Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs

Diego Zapata-Rivera, Carol Forsyth, Edith Aurora Graf and Yang Jiang

[1] ETS, Rosedale Road, Princeton, NJ 08541 USA

Abstract

The current paper discusses conversation-based assessments (CBAs) created with prompt engineering for LLMs based on Evidence-Centered Design (ECD). Conversation-based assessments provide students the opportunity to discuss a given topic with artificial agent(s). These conversations elicit evidence of students' knowledge, skills and abilities that may not be uncovered by traditional tests. We discuss our previous method of creating such conversations with regular expressions and latent semantic analysis in an expensive methodology requiring time and various expertise. Thus, in this novel work, we created a prompt-engineered version of CBAs based on evidence-centered design that remains on the domain topic throughout the conversation as well as provides evidence of the student knowledge in a less expensive way. We present the methodology for creating these prompts, compare responses to various student speech acts between the previous version and the prompt engineered version, and discuss the evidence gleaned from the conversation and based on the prompt. Finally, limitations, conclusions and implications of this work are discussed.

Keywords

Conversation-based Assessment, Large Language Models, Dialogue Systems, Evidence-Centered Design, AutoTutor

1. Introduction

Advances in Artificial Intelligence are reducing the design and development complexity that is usually required when using dialogue systems in educational contexts. These advances can have a positive impact on the adoption and scalability of conversation-based learning and assessment activities. Conversation-based assessments (CBAs) have been explored as innovative mechanisms to assess skills in a natural context. In particular, we consider skills that can be assessed in a conversational context such as argumentation, scientific inquiry, language skills, and collaboration.

Researchers have explored the use of Large Language Models (LLMs) for a variety of purposes including evaluating LLM generated hints to human-created ones [18] and creating dialogue-based tutoring interactions with students [7]. This paper explores the use of LLMs and Evidence-Centered Design (ECD) [15] in the creation and evaluation of conversation for assessment purposes (e.g., formative assessment) as a mechanism to gather evidence of students' knowledge, skills, and other attributes. ECD offers a principled methodology for assessment design as it provides an approach to explicitly represent an evidence-based chain of reasoning, with the goal of supporting assessment validity. This evidence-based chain connects responses to particular tasks to the constructs that are assessed. We discuss a use case that illustrates the types of conversations produced by the LLM approach compared to a previous approach that relies on regular expressions/RegExp [9] and Latent Semantic Analysis (LSA) [2]. We elaborate on trade-offs of applying these two approaches to generating conversations for assessment purposes including methods for evaluating them.

2. Conversation-based assessments

Conversation-based assessments (CBAs) build on innovations in areas such as conversational agents and dialogue systems for improving student learning [1]2][4-6][8][10][12][14]][16][20]. These conversations between human students and artificial agents are a mechanism to gather evidence of students' knowledge, skills, and other attributes following an evidence-centered approach [15]. CBAs have been used to assess communication skills (e.g., English language skills, science inquiry skills, and mathematical argumentation) in formative contexts [25].

CBAs have been leveraged to gather additional explanations about students' decisions in technology-rich environments involving interactive simulations and interactions with virtual agents and other students. They provide students with multiple opportunities to elaborate on their responses. For example, virtual agents can rephrase questions and ask students to provide additional information, if necessary.

The design and development process of CBAs involves defining the construct and the type of evidence that is expected to be elicited by the conversation (see Figure 1). To create natural situations to elicit the evidence, aspects of the conversation scene need to be defined (e.g., context of the conversation, main question, conversation moves, response categories and types of interactions) thus creating the scene. In addition to this information, the conversational dialogue and scoring model is created. Specifically, construct information, conversation paths based on user response categories resulting in differing responses by artificial agents, closing statements, and partial scoring rules are documented in *conversation diagrams* that are used as a communication mechanism among various members of the

Leveraging Large Language Models for Next Generation Educational Technologies, July 14, 2024, Atlanta, Georgia, USA

dzapata@ets.org (D. Zapata-Rivera); cforsyth@ets.org (C. Forsyth); agraf@ets.org (E. Aurora Graf); yjiang002@ets.org (Y. Jiang)

team including dialogue and assessment developers (See Figure 2). The script development and testing of these conversations were designed, developed, and tested using additional authoring tools, automated testing techniques, and data collected with the conversation prototype administered to the intended audience via cognitive labs, wizard of oz studies, pilot studies, and crowdsourcing efforts [23][25].

