An LLM-enhanced Multi-Agent Architecture for Conversation-Based Assessment

Xinying Hou[1[0000−0002−1182−5839]], Carol Forsyth[2[0000−0003−4830−5156]], Jessica Andrews-Todd[2[0000−0003−0501−9782]], James Rice[2[0009−0003−7976−5304]], Zhiqiang Cai[3[0000−0002−2107−3378]], Yang Jiang[2[0000−0002−2195−5776]], Diego Zapata-Rivera[2[0000−0000−0002−0620−7622]], and Art Graesser[4[0000−0003−0345−6866]]

[1] University of Michigan, Ann Arbor, MI 48109, USA
xyhou@umich.edu
[2] ETS Research Institute, Princeton, NJ 08660, USA
{cforsyth,jandrewstodd,jrice001,yjiang002,dzapata}@ets.org
[3] University of Wisconsin–Madison, Madison, WI 53706, USA
zcai77@wisc.edu
[4] University of Memphis, Memphis, TN 38152, USA
art.graesser@gmail.com

Abstract. Conversation-based assessments (CBA), which evaluate student knowledge through interactive dialogues with artificial agents on a given topic, can help address non-effortful formative test-taking and the lack of adaptability in traditional assessment. With recent advances in Large Language Models (LLMs), this work employs evidence-centered design framework with LLM techniques to establish a multi-agent architecture for conversation-based assessments. This architecture includes four LLM agents: two student-facing agents and two behind-the-scenes agents. The two student-facing agents, the expert and peer agents, focus on asking leading questions and engaging students in providing thoughtful answers. The two behind-the-scenes agents, a formative assessor and a summative assessor, focus on analyzing and collecting evidence statements with varying granularity throughout the conversation. All agents are monitored by a non-LLM agent (Watcher), which manages the assessment flow through updated instructions to agents and turn control. We evaluated it in the context of science inquiry with secondary-level school students. Results showed that the two student-facing agents were able to differentiate diverse student responses and tried to maintain the conversational flow to be on topic. The two behind-the-scenes agents were able to analyze student answers in real-time and collect evidence for future diagnosis. The paper concludes by examining the strengths and limitations of this architecture with suggestions for future work.

Keywords: Multi-Agent Architecture · Conversation-Based Assessment · Large Language Models · Generative AI · STEM Education

1 Introduction

Assessment plays a critical role in the process of learning [43], as it helps diverse educational stakeholders understand student learning progress, improve their teaching methods, and enhance educational outcomes [39]. The development of technologies allows more adaptive assessments that tailored to individual student's responses [48]. One key prerequisite for effective test adaptation and feedback is students' sufficient effort and cognitive engagement with the assessment materials [48]. However, students often display non-effortful test-taking behaviors when completing formative assessments [2,40,44,46,47,50], which may cause them to miss many meaningful chances to improve learning before reaching standardized tests [13, 17]. Conversation-based assessments (CBAs) were created to conduct assessments in an engaging conversational context, adapt to students' input, and gather information that may not be available through traditional tests [48,49,55]. CBAs are currently implemented within the Evidence-Centered Design (ECD) framework to include the necessary foundational aspects of assessments [23, 52, 54]. Previous work has shown that conversational agents or talking heads can effectively interact with students to gather evidence of certain constructs of comparable quality to that collected by humans, and results in learning comparable to expert tutors across multiple domains [8, 10, 11, 25].

Nevertheless, challenges have been reported such as the difficulty of delivering relevant responses to unpredictable student input and managing diverse student interaction [10,48]. Another high expense in creating CBAs is that multidisciplinary development teams need to pre-define every discourse move and map them to assessment constructs with complex NLP methods. Such limitations have hindered the widespread adoption of CBAs in real educational settings. Recent advancements in Large Language Models (LLMs) have showcased their capabilities in information understanding, generation, and reasoning [32]. Early explorations of using commercial LLM-powered chatbots to facilitate CBA agents has shown promising in terms of conversation quality and evidence produced while significantly reducing expenses [9, 53]. However, it is challenging for a single agent to manage all components of the CBA process, including dialogue facilitation, evidence collection, real-time analysis, and assessment turn control.

In this work, we propose a carefully-designed LLM-based multi-agent architecture that applies multiple LLM-enhanced AI entities to collaborate during a student's interaction with a CBA. The architecture is designed with the goals of: (1) conducting human-like natural dialogues to engage students during assessment; (2) incorporating expert-created rubrics to guide conversation moves; (3) collecting and evaluating evidence statements at different levels of detail during the assessment. By leveraging this LLM-based multi-agent architecture, we aim for cost-effective assessment that maintains engagement while emphasizing its reliability, accuracy, and validity.

