Using Machine Learning to Analyze Short-Answer Responses to Conceptually Challenging Chemical Engineering Thermodynamics 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.

Utilizing Machine Learning to Analyze Short-Answer Responses to Conceptually Challenging Chemical Engineering Thermodynamics Questions

Introduction

This paper describes the results of a study where generative Artificial Intelligence (AI) was used to analyze short-answer explanations to two conceptually challenging chemical engineering thermodynamics problems. This work comes from a collaboration between machine learning and engineering education researchers utilizing machine learning to analyze student narratives of understanding in short-answer explanations to conceptually challenging questions [1], [2].

Concept questions, sometimes called ConcepTests [3], are multiple-choice questions involving minimal calculations and give students experience applying conceptual knowledge. When utilized within active learning pedagogies, concept questions have been shown to improve student achievement, engagement, and have helped students develop conceptual understanding and problem-solving skills [4] - [14]. Additionally, when students are asked to write short-answer responses to explain their reasoning to concept questions, it has been observed to improve student performance, engagement, and prepare students for group discussion [15], [16]. These responses provide instructors and researchers with a wealth of information regarding student thinking [17]. Still, often, it is difficult for instructors and researchers to process all of this written information. Machine learning researchers have applied natural language processing (NLP) and large language models (LLMs) to automate the grading and scoring of textual responses from students and have shown that it has great potential to help instructors and researchers understand student thinking about complex concepts [18] - [24].

We utilized written responses from consenting students in the Concept Warehouse (CW) [25], a web-based online tool for active learning. Two related questions from chemical engineering thermodynamics were manually coded using emergent and inductive coding [26], [27], [28]: an enthalpy of mixing question (1396 responses) and an entropy of mixing question (1387 responses).

The written responses were then analyzed using LLM-based coding methods. We split each manually coded thermodynamics dataset into training, validation, and test sets. We used in-context learning for GPT-4 [29], where we prompted the model with the question, four in-context examples of answers, and the corresponding codes and instructed it to generate the code(s) for the new answer instance. The in-context examples for GPT-4 prompt are drawn from the training split of the manually-coded dataset. We finetuned the Mixtral of Experts (MoE) [30] model using input and target pairs derived from the manually-coded training datasets. This trained model was then prompted with new test inputs, and the model-generated coded sequence was evaluated against the manually coded target sequence. We evaluated both models on a test set of around 140 samples for each thermodynamics question. Using manual and language model-based coding, we aim to answer two research questions:

1. What aspects of student thinking are present in narratives of understanding constructed to justify answers to conceptual questions about the enthalpy and entropy of mixing ideal gases?
2. To what extent can we use Large Language Models to automate qualitative coding of student narratives of understanding?

Background

Conceptual Questions and Student Responses

Concept questions, sometimes called ConcepTests [3], are conceptually challenging multiple-choice questions involving minimal to no calculation. They allow students to identify and apply concepts to new scenarios. Concept questions are often used with active learning practices, like Peer Instruction (PI) [3]. Concept questions utilized within PI or other active learning strategies have been shown to improve student performance and help students develop conceptual understanding and problem-solving skills [4] - [14].

In addition to asking conceptual questions, instructors can ask students to write short-answer responses after asking conceptually challenging questions. Writing has been shown to improve critical thinking and learning because it is a way to organize one's thoughts and focus on understanding and communicating specific ideas [31]. Writing-to-learn (WTL) is one evidence-based learning strategy utilized in STEM classrooms where students write brief, low-stakes explanations where they can practice using content knowledge in writing. WTL has been shown to support the development of conceptual understanding and metacognition [31] - [34], and previous work where writing short-answer explanations in response to conceptually challenging questions in STEM classrooms has found that writing improves student confidence, chances of picking a correct answer and better prepare students for group and larger class discussions [15], [16], [35], [36]. Additionally, written responses give instructors insight into student thinking and how they utilize pieces of knowledge to construct explanations. Thus, these responses provide a wealth of information. However, it is often difficult for instructors and researchers to read and analyze large amounts of text to find information about trends and patterns in student thinking.

Machine Learning Applications to Education Research

Machine learning has been used in education for various text-based automated assessment tasks, such as automatic grading, automated text classification, automated feedback systems, and evaluating student writing in both short and long formats, like short constructed responses and essays [18] - [24].

Many of the earlier works used classical machine learning algorithms such as SVM, Naive-Bayes, Random Forest, and Logistic Regression [19], [37] - [41], while others used neural networks [41] - [45] to assess student-written responses. Few of the studies used transformer-based models [46] - [49] to analyze student textual narratives [24], [45], [50]. For example, some studies [51], [52], [53] have used BERT Field [43] to evaluate essays, while others have used BERT [24], [50] and RoBERTa [54] respectively, to conduct automatic grading of short-answers. Previously, we [1] finetuned T5 [48] and compared its results in assessing short student responses with GPT-3 [49]. However, we [1] only worked with one coded dataset. Most of these

studies focused on small encoder-only or sequence-to-sequence Transformer models. They did not train the state-of-the-art decoder-only Large Language Model's performance in assessing students' written explanations in science education.

The state-of-the-art decoder-only transformer models are multi-layer neural networks with attention mechanisms. These state-of-the-art models have billions of parameters and are trained on huge corpora of free text with the causal language modeling objective, which involves predicting the next word (or, more precisely, token) given the preceding context. These models can be used for in-context learning, where the model is queried with a prompt and is expected to generate a response to this prompt, or they can be finetuned on task-specific data. Bigger models (with billions of parameters) are more sample-efficient and require fewer manually-coded samples to perform better than smaller models (with a few hundred million parameters), making manual coding less laborious.

Earlier work used Transformer models like BERT, RoBERTa, or T5, which consist of less than one billion parameters, limiting our understanding of how well the new, bigger, sample-efficient state-of-the-art language model can help assess student responses. In our study, we leverage generative capabilities for larger decoder-only Transformer models to assess textual responses to conceptually challenging engineering questions written by students. Specifically, we used GPT-4 in-context learning [29] and finetuned Mixtral of Experts (MoE) [30] to automate the qualitative coding of the student narratives. Mixtral of Experts is a 47 billion-parameter model with eight distinct groups of parameters called “experts.” The model chooses two out of eight experts for every token and combines their output additively. This results in 13 billion active parameters for each token the model processes.