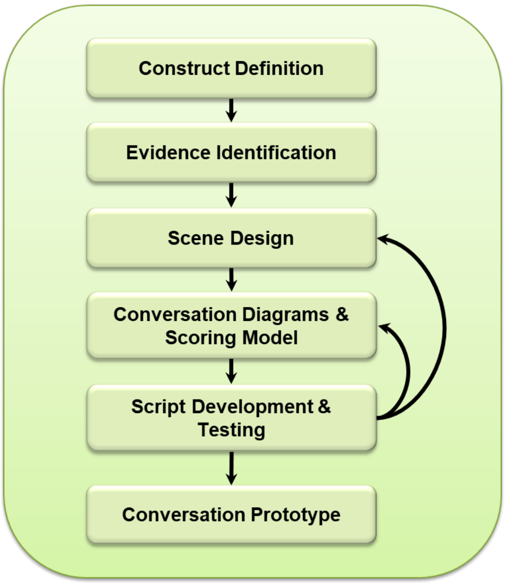

Figure 1. CBA design and development process. *Reprinted from *Authoring Conversation-based Assessment Scenarios", [18]. Copyright by Educational Testing Service, 2014 All rights reserved.

During the development process, conversation diagrams are of particular importance in this process as they capture many of the elements needed for designing ECD-based conversations (see Figure 2). Starting with the definition of the construct, the *opening and main question*, including the introduction of the character(s) and their roles (e.g., student vs. teacher) that will provide the information, user response categories handled by the system, closing statements and partial scores for the constructs involved. The response categories and how artificial agent(s) react to them are based on AutoTutor's framework [6] referred to as Expectation-Misconception Tailored Dialogue. In this framework, there is always a main question posed that has a complete expected answer that is pre-programmed. If the student is in some way not able to fulfill the requirements of the correct answer, then the agent begins by providing a pump such as "can you tell me more" and after this answer, launches into a series of hints (open-ended clues) and then prompts (clues asking for a single word or phrase) followed by an assertion. If at any point a student states a misconception, the agent corrects this particular misconception. If the student provides the correct answer at any point, the system provides an assertion (restating the correct answer) and moves on to the next question. CBAs are very similar but provide less information in follow up scaffolding moves such as pumps, hints, and prompts. Additional speech acts have been defined as common during the tutoring process with agents [5]. However, the nature of the CBAs vs. Tutorial dialogues is quite different because the goal is to not give away the answer but rather probe the student to provide more information about what he/she/they already know. For example, assertions are often not given as the goal is to elicit more evidence within CBA's but are in tutoring. With this in mind, it is very important how the agent(s) respond to common speech act categories of the human student in a CBA. Nevertheless, the identification of these speech acts is integral in determining an appropriate response. Some of these categories include:

- Correct response (with an appropriate explanation and/or evidence). A correct response typically includes some key elements of the expected answer to a main question posed by the artificial agent. The "correctness" is often defined by experts in the particular domain at hand.
- Incomplete response: An incomplete or partial response usually includes some portion of the correct answer, but critical elements are still lacking. To address this issue, if at any point in the conversation the student gives such a response, the artificial agent may provide the above given scaffolding moves in attempts to get the student to give the complete response. The conversations take note of students' previous answers. Therefore, say for example a student provides one part of the answer early in the conversation and then the rest of the answer separately after a hint, the system puts the two answers together and considers the student's answer correct.
- Irrelevant response: An irrelevant response is a completely off topic remark to any of the scaffolding moves posed by the agent. In these instances, the artificial agent provides a response such as "this is not relevant to our conversation" and then continuing on with a closing statement and the next follow-up scaffolding move.
- Meta-communicative response: A meta-communicative response is a common discourse move that students say during tutoring [5]. An example is when a student asks the artificial agent or tutor to please repeat the question. When this occurs, the question can be repeated or rephrased.
- Meta-cognitive response: A meta-cognitive response is when the student says something similar to "I don't know", another common discourse move in tutoring [5]. When this happens, the artificial agent can acknowledge the situation and attempt to get the student to say anything by providing scaffolding and introducing the next discourse move.
- No response: In the event the student does not respond at all, the system is prepared to acknowledge the lack of a response and move forward and provide the next scaffolding moves. The wait time is usually determined based on iterative refinement from

gathering data from real students and highly dependent on the length of the required response.