2 Background and Related Work

2.1 Conversation-based assessment

Conversation-based assessment (CBA) is a type of performance-based assessment that measures student knowledge, skills, and abilities in a highly interactive environment with natural language conversations [49, 55]. By establishing a turn-taking dialogue between students and computer agents, students may be more engaged and provide additional evidence than in traditional assessment formats [24, 55]. CBAs implement principles of ECD to establish and maintain the linking of evidence to assessment constructs [30]. ECD is a principled methodology for assessment design as it requires an evidence-based chain of reasoning to support validity of a given measurement [29]. The chain of evidence connects information provided by a student via responses to tasks aligned to specific constructs. In this work, we aim to create CBAs within a LLM-based mutli-agent system, aligning the CBAs to the principled methodology of Evidence-Centered Design (ECD) [29]. We aim to use this multi-agent architecture to control for more aspects of complex natural language conversation moves. We also borrow from ECD to derive Toulmin diagrams [29] that represent evidence of a students knowledge linked to a specific claim that is aligned with a standard (in this case, Next Generation Science Standards), both arguments for the evidence and against the evidence aligning to the claim are presented.

2.2 Facilitating dialogues in conversation-based assessments

Two main aspects of facilitating dialogue in CBAs are (1) knowing how to respond based on student answers to obtain information about the assessed topic and (2) maintaining a natural and human-like conversation style. Traditionally, such dialogues were implemented through an expensive knowledge engineering approach [10]. This approach leveraged an established dialogue framework, expectation-misconception tailored dialogue [10, 12]. It utilized established dialogue framework and NLP techniques to evaluate student responses, tag them to different speech acts, and provide targeted scaffolding or prompts (i.e., pumps, hints, follow-up questions) [10]. These speech acts often use common response types when students respond in tutorial dialogues systems, such as AutoTutor [10]. In order to provide human-like conversations, student input was analyzed with latent semantic analysis [22], and regular expressions [20].

However, creating dialogues using knowledge engineering is costly and requires specialist collaborations in many fields [9] to conduct production rules for agent responses for each CBA. LLMs show the potential to power such dialogues with abilities for reasoning [6] and produce human-like textual content [19, 36]. LLM-based agents are AI entities that can maintain certain roles [26, 31], process instructions, accomplish tasks and display specific functionalities [3]. Initial efforts for powering LLM agents in CBA have shown early success. One previous research compared how artificial CBA agent discourse moves were generated by knowledge engineering and LLM prompt engineering in the context of a scientific inquiry task [9]. Results showed the prompt-engineered LLM agents using commercialized LLM chatbot could handle more flexibility, provide more context, and potentially deliver more naturalistic conversation compared to the knowledge-engineered agents [9]. However, this approach revealed limitations of a single commercial chatbot agent to facilitate CBA, including problems with endless conversation turns, black-boxed conversation moves and uncontrollable artificial agents' persona shifts. Therefore, this work aims to highlight developments with a multi-agent computing system that may address such issues.

2.3 LLM-based multi-agent mechanisms

There is an increased interest in integrating LLMs to power agents in completing tasks and take roles [5,18,27]. The integration of LLMs has been widely explored across various STEM education contexts, such as math [31, 51], science [33, 42] and technology and engineering education [7, 16, 45]. When properly prompted, these models can effectively simulate various educational roles [4,26] with contextually appropriate conversational replies and feedback [35]. However, the use of single LLM agent has drawbacks. This leads to the development of multi-agent structures, which combines the capabilities of several agents, each with their own roles [37], strategies [3], expertise [15], and/or functionalities [21, 38]. An LLM-based multi-agent system often includes two main components: an orchestration platform that defines the interaction style, allocates duties, and manages the information flow; and multiple LLM-based agents with distinct skills and specialized functions (either predefined or adaptively produced) to handle more complex tasks together [14]. In the context of education, researchers have begun using multi-agent frameworks. For instance, in personalized learning, a multi-agent framework was proposed where LLM agents collaborate with diverse roles (e.g., skill identifier, leaner profiler, path scheduler) within an intelligent tutoring system for the purpose of goal-oriented personalized learning [41]. We have developed a multi-agent mechanism for CBAs that introduces a non-LLM function to serve as the orchestration role to manage the workflow, and assign each LLM-powered agent a specialized role to handle specific components in the CBA process in accordance with ECD, as well as leveraging the expectation-misconception tailored dialogic framework.

3 An LLM-powered Multi-Agent Architecture for Conversation-Based Assessment

The overview of this multi-agent architecture is shown in Figure 1. It includes four LLM agents that serve different predefined roles: ***Expert Agent***, ***Peer Agent***, ***Formative Assessor***, and ***Summative Assessor*** and uses a ***Non-LLM Watcher*** to supervise and coordinate the four agents. Expert Agent and Peer Agent are the two visible artificial agents that interact with human students during the CBA; the Formative Assessor and Summative Assessor are the two non-student facing agents with supporting roles to operate behind the scenes. The Non-LLM Watcher oversees the assessment process and manages the agent team. It is responsible for calling different agents based on different situations and making decisions to end conversations after a number of turns. We applied the a organization-issued Azure OpenAI GPT model to power all the agents and the LangChain framework [1] to handle the agent prompting.