Conceptual Framework

We frame student explanations to conceptually challenging chemical engineering thermodynamics questions as “narratives” because students use a combination of everyday and discipline-specific language to tell a story about a concept or a set of concepts. Thus, we take a resources-based approach to analyzing student thinking. This approach considers cognitive resources as “fine-grained knowledge elements that a student possesses, the activation of which depends on context” [55, p. 410]. As students write short answer responses, they formulate a narrative of understanding that shows the activation and application of pieces of knowledge that they regard as essential to explain a phenomenon. However broad or specific these pieces of knowledge may be, they are not isolated. Thus, we must contextualize all pieces of knowledge we find relative to one another [35], [55] - [58]. The connections between them are essential to understand, and we applied this thinking to our coding scheme so machine learning models could be trained effectively.

When using generative AI within discipline-based education research, we take a human-computer partnership approach. Both humans and computers can provide unique skills and input into the qualitative coding and analysis process, as seen by others who have implemented machine learning in various qualitative coding processes [59], [60], [61]. When human coders interact with computers as coding partners rather than as tools designed to automate the process completely, both can work towards bettering a machine learning model that enriches the

analytical process by improving scalability and abstraction [61], [62]. To foster this in our study, we will discuss the results of qualitative coding and machine learning coding to understand how machine learning can be a part of the qualitative coding process to analyze student narratives of understanding.

Methods

Research Design

Concept questions were delivered through the Concept Warehouse [25], an online-based active learning tool. Instructors used two questions, shown in Figure 1, as a pair as they asked about two concepts (enthalpy and entropy) in the same context. Concept questions assess students' understanding of enthalpy and entropy, which interest chemical engineering education researchers because they are two widely misunderstood concepts in thermodynamics courses. The abstract nature of both concepts and the multitude of formulas that can describe these quantities in different situations provide students with a challenging experience of balancing conceptual and procedural knowledge [63], [64]. Thus, we chose these concept questions to understand narratives of understanding in short answer responses and provide a large set of concepts to train our machine learning algorithms.

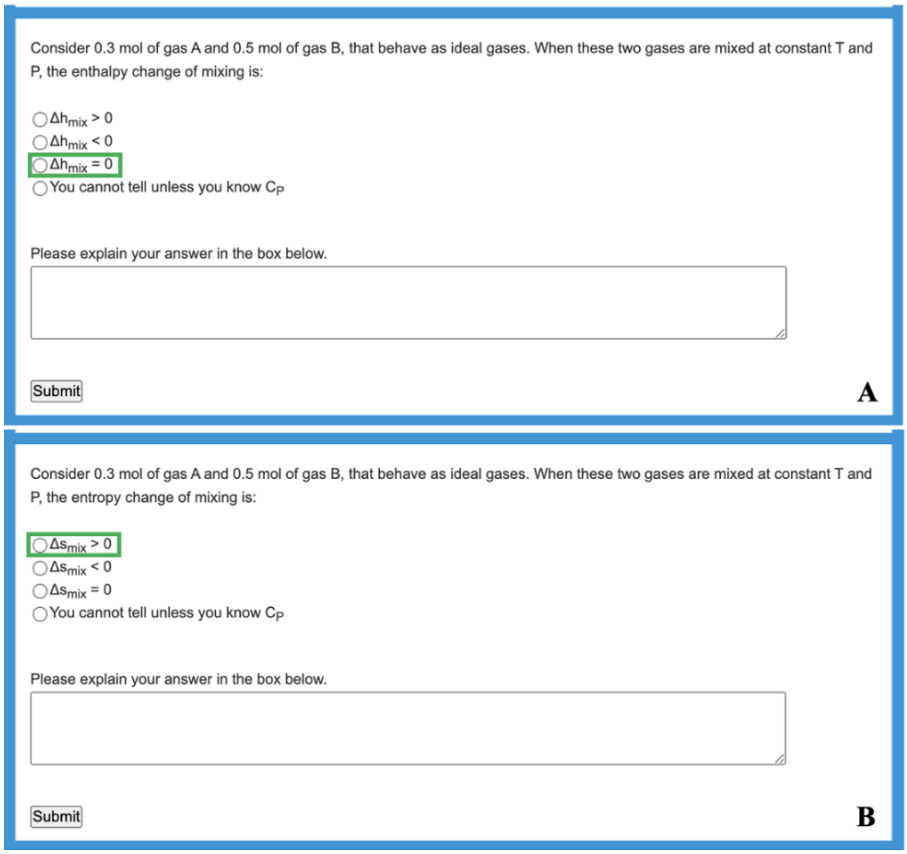

Figure 1. Student view of CT 1072 (A) and 1073 (B) on the Concept Warehouse. The image shows the multiple-choice question and the short-answer response field analyzed in this study. The correct answers are in the green boxes.

We can visualize the mixing process described in Figure 1 through the representation shown in Figure 2. The enthalpy change in concept question 1072 is zero because the gases are ideal, and the enthalpy of an ideal gas depends only on temperature; there are no intermolecular interactions except for elastic collisions. The change in entropy in concept question 1073 is greater than zero because even if the gases are ideal, mixing two gases increases the number of positions (configurations) available to the gas molecules of either gas, which increases entropy.

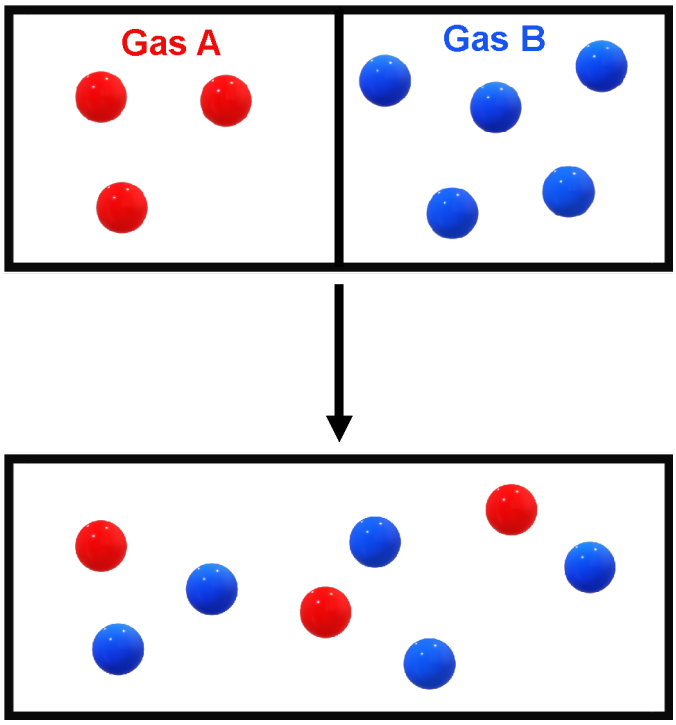

Figure 2. Visual representation of the process occurring in questions 1072 and 1073.