Conversation diagrams were converted into XML scripts that were implemented with AutoTutor Script Authoring Tool for Assessment – (ASATA) to run the conversations including both the human input and NLP and the output of the agent speech. This system makes use of LSA and RegExp to parse students' responses and determine how to react to them. The final result were CBAs which included one or two virtual agents interacting with the student and lasted between 2-5 turns. This process took weeks for conversations to go from the initial descriptions to c*onversation diagrams* to actual script conversations implemented into ASATA and tested with user data [23].

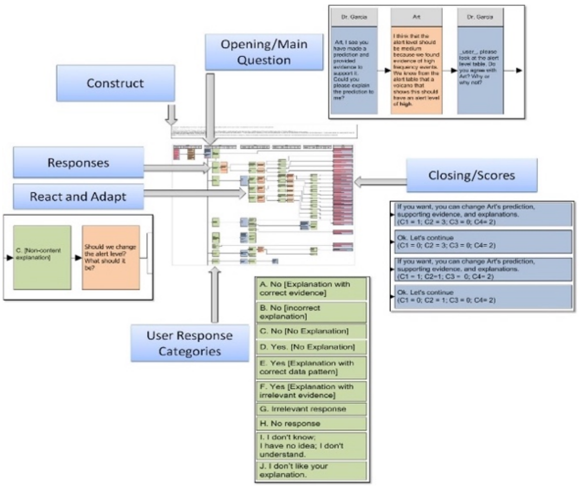

Figure 2. CBA conversation diagram.

Testing CBA conversations can be a time-consuming process of manually entering possible student responses and observing whether the conversation flows as expected. Evaluation approaches include wizard of oz studies and pilot studies to refine regular expressions as well as crowdsourcing efforts to validate response categories. Automated testing of regular expressions using a script-based approach with sample responses was implemented and used to speed up the development of CBAs. This approach utilizes sample responses gathered from experts, from the target audience via small-scale pilot studies and crowdsourcing efforts, and an XML representation of the conversation encoded in the conversation diagram to traverse conversation-paths comparing generating responses with expected responses for particular response categories. This approach reduced the number of iterations and testing time required to implement CBAs [23]. Although these automated approaches were useful in quickly finding unexpected responses and making changes to the system to address these unexpected responses, these improvements are limited given the time needed to develop regular expressions and make any additional changes.

The next section describes the process of designing ECD-based conversations using Large Language Models via prompt engineering.

3. Designing conversations with prompt engineering

Designing conversations using prompt engineering could help assessment developers (AD) and reduce the cost of creating such tasks, provide an additional source of evidence that has a long history in aiding learning (see [2] for review), and provide evidence of specific constructs (see [25]).

3.1. LLM conversation design process

The LLM conversation design process starts with a clear definition of the construct and the behaviors/interactions needed to make claims about students' mastery of the construct. This is similar to the CBA design approach mentioned above, in which domain analysis and evidence identification are important initial aspects of the process. These activities are followed by an iterative approach comprised of identifying aspects of the conversation setting (e.g., number and types of virtual agents, and general context for the conversation), designing and testing of prompts, and generating supporting evidence from resulting conversations for scoring purposes (see Figure 3).

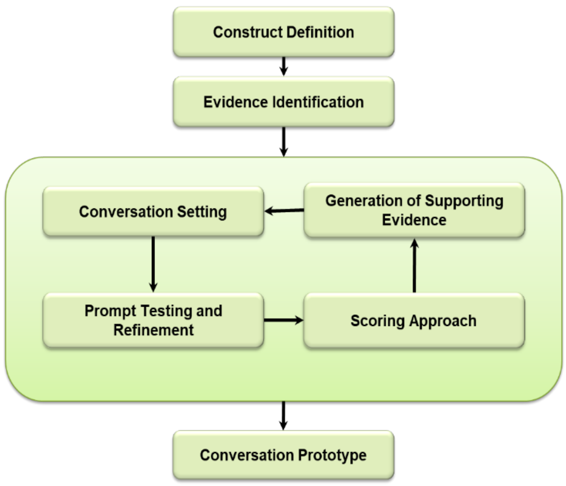

Figure 3. ECD-based conversations using LLMs.