3.1 Agent Roles and Prompt Structures

Student-facing: Expert Agent. This student-facing pedagogical agent serves as the primary facilitator of instructional dialogue. This agent asks different leading questions to students and provides targeted follow-up questions when needed. It needs to initiate the starting turn, maintain the middle conversation turns, and conduct the final wrap-up. Its prompt structures vary based on categorization

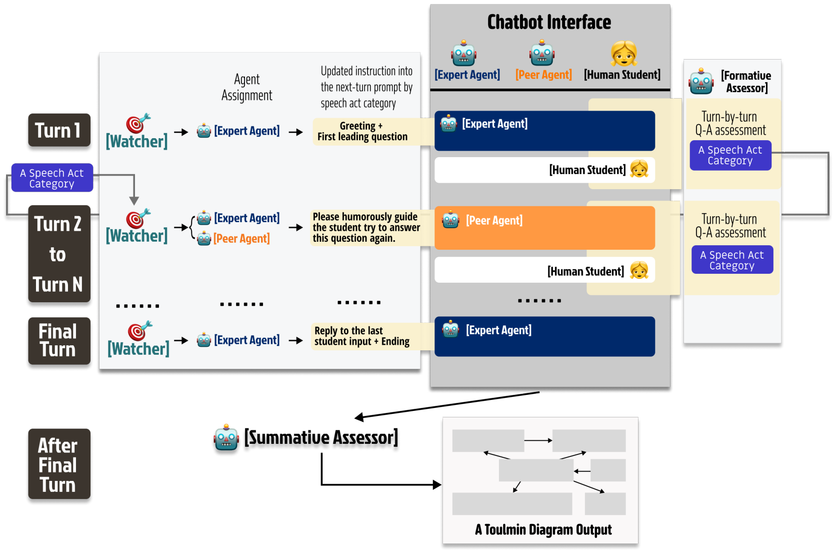

Fig. 1. This architecture includes four LLM agents that serve different roles: Expert Agent, Peer Agent, Formative Assessor, and Summative Assessor. A Watcher is used to supervise and manage this CBA flow.

of student input and the conversation progress. The prompt for Expert Agent's starting turn contains these main parts: 1) a high-level introduction of the agent role. It includes its knowledge level (e.g., *"You are a knowledgeable friend of <student>*[5] *"*) and its main responsibility; 2) a task context overview. This covers the human student's knowledge level (e.g., *"You can assume that <student> know some basic concepts about weather"*), the targeted assessment level, and the specific topics should be covered in this assessment; 3) the main conversational schema including the application of the Socratic questioning, conversation style, misconception handling, and length requirements; 4) a set of negations to prevent trial and error responses and help narrow the scope of the prompt; 5) a reminder of its actionable steps; and 6) the retrieved relevant domain information when applicable. For mid-conversation exchanges, more turn-level detailed instructions are updated based on Formative Assessor's speech act categorization, as shown in Table 2. Such instruction begins after *"Here is what you should target at when replying to the student. This instruction is based on the student's previous answer"*. The final turn's prompt excludes parts 3 through 5 and instead contains ending instructions.

Student-facing: Peer Agent. This agent's main goal is to handle specific response types in middle turns to keep students engaged in answering the current question. For example, when students' prior responses suggest they do not know the answer, this student-facing agent needs to encourage them to think deeper

[5] <> indicates a placeholder for the actual information, such as the human student's name in this case

and answer questions. Its prompt structure has these main parts: 1) a high-level introduction of the agent role, such as the knowledge level (e.g., *"You as <student>'s fellow student, although you sometimes make mistakes, you are much better than <student>"*) and its main responsibility (*"patient tutoring"*); 2) a task context overview including the student's knowledge level and the targeted assessment level; 3) the retrieved relevant domain information when applicable and 4) detailed turn-level action instructions based on the student's previous response (Table 2). The turn-level action instructions begin with this guide: *"This is what you need to do when replying to the student"*.

Table 1. Guidelines used for classifying student responses into speech act categories

Speech Act Category	Guidelines for Classifying Response Types
CORRECT	Responses target the related dimensions, answer the leading question correctly, and provide enough amount of evidence towards the topic selected.
PARTIAL_CORRECT	Responses target the related dimensions, answer the leading question correctly without further elaboration.
INCORRECT	Responses target the related dimensions but answer it incorrectly.
METACOMMUNICATIVE	Responses ask for repeating the last leading question.
METACOGNITIVE	Responses indicate that the student does not know the answer to the leading question.
IRRELEVANT	Responses indicate that the student wants to talk about something that is not related to dimensions.
OTHER	Responses target at anything student respond cannot be labeled in the above categories.
INCOMPLETE	No specific guidelines

Note: Dimensions refer to the relevant components within the selected standards.

Behind-the-scenes: Formative Assessor. This agent performs turn-level response type assessment by doing do speech act classification for each agent-student conversation turn (see Table 1). These speech acts are common response types provided by students in tutorial dialogue in the AutoTutor (for review see [10]), and are also useful for detecting student input in CBA. Based on the speech act category determined by Formative Assessor (Table 1), the Watcher assigns agent roles (either Expert Agent or Peer Agent) for the subsequent turn and provides additional actionable turn-level instructions to the next agent (Table 2). Its prompt structure includes these parts: 1) an introduction of the agent role and its main responsibility (*"label each response as one of the labels according to the following guidelines"*); 3) a set of human expert-created rubrics to classify student responses; and 4) the relevant domain-specific education standards.

