

Board 408: Toward Building a Human-Computer Coding Partnership: Using Machine Learning to Analyze Short-Answer Explanations to Conceptually Challenging Questions

Harpreet Auby, Tufts University

Harpreet is a graduate student in Chemical Engineering and STEM Education. He works with Dr. Milo Koretsky and helps study the role of learning assistants in the classroom as well as machine learning applications within educational research and evaluation. He is also involved in projects studying the uptake of the Concept Warehouse. His research interests include chemical engineering education, learning sciences, and social justice.

Namrata Shivagunde, University of Massachusetts, Lowell

Anna Rumshisky, University of Massachusetts, Lowell

Dr. Milo Koretsky, Tufts University

Milo Koretsky is the McDonnell Family Bridge Professor in the Department of Chemical and Biological Engineering and in the Department of Education at Tufts University. He is also co-Director of the Institute for Research on Learning and Instruction (IRLI). He received his B.S. and M.S. degrees from UC San Diego and his Ph.D. from UC Berkeley, all in chemical engineering.

Toward Building a Human-Computer Coding Partnership: Using Machine Learning to Analyze Short-Answer Explanations to Conceptually Challenging Questions

Introduction

This NSF Grantee Poster Session paper describes work on an NSF-funded collaboration between engineering education and machine learning researchers to automate the coding of short-answer explanations written by students to conceptually challenging questions in mechanics and thermodynamics [1], [2]. Concept questions, sometimes called ConcepTests [3], are challenging multiple-choice questions that allow students to practice utilizing conceptual knowledge in new scenarios. These questions have been used within multiple active learning strategies to promote conceptual understanding and student engagement [4] - [11]. Furthermore, students can be asked to write short-answer explanations justifying their answer choice. Written justifications have been shown to improve student engagement and understanding and better prepare students for small-group collaborative work and whole-class discussions [12], [13], [14]. Evaluating these responses is helpful for instructors and researchers to understand student thinking; however, the amount of information can be daunting.

Machine learning has been used in a variety of ways in education research. Work done to evaluate student-constructed responses has included automatic scoring, text classification, or pattern recognition of responses [15] - [20]. Various unsupervised and supervised learning techniques have been used to do this, but transformer models have not been widely used to analyze responses [21] - [26], even with their greater ability to analyze text. These methods have allowed for improved assessment of student responses and motivated our interest in using machine learning to analyze student explanations to concept questions.

To accomplish this goal, we collect written responses available from consenting students in mechanics and thermodynamics courses through the Concept Warehouse (CW) [27], a web-based online tool for active learning. These responses are then manually coded using emergent and inductive coding approaches [28], [29], [30]. Finally, the written responses are also analyzed using large language model (LLMs)-based coding methods like T5 (Text-to-Text Transfer Transformer) [31], OpenAI's GPT-3, GPT-4 [32], Mixtral of Expert (MoE) [33], and ATLAS.ti AI Coding [34].

In our overall project, we aim to answer the following research questions:

1. What ideas do students use to explain their reasoning when writing short answer responses to conceptually challenging questions?
2. How well do transformer-based machine learning models replicate the human-coded data?
3. For two isomorphic question pairs, how similar is the human coding of one question relative to the other? How well do the machine learning models trained on the first question's explanations perform on the second question?

Our end goal is to create a generative Artificial Intelligence (AI) tool that can supplement the CW and give instructors and researchers a way to understand patterns and trends in student

responses that reveal their conceptual thinking and reasoning. This poster paper will provide an overview of our current progress in manually coding student responses and fine-tuning LLMs.

Background

Conceptually Challenging Questions and Short-Answer Explanations

We use the term concept questions to describe qualitative, multiple-choice questions that require students to identify foundational concepts and then apply them in new situations. Concept questions are sometimes called “ConcepTests” [3] and are a common type of clicker question [35]. These concept questions are often used within active learning practices, like Peer Instruction [3], to help students process conceptual knowledge and develop conceptual understanding. Concept-based active learning has been shown to improve student performance and help students develop conceptual understanding and problem-solving skills [4], [7], [36], [37].

In addition to concept-based active learning, instructors can ask students to write short-answer justifications for their answers to these conceptually challenging multiple-choice questions. Writing short-answer responses has been shown to improve student confidence, chances of picking a correct answer, and better prepares students for group and larger class discussions [12], [13], [38], [39], [40]. Thus, asking students concept questions and writing short answer responses has shown to be very beneficial to their learning; however, the large amount of written data can be too much for instructors to manage effectively.

NLP in Education

Machine learning has been used in education research in a variety of ways [15] - [20], including analyzing student writing and dialogue [41]. Various unsupervised and supervised machine-learning methods have been used to assess student-constructed responses. For example, unsupervised support vector machines (SVM) and logistic regression have been used to classify text based on a human-coded rubric [15], [16], [42] - [45]. Additionally, supervised neural networks have been used to analyze texts [21], [46] - [49].

