Complex Conversations: LLM vs. Knowledge Engineering Conversation-based Assessment

Carol M. Forsyth, Diego Za-
pata-Rivera, Edith Aurora Graf,
& Yang Jiang
ETS Research Institute
cforsyth;dza-
pata;agraf;yjiang002@ets.org

ABSTRACT

This paper explores the differences between two types of natural language conversations between a student and pedagogical agent(s). Both types of conversations were created for formative assessment purposes. The first type is conversation-based assessment created via knowledge engineering which requires a large amount of human effort. The second type, which is less costly to produce, uses prompt engineering for LLMs based on Evidence-Centered Design to create these conversations and glean evidence about students' knowledge, skills, and abilities. The current work compares linguistic features of the discourse moves made by the artificial agent(s) in natural language conversations created by the two methodologies. Results indicate that more complex conversations are created by the prompt engineering method which may be more adaptive than the knowledge engineering approach. However, the affordances of prompt engineered, LLM generated conversation-based assessment may create more challenges for scoring than the original knowledge engineered conversations. Limitations and implications are discussed.

Keywords

Conversation-based assessment, large language models, evidence-centered design

1. INTRODUCTION

Conversation-based assessment (CBA) provides students with opportunities to converse with one or more artificial agents about a given domain (for review, see [11]). These conversations were created to help glean additional information about students' knowledge, abilities and skills that may not be captured by traditional assessments. However, these assessments have historically been generated based on a knowledge engineering approach which is costly to develop. There is an opportunity for researchers to harness the power of large language models (LLMs) via prompt engineering to create these conversations. In this paper, we compare data from the original knowledge engineering approach (KE) to the newly created CBAs via prompt engineering LLMs (PE) and explore differences in linguistic features between artificial agents(s) discourse generated by two methodologies.

1.1 Knowledge-Engineered CBA's

CBA's include natural language conversations between a human student and one or more artificial agents. The knowledge engineering approach leveraged the dialogic framework and associated natural language processing (NLP) in AutoTutor, an intelligent tutoring system with natural language conversations that has achieved learning gains reaching over an entire letter grade across decades in various domains and contexts [2]. The conversations are based on Expectation- Misconception tailored dialogue aligning to the Socratic method. The framework was derived from analyzing expert tutoring sessions. The main idea is that there is an expected answer to each question and a student goes through a series of scaffolding moves until this expected answer is given or the scaffolding moves are exhausted [for review see [2]). The NLP of student answers is conducted with regular expressions [4] and latent semantic analysis [5], which has yielded results comparable to human experts' ratings of the user input [see 2].

Both the conversational framework and the NLP are embedded in knowledge engineered CBAs but there are key differences. Specifically, CBAs are more constrained as they are for assessment purposes rather than learning. Over the years, KE CBAs were created across multiple domains and student responses were comparable to those provided to humans in the English Language Learning domain (see [11]). However, creating these conversations via knowledge engineering is extremely expensive requiring the work of experts in various areas. For example, to create these conversations, learning scientists had to first conduct a full domain analysis, which was then represented as a path diagram displaying adaptative paths with each discourse move within a path focused on the represented construct. The various paths made it possible for simple scoring. Next the conversations and associated regular expressions were written by a NLP expert and iteratively refined to account for human input, creating a large cost to developers. Authoring tools, crowdsourcing and an automated testing tool were developed to speed up the development and testing process of CBAs [see 11]. This large amount of time and expertise spurs the interest to create conversations with LLMs, particularly given the latest easily accessible innovations.

1.2 Prompt-Engineered LLM CBA's

In November 2022, Large Language Models (LLMs) came to the forefront for mass use likely due to the user-friendly applications and ability to write code in language (prompt engineering) with Open AI's ChatGPT after a series of other user-friendly tools. ChatGPT (GPT 3 and now often 4.5) is trained on a large number of corpora from the internet giving it a vast amount of text to draw from. Educational researchers and the EdTech industry have begun

exploring the power of using these highly trained models for creating natural language conversations for education. For example, Khan Academy and Duolingo have begun harnessing the power of LLMs for item development.

Researchers have focused on prompt engineering for a variety of purposes including natural language conversations. Notably, prompt engineering helped create AutoTutor-like conversations with perhaps more naturalistic components [3]. However, some research indicates that hints provided by LLMs are not comparable to human created hints which are far more relevant to the given domain [9].

In the present research, we incorporate Evidence-Centered Design [7] to prompt-engineered CBA tasks that glean evidence of a specific domain from the student in a careful linking between evidence and constructs. The major focus is to create conversations that elicit evidence that is needed to make inferences about what the student knows and can do in a particular domain. Therefore, we iteratively refined prompts over 50 times to create one that would provide such evidence. In the current work we compare how artificial agent discourse moves in CBAs are generated by the KE and PE approaches.