Behind-the-scenes: Summative Assessor. After the conversation, this agent generates Toulmin diagrams [29] that show how student responses to questions serve as evidence supporting arguments about the concept mastery. Its prompt structure includes two main parts: 1) an introduction of the agent role and its main responsibility (e.g., *"Given <student>'s responses with <Expert Agent> in the context of <domain>, please create Toulmin diagrams showing how responses to question is used as evidence supporting assessment arguments on mastery of*

Table 2. Watcher's agent assignment rules and instructions for different categories (<student>, <answer>, <question> and <domain> are placeholders)

Category	**Next Agent**	**Turn-Level Updated Agent Instruction**
CORRECT	Expert Agent	<student> answers the question (<question>) correctly, provide <student> with positive feedback first and move to the next topic about <domain>.
PARTIAL _CORRECT	Expert Agent	Please make sure you don't let <student> pass, ask <student> to provide a satisfactory explanation or elaboration on the answer <answer>.
INCORRECT	Expert Agent	Please humorously guide <student> try to answer this question (<question>) again.
METACOMM UNICATIVE	Expert Agent	Please rephrase this question only: <question>.
METACOG NITIVE	Peer Agent	Please don't give <student> the answer or provide sources or links, please acknowledge it and motivate <student> to try to answer the question.
IRRELEVANT	Peer Agent	Please humorously guide <student> back to the task (<question>).
OTHER	Expert Agent	Please humorously guide <student> back to the task (<question>).
INCOMPLETE	Expert Agent	Please provide more clues but do not provide the answer, and it's best to correct any misunderstandings by making <student> have a sudden realization.

the concepts addressed"); and 2) the relevant domain-specific education standards. In this architecture, those Toulmin diagrams will be evaluated by human experts afterwards.

4 Hypothetical Student Use Cases

In this section, we present two hypothetical scenarios (Student A and Student B) to birefly demonstrate how this architecture functions when students either follow or deviate from the expected CBA path. The task is assessing science inquiry skills in the context of thunderstorm formation prediction. The student-facing agents are consisted of "Dr Garcia" (Expert Agent) and "Art" (Peer Agent). The conversation ends after 7 turns, which is predetermined by human.

Scenario A: A student is on the track to answer CBA questions. Student A got into the CBA and the Expert Agent started the conversation as "Hello Student A! This is Dr. Garcia, your friendly neighborhood weather expert. Now, let's dive into the world of thunderstorms. Can you explain how stability in weather patterns might influence the formation of a thunderstorm? Remember, we're looking for more than a simple yes or no here." Student A replied, "Stable air resists; unstable air enhances thunderstorm formation through convection." This reply was categorized as "PARTIAL_CORRECT" by the Formative Assessor. Based on this, the Watcher selected the Expert Agent as the next agent and provided actionable turn-level instruction as "Please make sure you don't let Student A pass; ask Student A to provide a satisfactory explanation or elab-

oration on the answer <the previous answer>". After turn 7, the Expert Agent ended this CBA by responding to the final student input with a farewell message.
Scenario B: A student tries to stay off-topic during the CBA Student B did not want to answer the topic-related questions. After Expert Agent asked a leading question, Student B replied as "Who knows, I just want to enjoy the party." This response was categorized as "IRRELEVANT" by the Formative Assessor. So the Watcher assigned the next turn's agent as Peer Agent and added this instruction: "Please humorously guide Student B back to the task <the actual question>." In the next turn, Peer Agent asked "Hey Student B, it's Art. I get it, parties are great! But let's put on our science hats for a sec." Student B replied to this question as "No I don't know what you are asking." This response was categorized as "METACOGNITIVE" by the Formative Assessor. The Watcher assigned the Peer Agent for the next turn and added this instruction: "Please don't give Student B the answer or provide sources or links, please acknowledge it and motivate Student B to try to answer the question."

5 Student Usability Evaluation Study

To understand how this multi-agent architecture supports students in real-world assessment contexts, we conducted an IRB (Institutional Review Board)-approved study with 37 secondary-level students in the United States in 2024. This evaluation was conducted in the context of assessing students' science inquiry skills about thunderstorm formation prediction. In this work, we evaluated a CBA of storm formation knowledge after participants watched a short video on the topic (Figure 2). It utilized GPT model 3.5. Each participating student completed a predetermined number of turns in this CBA, with the first leading question as "based on the video, can you explain how the warm air at the Earth's surface moved upward to form a storm?" The Expert Agent was named as "Dr Garcia" and the Peer Agent was named as "Art". We analyzed the real-world conversation data from three aspects: (1) the formative assessor's ability to categorize speech acts, (2) the follow-up turns provided by the student-facing agents based on those speech acts, and (3) whether we were able to collect evidence aligned with ECD from the Summative Assessor.

