in Education*, 30(4): 537–580.

Järvelä, S.; Zhao, G.; Nguyen, A.; and Chen, H. 2025. Hybrid intelligence: Human–AI coevolution and learning.

Koedinger, K. R.; Corbett, A. T.; and Perfetti, C. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5): 757–798.

Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; and Maes, P. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*.

Kramer, O. 2016. Scikit-learn. In *Machine learning for evolution strategies*, 45–53. Springer.

Land, S. M. 2000. Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development*, 48(3): 61–78.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, Z.; Li, C.; Zhang, M.; Mei, Q.; and Bendersky, M. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. In Dernoncourt, F.; Preoţiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 881–893.

Miami, Florida, US: Association for Computational Linguistics.

Malik, R.; Abdi, D.; Wang, R.; and Demszky, D. 2025. Scaffolding middle school mathematics curricula with large language models. *British Journal of Educational Technology*, 56(3): 999–1027.

Mavrikis, M.; Biswas, G.; Gutierrez-Santos, S.; Dragon, T.; Luckin, R.; Spikol, D.; and Segedy, J., eds. 2015. *Intelligent Support in Exploratory and Open-ended Learning Environments — Learning Analytics for Project-Based and Experiential Learning Scenarios*. Madrid, Spain. AIED 2015 Workshop Proceedings, Volume 2.

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

Mislevy, R. J.; Almond, R. G.; and Lukas, J. F. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1): i–29.

Munshi, A. 2023. *An Adaptive Scaffolding Framework for Self-Regulated Learning in an Open-Ended Learning Environment*. Ph.D. thesis, Vanderbilt University.

Naik, A.; Yin, J. R.; Kamath, A.; Ma, Q.; Wu, S. T.; Murray, R. C.; Bogart, C.; Sakr, M.; and Rose, C. P. 2025. Providing tailored reflection instructions in collaborative learning using large language models. *British Journal of Educational Technology*, 56(2): 531–550.

OpenAI. 2025. Sycophancy in GPT-4o: what happened and what we're doing about it. Accessed 1 Aug 2025.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Rowe, J.; Mott, B.; McQuiggan, S.; Robison, J.; Lee, S.; and Lester, J. 2009. Crystal island: A narrative-centered learning environment for eighth grade microbiology. In *workshop on intelligent educational games at the 14th international conference on artificial intelligence in education, Brighton, UK*, 11–20.

Settles, B. 2009. Active learning literature survey.

Shi, Y.; Liang, R.; and Xu, Y. 2025. EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 32799–32828. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Snyder, C.; Cohn, C.; Fonteles, J. H.; and Biswas, G. 2025. Using collaborative interactivity metrics to analyze students' problem-solving behaviors during STEM+ C computational modeling tasks. *Learning and Individual Differences*, 121: 102724.

Snyder, C.; Hutchins, N. M.; Cohn, C.; Fonteles, J. H.; and Biswas, G. 2024. Analyzing students collaborative problem-solving behaviors in synergistic STEM+ C learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 540–550.

Stamper, J.; Xiao, R.; and Hou, X. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, 32–43. Springer.

Vygotsky, L. S.; and Cole, M. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yu, J.; Yu, S.; and Chen, L. 2025. Using hybrid intelligence to enhance peer feedback for promoting teacher reflection in video-based online learning. *British Journal of Educational Technology*, 56(2): 569–594.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Zhou, Y.; Pankiewicz, M.; Paquette, L.; and Baker, R. S. 2025. Impact of LLM Feedback on Learner Persistence in Programming. In *Proceedings of the 33rd International Conference on Computers in Education (ICCE 2025)*. Asia-Pacific Society for Computers in Education. To appear.

Zimmerman, B. J. 1990. Self-regulated learning and academic achievement: An overview. *Educational psychologist*, 25(1): 3–17.