Setting and Participants

Responses were collected from consenting students at a large, public research university. Two instructors used these questions from 2012 - 2023. We will refer to them as Instructor A and Instructor B. Instructor A utilized the questions in eight third-year chemical engineering thermodynamics courses, and Instructor B utilized them in seven second-year energy balances courses. The average course size amongst both instructors was 85 students, with a standard deviation of 44.

Data Collection

Concept questions 1072 and 1073 were collected through the Concept Warehouse [25], a web-based active learning tool. All responses used in this study were from participants who consented to participate in data collection. A total of 1396 responses for 1072 and 1387 for 1073 were collected. Descriptive statistics are provided in Table 1.

Table 1. Percent correct for each question by instructor

	1072 % Correct	**1073 % Correct**
Instructor A	83.5%	57.4%
Instructor B	84.3%	66.9%
Total	84.0%	66.5%

Qualitative Coding

An emergent coding approach was used to investigate narratives of understanding for the short answer explanations to the two questions shown in Fig. 1. This approach was similar to how we've conducted coding previously and allowed us to fully account for all aspects of student thinking in the data [1], [2]. Coding was conducted on ATLAS.ti by Author 1 using methods described by Miles, Huberman, & Saldaña [26], [27]. The first coding phase consisted of Author 1 familiarizing themselves with the data by reading responses and noting preliminary thoughts regarding the cognitive resources used by students to convey their understanding. The second coding phase consisted of Author 1 utilizing emergent ideas and creating formal codes that described the cognitive resources students used, as shown in Appendix A.1 Tables A1 and A2. Authors 1 and 4 discussed the coding process while creating and categorizing codes to ensure their reliability and validity. Within Phase 2, codes were grouped according to different cognitive categories, including identification, comparison, and inference. These larger categories are further described in the Results section.

Machine Learning Analysis

We formulate the ML problem as a sequence-labeling task, where the model takes an input prompt consisting of an instruction, a question, a student-written explanation, and the prediction target is a manually-coded student-written explanation. For manual coding, we used INCEpTION [65] to convert manually-coded text spans into TSV format, which we further processed to create inputs and targets for model training.

Models

We conducted our experiments by finetuning a large decoder-only language model Mixtral of Experts (MoE) on the manually-coded dataset using Huggingface's transformer library [66] and GPT-4 with in-context learning via OpenAI GPT-4 API [67]. MoE was finetuned using a language modeling objective where it is trained to predict the next token. We used the prompt shown in Appendix B.2 to train MoE. For GPT-4, we used four in-context examples in the prompt, as shown in Appendix B.1, to prompt the model and extract the coded response for a new test student's explanation using OpenAI GPT-4 API.

Dataset Split

For both thermodynamics datasets, we used 85% of the manually coded samples as a training set to train the MoE models, i.e., 1186 samples for Enthalpy and 1175 samples for the Entropy dataset. We used 5%, i.e., 69 samples for each dataset, as a validation set and finally tested our finetuned models on 10% of the datasets, i.e., 139 and 138 samples for Enthalpy and Entropy datasets, respectively. For GPT-4, we used four samples from the training set as in-context examples to prompt the model. GPT-4 was evaluated on the same set as MoE models were evaluated.

Training Configurations

We trained MoE three times, once on each of the thermodynamics datasets and once on both datasets. We evaluate all models on both datasets. For GPT-4, we prompt the model with either four only-enthalpy dataset samples or four only-entropy dataset samples, drawn from their training sets, as in-context examples. For GPT-4, no finetuning is involved. We show the prompt for both models in Appendix B. We also conducted coding using ATLAS.ti's built-in coding tool called ATLAS.ti Interactive Coding and compared our results with this method.

Hyperparameters

To train MoE, we used 4-bit quantization [45] with LoRa [68]. In LoRa, the parameter update for a weight matrix is decomposed into a product of two low-rank matrices. Using LoRa, we train 11.6% of the MoE's total parameters with a learning rate of 0.00041 and a batch size of 1 for 2 epochs. We save the model checkpoint where we get the least validation loss, which is used for inference on the test set. When the model is trained on a combined dataset, we pick the checkpoint after 1 epoch, as this leads to better performance. For GPT-4 generations, we used the greedy decoding strategy with the temperature set to 1.

Evaluation Metric

Model responses were evaluated using Exact Match. In Exact Match, we count 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 on both thermodynamics datasets. Precision is the percentage of correct model-generated codes relative to the total number of model-generated codes. Recall is the percentage of human codes that the model was able to generate correctly. The F1 score is the harmonic mean of the precision and recall [1]. We also performed qualitative analysis for model-generated codes for ten test instances for both thermodynamics datasets. We report the number of codes that are semantically relevant to the student's narrative but not an exact match under "misses but makes sense"), semantically irrelevant codes with "does not make sense," and the number of codes missed by the model with "code missed."

Researcher Positionality

Our strength as researchers improves as we acknowledge and reflect upon the backgrounds and experiences of ourselves and others in our team [69]. As this project is a collaboration between engineering education researchers and machine learning researchers, we can work together at the intersection of machine learning and discipline-based education research. During the qualitative coding process, we shared multiple perspectives on how students could discuss different concepts so that we could work towards making a more diverse codebook. When evaluating the codes generated by machine learning analysis alongside the results from manual coding, we discussed how to best work towards a better coding process to help train algorithms.

Limitations

This study did not factor in the differences between instructors and their context or instructional moves. For example, some instructors may emphasize the importance of written responses differently, impacting how much effort students put into this. If participation points were assigned to completing the question and short answer response, students may put in more effort and thus provide a detailed response that is more representative of their understanding. Additionally, interrater reliability could be strengthened with additional coders.

Our current model evaluation relies on an exact match metric, which doesn't always capture cases where the model predicts semantically similar but not exactly matching codes. To address this, we need a better evaluation metric that considers semantic similarity to accurately assess the model's performance.

Results

Overview

We characterize three cognitive processes students may use to construct their narratives of understanding when writing short-answer responses to conceptually challenging chemical engineering thermodynamics problems. These cognitive processes include identification, comparison, and inference. Students may use a combination of these different cognitive processes in their responses and may also use them in any order. We define these cognitive processes as the following:

- **Identification:** procedural or conceptual knowledge identified in response
- **Comparison:** comparison of a variable or construct before and after mixing
- **Inference:** conclusion about the state of the system based on identification and comparisons made in the response

We then used GPT-4, Mixtral of Experts (MoE), and ATLAS.ti to automate the coding of these responses. We will first discuss our qualitative findings from manual coding for questions 1072 and 1073 and then detail the results from automated coding.