2. METHODOLOGY

2.1 Materials

2.1.1 Science CBA task
The original knowledge-engineered CBA task focuses on scientific inquiry skills. Specifically, the students are asked to predict an alert level for a volcanic eruption. Throughout the task, students interact with two artificial agents, "Dr. Garcia" and "Art," as seen in Figure 1. Dr. Garcia provides guidance throughout the task. Art is a fellow student who helps to collect data from a volcano, analyzes the data, engages in argumentation, and finally determines the alert level.

Figure 1. Dr. Garcia and Art converse with the student

The task includes a simulation where students place seismometers on a volcano and collect data in a data table. The students can take notes on the data collection and compare notes with Art, the artificial peer agent. There are multiple mini-conversations between the student and artificial agents while the student engages with multiple sources of evidence (i.e., data table, notes). The final conversation focuses on explaining the alert level for a potential volcanic eruption and the basis for this claim with information from a data table.

2.1.2 Knowledge vs. Prompt-Engineered CBA
The PE conversations differ from the KE in that the PE only has conversations between the artificial agent and student in a chat window (vs the multiple sources in KE). The conversations are created following rules of a KE CBA including engaging in the Socratic

method but with constraints as to not give away the answer. The prompt includes information about the construct, roles, and information on what to do and what not to do to keep the conversation focused on the construct. Furthermore, the prompt for the PE version includes instructions on how to handle specific speech acts similarly to CBAs created with KE. For example, metacognitive responses such as "I don't know" are responded to with "Come on tell me something". However, even with these similarities, there are not multiple sources (e.g., simulations) as in the KE CBAs. Therefore, the PE conversations take the human through a variety of discourse moves to create the scene for determining an alert level for a volcanic eruption. After the stage is set, the conversations become more like the KE version and the artificial agent asks the student how to determine the alert level for a volcanic eruption. It's worth noting that evidence gleaned from these conversations currently includes domains aligned with several NGSS standards [8].

2.1.3 TextEvaluator
Text Evaluator is a computational linguistic tool that has hundreds of features which can be reduced into 8 main principal components [10] which have been correlated with the Common Core Text Complexity Model [1]. These dimensions include academic vocabulary (vs. non-formal vocabulary), argumentative text (requiring one to make inferences based on reasoning), cohesion (clearly connected text), word concreteness (vs. abstract language), interactivity (conversational language), narrativity (as opposed to expository text), syntactic complexity (level of complexity in sentence structure) and vocabulary difficulty. The resulting scores are standardized accounting for numerous features including word count resulting in a score from 0-100 per dimension. In the current research, we leverage this tool to compare artificial agent discourse between KE vs. PE CBAs to evaluate linguistic differences.

2.2 Research Question

RQ: How do the artificial agent discourse moves within the conversations differ on key linguistic metrics between the knowledge-engineered (KE) vs. prompt-engineered (PE) approaches?

2.3 Data Sources

Participants from the KE CBA (N=10) were randomly selected from a dataset of 105 middle-school students who completed this scientific inquiry task with IRB approval in 2014[6]. Although demographic information was collected at the time, this information was not available in the shared dataset. In the PE condition, simulated students (N=10) completed the interaction based on assigned roles within a secured version of Bing's Copilot accessing GPT 4.5. Therefore, the prompt was changed between each interaction to instruct the artificial agent to direct discourse moves towards human student roles which exemplified various levels of education (i.e., middle-school student vs. scientist). Although the conversations appeared different between roles, it's difficult to determine whether this was due to the roles vs. a lack of replicability within the LLM.

3. ANALYSES AND RESULTS

In the current analysis, we compared a subset of artificial agent(s) discourse focusing on the similar question about determining alert levels for a volcanic eruption from the KE CBAs vs. the discourse created via PE. Only a subset was included in the analyses to avoid biased comparisons because the PE conversations included information in the discourse leading up to the current topic that was represented by other sources (e.g. simulations) in the KE conversations. As simulated participants were used in the PE conversations, only the artificial agent discourse moves in the context of the same topic were analyzed for both conditions (e.g. KE and PE).

Therefore, the conversations were reduced to 10 subsets for the KE and 10 subsets for the PE including only the pedagogical agent(s) discourse moves for the topic of "alert level" thus totaling 20 subsets of pedagogical agent discourse for analysis.

