BOARD # 440: RFE: Machine Learning for Student Reasoning during Challenging Concept Questions - Year 2

Harpreet Auby, Tufts University

Harpreet (he/him) is pursuing a Ph.D. in Chemical Engineering at Tufts University under the guidance of Dr. Milo Koretsky. He earned a B.S. in Chemical Engineering from the University of Illinois at Urbana-Champaign in 2021, followed by an M.S. in STEM Education from Tufts University in 2023. Previously, he worked on studying shifts in learning assistant beliefs and the uptake of the Concept Warehouse. His current focus is analyzing short-answer explanations to statics, dynamics, and thermodynamics concept questions to understand student thinking and applications of LLMs to engineering education research. Harpreet's research interests encompass 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 received his B.S. and M.S. degrees from UC San Diego and his Ph.D. from UC Berkeley,

RFE: Machine Learning for Student Reasoning during Challenging Concept Questions - Year 2

Introduction

In this NSF Grantee Poster Session Paper, we outline the progress of a collaboration funded by NSF Research in the Formation of Engineers (RFC) 2226553 between engineering education researchers at Tufts University and machine learning researchers at University of Massachusetts Lowell to use Generative AI (GenAI) to automate qualitative coding and analysis of short-answer justifications to concept questions. Concept questions, sometimes called ConcepTests [1], [2], are single-right-answer multiple-choice questions that assess student understanding of recently learned challenging concepts. Instructors sometimes ask students to supply short-answer justifications to explain their answer choice reasoning. These instructional practices have been shown to improve student outcomes and conceptual understanding [2]-[4]. Analysis of these justifications provides insight into student thinking but can be laborious and time-consuming for instructors and researchers. Machine learning (ML) has been used for adaptive learning experiences, lesson planning, real-time tutoring, grading, and analysis of short- and long-answer student text [5]-[7]. However, too often, ML approaches in education research are focused on the *products* of learning rather than the *processes* of learning. Here, we explore the use of state-of-the-art (SOTA) large language models (LLMs) to automate the coding and analysis of student thinking within short-answer justifications to concept questions collected through an educational technology tool.

Background

Concept Questions and Short-Answer Justifications

Concept questions [1], [2] are single-right-answer multiple-choice questions that assess students' understanding of recently learned challenging concepts. Questions are designed to help instructors enact social, cognitive, and epistemological goals around teaching and learning [8]. Researchers have observed that using concept questions within active learning pedagogies has improved student outcomes, promoted conceptual understanding, and encouraged engagement in the classroom [2]. Instructors sometimes pair concept questions with a short-answer justification, a low-stakes writing task that asks students to explain their answer choice reasoning. Work has shown that writing justifications promotes conceptual understanding and prepares students for in-class discussions [3], [4]. Thus, analysis of justifications can give insight into student thinking, but it can require a lot of time and resources, prompting our motivation to use GenAI to supplement analysis.

GenAI in Education Research

ML in education has been implemented to provide adaptive learning experiences, lesson planning, real-time tutoring, grading, and analysis of short- and long-answer student text [9]. The emergence of transformer-based generative LLMs [10], [11] have emerged as state-of-the-art in understanding and generating natural language text. The use of LLMs to analyze student text is emergent, but work that has utilized GPT-3 [10], GPT-4 [11], and Llama-2 [12] show promise in their ability in grading and rubric-based analysis tasks [13], [14]. We also aim to take a *human-*

centered AI [15] approach, as these tools can provide assistance with time-consuming tasks and provide another perspective on qualitative coding and analysis.

Methods

Data Collection

Short-answer explanations were collected through the Concept Warehouse (CW) [16], a free, web-based active learning tool and content repository, between 2012 and 2024. Students are from a diverse array of two- and four-year institutions. Instructors delivered concept questions in the way they deemed fit for their classes. Active data collection occurred for statics and dynamics from 2021 to the present, while historical data from the CW was used for thermodynamics questions. Questions ranged from 49-80% correctness; further details are provided in Table I and Figures 1 and 2.