The use of Transformer-based machine learning models [31], [50], [51], [52] in education research is an emerging method and even more novel for analyzing short answer responses [21] - [26]. For example, researchers have used BERT and RoBERTa [53] to automatically grade short answers [25], [26]. These models have been used to critique arguments in student essays and conduct essay scoring [22], [54]. Most of the earlier studies were focused on small encoder-only Transformer models, and they did not experiment with sequence-to-sequence and state-of-the-art decoder-only Large Language Models to assess students' written explanations in science education. Based on this, we identify a need to apply Transformer-based machine-learning models to automate coding and analysis of short answer explanations to conceptually challenging questions. The benefits of automated coding would provide researchers and instructors a more efficient way to analyze student responses. This work can also provide machine learning researchers with a further understanding of handling limited labeled data.

Below, we describe how we have leveraged the generative capabilities for sequence-to-sequence and larger decoder-only Transformer models to assess textual responses to conceptually challenging engineering questions written by students. Specifically, we used GPT-3 [50] and GPT-4 [32] via in-context learning and finetuned T5 [31] and Mixtral of Experts (MoE) [33] on a manually coded dataset to automate the qualitative coding of the student narratives of understanding.

Methods

Data Collection

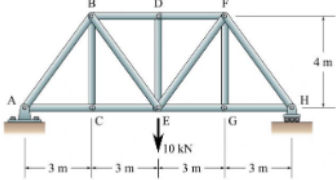
Participants in this study are students who consented to have their responses to short-answer concept questions used in research. Students are from a diverse array of two- and four-year institutions, which include minority-serving institutions, community colleges, teaching-centered universities, and R1 universities. Participating instructors are in varying research- and teaching-focused faculty positions. Enrollment in these courses varies from 25 - 100 students.

All data was collected through the Concept Warehouse (CW) [27], a web-based active learning tool. The CW serves as a content repository, a classroom response system to deliver content and collect student responses, and a learning analytics tool that provides data to instructors and researchers. We have collected and analyzed data on two different topics: mechanics and thermodynamics. We are actively collecting data in mechanics, while the analysis of thermodynamics responses comes from historical data collected in the tool. For the former source, eight common statics and dynamics concept questions were selected to ask across all institutions. The current common statics questions are related to the following topics:

- Q1: Moment of Force
- Q2: Trusses
- Q3: Static Friction
- Q4: Frames and Machines
- Q5: Forces
- Q6: 2-D Moments Concepts
- Q7: 3-D Moments Concepts
- Q8: Moment of Inertia

Instructors choose their preferred method for question delivery and often include, but aren't limited to, pre-class assignments, homework assignments, or in-class group work. In addition to the question, instructors also ask students three follow-ups: short answer justification, confidence rating, and question effectiveness, as shown in Figure 1. In the work involved in this project, we focus on the analysis of the short answer justification follow-up to understand how students utilize ideas to form narratives of understanding.

Which of the following best describes the force carried by the bar ED ?



☐ 10 kN
☐ about one third of 10 kN
☐ about one fifth of 10 kN
☐ approximately zero

Explanation

Please explain your answer in the box below.

Confidence

Please rate how confident you are with your answer.

substantially unsure moderately unsure neutral moderately confident substantially confident
☐ ☐ ☐ ☐ ☐

Question Effectiveness

Please help us assess the effectiveness of this question by answering the items below:

I understood what this question was asking.

strongly disagree moderately disagree neutral moderately agree strongly agree
☐ ☐ ☐ ☐ ☐

Explain your response to the item above.

Trying to answer this question made me think deeply about course material.

strongly disagree moderately disagree neutral moderately agree strongly agree
☐ ☐ ☐ ☐ ☐

Explain your response to the item above.

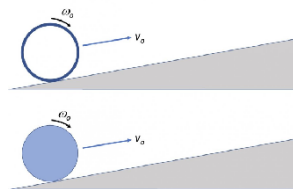
Figure 1. Example of a student's view of a question asked in this study. The question text and figure are provided along with the multiple-choice options. Additionally, instructors utilized the explanation, confidence, and question-effectiveness follow-ups.

For the mechanics data collection, we organized a Community of Practice [55], [56], which brought together participating instructors twice a term to discuss the use of the common

questions and the implementation of the CW in their classrooms more broadly. During these meetings, it was often decided if concept questions should be revised or if we wanted to focus on a different concept question. Figure 2 shows an example of an activity done in a Community of Practice session with instructors where short-answer responses written by students were discussed amongst the group to evaluate what instructors and researchers could learn from them.

CW CT 6141 - Example Response #2

Each of the objects - the pipe and the solid cylinder - is rolling uphill along a rough surface with the same velocity v_0 and the same angular velocity. The cylinders have the same mass and radius but different cross-sectional areas. Compare the distance d that each object will travel before stopping.



- ☐ The pipe travels farther up the hill
- ☐ The solid cylinder travels farther up the hill
- ☐ The pipe and cylinder travel the same distance up the hill

Pipe has a larger I , lose more speed in rotation.

partially understood the problem statement - only operate on a portion of it

got the key concept (MMOI), left out all kinds of facts though

Newton's 2nd Law? --> Connect to I --> connect to α

Language (terminology) issue or concept issue?

I (stiffness and strength), MMOI (in motion) --> are they differentiating this?

Losing more speed over time? distance?

Figure 2. Screenshot of interactive activity done during a Community of Practice meeting.

