Evaluation of mathematical questioning strategies using data collected through weak supervision

Debajyoti Datta
SEAS, University of Virginia
Charlottesville, VA 22903
USA
dd3ar@virginia.edu

Maria Phillips
SEAS, University of Virginia
Charlottesville, VA 22903
USA

James P Bywater
James Madison University
Charlottesville, VA 22903
USA

Jennifer Chiu
School of Education and Human Development
University of Virginia
Charlottesville, VA 22903
USA

Ginger S. Watson
School of Education and Human Development
University of Virginia
Charlottesville, VA 22903
USA

Laura E. Barnes
SEAS, University of Virginia
Charlottesville, VA 22903
USA

Donald E Brown
SEAS, University of Virginia
Charlottesville, VA 22903
USA

Abstract

A large body of research demonstrates how teachers' questioning strategies can improve student learning outcomes. However, developing new scenarios is challenging because of the lack of training data for a specific scenario and the costs associated with labeling. This paper presents a high-fidelity, AI-based classroom simulator to help teachers rehearse research-based mathematical questioning skills. Using a human-in-the-loop approach, we collected a high-quality training dataset for a mathematical questioning scenario. Using recent advances in uncertainty quantification, we evaluated the conversational agents for usability and analyzed the practicality of human-in-the-loop for data collection and system evaluation for mathematical questioning conversations.

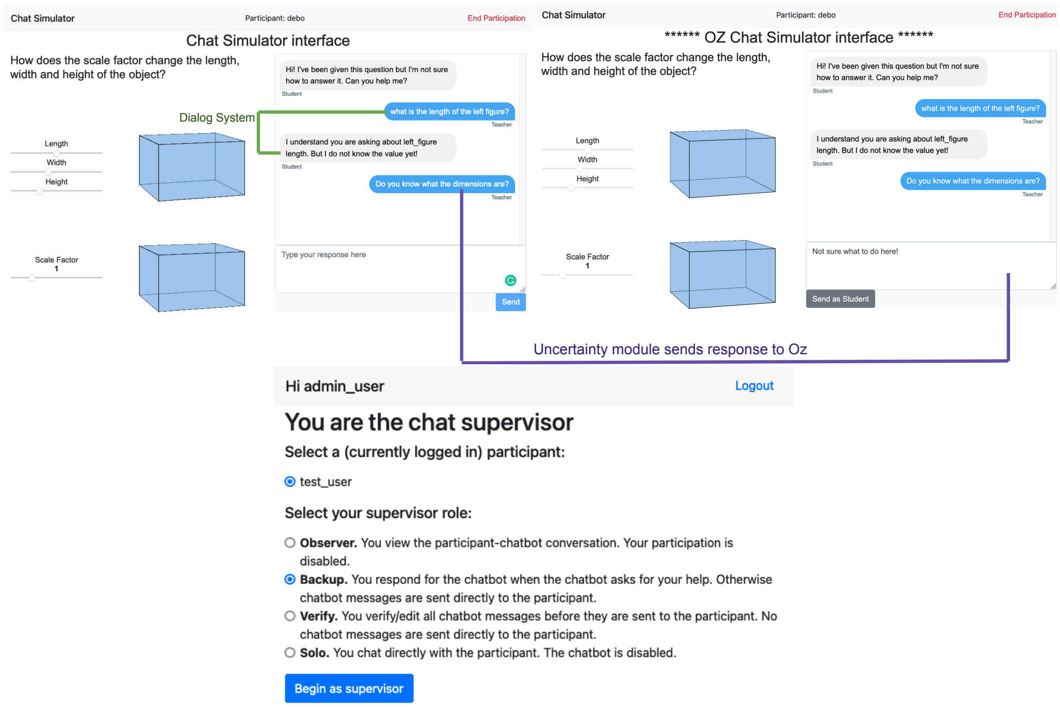
1 Introduction

Real-world applications of deep learning models require hand-labeled training data specific to the domain. While the natural language processing community has divested a significant effort in collecting datasets for popular tasks like sentiment analysis or textual entailment, domain-specific datasets like teacher questioning often lack high-quality labeled datasets. In this work, we use weak supervision to collect a teacher questioning dataset for a mathematical training scenario and use an expert-in-the-loop system to deploy and collect more training data in the process.