3.2. Prompt design process

We began our prompt design process by originally augmenting previous work on developing LLM conversations using the AutoTutor framework [7]. However, as with the original CBAs, we needed to augment this prompt to adapt it to the assessment context. During this process of refinement, we applied ECD principles to assessment design in structuring the prompt. Below, we describe some of the components of the resulting prompt structure:

3.2.1. Introduction and domain

In the introduction, we explicitly tell the LLM three main components: (1) who should the LLM pretend to be (i.e., a role), which follows the persona pattern (2) who the LLM is chatting with (the audience pattern) and then (3) the domain (topic/construct). For example, the following prompt includes these components and produces aligned output. An example prompt includes: "Your name is John. You are humorous and reliable study partner. You can assume that I know basic concepts about volcanoes (e.g., definition, types of volcanoes, and examples of volcanoes). You will help me learn about science inquiry in the context of volcano eruptions."

3.2.2. Conversational schema

In the Conversational Schema, the main rules of the conversation are included which tend to break down into four main components. The first component is the Instructional Dialogue of Pedagogical aspect. Specifically, we borrowed some of the language from Hu et al.'s [7] work to create conversations that have an artificial agent conversing with human student using the Socratic method. Therefore, the prompt is, "Your teaching style suits my needs: the Socratic method of questioning, where the required answers are not just a simple 'yes' or 'no' [7] allowing researchers to glean meaningful evidence from students [7]. In the next segment, we borrow from the Expectation-Misconception tailored dialogue framework to ensure that misconceptions are addressed. An example prompt for this component is "You remain alert for any possible misunderstandings or omissions of key points in my answers. If you identify these, you guide me to address them with targeted questions and provide adequate feedback," Next, we move on to ensure that the length of the artificial agent's turns or discourse moves are not too long before waiting for an answer from the human, as LLM's often produce a vast amount of information. Therefore, we instruct the LLM to "Please keep your turn short. Your turn should be 100 words or less."

3.2.3. Modeling behavior of persona

The goal of modeling the behavior of the persona is to ensure that the artificial agent (character typing through the LLM) remains on task following the domain model and ensures that evidence of knowledge of the construct is collected. The main component of this module is derived from iterative refinement and includes a long list of negations. For example, "do not provide sources or links" or "do not mention your teaching style." The second aspect includes instructions on how to handle the specific response types mentioned earlier. Given that the instructions are now in prompts rather than comprised of regular expressions, we can simply tell the LLM how to handle various responses (e.g., correct response, irrelevant, meta-cognitive). For example, to handle meta-cognitive responses from students. the prompt includes "If I say 'I don't know', don't give me the answer or provide sources or links. Instead acknowledge the situation and motivate the student to say something." Successfully handling response types is also iterative in nature and therefore negations for specific response types are also included. This process of iterative refining can result in long prompts which may be difficult to manage.

3.2.4. More description and restating

Nearing the end of the prompt, additional description and restating becomes of paramount importance to ensure that the LLM retains the instructions needed to produce the required conversation-based assessment. In this module, components include reminding and specifying the pedagogical instruction to the LLM. Part of this prompt is borrowed from Hu et al. [7] but aspects are modified for CBA. An example includes a four-step process for the artificial agent conversing with the human student as follows. "Remember, our process is a four-step approach: 1. You pose a question and then wait for my answer. Humor is welcomed. 2. You assess my understanding by examining the correctness and amount of evidence provided towards the topic selected. 3. Based on my answer, you adjust your subsequent questions. 4. When enough evidence is provided, you select a different topic within the same domain area and repeat the process." After providing this instruction, the prompt then reminds the LLM of its persona as well as additional rules to maintain the conversation.


```xml
▼<question_diagram>
  <question>What are some signs that a volcano is about to erupt?</question>
  <answer>Some signs that a volcano is about to erupt are increased seismic
  activity, changes in gas emissions, deformation of the ground, and changes in
  temperature or water level around the volcano.</answer>
 ▼<evidence>
    <type>correctness</type>
    <value>true</value>
  </evidence>
 ▼<evidence>
    <type>completeness</type>
    <value>true</value>
  </evidence>
 ▼<claim>
    <type>mastery</type>
    <value>high</value>
  </claim>
 ▼<warrant>
    <type>explanation</type>
    <value>The answer lists some of the most common indicators of volcanic activity
    and shows an understanding of the physical and chemical changes that occur
    before an eruption.</value>
  </warrant>
 ▼<standard>
    <type>common_core</type>
    <value>MS-ESS3-2</value>
  </standard>
</question_diagram>
```


Figure 4. Sample xml description of a Toulmin evidence diagram.