Our work has recently expanded to analyze short-answer explanations to conceptually challenging questions in engineering thermodynamics. These questions test students on enthalpy and entropy, two commonly challenging concepts [57].

Data Analysis: Qualitative Coding

Coding approaches have evolved throughout this project; however, the basis of our processes has utilized a combination of *a priori* and emergent approaches [28], [29], [30]. Coding involved generating an “ideal” response to implement aspects of *a priori* coding and thinking about how students may use concepts in the question of interest. These preliminary ideas and other emergent codes from written explanations were then iteratively refined to create a stable codebook that described the resources students used to formulate a narrative of understanding. We grouped these codes into three categories:

- **Identification:** The student identifies a concept or other piece of information.
- **Comparison:** The student compares a concept across two different system states.
- **Inference:** The student concludes about the system's state based on the information in their response.

Data Analysis: Machine Learning

Analyzing short-answer responses was defined as a sequence labeling problem where the spans of the students' responses were coded with manually coded labels. Instead of training the large language model from scratch, we leveraged transfer learning via fine-tuning, and in-context learning. In fine-tuning, we use a pre-trained model and train it further on the coded responses. The pre-trained model is a language model that initially has undergone training on a large corpus of free text. In in-context learning, we prompt the model with a few samples and task it to generate the coded response for a new student response instance. We do not train the model in in-context learning. We've utilized the following models throughout this work:

- **T5 (Text-to-Text Transfer Transformer):** A sequence-to-sequence model that was used to formulate a task into a text-to-text format and fine-tuned T5-base (220M parameters) and T5-large (770M parameters) [31] with 20 to 240 manually coded responses.
- **GPT-3 and GPT-4:** A transformer decoder model with 175B and more parameters trained using a “causal” language modeling approach. We present the model with a prompt consisting of an instruction, a few examples, and a new set of inputs. It then outputs a coded response. GPT-4 [32] is an advanced version of GPT-3 [50] that is better able to understand and generate natural language text.
- **Mixtral of Expert (MoE):** A 47B parameter model with eight distinct groups of parameters called “experts.” For every token, the model chooses two out of eight experts and combines their output additively. This results in 13 billion active parameters for each token the model processes. It is a large decoder-only transformer-based language model that we finetuned on the manually-coded dataset using Huggingface’s transformer library [33], [58].
- **ATLAS.ti AI Coding:** An automated coding feature on the ATLAS.ti qualitative data software that uses OpenAI to prompt qualitative coding [34].

To understand the effectiveness of the machine learning models, we compare model-generated codes to human-written codes. We use an Exact Match metric to compare the model-predicted coded response to the ground truth response, which involves counting the number of codes in the model-generated responses that match exactly with the codes in manually coded responses. We also compute Precision, Recall, and F1 scores for each model. Precision is the percentage of correct model-generated codes relative to the total number. Recall is the percentage of human codes that the model could generate correctly. The F1 score is the harmonic mean of precision and recall. Additionally, since some models generate new codes to apply to responses, we analyze those newly generated codes to see if they are reasonable to include in the codebook or not applicable.

Qualitative Coding Challenges and Limitations

Manual coding to train models takes a substantial amount of time, and to improve credibility, additional coders could be involved for the data described below. Additionally, students are on different paths toward utilizing disciplinary concepts when writing responses, so some students describe concepts with language closer to their everyday language. It is still important to capture

this within manual and automated coding as it can help instructors and researchers learn about the cognitive resources and associated language used by students to describe challenging science and engineering concepts. This is a challenge as human coders need to code all instances of everyday language and disciplinary language associated with the same concept, and there are usually a small number of samples within an already limited data set that have instances of everyday language to describe disciplinary concepts.

Machine Learning Challenges and Limitations

In our study, we used LLMs, which are multi-layer neural networks with billions of parameters trained on large amounts of free text. These models learn to predict the next word based on the context, and for this reason, they also pick up biases present in the text on which they are trained. For example, they might favor certain writing styles seen during training, potentially affecting how they annotate student narratives. No identifiable information or protected attributes such as gender or race are included in our training data, precluding the introduction of additional biases. However, the biases associated with LLMs remain an issue to address. Rather than looking at the machine as an authority, we look at it as a partner. That puts us in a better position to evaluate biases, but with any collaborative project, there are some things that we will not be able to attend to.

In addition to biases, creating effective prompts was also a challenge. Each machine learning model requires a different input and output format for optimal performance. Therefore, one challenge was to methodically design the input and output prompts, through repeated testing and adjustments, specific to a given machine learning model.

Findings

In this section, we describe the findings of our work based on our activities as mentioned above.

ASEE 2022

Our first ASEE collaboration [1] investigated the use of a Text-to-Text Transformer (T5) [31] and GPT-3 [50] to automate the coding of 290 short-answer explanations to a statics conceptual question. This conceptual question, shown in Figure 3, asked students to calculate the force of friction on a block after a pushing force was applied. A combination of *a priori* and emergent coding methods was used to manually code the responses, where coders identified cognitive resources students used to construct their narratives of understanding for this question. These responses were then automatically coded by two large pre-trained generative sequence language models: T5 [31] and GPT-3 [50].

Force $P = 10 \text{ N}$ is applied to the block of mass $m = 5 \text{ kg}$ on a horizontal rough surface with $\mu_s = 0.3$ and $\mu_k = 0.25$.