1.1 Scenario

The conversational agent (CA) is a student who is learning the concept of scale factor, and the user of the system is a pre-service teacher in training. In traditional classroom scenarios, teachers often use visual aids to explain various mathematical concepts. We developed an HTML5 based

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Workshop on Math AI for Education (MATHAI4ED).



How does the scale factor change the length, width and height of the object?

Hi! I've been given this question but I'm not sure how to answer it. Can you help me?

what is the length of the left figure?

I understand you are asking about left_figure length. But I do not know the value yet!

Do you know what the dimensions are?

Type your response here

How does the scale factor change the length, width and height of the object?

Hi! I've been given this question but I'm not sure how to answer it. Can you help me?

what is the length of the left figure?

I understand you are asking about left_figure length. But I do not know the value yet!

Do you know what the dimensions are?

Not sure what to do here!

Hi admin_user

You are the chat supervisor

Select a (currently logged in) participant:

● test_user

Select your supervisor role:

○ **Observer.** You view the participant-chatbot conversation. Your participation is disabled.
● **Backup.** You respond for the chatbot when the chatbot asks for your help. Otherwise chatbot messages are sent directly to the participant.
○ **Verify.** You verify/edit all chatbot messages before they are sent to the participant. No chatbot messages are sent directly to the participant.
○ **Solo.** You chat directly with the participant. The chatbot is disabled.

Figure 1: **The expert user has access to the Supervisor Interface to the right. The dialogue system responds when it is certain about the dialogue acts and entities. When the uncertainty thresholds are met, it sends the prompt to the expert user. The expert user can then type in the response as a student, and it will appear on the prompt (left). This prevents conversations from breaking because of the failure of ML pipelines and dialogue system components.**

interactive widget that the pre-service teacher can use to engage in mathematical discussion with the virtual students (Figure 1). One main goal of this dialogue system is to help pre-service teachers develop mathematical question-asking strategies through rehearsal with a CA. To evaluate pre-service teachers' dialogue and provide feedback on performance, the CA uses an adapted subsection of the Instructional Quality Assessment (IQA) focused on instructional questions [Junker et al., 2005]. In our adapted IQA rubric, questions are categorized into probing or exploring, factual or recall, expository or cueing, and others (Table 1). Table 2 provides specific details of the data collected by two annotators using weak supervision for the dialogue system components.

2 Challenges in data collection and evaluation

Effective questioning strategies from instructors in mathematics classrooms can improve student learning outcomes [Kilgo et al., 2015, Ellis, 1993, Cotton, 1988, Wilen and Clegg Jr, 1986], but practicing such questions systematically and deliberately during teacher education programs and opportunities can be difficult. For example, in a traditional conversational agent (CA) for flight booking, the CA might only have to answer very domain-specific questions like "Where are you flying to?", but in a mathematics classrooms questions are generally more open-ended like "How did you get that answer?". Similar probing questions, as in Table 1 are often challenging to evaluate because there is a subjective component that depends significantly on the context and the previous utterances. In education, CAs used with students directly have resulted in a variety of learning outcomes, [D'Mello et al., 2014] including improved learning of mathematics concepts like Mathbot [Graesser et al., 2014, Grossman et al., 2019] and writing skills [Li and Graesser, 2021]. However, very few of these systems focus on pre-service teachers, and none to our knowledge incorporate a human-centered approach for facilitating a task-specific conversation.

Weak supervision for data labeling: Weak supervision is a machine learning paradigm that trains models by incorporating noisy labeled data. These noisy labels are either crowd-sourced [Dawid