Qualitative Findings: 1072 – Enthalpy
Identification

Identification of concepts was primarily associated with ideal gases and their behavior. This included identifying important kinetic molecular theory (KMT) assumptions, such as ideal gases lack interaction, all collisions between gases are elastic, and temperature is directly proportional to the kinetic energy. The most widely utilized assumption was that ideal gases don't have interactions or intermolecular forces.

Additionally, students identified rules related to specific problems or situations in thermodynamics and applied them broadly without considering how those rules may differ depending on the problem. Some students identify that if the enthalpy of mixing was negative, zero, or positive in one case, it must also be in this case. For example, Students 1 and 2 used the rule that the enthalpy of mixing should always be positive.

> Student 1: Enthalpy of mixing is always positive, as far as I can tell. I can't remember any instances it's not.

> Student 2: Hmix is always positive, it can never be negative - and since the molar quantities are not even the change in enthalpy will be greater than 0.

They identified these rules from memory due to previous material they learned in class and applied it to this scenario.

Finally, to explain their reasoning, students typically identified the first law of thermodynamics ($\Delta U = q + w$) and the auxiliary function for the change in enthalpy ($\Delta H = \Delta U + P\Delta V$) within this problem. See Tables C1 and C2 in Appendix C.1 for information on variables used within these equations. For example, Student 3 stated the following to begin their explanation:

> Student 3: Enthalpy is internal energy plus work done (H=U+PV). Therefore, dH=dU+PdV+VdP.

Comparison

Students compared variables and properties of the system (i.e., volume, internal energy, etc.) and how the mixing process has impacted them. For example, Student 3, who identified the enthalpy equation, continues their explanation by saying:

> Student 3: … it is stated that there is no change in temperature or pressure, so dP=0. Also, the total volume will also not change, making dV=0, since it is dependent only on the intensive variables (P,T). U is dependent on T for an ideal gas, so dU=0. Using all these ideas, the equation simplifies to dH=0.

They utilized the equation to guide their response and explain their reasoning. They also compared the volume and internal energy at the process's beginning and end.

Inference

Utilizing different pieces of identified information and comparisons, students made conclusions about the change in enthalpy of the process. For example, Student 4 states the following:

> Student 4: Enthalpy is equal to the internal energy, U, with addition with the pressure multiplied by the volume. If we know that the change in volume of the mixture is 0 and there is no change in temperature or pressure, the extensive property H of mixing also has to be 0. This concludes that the intensive property Hmix is also 0.

Student 4 first identifies the definition of enthalpy and compares volume, temperature, and pressure. This leads them to make an inference about the enthalpy of mixing.

Qualitative Findings: 1073 – Entropy
Identification

Students utilized similar resources to 1072 regarding ideal gases; however, they also identified different ways to characterize the concept of entropy, including entropy as a measure of chaos, disorder, and microstates. Some students discussed definitions of entropy that are not typically seen as canonical definitions:

> Student 5: You can't unmix the two gasses and entropy increases in the universe. As entropy is the arrow of time it has to increase after the mixture.

Student 5 discusses entropy within the concept of irreversibility (i.e., all real processes within our universe occur irreversibly, as the entropy of a process must be greater than or equal to zero).

Similarly to 1072, students used rules like "entropy always increases" in their responses. Finally, students identified the first law of thermodynamics ($\Delta U = q + w$), entropy's relation to microstates ($S = k_b ln\, \Omega$), the entropy of mixing equation ($\Delta S = -R[y_a ln\, y_a + y_b ln\, y_b]$), and the auxiliary function for the change in Gibb's Free Energy ($\Delta G = \Delta H - T\Delta S$) within this problem. See Tables C1 and C2 in Appendix C.1 for information on variables used within these equations.

Comparison

Students compared variables and properties of the system similarly to 1072; however, in this question, written responses also emphasized changes in the system's disorder. We must note that this is different from identifying a *definition of entropy* and is instead *a comparison* of the disorder, number of microstates, or "chaos" from before and after mixing. For example:

> Student 6: Entropy of mixing represents the number of possible molecular configurations of a substance, so adding more substance will increase the possible molecular configurations.

In the first half of Student 6's response, they defined entropy, while in the second half, they compared the molecular configurations from the beginning to the end of the mixing process.

Inference

Utilizing different pieces of identified information and comparisons, students made conclusions about the change in entropy of the process. For example:

> Student 7: There are more random microconfigurations that the molecules of gas could be found in once they are mixed, and therefore, the entropy of the system will have increased.

Student 7 compared "microconfigurations" before and after mixing, which allowed them to infer about the system's entropy change.

Machine Learning: Automated Coding Using LLMs

Tables 2 and 3 show the evaluation results for finetuned MoE and GPT-4 (with different in-context examples in the prompt) on Enthalpy and Entropy test datasets, respectively. The ground truth includes 600 codes for Enthalpy and 644 codes for Entropy. MoE, when trained on the combined dataset, which includes Entropy and Enthalpy training samples, achieves the highest F1 score of 66% for Enthalpy (see Table 2, row "MoE trained on both datasets") and 59% for Entropy (see Table 3, row "MoE trained on both datasets").

Table 2. Comparison of ground truth and model-generated responses for Enthalpy dataset. The highest value is in bold.

Model	No. of Correct Codes	No. of Codes	Precision	Recall	F1
Ground truth	**600**				
MoE trained on both datasets	**458**	783	0.58	**0.76**	**0.66**
MoE trained on Enthalpy dataset	435	864	0.5	0.72	0.59
MoE trained on Entropy dataset	438	**1250**	0.35	0.73	0.47
GPT-4 (Enthalpy examples as in-context examples)	298	491	**0.61**	0.49	0.54
GPT-4 (Entropy examples as in-context examples)	293	570	0.51	0.48	0.49
ATLAS.ti AI Interactive Coding	37	1104	0.03	0.06	0.04

Table 3. Comparison of ground truth and model-generated responses for Entropy dataset. The highest value is in bold.