If $g = 9.81 \text{ m/s}^2$, what is the force of friction on the block?



- ☐ 10 N
- ☐ 12.26 N
- ☐ 14.7 N
- ☐ 45.1 N

Please explain your answer in the box below.

Please rate how confident you are with your answer.

- | | | | | |
|-------------------------|-----------------------|-----------------------|-------------------------|----------------------------|
| substantially
unsure | moderately
unsure | neutral | moderately
confident | substantially
confident |
| ☐ | ☐ | ☐ | ☐ | ☐ |

Submit

Figure 3. Concept question 5703 asks students to think about the force of friction on a block.

We found that T5 performed better than GPT-3, as the former would produce new codes not present in the training examples, as shown in Table 1. Through this preliminary work, we found potential for analyzing short-answer explanations using pre-trained text models like T5 [31] and GPT-3 [50]. Table 1 shows the results of this work.

Table 1. GPT-3 and T5 Model Performance on Concept Question 5703. Table reproduced from [1].

	# correct codes	No. of codes	Precision	Recall	F1	Incorrect but makes sense	Does not make sense	Codes missed
Ground truth	175	NA	NA	NA	NA	NA	NA	NA
t5-base-f020	0	0	0	0	0	0	0	175
t5-base-f050	40	49	0.82	0.23	0.36	2	7	126
t5-base-f100	60	90	0.67	0.34	0.45	14	16	85
t5-base-f150	80	92	0.87	0.46	0.6	7	5	83
t5-base-f200	93	126	0.74	0.54	0.62	19	14	49
t5-base-f240	105	133	0.79	0.6	0.68	14	14	42
t5-large- f150	107	118	0.91	0.61	0.73	6	5	57
gpt3- davinci- Instruct	89	189	0.47	0.51	0.49	52	48	-14

ASEE 2023

Our second ASEE collaboration [2] utilized unsupervised machine learning techniques to analyze 160 short-answer responses to a mechanics conceptual question. As shown in Figure 4, this question asked students to determine if a solid or hollow cylinder would make it to the top of the ramp first. Similar manual coding processes were conducted as previously described [1]. Then, we used text summary, text modeling, and a Naïve Bayes Classifier to understand how common unsupervised machine learning techniques could be used to understand student narratives of understanding in short-answer responses. Within this work, we began to integrate principles of Linguistic Justice in our coding and machine learning processes. Linguistic justice is a conceptual framework that aims to ensure that all have equitable access to political or social life through language [59], [60]. To promote linguistic justice in our work, we established ideas of involving a human-computer partnership that can work together to analyze student responses.

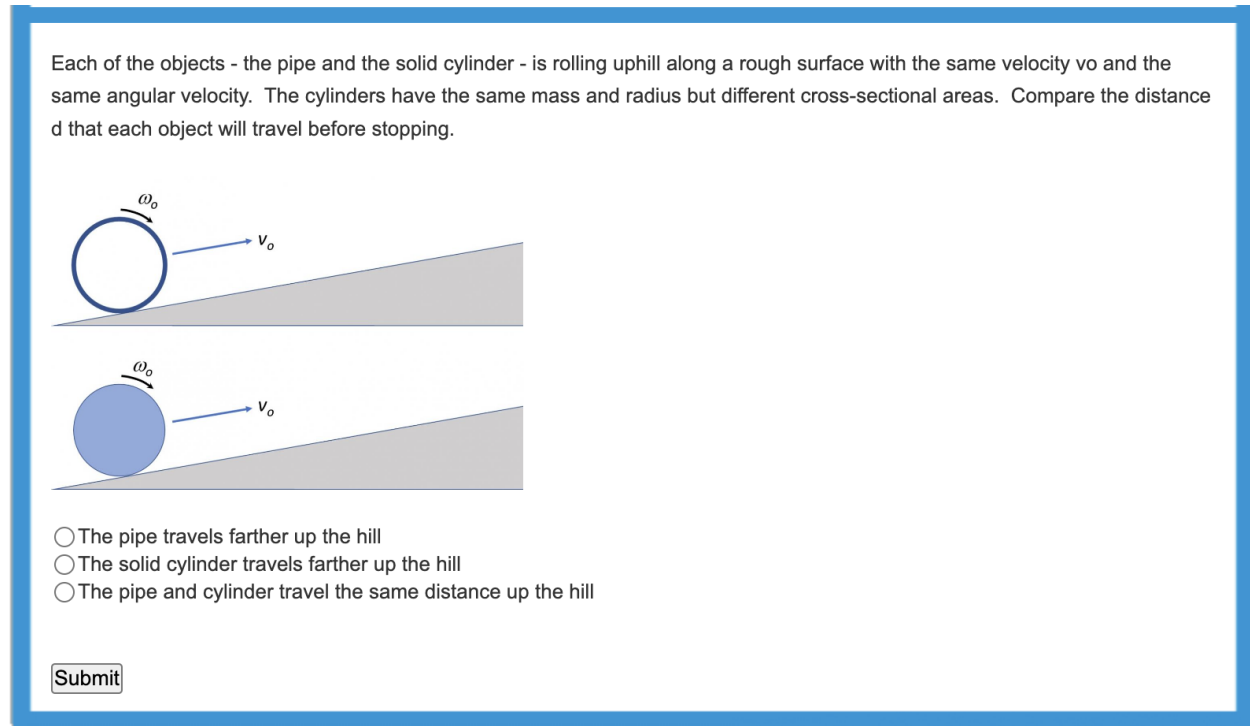


Figure 4. The concept question was used in the preliminary work of this study.