Question label	Description	Examples	Weak Supervision Pattern Example
Probing or exploring mathematical meanings and relationships	Question clarifies student thinking, enables students to elaborate their own thinking for their own benefit and the class; Points to underlying mathematical relationships and meanings; Makes links among mathematical ideas	How did you get that answer? Explain to me how you got that expression? What does n represent in terms of the diagram? Why is it staying the same?	Includes words/phrases “how”, “how did you”, “why did you”, “what is staying the same”, “explain to me”, “how could you”
Factual or recall	Elicits a mathematical fact; Requires a single response answer; Requires the recall of a memorized fact or procedure, can be a yes/no answer but for a specific mathematical question	What is 3x5? Does this picture show ½ or ¼? What do you subtract first?	Includes words/phrases: “what is this”, “what is next”, “what is [fact]”, “what would you do next”
Expository or cueing	Provides mathematical cueing or mathematical information to students, tells them to look at specific information without engaging students’ ideas	Rhetorical questions (“The answer is three, right?”) Clarifying statements “Between the 2?” Look at this diagram	Includes words/phrases: “right?”, “and then you”, “then I”, “this is”
Other	Non-academic behavioral talk; General classroom management; everything else.	Sit down Close your books	Semantic similarity with a collection of examples labeled as “Others”

Table 1: Weak supervision provides an effective way to generate pre-filled labels that can be used with model-assisted labeling to increase annotators’ speed.

and Skene, 1979] or created through machine learning models and labeling functions [Ratner et al., 2016]. After using weak supervision for the initial data labeling stages, we use model-assisted labeling [Tkachenko et al., 2020] to speed up the annotators’ labeling process. Model-assisted labeling significantly benefits the annotation pipeline since expert annotators take 70% less time to label data. One advantage of using weak supervision with expert labelers is that they can effectively develop nuanced labeling functions with higher coverage that speed up annotation in the model-assisted labeling stage. It is worth noting that when dataset sizes are large weak supervision achieves accuracies similar to supervised labeled datasets [Ratner et al., 2016]. Since the dataset size is small in our context, it is used as part of the data labeling pipeline and not by itself.

3 Uncertainty based dialogue system

Dialogue Act Module: The intent classification stage quantifies the uncertainty between the different categories of dialogue acts. For uncertainty measurement, we use Active Dropout [Gal and Ghahramani, 2016] since it is easy to adapt to existing models with minimal change. The dialogue acts are shown in Table 1. Based on the module-specific threshold of the uncertainty as shown in Figure 1, we determine if the response should be sent to the Supervisor. This can be a complex query, a failure to match an intent or an out-of-domain query like “Have you seen the recent Batman movie?”. If the uncertainty module gets triggered by the threshold, the Supervisor intervenes and states the system’s limitations or answers the question if it was because of a failure of one of the components. The average **F1 score was 0.71** for the dataset collected through weak supervision. The model used was DistilBERT [Sanh et al., 2019].

Entity Recognition Module: The entity recognition module was a NER (Named Entity Recognition Module) module trained with Spacy [Honnibal et al., 2020]. Spacy’s rule-based pattern ‘matchers’ can infer complex patterns from the text by using linguistic properties of words (like part of speech tags) and regular expressions. The rule-based matcher was then used to create the dataset, which was subsequently trained on a logistic regression classifier similar to the approach by [Bar, 2016]. We used pre-trained BERT sentence representations to train the logistic regression classifier (**precision: 0.84, recall: 0.82, f1: 0.83**). The entity extraction task was framed as shown in Table 4 (In Appendix).

Turn-Taking Module: Turn coherence and turn-based uncertainty are critical in conversational agents since conversations can derail with an out-of-context response [Lin et al., 2019] or an inappropriate response [Young et al., 2018]. To find if users are asking semantically equivalent questions in consecutive turns, we match the semantic similarity of the previous user question and the new user question using Sentence-BERT [Reimers and Gurevych, 2019]. The cosine similarity between the embeddings gives a score we use to compute if the utterance is very similar to the previous utterance.

4 Experimental Setting

The evaluation included eight users with teaching experience (ages 24-60), each conversing with the system twice. The analysis of the user utterances are shown in Table 3. Certain utterances (Always

Technique		Annotator	n	Annotator Accuracy	Agreement (kappa)	Model-Assisted Agreement	Time	
							M(SD) (seconds)	*p-value*
A	Classical Labeling	Both	1730	0.82	0.52	-	15.2(42.1)	<0.001
		A1	864	0.89		-	13.2(37.8)	
		A2	866	0.74		-	17.3(45.9)	
B	WS - MAL	Both	3983	0.84	0.61	0.70	10.4(32.7)	
		B1	1994	0.89		0.80	7.1(18.3)	
		B2	1989	0.79		0.60	13.6(42.3)	

Table 2: Performance comparison of annotators between traditional supervised labeling and weak-supervision + model-assisted labeling approach. The gold accuracy table refers to anonymized annotators accuracy compared to gold labels generated by expert teachers with significant experience in evaluating IQA metrics.