Model	No. of Correct Codes	No. of Codes	Precision	Recall	F1
Ground truth	**644**				
MoE trained on both datasets	**473**	963	**0.49**	**0.73**	**0.59**
MoE trained on Enthalpy dataset	347	806	0.43	0.54	0.48
MoE trained on Entropy dataset	464	1209	0.38	0.72	0.50
GPT-4 (Enthalpy examples as in-context examples)	224	490	0.45	0.35	0.45
GPT-4 (Entropy examples as in-context examples)	277	576	0.48	0.43	0.45
ATLAS.ti AI Interactive Coding	184	**1990**	0.09	0.29	0.14

Results on In-Distribution Test Sets: When MoE is trained specifically on the Enthalpy dataset and evaluated on its own held-out Enthalpy test set (see Table 2, row "MoE trained on Enthalpy"), it achieves an F1 score of 59%. On the other hand, when MoE is trained and evaluated on the Entropy data splits, the F1 score is 50%, a 9% drop, as compared to the Enthalpy test set (refer to Table 3, row "MoE trained on Entropy"). We observe a similar pattern with GPT-4 (refer to Table 2, row "GPT-4 with Enthalpy examples as in-context examples" and Table 2, row "GPT-4 with entropy examples as in-context example"). GPT-4 performs better on enthalpy with an F1 score of 54% than Entropy with an F1 score of 45% (a drop of 9%) when prompted with their respective in-context examples. This indicates that Entropy presents a more challenging dataset for the MoE and GPT-4 models than Enthalpy. The ATLAS.ti AI Interactive Coding shows the lowest performance on Enthalpy and Entropy test sets, yielding an F1 score of just 4% on Enthalpy and 10% on Entropy. In contrast to the MoE and GPT-4 data, Entropy had fewer matching codes with ATLAS.ti AI.

Results on Out-Of-Distribution Test Sets: When MoE is trained on the Enthalpy dataset but evaluated on the Entropy test set, we observe recall of 54%, a 19% drop from the best performance on the Entropy test set (refer to Table 3, row "MoE trained on Enthalpy"). On the other hand, when MoE is trained on the Entropy dataset and evaluated on the Enthalpy test set, the recall is 73%, only a 3% drop from the best performance on the Entropy test set. This suggests that training MoE on more challenging datasets (Entropy in this case) helps the model with better generalization capability. Our experiments with GPT-4 show a similar trend. The gap between recall, when GPT-4 is prompted with four in-distribution training examples versus out-of-distribution training examples on the entropy test set is 8%, whereas, for Enthalpy, it is just 1%, suggesting that GPT-4 requires in-context examples from challenging datasets for better generalization performance.

Table 4 summarizes the overall model performance on a combined test set when trained on a combined training set. MoE, when trained on the combined thermodynamics datasets, leads with

an F1 score of 62% on the combined test set. GPT-4 has the highest F1 score of 48% on the combined test set, with Entropy in-context examples. ATLAS.ti AI Interactive coding scores lowest at an F1 score of 10%.

Table 4. 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	**1244**				
MoE trained on both datasets	**931**	1746	**0.53**	**0.75**	**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
GPT-4 (Enthalpy examples as in-context examples)	522	981	**0.53**	0.42	0.47
GPT-4 (Entropy examples as in-context examples)	570	1146	0.50	0.46	0.48
ATLAS.ti AI Interactive Coding	221	**3094**	0.07	0.18	0.10

Table 5 shows the results of the qualitative analysis of the model prediction compared to manual codes. Overall, MoE models were better at generating codes close to manually written codes. We see the least percentage of codes missed or diverged by the MoE model, trained on both thermodynamic datasets, with “codes missed but makes sense” as 8% and “% codes do not make sense” as 14%. MoE models also show the least “% codes missed by the model,” ranging between 58%-64%, at least 20% below other models. ATLAS.ti AI Interactive Coding generated 49% of codes irrelevant to the responses, the highest “% codes missed by model” among all models. The tool also has the highest “% code the model missed,” which is 88% when compared to ground truth. In the case of GPT-4, when we prompt the model with Entropy in-context samples, we found that the GPT-4 had 42% codes missed but relevant to student responses, which is 4% higher than when GPT-4 is prompted with Enthalpy in-context examples. Interestingly, the Entropy prompt results in a lower "codes do not make sense" (-4%) and "codes missed by model" (-1%) compared to Enthalpy prompts. These findings support that challenging dataset samples enhance the model's generalization capabilities.

Table 5. Qualitative analysis of model-generated codes on a combined test set. The highest value is in bold.

	MoE trained on both datasets	**MoE trained on Enthalpy**	**MoE trained on Entropy**	**GPT-4 (Enthalpy examples as in-context examples)**	**GPT-4 (Entropy examples as in-context examples)**	**ATLAS.ti AI Interactive Coding**
% codes missed but makes sense	0.08	0.19	0.16	0.38	**0.42**	0.38
% codes do not make sense	0.14	0.19	0.29	0.21	0.17	**0.49**
% codes missed by model	0.58	0.64	0.58	0.85	0.84	**0.88**

In summary, we found that MoE trained on a combined dataset achieved the highest F1 score on both thermodynamics datasets. We show 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.

Discussion

Research Question 1

Students utilized different cognitive resources, which we categorized under the larger categories of identification, comparison, and inference. The different pieces of knowledge that students identify and compare to make inferences to formulate their response is a way to understand their emerging understanding of concepts as novices [35], [36], [56], [57], [58]. For example, in concept question 1073, we see difficulty with students and their understanding of entropy. Students who have not fully developed a set of resources that can be applied productively to thermodynamics problems might stumble into an issue when applying entropy because pressure and temperature are constant. In written responses, some students identify definitions of entropy that are more closely aligned with canonical language and relate it to "microstates" or "disorder," while some relate it to concepts closer to their everyday lives. Generally, these conceptions of entropy allow students to get to the right answer and explain how mixing increases entropy. However, we see some students state or describe an equation for entropy as a function of temperature and pressure, who then apply this resource and say that entropy will not increase in this system because both are held constant in the problem statement. Neither identification (conceptual nor formulaic) is wrong, but using the equation as a productive resource is not always fully present in responses.

Furthermore, in chemical engineering thermodynamics coursework, much of the content assesses students and their understanding of the system, surroundings, and universe before and after a process, and early content in thermodynamics is a stepping stone to later content in thermodynamics [63]. Thus, for instructors and engineering education researchers, analyzing

short-answer responses allows for a unique picture of students' understanding of thermodynamics concepts at different points of the course, which can help promote instructional change as they seek to tie these concepts together. As this is often paired with active learning instructional strategies, like Peer Instruction [3], [4], asking for short-answer responses after group discussion can also offer insight into how students have integrated new information into their mental models [6], [10], [11], [13], [14], [70]. Finally, when considering the formation of engineers, students need to build a repertoire of diverse cognitive resources and ways to productively use these resources to communicate with their peers. Professional engineers often explain their work in written formats (e.g., technical reports) and in presentations, so building writing skills that can effectively convey this knowledge to their colleagues is immensely helpful and a goal in science and engineering education work. This framework of understanding short-answer responses allows us to gain insight into student thinking, which can help instructors and researchers construct evidence-based instructional strategies to improve writing skills, conceptual understanding, and benefit the formation of engineers.