ASEE 2024

In this study, we shifted our focus to analyzing two related engineering thermodynamics concept questions and using new LLMs. GPT-4, Mixtral of Expert (MoE), and ATLAS.ti were used to analyze responses. Questions were manually coded and included an enthalpy of mixing questions (1396 responses) and an entropy of mixing questions (1387 responses), shown in Figure 5. We utilized coding processes similar to the previous two years to code the responses manually; however, we began to integrate a resources-based framework into the overall analytical framework [61], [62], [63]. Through comparison of these LLMs, we achieved an F1 score of 62% on the thermodynamics test set when MoE was trained on the thermodynamics training set. Table 2 summarizes the various model performances on the thermodynamic combined test set (which includes both enthalpy and entropy-balanced test samples) when trained on a combined training set. GPT-4 achieved its highest F1 score of 48% on the test set, with entropy in-context examples. When we trained MoE on the statistics training dataset and evaluated on the thermodynamics test set, we observed an F1 score of 32%. ATLAS AI Interactive coding scores lowest at an F1 score of 10%.

Consider 0.3 mol of gas A and 0.5 mol of gas B, that behave as ideal gases. When these two gases are mixed at constant T and P, the enthalpy change of mixing is:

☐ $\Delta h_{\text{mix}} > 0$
☐ $\Delta h_{\text{mix}} < 0$
☒ $\Delta h_{\text{mix}} = 0$
☐ You cannot tell unless you know C_p

Please explain your answer in the box below.

A

Consider 0.3 mol of gas A and 0.5 mol of gas B, that behave as ideal gases. When these two gases are mixed at constant T and P, the entropy change of mixing is:

☒ $\Delta s_{\text{mix}} > 0$
☐ $\Delta s_{\text{mix}} < 0$
☐ $\Delta s_{\text{mix}} = 0$
☐ You cannot tell unless you know C_p

Please explain your answer in the box below.

B

Figure 5. Thermodynamics concept questions that were used in this study.

Table 2. Comparison of ground truth and model-generated responses on enthalpy and entropy combined test set. The highest value is in bold.

Model	No. of correct codes	No. of codes	Precision	Recall	F1
Ground truth No. of codes	1244				
MoE trained on Enthalpy+Entropy datasets	931	1746	0.53	0.75	0.62
MoE trained on Enthalpy+Entropy+Statics	917	1719	0.53	0.74	0.62
MoE trained on Enthalpy dataset	782	1670	0.47	0.63	0.54
MoE trained on Entropy dataset	902	2459	0.37	0.73	0.49
MoE trained on Statics dataset	383	1176	0.33	0.31	0.32
GPT4 (enthalpy examples as in-context examples)	522	981	0.53	0.42	0.47
GPT4 (entropy examples as in-context examples)	570	1146	0.50	0.46	0.48
ATLAS AI Interactive Coding	221	3094	0.07	0.18	0.10

In summary, this work found that MoE trained on a thermodynamics dataset achieved the highest F1 score on both datasets. We also found that the entropy dataset is more challenging for MoE and GPT-4 than the enthalpy dataset. Additionally, our study shows that the model can tackle other tasks better when trained or prompted with examples from a more challenging dataset.

Implications and Future Work

This work contributes to machine learning in education research by showing that LLMs can be utilized to analyze short-answer responses in the few-shot approach. As we plan to form a human-computer partnership to create an AI assistant tool for the CW, we want to iterate our qualitative coding and use of machine learning tools before we create and launch our final tool. Regarding our qualitative coding, we have begun to integrate a resources-based framing [61], [62], [63] into the coding scheme, which can help us further investigate how students use pieces of knowledge in specific contexts. This will require more manual annotation of a few thousand samples and fine-tuning a large language model (LLM) on this data. Regarding machine learning, we formulated the problem in our study as a sequence labeling problem, where the spans of the student responses are manually coded with labels. The Exact Match metric provides some insight into the model's performance. However, as expected – and as our qualitative analysis confirms – this metric falls short in cases where the model predicts codes that are semantically similar, but not exact matches. In our study, we performed a manual qualitative analysis to gain a better understanding of the models' capabilities. In follow-up work, we expect to shift to model-based evaluation metrics such as BERTScore [64] that can account for lexical variation. The work highlighted above showed that LLMs can generalize to new student responses to the same questions. We aim to extend this work to ensure that models can generalize to new questions and generate response summaries.