The two sets of subsets were then analyzed with TextEvaluator resulting in standardized scores for each of the principal components. Indeed, the standardization accounted for the high variance in word count. First, we calculated descriptive statistics for each score by condition. Next, we conducted independent t-tests with Bonferroni corrections to compare the scores across the two conditions (KE vs. PE). When equal variances could not be assumed, we reported the t-test conducted with unpooled variance. Note that the sample size clearly does not meet the threshold for these inferential statistics as a power analysis indicates an N of 393 per group to detect a small effect. With this limitation, the reported inferential statistics are meant only as a preliminary guide to help researchers understand the text without claiming any real statistical significance.

3.1 Results

3.1.1 TextEvaluator Results

Results indicate differences between the two sets of conversations which were not surprising (see Figure 2).

Overall, an aggregate measure of text complexity shows higher complexity for the PE vs. the KE pedagogical agent(s) discourse moves (t(18)=-17.26, p<.001, MD= -76, SE = 43.86). To better understand these results, refer to the sample of discourse moves in Table 1.
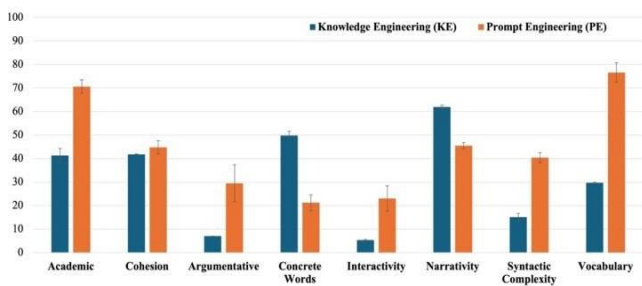

Figure 2. Comparison of features between PE and KE

In viewing the knowledge engineered (KE) vs. prompt-engineered (PE) generated conversation, one can easily see differences between the two different sets of agent discourse. Specifically, there is more of a conversational nature in the PE generated agent discourse represented in the dimension of Interactivity (i.e., conversational in nature) which was higher for PE than KE, but is marginal in nature given the Bonferroni correction yielding a threshold of *p* = .006, (*t* (9.06) = -3.28, *p*=.009, *MD*= -17.7, *SE* =5.40). A difference is also quite clear between the amount of academic vocabulary wherein the PE generated agent discourse has more scientific verbiage than the KE, ("Academic" = *t*(18) = -7.09,p<.001, *MD*= -29.3, *SE*= 4.14) as well as difficult vocabulary ("Vocabulary",= *t* (9.07) = -11.35 *p*<.001, *MD* = -46.9, *SE* = 4.13) and higher in language found in expository versus narrative texts ("Narrativity", *t(*18)= 10.7, *p*<.001, *MD*=16.4, *SE* =1.54), which may be an indication of the agent discourse moves going into greater depth about the topic given training and lack of additional context beyond the conversations which is available in the KE CBAs.

This may also explain the higher complexity in the sentence structure ("Syntactic Complexity, *t* (18) = -9.52, *p*<.0.001, *MD* = - 25.3, *SE*= 2.66) as the PE generated agent discourse moves provide more context and potentially naturalistic conversation in complex

sentences. This finding could also be related to an aspect of the prompt which asks the LLM to keep the turns to a set number of characters, thus requiring more complex sentences to capture all of the information to be delivered given this constraint. That said, it's interesting that concrete vs. abstract words is higher for KE generated agent moves as ("Concrete Words", *t* (18) = 7.42 *p*<.001, *MD*= 28.6, *SE* =3.86) which makes sense as more abstract language was in the PE conversations such as analogies which will be seen in the qualitative analysis.

Table 1. Sample of subset of discourse moves

KE Conversation Fragment	PE Conversation Fragment
Professor: [insert name] can you compare these two notes? Which one do you think we should keep for making predictions later?	Copilot: That's a good start! Seismic activity caused by tectonic plates is usually deeper and less frequent, while seismic activity caused by magma movement is shallower and more frequent. Now, how do seismologists use this information to determine the alert level for a potential volcanic eruption?
Human: Mine, because it has more data.	
Professor: [insert name], please look at the alert level table. Do you agree with Art? Why or why not?	Human: They watch for changes in the seismometer to see if they need a higher alert level.
Human: Yes.	
Professor: Why?	
Human: Because art is a computer and knows everything.	Copilot: That's correct! Seismologists monitor changes in seismic activity, including the frequency and intensity of earthquakes. If these increase or change in a way that suggests magma movement, the alert level for a potential volcanic eruption may be raised. Now, can you tell me what types of seismic waves are typically associated with volcanic activity?

*Note that there are many more turns possible. This is just a snippet to show domain equivalence. Only agent discourse moves (i.e., professor or copilot) are included in the analysis.

3.1.2 Qualitative Analysis

To better understand these results as well as interpret other differences between the conversations, we did a base level qualitative analysis. We discovered that the PE generated conversations certainly have more flexibility than the knowledge engineered CBAs.