Research Question 2

State-of-the-art Large Language Models, such as Mixtral of Experts (MoE) and GPT-4, present a promising avenue for automating qualitative coding in the analysis of student narratives. Our study demonstrates that MoE, when trained on combined thermodynamics datasets, leads with the highest F1 score of 62% on the combined test set. GPT-4 shows its highest F1 score of 48% on the combined test set, specifically with entropy in-context examples. These results surpass the performance of ATLAS.ti AI Interactive coding, which achieves an F1 score of only 10% on the combined test set.

It's worth noting that GPT-4, being a larger model, exhibits higher sample efficiency compared to MoE. With just four training samples, GPT-4 achieves an F1 score of 48%, whereas MoE, a smaller model, requires a larger dataset with hundreds of training samples to achieve an overall F1 score of 62%. However, it's important to consider the resources required for both approaches as prompting GPT-4 is expensive, whereas manually coding hundreds of training samples for MoE is laborious and time-consuming. Our study shows the performance of both approaches, offering the research community insights on choosing different models for automating the assessment of student responses.

In the future, we hope to:

- Further integrate a resources-based conceptual framework into our overall methodology.
- Improve the model's ability to automate the coding of student narratives by input and target format while training models, experimenting with different decoding strategies, and additional hyperparameter tuning.
- Work towards developing a generative AI tool that can automate analysis of short-answer responses and help instructors and researchers understand patterns and trends in student thinking.

Conclusion

In this paper, we analyzed student explanations to conceptually challenging chemical engineering thermodynamics questions about enthalpy and entropy using manual coding and machine learning coding processes. Through emergent coding and a resources-based lens, we found that students use identification, comparison, and inference to explain their reasoning when writing short-answer explanations. GPT-4, Mixtral of Experts (MoE), and ATLAS.ti were used to automate coding, and it was found that GPT-4 generated codes with an F1 score of 54% for the enthalpy of mixing questions. For the entropy of the mixing question, the model has an F1 score of 45%. The MoE model, when trained on both thermodynamics datasets, outperforms GPT-4 performance on both enthalpy and entropy test datasets. MoE (when trained on both datasets) achieves an F1 score of 66% on enthalpy, outperforming GPT-4 by 12%. On the entropy dataset, it has an F1 score of 59%, which is 14% higher than GPT-4. We also show that the entropy dataset is more challenging for both MoE and GPT-4 models than the enthalpy dataset. Finally, our study shows that the model can tackle other tasks better when trained or prompted with examples from a more challenging dataset. This shows that machine learning models have tremendous potential to analyze short-answer explanations. With this knowledge, we hope to develop a generative AI tool for the CW to aid instructors and researchers in their pursuit to evaluate student understanding.

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] 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.

[5] 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.

[6] 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.

[7] N. Yannier *et al.*, "Active learning: 'Hands-on' meets 'minds-on,'" *Science*, vol. 374, no. 6563, pp. 26–30, Oct. 2021, doi: 10.1126/science.abj9957.
[8] 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.
[9] N. Joshi, S.-K. Lau, M. F. Pang, and S. S. Y. Lau, "Clickers in class: Fostering higher cognitive thinking using ConcepTests in a large undergraduate class," *Asia-Pac. Educ. Res.*, vol. 30, no. 5, pp. 375–394, Oct. 2021, doi: 10.1007/s40299-020-00525-x.
[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] N. Lasry, E. Mazur, and J. Watkins, "Peer Instruction: From Harvard to the two-year college," *Am. J. Phys.*, vol. 76, no. 11, pp. 1066–1069, Nov. 2008, doi: 10.1119/1.2978182.
[13] J. Schell and E. Mazur, "Flipping the chemistry classroom with Peer Instruction," in *Chemistry Education*, John Wiley & Sons, Ltd, 2015, pp. 319–344. doi: 10.1002/9783527679300.ch13.
[14] J. G. Tullis and R. L. Goldstone, "Why does peer instruction benefit student learning?," *Cogn. Res. Princ. Implic.*, vol. 5, no. 1, p. 15, Apr. 2020, doi: 10.1186/s41235-020-00218-5.
[15] 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.
[16] 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.
[17] 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
[18] 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.
[19] 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.
[20] 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.
[21] X. Zhai, K. C. Haudek, L. Shi, R. H. Nehm, and M. Urban-Lurain, "From substitution to redefinition: A framework of machine learning-based science assessment," *J. Res. Sci. Teach.*, vol. 57, no. 9, pp. 1430–1459, 2020, doi: 10.1002/tea.21658.
[22] 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

[23] 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

[24] 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.

[25] 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.

[26] J. Saldaña, *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2021.

[27] M. B. Miles, A. M. Huberman, and J. Saldana, *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, 2018.

[28] J. W. Creswell and C. N. Poth, *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, 4th ed. SAGE, 2018.

[29] OpenAI, "GPT-4 Technical Report." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.

[30] A. Q. Jiang *et al.*, "Mixtral of Experts." arXiv, Jan. 08, 2024. doi: 10.48550/arXiv.2401.04088.

[31] L. O. P. Rivard, "A review of writing to learn in science: Implications for practice and research," *J. Res. Sci. Teach.*, vol. 31, no. 9, pp. 969–983, 1994, doi: 10.1002/tea.3660310910.

[32] 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.

[33] J. A. Reynolds, C. Thaiss, W. Katkin, and R. J. Thompson, "Writing-to-Learn in Undergraduate Science Education: A Community-Based, Conceptually Driven Approach," *CBE Life Sci. Educ.*, vol. 11, no. 1, pp. 17–25, 2012, doi: 10.1187/cbe.11-08-0064.

[34] J. Nicholes, "How exposure to and evaluation of Writing-to-Learn activities impact STEM students' use of those activities," vol. 29, pp. 189–206, Dec. 2018, doi: 10.37514/WAC-J.2018.29.1.09.

[35] Y. Cao and M. D. Koretsky, "Shared resources: Engineering students' emerging group understanding of thermodynamic work," *J. Eng. Educ.*, vol. 107, no. 4, pp. 656–689, 2018, doi: 10.1002/jee.20237.

[36] 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.

[37] 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.

[38] L. Mao *et al.*, "Validation of automated scoring for a formative assessment that employs scientific argumentation," *Educ. Assess.*, vol. 23, no. 2, pp. 121–138, 2018.

[39] B. J. Yik, A. J. Dood, D. C. R. de Arellano, K. B. Fields, and J. R. Raker, "Development of a machine learning-based tool to evaluate correct Lewis acid–base model use in

written responses to open-ended formative assessment items," *Chem. Educ. Res. Pract.*, vol. 22, no. 4, pp. 866–885, 2021.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[47] 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.