Furthermore, we aim to create an AI assistant tool for the CW, which will be offered as a plugin or a separate interface to the existing CW platform. The tool will annotate student responses, capturing the student's thinking process. The tool will allow the instructors to consolidate the insights from student responses, generating reports and graphs to represent differences in students' thinking around a given concept question, which they can then use to inform their teaching practices. Additionally, through this tool, researchers can further understand student thinking by having coded student responses on a scale that is not possible with manual coding. Through tool development, we aim to ensure that our qualitative coding and ML processes account for disciplinary and everyday language in students' responses. This can help us make the tool a more inclusive generative AI tool that understands the various language students may use to explain their thinking. In turn, instructors and researchers will be more aware of the diverse language and thought patterns students use to wrestle with challenging concepts in the discipline.

Acknowledgments

We acknowledge the support from the National Science Foundation (NSF) through grant EEC 2226553. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

References

- [1] H. Auby, N. Shivagunde, A. Rumshisky, and M. Koretsky, "WIP: Using machine learning to automate coding of student explanations to challenging mechanics concept questions," in *Proceedings of the 2022 American Society of Engineering Education Annual Conference & Exposition*, Jun. 2022. [Online]. Available: <https://peer.asee.org/40507>
- [2] H. Auby and M. Koretsky, "Work in progress: Using machine learning to map student narratives of understanding and promoting linguistic justice," in *Proceedings of the 2023 American Society of Engineering Education Annual Conference & Exposition*, Jun. 2023.
- [3] E. Mazur, *Peer Instruction: A User's Manual*. in Series in Educational Innovation. Prentice Hall, 1997.
- [4] T. Vickrey, K. Rosploch, R. Rahmanian, M. Pilarz, and M. Stains, "Research-based implementation of peer instruction: A literature review," *CBE—Life Sci. Educ.*, vol. 14, no. 1, p. es3, Mar. 2015, doi: 10.1187/cbe.14-11-0198.
- [5] S. Freeman *et al.*, "Active learning increases student performance in science, engineering, and mathematics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 23, pp. 8410–8415, Jun. 2014, doi: 10.1073/pnas.1319030111.
- [6] D. C. Haak, J. HilleRisLambers, E. Pitre, and S. Freeman, "Increased structure and active learning reduce the achievement gap in introductory biology," *Science*, vol. 332, no. 6034, pp. 1213–1216, 2011.
- [7] C. H. Crouch and E. Mazur, "Peer Instruction: Ten years of experience and results," *Am. J. Phys.*, vol. 69, no. 9, pp. 970–977, Sep. 2001, doi: 10.1119/1.1374249.
- [8] A. P. Fagen, C. H. Crouch, and E. Mazur, "Peer Instruction: Results from a range of classrooms," *Phys. Teach.*, vol. 40, no. 4, pp. 206–209, Apr. 2002, doi: 10.1119/1.1474140.
- [9] I. dos Santos Belmonte, A. V. Borges, and I. T. S. Garcia, "Adaptation of physical chemistry course in COVID-19 period: Reflections on Peer Instruction and team-based learning," *J. Chem. Educ.*, vol. 99, no. 6, pp. 2252–2258, Jun. 2022, doi: 10.1021/acs.jchemed.1c00529.
- [10] T. Gok and O. Gok, "Peer Instruction in chemistry education: Assessment of students' learning strategies," *Learn. Strateg.*, vol. 17, no. 1, 2016.
- [11] M. F. Golde, C. L. McCreary, and R. Koeske, "Peer Instruction in the general chemistry laboratory: Assessment of student learning," *J. Chem. Educ.*, vol. 83, no. 5, p. 804, May 2006, doi: 10.1021/ed083p804.
- [12] M. D. Koretsky, B. J. Brooks, R. M. White, and A. S. Bowen, "Querying the questions: Student responses and reasoning in an active learning class," *J. Eng. Educ.*, vol. 105, no. 2, pp. 219–244, 2016, doi: 10.1002/jee.20116.
- [13] M. D. Koretsky, B. J. Brooks, and A. Z. Higgins, "Written justifications to multiple-choice concept questions during active learning in class," *Int. J. Sci. Educ.*, vol. 38, no. 11, pp. 1747–1765, Jul. 2016, doi: 10.1080/09500693.2016.1214303.
- [14] S. A. Finkenstaedt-Quinn, M. Petterson, A. Gere, and G. Shultz, "Praxis of Writing-to-Learn: A model for the design and propagation of Writing-to-Learn in STEM," *J. Chem. Educ.*, vol. 98, no. 5, pp. 1548–1555, May 2021, doi: 10.1021/acs.jchemed.0c01482.
- [15] X. Zhai, Y. Yin, J. W. Pellegrino, K. C. Haudek, and L. Shi, "Applying machine learning in science assessment: a systematic review," *Stud. Sci. Educ.*, vol. 56, no. 1, pp. 111–151, Jan. 2020, doi: 10.1080/03057267.2020.1735757.