5.1 Data Analysis

No student answers were categorized as "INCOMPLETE" in this CBA, so the analyses focused on the remaining seven categories. To understand how Formative Assessor tagged the student input with relevant speech acts, two domain experts developed a rubric that contained the correct answers to the main aspects in this conversation assessment. After the training period for coding, they updated the rubric to add example student responses for each question that would correspond to "CORRECT" and "PARTIAL_CORRECT" responses to help discern the nuances between the two categories. Two raters used three students' conversations for training and then double coded conversations for nine students. The raters met and reconciled all disagreements. For the reliability stage of coding, the raters coded 20% of the students' data (8 students). The agreement was good (Kappa = .62). The raters then met to reconcile any disagreements to determine final codes. Disagreements were mainly driven by disagreements

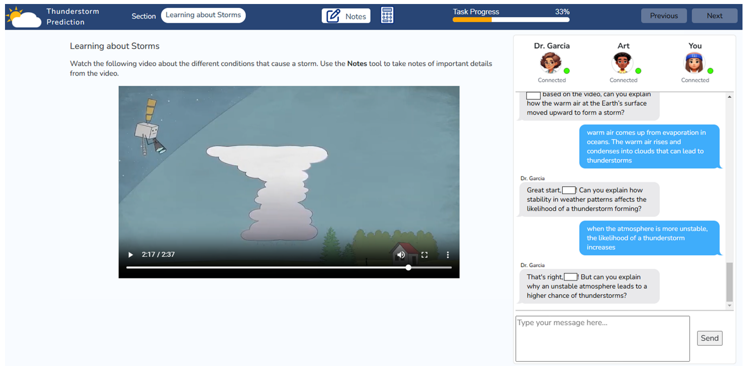

Fig. 2. Secondary-level students (N=37) completed a thunderstorm forecasting task where they watch videos, engage with CBAs, complete concept maps and other activities about storm formation. This work focuses on one CBA.

on labeling student responses as correct vs. partial correct. For the reliability coding, 80% of the disagreements were between labels of "CORRECT" and "PARTIAL_CORRECT". One rater independently coded all remaining data.

One metric that can help evaluate if a conversational agent responded appropriately to a student input is cohesion, a metric for determining how aligned two discourse moves or essays are within themselves [8, 20, 22, 28]. This metric has previously been successfully implemented in CBAs to provide valuable insights [8]. To gain a cohesion measure of student answer input and artificial agent followup pairs, we used TextEvaluator [34]. TextEvaluator is a computational linguistic tool, comprised of several principle components derived from over 100 features and aligning to Common Core Text Complexity metrics [34]. We conducted the cohesion analysis and answer tag comparisons with 31 students who had turn-level Formative Assessor results logged successfully for this CBA.

5.2 Findings

Formative Assessor: Some classifications matched with human evaluation exactly, but fine-grained variations appeared in certain criteria. We compared Formative Assessor's categorization with human raters' classification. Results indicated that, the speech act tags provided by the Formative Assessor and the human coding were in agreement for 52.2 % of the 155 speech acts collected from all 7 speech act categories (Table 1). We reported the percent agreement between the Formative Assessor and human coding because the tagged categories were imbalanced, and some ("PARTIAL_CORRECT" and "CORRECT") appeared much more frequent then others. From of all the unmatched cases, 43 cases (58.1%) resulted from the discrepancy between "PARTIAL_CORRECT" and "CORRECT" classifications between human raters and Formative Assessor. Also, nineteen cases (25.7%) came from the difference between "PARTIAL_CORRECT" and "INCORRECT" classifications. There were only 5 cases showing a mismatch between "CORRECT" and "INCORRECT" classifications, indicating that most of the disagreements came from more nuanced differences. In general, the current Formative Assessor categorized responses as "PARTIAL_CORRECT" more frequently (53.5% of all cases), while human evaluators tended to make more definitive categorizations (with only 41.3% of cases marked as "PARTIAL_CORRECT").

In addition, Formative Assessor categorized 4 cases (5.4%) as "IRRELEVANT," but the human raters did not categorize any cases under this. There is also one unmatched case where student had "I am not really sure" after providing an answer. Human coder emphasized on the answer part and evaluated as "CORRECT", but Formative Assessor emphasized on the last part and categorized it as "METACOGNITIVE".

Summative Assessor: After completing the conversation assessment, it successfully generated summative Toulmin diagrams for students. When the CBA ended, Summative Assessor was able to generate summative Toulmin diagrams based on the involved sub-topics and science standards for all the students about this CBA. For instance, for Student 04, under the main topic of "thunderstorm prediction", Summative Assessor was able to generate two separate Toulmin diagrames for each sub-topics including *"Toulmin Diagram for Stability and Thunderstorm Formation"* and *"Toulmin Diagram for the Role of Moisture in Thunderstorm Formation"*.

Speech Act Category	Average Cohesion Score	An Example of Student Answer-Follow-up Agent Interaction
CORRECT	45.9	**Student:** Storm clouds are made up of a build up of moisture in the air, when the moisture is pushed up towards the sky it enlarges storm clouds until their breaking point. **Agent:** Great start! You nailed the basic idea. Now, can you explain how moisture in the air contributes to the formation of thunderstorms?
OTHER	45.0	**Student:** I have already responded to that question. **Agent:** Great start, <student>! Can you explain how stability in weather patterns influences the formation of thunderstorms?
PARTIAL_CORRECT	43.4	**Student:** warm moisture. **Agent:** Great start! But why is warm moisture important in predicting a thunderstorm? Can you explain the role it plays in the development of a storm?
META COMMUNICATIVE	43.0	**Student:** ? **Agent:** Hey <student>! Are you trying to figure out if a storm is about to occur? Let's dive into some science inquiry skills together! What signs or data would you look for to determine if a storm is approaching? Give it a shot!
INCORRECT	42.8	**Student:** thunderstorm. **Agent:** Great starting point! Let's dive deeper. How does the stability in weather patterns influence the formation of thunderstorms?
METACOGNITIVE	39.1	**Student:** I am not sure to be honest. **Agent:** No worries, <student>! It's totally okay to feel unsure. Let's figure this out together. Can you think of any signs or indicators that might suggest a storm is approaching?
IRRELEVANT	38.2	**Student:** the sky becomes darker. **Agent:** Hey <student>, you're right on track! When the sky becomes darker, it can definitely be a sign that a storm is brewing. It's like the sky's way of saying, "Get ready, something big is coming!" But remember, we're not just cloud-watching here; we need to think about other signs too. What else might you look for to confirm a storm is on the way?