[48] 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

[49] 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

[50] 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

[51] T. Alhindi, A. Alabdulkarim, A. Alshehri, M. Abdul-Mageed, and P. Nakov, "AraStance: A Multi-Country and Multi-Domain Dataset of Arabic Stance Detection for Fact Checking," in *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, A. Feldman, G. Da San Martino, C. Leberknight, and P. Nakov, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 57–65. doi: 10.18653/v1/2021.nlp4if-1.9.

[52] 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.

[53] 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.

[54] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach." arXiv, Jul. 26, 2019. doi: 10.48550/arXiv.1907.11692.

[55] 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.

[56] D. Hammer, "Misconceptions or P-Prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions," *J. Learn. Sci.*, vol. 5, no. 2, pp. 97–127, Apr. 1996, doi: 10.1207/s15327809jls0502_1.

[57] 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.

[58] 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

[59] M. S. González Canché, "Latent code identification (LACOID): A machine learning-based integrative framework [and open-source software] to classify big textual data, rebuild contextualized/unaltered meanings, and avoid aggregation bias," *Int. J. Qual. Methods*, vol. 22, p. 16094069221144940, Jan. 2023, doi: 10.1177/16094069221144940.

[60] S. Hilbert *et al.*, "Machine learning for the educational sciences," *Rev. Educ.*, vol. 9, no. 3, p. e3310, 2021, doi: 10.1002/rev3.3310.

[61] J. L. Feuston and J. R. Brubaker, "Putting tools in their place: The role of time and perspective in Human-AI collaboration for qualitative analysis," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, p. 469:1-469:25, Oct. 2021, doi: 10.1145/3479856.

[62] G. Tokadlı and M. C. Dorneich, "Interaction paradigms: From human-human teaming to human-autonomy teaming," in *2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC)*, Sep. 2019, pp. 1–8. doi: 10.1109/DASC43569.2019.9081665.

[63] 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.

[64] H. Saricayir, S. Ay, A. Comek, G. Cansiz, and M. Uce, "Determining students' conceptual understanding level of thermodynamics," *J. Educ. Train. Stud.*, vol. 4, no. 6, pp. 69–79, Mar. 2016, doi: 10.11114/jets.v4i6.1421.

[65] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. Gurevych, "The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation," Jul. 2018.

[66] 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

[67] "OpenAI Platform." [Online]. Available: https://platform.openai.com

[68] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ArXiv Prepr. ArXiv210609685*, 2021.
[69] S. Secules *et al.*, "Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community," *J. Eng. Educ.*, vol. 110, no. 1, pp. 19–43, 2021, doi: 10.1002/jee.20377.
[70] 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.

Appendix A.

1. Qualitative Coding

Table A1. Qualitative Coding for Conceptual Question 1072

Category	Code	Description
Identification	Chemical reaction	Gas A and Gas B undergo a chemical reaction in which intramolecular bonds are broken and formed, resulting in a product.
	Collisions	Gas A and Gas B will collide with the container's walls, between themselves, or with one another.
	Heat capacity	Mathematical or conceptual definition of C_p
	Ideal gas	Gas A and/or Gas B follow properties of an ideal gas
	Intermolecular forces	Gas A and Gas B have molecular interactions amongst themselves or between one another
	No interactions	There are no molecular interactions between Gas A and Gas B
	ΔH formula or dependencies	Enthalpy has a dependence or is independent of specific state variables. This can be written as a formula or in words.
Comparison	Chemical composition	Comparison of the chemical composition of the system before and after mixing.
	Gibbs free energy	Changes of Gibbs Free Energy before and after mixing.
	Phase	Comparison of the phase of the system before and after mixing.

	Pure and partial Species	Comparison between the behavior of Gas A or Gas B as a pure and partial species in a mixture.
	Gas A and gas B	Comparison of properties between gas A and gas B.
	Quantity of gas	Comparison of the amount of gas before and after mixing (can be in moles, mass, partial molar fractions, etc.)
	Pressure	Comparison of the pressure of the system before and after mixing.
	Temperature	Comparison of the temperature of the system before and after mixing.
	Volume	Comparison of the volume of the system before and after mixing.
	First law: q	Heat transfer processes
	First law: w	Whether work is done on the system and surroundings.
	First law: U	Changes in internal energy.
	Second law: entropy	Changes in entropy.
Inference	No change in enthalpy	There is no change in the enthalpy or the $\Delta H_{mix} = 0$.
	Nonzero change in enthalpy	There is a nonzero change in the enthalpy or the $\Delta H_{mix} > \text{or} < 0$.
Uncertainty	Uncertainty	The student is uncertain about the concept

Table A2. Qualitative Coding for Conceptual Question 1073

Category	Code	Description
Identification	Chemical reaction	Gas A and Gas B undergo a chemical reaction in which intramolecular bonds are broken and formed, resulting in a product.

	Collisions	Gas A and Gas B will collide with the container's walls, between themselves, or with one another.
	Heat capacity	Mathematical or conceptual definition of C_p
	Ideal gas	Gas A and/or Gas B follow properties of an ideal gas
	Intermolecular forces	Gas A and Gas B have molecular interactions amongst themselves or between one another
	No interactions	There are no molecular interactions between Gas A and Gas B
	ΔH formula or dependencies	Enthalpy has a dependence or is independent of specific state variables. This can be written as a formula or in words.
	No change in enthalpy	There is no change in the enthalpy or the $\Delta H_{mix} = 0$.
	Nonzero change in enthalpy	There is a nonzero change in the enthalpy, or the $\Delta H_{mix} > \text{or} < 0$.
	Reversibility	A process occurs without any net impacts on the system or surroundings.
Comparison	Chemical composition	Comparison of the chemical composition of the system before and after mixing.
	Gibbs free energy	Changes of Gibbs Free Energy before and after mixing.
	Phase	Comparison of the phase of the system before and after mixing.
	Pure and partial Species	Comparison between the behavior of Gas A or Gas B as a pure and partial species in a mixture.
	Gas A and gas B	Comparison of properties between gas A and gas B.

	Quantity of gas	Comparison of the amount of gas before and after mixing (can be in moles, mass, partial molar fractions, etc.)
	Pressure	Comparison of the pressure of the system before and after mixing.
	Temperature	Comparison of the temperature of the system before and after mixing.
	Volume	Comparison of the volume of the system before and after mixing.
	First law: q	Heat transfer processes
	First law: w	Whether work is done on the system and surroundings.
	First law: U	Changes in internal energy.
	Disorder	Changes in the disorder of the system.
	Microstates or configurations	Changes in the microstates or configurations of the system
Inference	No change in entropy	There is no change in the entropy or the $\Delta S_{mix} = 0$.
	Increase in entropy	There is an increase in entropy or the $\Delta S_{mix} > 0$.
	Decrease in entropy	There is a decrease in entropy or $\Delta S_{mix} < 0$.
Uncertainty	Uncertainty	The student is uncertain about the concept

Appendix B.
The prompt format consists of an instruction in *purple*, the question in orange, four training samples in red and a new test instance answer in blue for which we want the model to generate codes. Model output is indicated in green.