- [16] X. Zhai, K. C. Haudek, M. A. Stuhlsatz, and C. Wilson, "Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment," *Stud. Educ. Eval.*, vol. 67, p. 100916, 2020.
- [17] X. Zhai, L. Shi, and R. H. Nehm, "A meta-analysis of machine learning-based science assessments: factors impacting machine-human score agreements," *J. Sci. Educ. Technol.*, vol. 30, no. 3, pp. 361–379, 2021.
- [18] J. Burstein *et al.*, Eds., *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA → Online: Association for Computational Linguistics, 2020. [Online]. Available: <https://aclanthology.org/2020.bea-1.0>
- [19] J. Burstein *et al.*, Eds., *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. Online: Association for Computational Linguistics, 2021. [Online]. Available: <https://aclanthology.org/2021.bea-1.0>
- [20] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. Srinivasa, "Automatic assessment of text-based responses in post-secondary education: A systematic review," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100206, Jun. 2024, doi: 10.1016/j.caeai.2024.100206.
- [21] Y. Liu, J. Han, A. Sboev, and I. Makarov, "GEEF: A neural network model for automatic essay feedback generation by integrating writing skills assessment," *Expert Syst. Appl.*, vol. 245, p. 123043, 2023, doi: 10.1016/j.eswa.2023.123043.
- [22] T. Alhindi, B. McManus, and S. Muresan, "What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, H. Li, G.-A. Levow, Z. Yu, C. Gupta, B. Sisman, S. Cai, D. Vandyke, N. Dethlefs, Y. Wu, and J. J. Li, Eds., Singapore and Online: Association for Computational Linguistics, Jul. 2021, pp. 380–391. doi: 10.18653/v1/2021.sigdial-1.40.
- [23] E. Mayfield and A. W. Black, "Should you fine-tune BERT for automated essay scoring?," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, Eds., Seattle, WA, USA → Online: Association for Computational Linguistics, Jul. 2020, pp. 151–162. doi: 10.18653/v1/2020.bea-1.15.
- [24] P. U. Rodriguez, A. Jafari, and C. M. Ormerod, "Language models and automated essay scoring." *arXiv*, Sep. 18, 2019. doi: 10.48550/arXiv.1909.09482.
- [25] X. Zhu, H. Wu, and L. Zhang, "Automatic short-answer grading via BERT-based deep neural networks," *IEEE Trans. Learn. Technol.*, vol. 15, no. 3, pp. 364–375, 2022, doi: 10.1109/TLT.2022.3175537.
- [26] M. A. Sayeed and D. Gupta, "Automate descriptive answer grading using reference based models," in *2022 OITS International Conference on Information Technology (OCIT)*, Bhubaneswar, India: IEEE, Dec. 2022, pp. 262–267. [Online]. Available: 10.1109/OCIT56763.2022.00057
- [27] M. D. Koretsky *et al.*, "The AICHe Concept Warehouse: A web-based tool to promote concept-based instruction," *Adv. Eng. Educ.*, vol. 4, no. 1, p. 27, 2014.
- [28] M. B. Miles, A. M. Huberman, and J. Saldana, *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, 2018.
- [29] J. Saldaña, *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2021.

- [30] J. W. Creswell, *Qualitative inquiry and research design: choosing among five approaches*, 3rd ed. Los Angeles: SAGE Publications, 2013.
- [31] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified Text-to-Text Transformer.” arXiv, Jul. 28, 2020. Accessed: Apr. 03, 2023. [Online]. Available: <http://arxiv.org/abs/1910.10683>
- [32] OpenAI, “GPT-4 Technical Report.” arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.
- [33] A. Q. Jiang *et al.*, “Mixtral of Experts.” arXiv, Jan. 08, 2024. doi: 10.48550/arXiv.2401.04088.
- [34] “AI Coding powered by OpenAI,” ATLAS.ti. [Online]. Available: <https://atlasti.com/ai-coding-powered-by-openai>
- [35] D. Duncan, “Clickers: A new teaching aid with exceptional promise,” *Astron. Educ. Rev.*, vol. 5, no. 1, pp. 70–88, 2006.
- [36] G. Mora, “Peer Instruction and lecture tutorials equally improve student learning in introductory geology classes,” *J. Geosci. Educ.*, vol. 58, no. 5, pp. 286–296, Nov. 2010, doi: 10.5408/1.3559693.
- [37] M. K. Smith *et al.*, “Why peer discussion improves student performance on in-class concept questions,” *Science*, vol. 323, no. 5910, pp. 122–124, Jan. 2009, doi: 10.1126/science.1165919.
- [38] M. D. Koretsky and A. J. Magana, “Using technology to enhance learning and engagement in engineering,” *Adv. Eng. Educ.*, 2019, Accessed: Oct. 21, 2023. [Online]. Available: <https://eric.ed.gov/?id=EJ1220296>
- [39] S. Brown, D. Montfort, N. Perova-Mello, B. Lutz, A. Berger, and R. Streveler, “Framework theory of conceptual change to interpret undergraduate engineering students’ explanations about mechanics of materials concepts,” *J. Eng. Educ.*, vol. 107, no. 1, pp. 113–139, 2018, doi: 10.1002/jee.20186.
- [40] A. T. Kararo, R. A. Colvin, M. M. Cooper, and S. M. Underwood, “Predictions and constructing explanations: an investigation into introductory chemistry students’ understanding of structure–property relationships,” *Chem. Educ. Res. Pract.*, vol. 20, no. 1, pp. 316–328, 2019, doi: 10.1039/C8RP00195B.
- [41] Y. Meng, A. Rumshisky, and F. Sullivan, “Automatic labeling of problem-solving dialogues for computational microgenetic learning analytics,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. Accessed: Apr. 15, 2024. [Online]. Available: <https://aclanthology.org/L18-1639>
- [42] K. C. Haudek and X. Zhai, “Examining the effect of assessment construct characteristics on machine learning scoring of scientific argumentation,” *Int. J. Artif. Intell. Educ.*, Dec. 2023, doi: 10.1007/s40593-023-00385-8.
- [43] K. J. Kim, D. S. Pope, D. Wendel, and E. Meir, “WordBytes: Exploring an intermediate constraint format for rapid classification of student answers on constructed response assessments,” *J. Educ. Data Min.*, vol. 9, no. 2, pp. 45–71, Dec. 2017, doi: 10.5281/zenodo.3554722.
- [44] L. N. Jescovitch *et al.*, “Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science