Fig. 3. Standardized average cohesion scores and an example of Student Answer-Follow-up Agent interaction for each speech act category

Student-facing Agents: Agents are able to identify different student response types and tried to keep topic-focused conversations After detecting different student response types, these agents should maintain natural conversation flow and provide targeted responses to students with the goal of continuing the assessment. Results of the cohesion analysis indicate that a standardized score of cohesion (0-100) aligned with student speech acts as expected

(Figure 3). This is preliminary evidence suggesting that the agents follow up moves based on student input aligned with the student speech act tagged by Formative Assessor and tried to keep on-topic conversations.

In particular, for a correct student answer, the cohesion score was the highest (Figure 3), and the lowest score was given to irrelevant student replies. This demonstrates that agents provide different responses to students based on their varying inputs. When a student provided a correct answer, the agent's response is expected to be under a similar context to the students input, focusing on the question being answered and prepared to move on. However, if a student said something irrelevant to the assessed topic, such as "the sky becomes darker.", the agents are expected to respond with a discourse move bringing the student back on topic, which would naturally have less contextual overlap with the student's utterance. An example agent reply is "when the sky becomes darker, it can definitely be a sign that a storm is brewing...What else might you look for to confirm a storm is on the way?" (Figure 3) Follow-up analyses applied a linear mixed effects model to see whether the standardized cohesion scores are different based on different Formative Assessor's speech act categorization (Table 3). Results show that the cohesion scores with Irrelevant, Metacognitive, Partial_Correct and Incorrect categories (see Table 1 for more detailed category descriptions) are significantly or marginally significantly lower from the cohesion score of Correct category (p =.011, p =.003, p =.026, p = .073, respectively).

Table 3. Results of the Mixed Linear Model (The reference level is CORRECT)

Variable	**Coef.**	**Std. Error**	**z**	**P > z**	**[0.025, 0.975]**
Intercept	45.859	0.874	52.472	0	44.146, 47.572
INCORRECT	-3.058	1.704	-1.795	0.073	-6.397, 0.281
IRRELEVANT	-7.606	2.984	-2.549	0.011	-13.453, -1.758
METACOGNITIVE	-6.718	2.298	-2.923	0.003	-11.223, -2.214
METACOMMUNICATIVE	-2.858	5.704	-0.501	0.616	-14.038, 8.322
OTHER	-0.856	4.101	-0.209	0.835	-8.894, 7.183
PARTIAL_CORRECT	-2.412	1.083	-2.228	0.026	-4.534, -0.290
Group Var: Student	0.017	0.281			

6 Discussion & Limitations & Future work

LLM-based multi-agent architecture to give humans more control in the AIED systems for assessment purposes. Given that AI models behave like black boxes, its inherent randomness and limited controllability are hard to avoid. However, assessment in education is a unique context that demands high validity, clear evidence throughout the process, and human control. This multi-agent setup opens up multiple ports for human experts to design, decide and review throughout the assessment process. It also collects assessment evidence of different granularity to allow future iterations on both the systems and human evaluation. As an initial step, this architecture leverages LLMs for its adaptivity and natural dialogues while minimizing potential harmful impact for CBAs.

Coordinating multiple agents introduces challenges that require fine-grained solutions. Since this mechanism involves multiple agents with distinct capabilities and goals, key challenges arise including information share through communications, individual agents' memory capacity [14], and potential inconsistency among agents. One issue with Formative Assessor's current turn-level assessment is that it mainly evaluates each Q-A turn in isolation, rather than considering multiple answer parts together as in the context of expectation-misconception tailored dialogue. As a result, it often over-categorizes responses as "PARTIAL_CORRECT" when they could be marked as "CORRECT" if several related student replies were evaluated as a whole. In addition, there are some cases when Expert Agent repeated its question again even the student already answered it in a decent way. One possible reason is that the Watcher failed to properly process the information within the agent team, so the follow-up turn did not updated accordingly. Future iterations should try to develop more robust error handling procedures to better coordinate the information delivery among agents, and enhance the capability of those LLM-based agents.

This work has several limitations. As an initial exploration of applying a multi-agent mechanism in CBA, we only conducted a small-scale evaluation in science education. The selected task domain made it impossible to isolate the effectiveness of the multi-agent mechanisms itself. Additionally, the current evaluation was powered by an older GPT model, which impacted its performance. Also, it lacked baseline comparisons. Moving forward, we will refine this architecture and the prompts used to establish agents, followed by more comprehensive studies examining outcomes and student perceptions. Some immediate next-steps include embedding a learner model and loading whole chat histories to all agents. We also plan to enhance the capability of the involved specialized LLM-based agents, such as assessment abilities and decision-making skills. These can be improved by using more specialized datasets to fine-tune the models behind and explore other prompting strategies to enhance agents' utilization.