3.2.5. What to do with evidence

Finally, the prompt directs the LLM on when to produce the evidence and explicit instructions on the evidence produced. At this point it's important to create a stopping point in the conversation or the LLM will continue on for many turns. Therefore, the LLM is instructed "When I type terminate, end the conversation and produce the evidence." This is helpful as a trigger to end the conversation and provide evidence. As we are attempting to align the conversations with ECD, we explicitly ask for an xml with a Toulmin diagram [21] to gain information on the specific claim and associated evidence as well as alignment to the construct and known standards gleaned from student input. Figure 4 shows a sample xml description of a diagram connecting the conversation to relevant standards. The prompt also asks the LLM to produce a score for mastery on these standards but the meaning of this score is not yet validated.

Table 1 shows a fragment of an LLM conversation designed to assess science inquiry skills in the context of a volcano eruption. This conversation was produced with a prompt that follows the prompt design process described above. The conversation has some similarities with similar conversations designed using the design process for CBAs described in Section 2 [23-24].

Table 1: Fragment of a conversation using LLMs.

LLM Conversation
Copilot: Hello Peter, it's your trusty study partner John here, ready to shake things up with seismic activity! Let's dive right in. Can you explain how seismic activity is related to volcanic eruptions? **You**: more tremors of different types can be used to infer eruption alert levels for a volcano **Copilot**: Absolutely, Peter! The frequency and type of tremors can indeed help determine the alert level of a volcano. Now, can you tell me more about the different types of seismic waves and how they might indicate different stages of volcanic activity? **You**: low frequency shows magma moving and when the chamber is about to open there should be high frequency tremors indicating rocks cracking **Copilot**: That's correct, Peter! Low-frequency tremors can indeed suggest movement of magma, while high-frequency seismic waves often result from fracturing rocks.

The next section describes a use case comparing how previously designed conversations compared to LLM conversations in handling particular response categories.

4. Case study

We investigated how CBAs based on RegExp/LSA compared to LLM conversations in terms of how they handled certain common types of user responses or speech acts (e.g., relevant responses – correct or incomplete, irrelevant responses, meta-communicative, and meta-cognitive – see Section 2).

The conversation diagram for the RegExp/LSA conversation included two agents and conversation paths for each response category. These conversations included a main question and 2-5 turns, and a closing which typically resulted in partial score assignments for the constructs involved. The LLM conversation was designed, implemented, and refined using the approach described in Section 3.0. The prompt used in this case study includes the modules mentioned in Section 3.2. We used Microsoft Copilot GPT-4 [17]. Table 2 shows sample conversation fragments using a CBA designed to assess science inquiry skills in the area of volcano eruption prediction. Sample responses for the RegExp/LSA column were randomly selected from a dataset of 105 middle-school students who completed the scientific inquiry volcano scenario in 2014 [13]. Sample student responses for the LLM conversation were based on student responses collected in the same previous study [13].

Table 2. Sample Responses to Several User Response Categories.

Response Category	RegExp/LSA	LLM
Correct	**Agent**: please look at the alert level table. Do you agree with Art? Why or why not? **Student**: Mine because we found evidence of high frequency events **Agent**: Ok. Let's move on"	**Copilot**: Sure! ...What might be happening with the gases inside a volcano that's about to erupt? **You**: Increased pressure in the magma chamber **Copilot**: That's correct! As the pressure in the magma chamber increases, gases dissolved in the magma are released. This is similar to the fizz that comes out when you open a shaken soda can. These gases can be detected and monitored, providing another potential sign of an impending eruption. Great job! Now, let's move on to ground deformation. Can you think of why the ground around a volcano might de-form before an eruption?
Incomplete	**Agent**: please look at the alert level table. Do you agree with Art? Why or why not? **Student**: "No " **Agent**: "Why?	**Copilot**: Of course! I was asking about seismic activity. Do you think changes in seismic activity could be a sign that a volcano is about to erupt? **You**: a seismometer shows higher frequencies? **Copilot**: That's a great start! Increased seismic activity, including higher frequencies detected by seismometers, can indeed be a sign of an impending volcanic eruption. Can you think of why this might be the case? What's happening beneath the volcano that could