For example, PE generated conversations can produce analogies on the fly which explains the increased complexity and abstract nature of the texts:

> *You: Can you give me a hint?*
>
> *Copilot: Sure! Think about a drumroll. As a big moment approaches, the drumroll gets faster and faster. Now, how might this relate to the frequency of earthquakes as a volcano gets closer to erupting?*

Unless preprogrammed, this type of analogy was not often

represented in knowledge engineered CBAs.

Another interesting component is that the PE conversations included additional motivational comments, which were not included in a KE conversation. For example, in the PE CBA, after a series of discourse moves by the student displaying uncertainty and finally agreeing to try to answer a question posed by the artificial agent, the artificial agent says, "That's the spirit". Moreover, perhaps one of the most notable differences is that there is truly mixed-initiative dialogue in PE conversations, which has been a challenge with knowledge-engineered CBAs. For example, the student can redirect the conversation with "Give me a hint". These types of discourse moves would yield only a preplanned response to maintain the illusion of intelligence with KE but provide a novel hint in the PE generated conversations. However, this flexibility may create issues for scoring as it creates the potential unplanned paths.

4. CONCLUSIONS

Overall, we found differences between the language in the original knowledge-engineered conversation-based assessments and newly created conversations with prompt engineering. Specifically, the conversations are more complex in the PE generated conversations. Furthermore, these conversations allow for motivational components, analogies, and mixed-initiative dialogue that are not available in the knowledge engineered CBAs. However, as we are attempting to create conversations based on ECD, the potential number of conversational paths and turns is exponential with PE and therefore present a challenge for scoring that did not exist in the original KE CBA's.

5. LIMITATIONS

Limitations of this study include a small sample size, simulated participants, and differences in the conversational flow. Specifically, the PE conversations were based on creating roles such as Middle School Teacher, but this was not reinforced throughout the prompt and therefore it is unclear the amount of influence each role had on the conversation. Though the conversations were visibly different, the differences could have simply been due to the LLM's lack of consistency. Another issue is drawing conclusions from such a small sample. Finally, there are differences between the two environments as the original KE version included simulations and other sources of evidence whereas the PE conversations did not. Another limitation is that there was one artificial agent in the PE vs. two in the KE. A final limitation is to consider the possibility of the PE conversations differing across domains given amount of information about the constructs in the training data.

6. FUTURE DIRECTIONS & DISCUSSION

In future directions, we hope to have more participants complete the conversations in a PE CBA that includes multiple sources similar to the KE CBA. Furthermore, we may need to add additional layers between the student and artificial agent to ensure reliability in the PE CBA.We must also determine a method for scoring these conversations.

The current work has made great strides towards producing conversation- based assessments in an inexpensive manner via prompt engineering that can hopefully be applied to multiple domains. Although there are differences in the type of language used, the material surrounding the domain itself is covered in both conversations. Therefore, the aim is to create natural language conversations at scale utilizing this methodology to have flexible and adaptive conversations eliciting evidence of a given domain at a fine-grained level. Currently, these conversations generate

evidence such as Toulmin diagrams that will be evaluated by human experts. Additional testing is needed to make sure the conversations are suitable for assessment purposes.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] Common Core State Standards Initiative (CCSSI). 2010. *Common Core State Standards for English Language Arts And Literacy In History/Social Studies, Science And Technical Subjects*. CCSSO and National Governors Association, Washington, DC.

[2] Graesser, A.C. 2016. Conversations with AutoTutor help students learn. *Int. J. of Artificial Intelli. in Edu.* 26 (2016), 124- 132.

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

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

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

[6] 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.

[7] Mislevy, R.J., Steinberg, L.S. and Almond, R.G. 2003. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*. 1 (2003), 3–62.

[8] NGSS Lead States. 2013. *Next Generation Science Standards: For States, By State*s.The National Academies Press, Washington, DC.

[9] Pardos, Z. A., and Bhandari, S. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv[Cs.CY]*.Retrieved from http://arxiv.org/abs/2302.06871.DOI:https://doi.org/10.486 60.arXiv.2302.06871

[10] Sheehan, K. M., Kostin, I., Napolitano, D., and Flor, M. 2014. The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Eleme. School J.*, 115 (2014), 184–209.

[11] Zapata-Rivera, D., Sparks, J.R., Forsyth, C., Lehman, B., 2023. Conversation-Based assessment: Current findings and future work. In *International Encyclopedia of Education 4th Edition*, R. J. Tierney, F. Rizvi, and K. Ercikan Eds. Elsevier. Amsterdam,Netherlands, 503-518.