Table 3: **Certain utterances were always sent to the supervisor. However for most cases, the NLP components could understand the semantic meaning of the utterances**

Collaboration Approach	Definitions and Examples	Number of Occurances	Varied Based On Turn
Always Supervisor	Chit-Chat type questions "What is your name?"	6	No
	Detailed explanation questions "Can you explain why the volume is 25?"	8	No
	Out of domain questions "Can you calculate it for a sphere?"	3	No
	Overly complex descriptions "If all the dimensions of the left box is 5 and you change the scale factor by 2, and all other dimensions remain unchanged what is the volume of the right box?	5	Yes
Always AI	Question about dimensions "What is the right figure volume?" "What is the scale factor?"	14	No
	Greetings "Hi, How are you?"	5	No
	Simple Explanations "Do you know how to calculate the volume?"	16	No
Mix of AI and Supervisor	Definitions "How did you calculate the scale factor?" "Would the volume have changed if the scale factor was 1?"	7	No
	Acknowledgment "Great job!", "Oh nice!", "Very good!"	8	No
	Task Status "So do you think you understand the task at hand?"	2	No

AI and Mix of AI and Supervisor) benefit from weak-supervision-based data labeling approaches. Complex queries that always go to the supervisor can be used for future data labeling scenarios.

5 Conclusion and Future Work

Our goal in this paper was to build a conversational agent using weak supervision for a mathematical training scenario. Our analysis showed that conversational agents could be built for mathematical scenarios that do not have large datasets by combining weak supervision and accounting for model failures. Uncertainty quantification and training with noisy labels like weak supervision is an active area of research, and more robust training paradigms will help build newer scenarios with minimal data. This is especially useful in educational research where data collection predominantly relies on video and audio transcriptions of classroom recordings. In the future, we plan to incorporate other modalities like speech into the system to understand and evaluate limitations of uncertainty in other modalities other than text.

References

Nadav Bar. Training a machine learning classifier for relation extraction from medical literature. 2016. URL `https://devblogs.microsoft.com/cse/2016/09/13/training-a-classifier-for-relation-extraction-from-medical-literature/`.

Kathleen Cotton. Classroom questioning. *School improvement research series*, 5:1–22, 1988.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170, 2014.

Kathleen Ellis. Teacher questioning behavior and student learning: What research says to teachers. 1993.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Arthur C Graesser, Haiying Li, and Carol Forsyth. Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, 23(5):374–380, 2014.

Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph J Williams, and Sharad Goel. Mathbot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence*, 2019.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

Brian William Junker, Yanna Weisberg, Lindsay Clare Matsumura, Amy Crosson, Mikyung Wolf, Allison Levison, and Lauren Resnick. *Overview of the instructional quality assessment*. Regents of the University of California, 2005.

Cindy A Kilgo, Jessica K Ezell Sheets, and Ernest T Pascarella. The link between high-impact practices and student learning: Some longitudinal evidence. *Higher Education*, 69(4):509–525, 2015.

Haiying Li and Arthur C Graesser. The impact of conversational agents' language on summary writing. *Journal of Research on Technology in Education*, 53(1):44–66, 2021.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, 2019.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020. URL `https://github.com/heartexlabs/label-studio`. Open source software available from https://github.com/heartexlabs/label-studio.

William W Wilen and Ambrose A Clegg Jr. Effective questions and questioning: A research review. *Theory & Research in Social Education*, 14(2):153–161, 1986.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A Appendix

Table 4: **Entity extraction task for uncertainty modeling**

Text	**Entities**	**Relation**	**Label**
The length of the object is 5, what is the width?	length, 5, width	(length, 5)	True
What is the scale factor?	scale factor	(scale factor, __)	False
No, the length is not 5, the width is.	length, 5, width	(width, 5)	True
No, the length is not 5, the width is.	length, 5, width	(length, 5)	False