1. **GPT-4 Sample Prompt**

   We show the input we used to prompt GPT-4. This prompt is when we use Enthalpy in-context examples to generate responses on both thermodynamics datasets.

"*Answer annotation task: We have a question that was presented to the students during their test. We collected answers from multiple students and annotated the span of these answers. Each span is annotated with two-level annotations. The first level identifies the sequential cognitive processes and we annotate the spans by either identification, comparison, or inference. The same spans are annotated once again with more fine-grained insights about the answers. These second-level annotations are not pre-defined and you need to identify them for each of the answers. We provide a few samples of how annotations of the answers should be done. Note: The annotation should be included within <> brackets.*
Here is the question which was presented to the students.
Question: 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: A. delta_h_mix > 0, B.delta_h_mix < 0, C. delta_h_mix = 0, D. You cannot tell unless you know C_p.
###
First level annotations can be <Identification>, <Comparison>, or <Inference>. Second level annotations can be (but are not limited to) <Chemical Composition>, <No interactions>, <Volume>, <Pressure>, <Uncertainty>.
###
Answer: Ideal gases do not react and should not increase the enthalpy or decreases
Annotated Answer: Ideal gases <Identification> <Ideal gas> do not react and <Identification> <Chemical Reaction> should not increase the enthalpy or decreases <Inference> <no change heat or enthalpy>
###
Answer: IMF are minimized since they are ideal
Annotated Answer: IMF are minimized <Identification> <Intermolecular Forces> they are ideal <Identification> <Ideal gas>
###
Answer: The internal energy is greater when the gasses are together rather than when they are alone.
Annotated Answer: internal energy is greater <Comparison> <First law U> when the gasses are together rather than when they are alone <Comparison> <Pure Partial Species>
###
Answer: enthalpy of ideal gasses is 0.
Annotated Answer: enthalpy of ideal gasses is 0 <Inference> <no change heat or enthalpy>
###
Because they are ideal gases, they don't interact at all, so no heat change is associated with mixing them.”

Model output:
Because they are ideal gases, <Identification> <Ideal gas> they don't interact at all, <Identification> <No interactions> so no heat change is associated with mixing them. <Inference> <no change heat or enthalpy>

2. Mixtral of Experts Input and Target

The model was finetuned on Input + Target whereas models were evaluated with Inputs and the model generated are compared against the target.

Model Input

Instruction: Text annotation task : The task involves annotating student answers to conceptually challenging science questions with a focus on identifying aspects of student thinking. Two levels of annotations are required: the first level involves categorizing spans into identification, comparison, or inference, while the second level involves providing more fine-grained insights into the answers. Each span of text must have two annotations, one of each level. The output should be the span followed by the annotation enclosed in <> brackets. The order of the annotation is important. First annotation should be level1 e.g. <Identification> or <Comparison> or <Inference>, followed by level2 annotation e.g. <Gravity> or <Newtons law>. e.g. span of text<annotation level1><annotataion level2> another span of text<annotation evel1><annotataion level2> Given a question and student's answer to the question, generate the annotated answer.

Question: Question: 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: A. delta_h_mix > 0, B.delta_h_mix < 0, C. delta_h_mix = 0, D. You cannot tell unless you know C_p.

Answer: Ideal gases do not react and should not increase the enthalpy or decreases

Annotated Answer:

Target

Ideal gases <Identification> <Ideal gas> do not react and <Identification> <Chemical Reaction> should not increase the enthalpy or decreases <Inference> <no change heat or enthalpy>

3. ATLAS.ti

Below is the intention to prompt ATLAS.ti to generate codes for questions 1072 and 1073.

Research Question: What aspects of student thinking are seen within short answer responses to conceptual challenging chemical engineering thermodynamics problems? Context: We utilize three categories to classify sub-codes, which include: identification - student identifies concept to use in short answer response; comparison: student compares the state of the system before and after some process; inference: student makes a conclusion

ATLAS.ti then created a set of questions and associated code groups to prompt the model to generate codes. For question 1072, 6 questions (in red) and 6 associated code groups (in green) were generated. All were selected to generate codes.

What aspects of student thinking are observed in the identification of concepts used in short answer responses?
What aspects of student thinking are observed in the comparison of the state of the system before and after a process in short answer responses?
What aspects of student thinking are observed in the inference and conclusions made in short answer responses?
How do students determine the correct concept to use in their short answer responses?
What strategies do students employ to compare the state of the system before and after a process in their short answer responses?
How do students draw conclusions and make inferences based on the information provided in the short answer problems?

Concept Identification
Comparison of System States
Inference and Conclusions
Concept Selection
Comparison Strategies
Inference Strategies

For question 1073, 6 questions (in red) and 6 associated code groups (in green) were generated. All were selected to generate codes.

What concepts do students identify in their short answer responses to challenging statics and mechanics problems?
How do students compare the state of the system before and after a process in their short answer responses to challenging statics and mechanics problems?
What conclusions do students draw in their short answer responses to challenging statics and mechanics problems?
Can the identified concepts in student responses be categorized into sub-categories based on their level of understanding? (e.g., basic, intermediate, advanced)
What specific processes or events do students compare when making state comparisons in their short answer responses?
Are there any patterns or trends in the types of conclusions that students draw in their short answer responses?

Concept Identification
State Comparison
Conclusions Inference
Concept Categorization
Process Comparison
Conclusion Patterns

Appendix C

1. Thermodynamic Parameters

Below is the key for variables, variable names, and constants for equations mentioned in the Results.

Table C1. Variable and variable names for equations mentioned in the Results.

Variable	Variable Name
P	Pressure
T	Temperature
V	Volume
ΔU	Change in Internal Energy
ΔH	Change in Enthalpy
ΔG	Change in Gibb's Free Energy
ΔS	Change in Entropy
q	Heat transfer
w	Mechanical work
Ω	Number of microstates
y_i	Mole fraction of gas

Table C2. Constants and associated values for equations mentioned in the Results.

Constant	Constant Name and Value
k_b	Boltzmann's constant (1.380649×10^{-23} J K^{-1})
R	Ideal gas constant (8.314 J K^{-1} mol^{-1})