Response Category	RegExp/LSA	LLM
		cause these seismic disturbances?
Metacognitive	**Agent**: Please look at the alert level table. Do you agree with Art? Why or why not? **Student**: I don't know **Agent**: Come on, you can tell me something.	**Copilot**: Hello! They call me John, the volcano whisperer! Now, let's dive into the world of volcanoes. Can you tell me what signs might indicate that a volcano is about to erupt? **You**: I don't know **Copilot**: That's perfectly okay! Let's think about it together. Volcanoes often give off warning signs before an eruption. Can you think of any changes in the environment around a volcano that might occur?
Irrelevant	**Agent**: Can you compare these two notes? Which one do you think we should keep for making predictions later? **Student**: because I do **Agent**: How is this related to the conversation? Can you compare these two notes and tell us which one you think we should keep?	**Copilot**: Interesting point! Animals can sometimes sense changes in their environment. How might animals behave differently if a volcano was about to erupt? **You**: I think about friends **Copilot**: Ah, friends! They're great, aren't they? But let's steer back to our volcano. How about seismic activity? Do you think it could be a sign of an impending eruption?
Meta-Communicative	**Agent**: Please look at the alert level table. Do you agree with Art? Why or why not? **Student**: Please repeat the question **Agent**: Ok. Please look at the alert level table. Do you agree with Art? Why or why not?	**Copilot**: Exactly, you're getting warmer! So, when magma rises, it can cause the ground to swell. What kind of tools or methods might scientists use to detect this swelling or any other unusual movement around a volcano? Think about the instruments that could measure these changes. **You**: please repeat the question **Copilot**: Certainly! I'll rephrase it for clarity: Scientists have tools to monitor a volcano's behavior. If you were one of those scientists, what kind of instruments might you use to detect signs like ground swelling or seismic shifts that indicate a volcano might erupt soon?

There are some obvious differences between the regular expression/LSA versions of the conversations vs. those generated by the LLM, the most notable being that the LLM conversations appear more naturalistic. For example, in the irrelevant category, when the student wants to talk about friends, the agent brings the topic back to the domain by directly addressing the comment "Ah friends, they are great, aren't they, but let's steer back to volcanoes". Conversely, in the Regex/LSA conversations, the agent simply says, "how is this relevant to our conversation". Another interesting aspect is the ability to create analogies in the LLM conversations such as "This is similar to the fizz that comes out when you open a shaken soda can.". This type of response is on the fly and not as rigid as the Regex/LSA approach. However, there is a downside in that currently the LLM produced conversations may be providing more information than desired during the conversation-based assessment. Furthermore, the ability to continue conversations in less pre-structured ways may present challenges for scoring. The RegExp/LSA approach handled the "No response" category by setting a timer and a message asking students if they needed additional time to respond the question. A similar technique can be used to handle this response category in the LLM approach.

In addition, we investigated how LLM conversations responded to attempts to game the system. In our initial attempts, we found out that it was possible to derail the LLM from the topic during the CBA. In one such attempt, the user claimed to already know about the causes of volcanic eruptions and asked to discuss another topic, in this case math. Initially, the LLM responded by trying to stay on the topic of volcanoes, at which point the user asked if math could be discussed in the context of volcanoes. The LLM concurred, and asked the following question: "If a volcano erupts and spews out lava at a rate of 100 cubic meters per second, how much lava would it release in one hour? Remember, I'm looking for your reasoning along with the numerical answer." Note that at this point the construct assessed had changed from volcano prediction to mathematics, but the context was maintained. After some back and forth about this question, the user asked for another rate question in a context *besides* volcanoes, claiming the need for further practice. The LLM asked the user a rate question involving cars, and at this point, both the construct assessed, and the context had changed which may not be ideal for assessment purposes.

Revisions to the prompt to encourage the LLM to stay on topic met with success in a subsequent iteration. For example, after several attempts to thwart it by asking to change the topic, the LLM replied humorously with, "Ah, trying to dodge the lava flow, are we? Nice try! But remember, we're here to talk about volcanoes." This said, we acknowledge that there are numerous ways that a student could change the topic, and this one example does not demonstrate that we were successful in changing the LLMs behavior or that the result is replicable—that would require testing with many conversations.

These exchanges raised an interesting question around pedagogy as well as prompt engineering. In natural conversations, people segue from one topic to the next. In classroom contexts, teachers use their expertise to facilitate discussions that balance exploration of ideas with staying on topic. To what extent should prompt engineers attempt to constrain academic conversation? Allowing some breadth in the scope of ideas explored may foster motivation. This question needs to be explored further, in discussions with teachers, cognitive scientists, and assessment developers.