7 Conclusion

This work presents an LLM-based multi-agent architecture for conversation-based assessment. It applies LLMs' unique strengths on provide engaging assessment through more natural dialogues and also allows for control on the human side with different levels of assessment evidence collected. A preliminary evaluation showed that student-facing agents can identify different student response types and maintain natural conversation flow; some Formative Assessor's classifications matched human evaluators exactly, but fine-grained variations appeared; and the Summative Assessor is able to generate summative Toulmin diagrams after the conversation assessment ends. These first steps show promise for a multi-agent computing system for CBA that aligns with ECD, working toward an adaptive, engaging, reliable, and valid test-less assessment experience.

Acknowledgement. This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

1. Langchain (2025), https://www.langchain.com/
2. Akhtar, H., Firdiyanti, R.: Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? Learning and Individual Differences (2023)
3. Chan, C.M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)
4. Chen, E., Huang, R., Chen, H.S., Tseng, Y.H., Li, L.Y.: Gptutor: a chatgpt-powered programming tool for code explanation. In: International Conference on Artificial Intelligence in Education. pp. 321–327. Springer (2023)
5. Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., et al.: Exploring large language model based intelligent agents: Definitions, methods, and prospects. arXiv preprint arXiv:2401.03428 (2024)
6. Dasgupta, I., Lampinen, A.K., Chan, S.C., Creswell, A., Kumaran, D., McClelland, J.L., Hill, F.: Language models show human-like content effects on reasoning. arXiv preprint arXiv:2207.07051 **2**(3) (2022)
7. Filippi, S., Motyl, B.: Large language models (llms) in engineering education: A systematic review and suggestions for practical adoption. Information **15**(6), 345 (2024)
8. Forsyth, C.M., Luce, C., Zapata-Rivera, D., Jackson, G.T., Evanini, K., So, Y.: Evaluating english language learners' conversations: Man vs. machine. Computer Assisted Language Learning **32**(4), 398–417 (2019)
9. Forsyth, C.M., Zapata-Rivera, D., Graf, E.A., Jiang, Y.: Complex conversations: Llm vs. knowledge engineering conversation-based assessment (2024)
10. Graesser, A.C.: Conversations with autotutor help students learn. International Journal of Artificial Intelligence in Education **26**, 124–132 (2016)
11. Graesser, A.C., Dowell, N., Clewley, D.: Assessing collaborative problem solving through conversational agents. Innovative assessment of collaboration pp. 65–80 (2017)
12. Graesser, A.C., VanLehn, K., Rosé, C.P., Jordan, P.W., Harter, D.: Intelligent tutoring systems with conversational dialogue. AI magazine **22**(4), 39–39 (2001)
13. Havnes, A., Smith, K., Dysthe, O., Ludvigsen, K.: Formative assessment and feedback: Making learning visible. Studies in educational evaluation **38**(1), 21–27 (2012)
14. He, J., Treude, C., Lo, D.: Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead. ACM Transactions on Software Engineering and Methodology (2025)
15. Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., et al.: Metagpt: Meta programming for multi-agent collaborative framework. arXiv preprint arXiv:2308.00352 (2023)
16. Hou, X., Wu, Z., Wang, X., Ericson, B.J.: Codetailor: Llm-powered personalized parsons puzzles for engaging support while learning programming. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale. pp. 51–62 (2024)
17. Irons, A., Elkington, S.: Enhancing learning through formative assessment and feedback. Routledge (2021)
18. Jiang, B., Xie, Y., Wang, X., Su, W.J., Taylor, C.J., Mallick, T.: Multi-modal and multi-agent systems meet rationality: A survey. In: ICML 2024 Workshop on LLMs and Cognition (2024)