TABLE I
CONCEPT QUESTIONS ANALYZED IN PROJECT

Domain	Question ID	Topic	No. Responses
Statics	4975	3-D Moments	54
	4976	3-D Moments	53
Dynamics	5703	Friction	240
	6141	Moment of Inertia	106
Thermodynamics	1072	Enthalpy of mixing ideal gases	1396
	1073	Entropy of mixing ideal gases	1387
	1287	Enthalpy of mixing two-non-ideal liquids	904

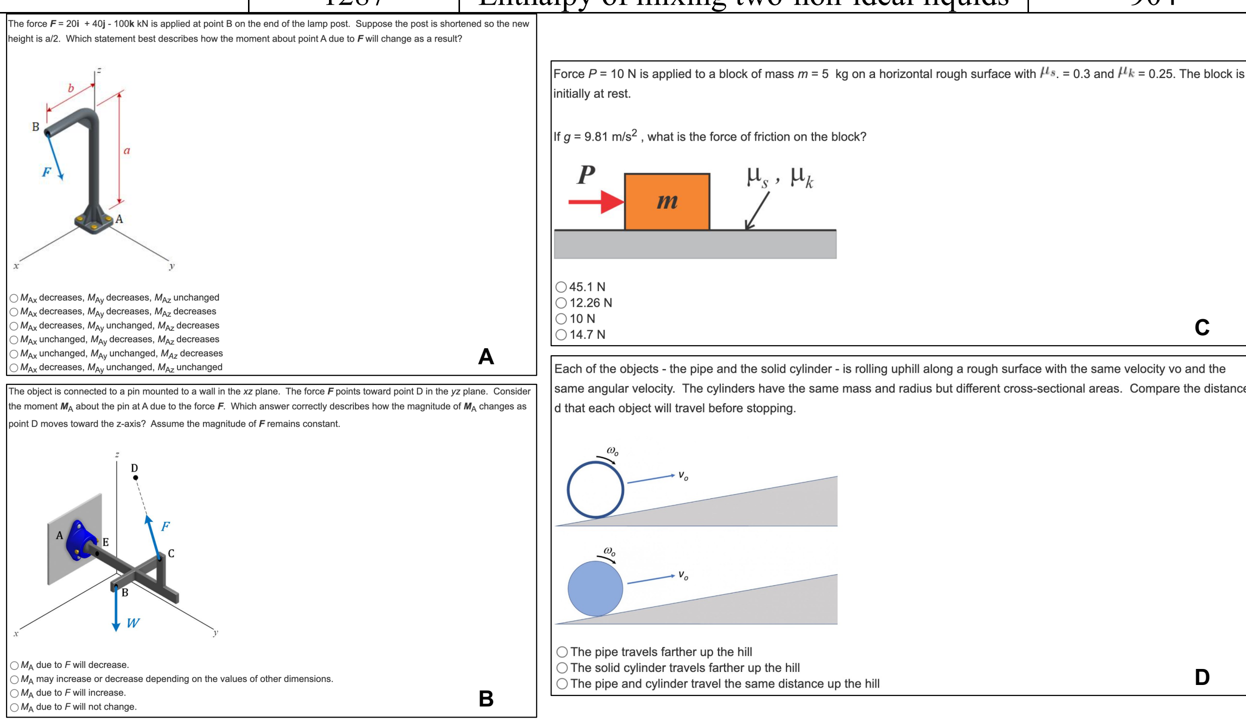

Fig. 1. Mechanics Concept Questions (A) 4975, (B) 4976, (C) 5703, and (D) 6141

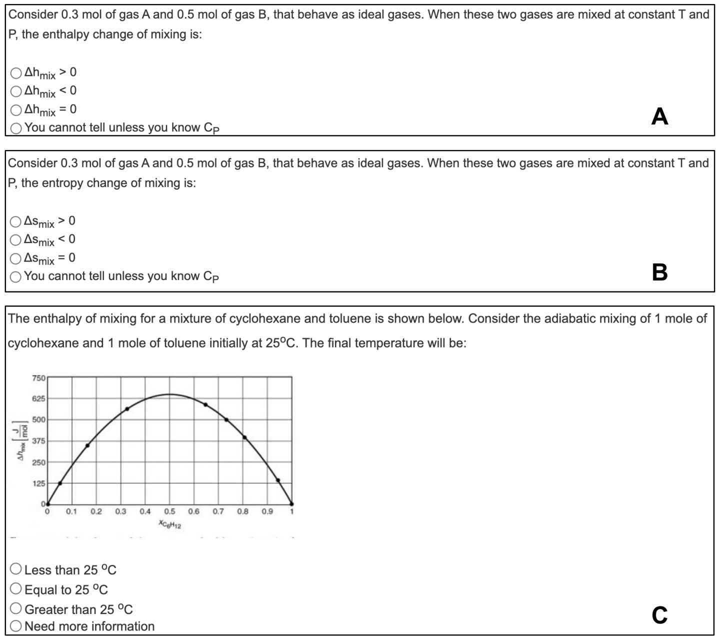

Fig. 2. Thermodynamics Concept Questions (A) 1072, (B) 1073, and (C) 1287.

Data Analysis

Qualitative

Only students who consented had their responses analyzed. Qualitative coding of responses was done in a two-stage coding cycle [17]. The first cycle consisted of emergent coding that looked for cognitive resources in their responses. The second cycle involved iterating and refining these emergent codes and then generating salient themes. Coding practices were discussed amongst the team to promote reliability.

Machine Learning

This task was treated as a sequence labeling problem where the machine attaches a label to spans of student text. We've utilized various models in this project, including Text-to-Text Transformer (T5)-base, T5-large [18], Mixtral of Experts (MoE) [19], GPT-3 [10], GPT-4 [11], GPT-4o-mini [20], Llama-3-8B [21], and Phi-3.5-mini [22]. Transfer learning via fine-tuning and in-context learning were used to simplify the training process. T5, MoE, Llama-3-8B, and Phi-3.5-mini utilized transfer learning via fine-tuning, where the models are pre-trained on large amounts of text and then further fine-tuned using our datasets. GPT-3, GPT-4, and GPT-4o-mini used in-context learning where the model is prompted using a few samples and asked to code responses. There was no training done for models using in-context learning. An Exact Match metric was used to compare the machine-coded responses to human-coded responses, evaluating how many codes were semantically identical. Precision, recall, and F1 scores were also calculated.