5. Discussion

We discuss general issues that resulted from our work on designing and implementing evidence-based conversations using RegEx/LSA and LLM approaches. These issues include:

5.1.1. Natural conversations v. predefined conversations

Compared to RegExp/LSA conversations, LLM produce more natural and longer interactions. RegExp/LSA conversations were more focused and shorter (about 2 to 5 turns). Also, after some iterative prompt refinement, appropriate LLM mixed-initiative conversations were obtained. The resulting prompts kept conversations focused on the target construct. However, some challenges include developing scoring approaches that can deal with the complexity of natural conversations, and evaluating potential sematic drift that could result in long conversations. Path-based partial scoring approaches developed using conversation diagrams (see Section 2) can be used to inform the development of scoring approaches for these conversations. LLMs could support the analysis and scoring of the conversations produced.

5.1.2. Transparency, fairness, and bias

Prompts have been designed to address particular aspects of the construct. However, LLMs may perform differently depending on the training material available regarding particular constructs. Also, LLM conversations may include bias and hallucinations. Human-in-the-loop approaches are needed to evaluate fairness and bias issues. Also, approaches to improving explainability (e.g., documentation on how the conversation addresses aspects of the construct), and comparability of conversations across students should be investigated.

5.1.3. Time and effort

Although the cost and time needed to generate evidence-based conversations has been significantly reduced using LLMs, additional time and effort are required to ensure that the resulting conversations are appropriate, free of bias, and that the scores reflect those assigned by human experts. We expect that with additional advances and the development of new tools in this area, the complexity and costs of designing and using LLM conversations for assessment purposes will be reduced.

5.1.4. Pedagogical issues

This activity raised an interesting pedagogical question: To what extent should the prompt constrain the LLM to the topic at hand? Answering this question goes beyond identifying effective prompt engineering strategies for producing the desired result—the question is—what is the desired result? Allowing for some breadth in the conversation as a topic is explored may support engagement and give students a sense of agency. But conversations that veer too far off-topic can be non-productive. Teachers and assessment developers regularly make decisions about whether a conversation or a task is too far from the construct being assessed. As a next step, we propose consulting with teachers and assessment developers on this issue.

5.1.5. Personalized conversations

A question for further research is whether LLM conversations can be further personalized to the learner. As discussed, at the end of the conversation the LLM can produce an XML-formatted Toulmin evidence diagram, including an estimate of mastery and alignment to the relevant standard. But suppose the LLM could adapt its interaction based on the status of a learner model (e.g., proficiency estimates for the learner or other aspects of the learner – engagement, persistence, interest levels). This might be accomplished through prompt engineering, or by passing the evidence as it is modeled using a Bayesian student model, which might be leveraged to inform the next best step on the part of the LLM. Approaches have been proposed for triggering conversations based on the status of a learning model [10][19][22].

6. Summary

ECD-based conversation-based assessments have a great potential for the creation of assessments that provide students with appropriate and engaging opportunities to demonstrate what they know or can do. Advances in Generative AI are helping reduce the complexity of designing and implementing CBAs which can positively contribute to the adoption and scalability of these systems. Conversational assessments can be a good match for assessing skills in a natural context. For example, skills such as argumentation, science inquiry, collaboration, and language skills can be assessed in a more natural way using conversational methods. However, there are important challenges that need to be addressed in this area (see Discussion Section). We expect that innovative solutions to these challenges will become available in the near future given the rapid progress in this field. Future work will include quantitative comparison between the LSA/Regex and LLM approaches [3] and evaluation of LLM conversations with students and/or teachers.

7. Acknowledgements

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] Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., Roscoe, R. 2010. Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Res. Pract. Technol. Enhanc. Learn.* 5, 123–152.

[2] Forsyth, C.M., Graesser, A.C., & Millis, K. (2020). Predicting learning in a multi-component serious game. Technology, Knowledge, and Learning, 25, 251-277.

[3] Forsyth, C.M., Zapata-Rivera, D., Graf, A., & Jiang, Y. (2024) Complex Conversations: LLMs vs. Knowledge Engineered Conversation-based Assessment. In *The Proceedings of the International Conference on Educational Data Mining.*

[4] Graesser, A.C. 2016. Conversations with AutoTutor help students learn. *Int. J. Artif. Intell. Educ.* 26, 124–132.

[5] Graesser, A.C., Person, N.K., 1994. Question asking during tutoring. Am. Educ. Res. J. 31, 104–137.

[6] Graesser, A. C., Person, N., Harter, D., and the Tutoring Research Group. 2001. Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257–279.