- learning progression,” *J. Sci. Educ. Technol.*, vol. 30, no. 2, pp. 150–167, Apr. 2021, doi: 10.1007/s10956-020-09858-0.
- [45] M. Shiroda, J. D. Uhl, M. Urban-Lurain, and K. C. Haudek, “Comparison of computer scoring model performance for short text responses across undergraduate institutional types,” *J. Sci. Educ. Technol.*, vol. 31, no. 1, pp. 117–128, Feb. 2022, doi: 10.1007/s10956-021-09935-y.
 - [46] H. Luan and C.-C. Tsai, “A review of using machine learning approaches for precision education,” *Educ. Technol. Soc.*, vol. 24, no. 1, pp. 250–266, 2021.
 - [47] R. Jiang, J. Gouvea, D. Hammer, E. Miller, and S. Aeron, “Automatic coding of students’ writing via Contrastive Representation Learning in the Wasserstein space.” arXiv, Dec. 01, 2020. doi: 10.48550/arXiv.2011.13384.
 - [48] J. M. Rosenberg and C. Krist, “Combining machine learning and qualitative methods to elaborate students’ ideas about the generality of their model-based explanations,” *J. Sci. Educ. Technol.*, vol. 30, no. 2, pp. 255–267, 2021.
 - [49] N. Yeruva, S. Venna, H. Indukuri, and M. Marreddy, “Triplet loss based Siamese networks for automatic short answer grading,” in *Proceedings of the 14th annual meeting of the forum for information retrieval evaluation*, Kolkata, India, 2023. doi: 10.1145/3574318.3574337.
 - [50] T. B. Brown *et al.*, “Language models are few-shot learners.” arXiv, Jul. 22, 2020. Accessed: Apr. 03, 2023. [Online]. Available: <http://arxiv.org/abs/2005.14165>
 - [51] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
 - [52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
 - [53] Y. Liu *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach.” arXiv, Jul. 26, 2019. doi: 10.48550/arXiv.1907.11692.
 - [54] D. Ghosh, B. B. Klebanov, and Y. Song, “An Exploratory Study of Argumentative Writing by Young Students: A Transformer-based Approach,” *ArXiv200609873 Cs*, Jun. 2020, Accessed: Apr. 05, 2022. [Online]. Available: <http://arxiv.org/abs/2006.09873>
 - [55] J. Lave and E. Wenger, *Situated learning: Legitimate peripheral participation*. New York, NY, US: Cambridge University Press, 1991.
 - [56] E. Wenger, *Communities of practice: Learning, meaning, and identity*. New York, NY, US: Cambridge University Press, 1998.
 - [57] K. Bain, A. Moon, M. R. Mack, and M. H. Towns, “A review of research on the teaching and learning of thermodynamics at the university level,” *Chem. Educ. Res. Pract.*, vol. 15, no. 3, pp. 320–335, Jul. 2014, doi: 10.1039/C4RP00011K.
 - [58] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing.” Association for Computational Linguistics, pp. 38–45, Oct. 2020. Accessed: Feb. 07, 2024. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
 - [59] J. Nee, G. M. Smith, A. Sheares, and I. Rustagi, “Linguistic justice as a framework for designing, developing, and managing natural language processing tools,” *Big Data Soc.*, vol. 9, no. 1, p. 20539517221090930, Jan. 2022, doi: 10.1177/20539517221090930.

- [60] J. Nee, G. Macfarlane Smith, A. Sheares, and I. Rustagi, “Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology,” in *Equity and Access in Algorithms, Mechanisms, and Optimization*, New York, NY, US: ACM, Oct. 2021, pp. 1–9. doi: 10.1145/3465416.3483301.
- [61] D. Hammer, “Student resources for learning introductory physics,” *Am. J. Phys.*, vol. 68, no. S1, pp. S52–S59, Jul. 2000, doi: 10.1119/1.19520.
- [62] A. Elby and D. Hammer, “Epistemological resources and framing: a cognitive framework for helping teachers interpret and respond to their students’ epistemologies,” in *Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice*, F. C. Feucht and L. D. Bendixen, Eds., Cambridge: Cambridge University Press, 2010, pp. 409–434. doi: 10.1017/CBO9780511691904.013.
- [63] M. C. Wittmann, “Research in the resources framework: Changing environments, consistent exploration.” arXiv, Jan. 29, 2018. Accessed: Nov. 07, 2023. [Online]. Available: <http://arxiv.org/abs/1801.09592>
- [64] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT.” arXiv, Feb. 24, 2020. doi: 10.48550/arXiv.1904.09675.