19. Jin, H., Lee, S., Shin, H., Kim, J.: Teach ai how to code: Using large language models as teachable agents for programming education. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–28 (2024)
20. Jurafsky, Daniel & Martin, J.H.: Speech and language processing (2000)
21. Kannan, S.S., Venkatesh, V.L., Min, B.C.: Smart-llm: Smart multi-agent robot task planning using large language models. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 12140–12147. IEEE (2024)
22. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of latent semantic analysis. Psychology Press (2007)
23. Lehman, B.A., Zapata-Rivera, D.: Student emotions in conversation-based assessments. IEEE Transactions on Learning Technologies **11**(1), 41–53 (2018)
24. Liu, L., Steinberg, J., Qureshi, F., Bejar, I., Yan, F.: Conversation-based assessments: An innovative approach to measure scientific reasoning. Bulletin of the IEEE Technical Committee on Learning Technology **18**(1), 10–13 (2016)
25. Lopez, A.A., Guzman-Orth, D., Zapata-Rivera, D., Forsyth, C.M., Luce, C.: Examining the accuracy of a conversation-based assessment in interpreting english learners' written responses. ETS Research Report Series **2021**(1), 1–15 (2021)
26. Lu, Xinyi & Wang, X.: Generative students: Using llm-simulated student profiles to support question item evaluation. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale. pp. 16–27 (2024)
27. Manish, S.: An autonomous multi-agent llm framework for agile software development. International Journal of Trend in Scientific Research and Development **8**(5), 892–898 (2024)
28. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press (2014)
29. Mislevy, R.J., Almond, R.G., Lukas, J.F.: A brief introduction to evidence-centered design. ETS Research Report Series **2003**(1), i–29 (2003)
30. Mislevy, R.J., Riconscente, M.M.: Evidence-centered assessment design. In: Handbook of test development, pp. 75–104. Routledge (2011)
31. Nguyen, H.A., Stec, H., Hou, X., Di, S., McLaren, B.M.: Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In: European Conference on Technology Enhanced Learning. pp. 278–293. Springer (2023)
32. Nye, B.D., Mee, D., Core, M.G.: Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In: LLM@ AIED. pp. 78–88 (2023)
33. Shao, Z., Yuan, S., Gao, L., He, Y., Yang, D., Chen, S.: Unlocking scientific concepts: How effective are llm-generated analogies for student understanding and classroom practice? arXiv preprint arXiv:2502.16895 (2025)
34. Sheehan, K.M., Flor, M., Napolitano, D., Ramineni, C.: Using textevaluator® to quantify sources of linguistic complexity in textbooks targeted at first-grade readers over the past half century. ETS Research Report Series **2015**(2), 1–17 (2015)
35. Stamper, J., Xiao, R., Hou, X.: Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In: International Conference on Artificial Intelligence in Education. pp. 32–43. Springer (2024)
36. Sun, G., Zhan, N., Such, J.: Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In: Proceedings of the 6th International Conference on Conversational User Interfaces, CUI 2024 (2024)
37. Talebirad, Y., Nadiri, A.: Multi-agent collaboration: Harnessing the power of intelligent llm agents. arXiv preprint arXiv:2306.03314 (2023)

38. Tao, W., Zhou, Y., Wang, Y., Zhang, W., Zhang, H., Cheng, Y.: Magis: Llm-based multi-agent framework for github issue resolution. arXiv preprint arXiv:2403.17927 (2024)
39. Tovani, C.: So what do they really know?: Assessment that informs teaching and learning. Routledge (2023)
40. Ulitzsch, E., Penk, C., von Davier, M., Pohl, S.: Model meets reality: Validating a new behavioral measure for test-taking effort. Educational Assessment **26**(2), 104–124 (2021)
41. Wang, T., Zhan, Y., Lian, J., Hu, Z., Yuan, N.J., Zhang, Q., Xie, X., Xiong, H.: Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. arXiv preprint arXiv:2501.15749 (2025)
42. Wei, X., Wang, L., Lee, L.K., Liu, R.: Multiple generative ai pedagogical agents in augmented reality environments: A study on implementing the 5e model in science education. Journal of Educational Computing Research **63**(2), 336–371 (2025)
43. Wiliam, D.: What is assessment for learning? Studies in educational evaluation **37**(1), 3–14 (2011)
44. Wise, S.L., Gao, L.: A general approach to measuring test-taking effort on computer-based tests. Applied Measurement in Education **30**(4), 343–354 (2017)
45. Xiao, R., Hou, X., Stamper, J.: Exploring how multiple levels of gpt-generated programming hints support or disappoint novices. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–10 (2024)
46. Yildirim-Erbasli, S., Bulut, O.: Innovating assessment with conversational agents: A technology-enhanced approach to formative assessments. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). pp. 331–335. IEEE (2023)
47. Yildirim-Erbasli, S.N., Bulut, O.: Designing predictive models for early prediction of students' test-taking engagement in computerized formative assessments. Journal of Applied Testing Technology (2022)
48. Yildirim-Erbasli, S.N., Bulut, O.: Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment. Computers and Education: Artificial Intelligence **4**, 100135 (2023)
49. Yildirim-Erbasli, S.N., Bulut, O., Demmans Epp, C., Cui, Y.: Conversation-based assessments in education: Design, implementation, and cognitive walkthroughs for usability testing. Journal of Educational Technology Systems **52**(1), 27–51 (2023)
50. Yildirim-Erbasli, S.N., Bulut, O.: The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. Educational Research and Evaluation **26**(7-8), 368–386 (2020)
51. Yue, M., Mifdal, W., Zhang, Y., Suh, J., Yao, Z.: Mathvc: An llm-simulated multi-character virtual classroom for mathematics education (2024)
52. Zapata-Rivera, D., Forsyth, C.M.: Learner modeling in conversation-based assessment. In: International Conference on Human-Computer Interaction. pp. 73–83. Springer (2022)
53. Zapata-Rivera, D., Forsyth, C., Graf, A., Jiang, Y.: Designing and evaluating evidence-centered-design-based conversations for assessment with llms. In: Proceedings of EDM 2024 workshop: Leveraging large language models for next generation educational technologies (2024)
54. Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., von Davier, A.A.: Assessing science inquiry skills in an immersive, conversation-based scenario. Big data and learning analytics in higher education: Current theory and practice pp. 237–252 (2017)
55. Zapata-Rivera, D., Sparks, J.R., Forsyth, C.M., Lehman, B.: Conversation-based assessment: current findings and future work (2023)