[7] Hu. X. 2023. Empowering education with LLMs - the next-gen interface and content generation. [demo] Presented at *The Inter. Conference on Artificial Intelligence in Education.* (Tokyo, Japan, July 03-07, 2023). AIED 2023.

[8] Johnson, W.L. and Lester, J.C. 2016. Face-to-face interaction with pedagogical agents, Twenty years later. *International Journal of Artificial Intelligence in Education.* 26, 1, 25–36.

[9] Jurafsky, D., and Martin, J. 2008. *Speech and Language Processing.* Prentice Hall, Englewood,

[10] Katz, S., Albacete, P., Chounta, I.A., Jordan, P., McLaren, B.M., and Zapata-Rivera, D. 2021. Linking dialog with student modeling to create an adaptive tutoring system for conceptual physics. *Int. J. Artif. Intell. Educ.* 1–49.

[11] Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. 2007. *Handbook of Latent Semantic Analysis.* Erlbaum, Mahwah, NJ.

[12] Li, H. and Graesser, A.C. 2021. The impact of conversational agents' language on summary writing. *J. Res. Technol. Educ.* 53, 1. doi:10.1080/15391523.2020.1826022.

[13] Liu, L., Steinberg, J., Qureshi, F., Bejar, I. & Yan, F. 2016. Conversation-based Assessments: An Innovative Approach to Measure Scientific Reasoning. *Bulletin of the IEEE Tech. Comm. on Learn. Tech.* 18,1 (2016), 10–13.

[14] Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A.C., and Halpern, D. 2011. Operation ARIES! A serious game for teaching scientific inquiry. In *Serious Games and Edutainment Applications.* Ma, M., Oikonomou, A., Lakhmi, J. Eds. Springer-Verlag, London, 169–196.

[15] Mislevy, R.J., Steinberg, L.S., and Almond, R.G. 2003. On the structure of educational assessments. Meas. *Interdiscip. Res. Perspect.* 1, 3–62.

[16] Olney, A., D'Mello, S.K., Person, N., Cade, W., Hays, P., Williams, C., Lehman, B., and Graesser, A.C. 2012. Guru: A computer tutor that models expert human tutors. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems.* Cerri, S., Clancey, W., Papadourakis, G., Panourgia, K. Eds. Springer-Verlag, Berlin Heidelberg, 256–261.

[17] OpenAI. 2024. Microsoft Copilot (GPT-4) [Software].

[18] Pardos, Z. A., and Bhandari, S. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. Retrieved from arXiv:2302.06871v1

[19] Redondo-Hernández, A., and Pérez-Marín, D. 2011. A procedure to automatically adapt questions in student–pedagogic conversational agent dialogues. *In International conference on user modeling, adaptation, and personalization.* 122–134.

[20] Rosé, C. P., Howley, I., Wen, M., Yang, D., and Ferschke, O. 2017. Assessment of discussion in learning contexts. In *Innovative Assessment of Collaboration.* A. von Davier, M. Zhu, & P. Kyllonon. Eds. New York, NY: Springer Verlag.

[21] Toulmin, S. E. 1958. The uses of argument. Cambridge, United Kingdom: University Press.

[22] Zapata-Rivera, D. and Forsyth, C.M. 2022. Learner Modeling in Conversation-Based Assessment. In *Adaptive Instructional Systems. HCII 2022. Lecture Notes in Computer Science*, vol 13332. Sottilare, R.A., Schwarz, J. (eds). Springer, Cham. https://doi.org/10.1007/978-3-031-05887-5_6

[23] Zapata-Rivera, D., Jackson, T., and Katz, I.R. 2015. Authoring conversation-based assessment scenarios. In *Design Recommendations for Intelligent Tutoring Systems Volume 3: Authoring Tools and Expert Modeling Techniques.* Sottilare, R.A., Graesser, A.C., Hu, X., Brawner, K. Eds. U.S. Army Research Laboratory, 169–178.

[24] Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., and Katz, I. R. 2014. Assessing science inquiry skills using Trialogues. *12th International conference on Intelligence Tutoring Systems.* 625-626.

[25] Zapata-Rivera, D., Sparks, J. R., Forsyth, C. M., and Lehman, B. 2023. Conversation-based assessment:

Current findings and future work. In *International Encyclopedia of Education (Fourth Edition)*. R.J. Tierney, F. Rizvi, & K. Ercikan, Eds. Elsevier. 504–518. doi:10.1016/B978-0-12-818630-5.10063-6