Findings

Year 1

Building on our previous work [23], we automated coding using GPT-4 [11], MoE [18], and ATLAS.ti's Interactive Coding tool powered by OpenAI [24] on thermodynamics questions about the entropy and enthalpy of mixing ideal gases (QIDs 1072 and 1073) [25], [26]. The manual analysis found that students use three main cognitive processes to formulate their responses: identification, comparison, and inference. Within these main cognitive processes, we group smaller cognitive resources, or ideas, that further describe the qualities of these processes. For MoE, the highest F1 score of 62% was achieved using a combined training set (enthalpy and entropy-coded responses). For GPT-4, the highest F1 score of 48% was achieved with enthalpy in-context examples. Finally, ATLAS.ti achieved an F1 score of 10%.

Year 2

To further investigate the ability of SOTA LLMs to automate the coding of short-answer justifications, we analyzed student thinking in all concept questions mentioned above. We then compared the ability of dense (GPT-4, GPT-4o-mini, Llama-3-8B, Phi-3.5-mini) and sparse (MoE) LLMs to automate the coding of cognitive resources within the same question, within the domain (e.g., train on 1287 in thermodynamics and test on 1073 in the thermodynamics test set), and across domains (e.g., train on thermodynamics question and test on a statics or dynamics question). This study revealed that MoE and Llama-3 performed the best with in-domain coding tasks, while GPT-4 and GPT-4o-mini generally performed better for cross-domain tasks.

Implications and Future Directions

This work contributes to the body of work implementing GenAI in education research. We aim to develop an AI assistant for the CW, which automates coding and reports on patterns and trends within justifications. This tool could supplement analysis to allow instructors to gain insight from responses through patterns and trends, and give researchers access to coded responses on a scale not feasible with manual coding. This will require the design of a user interface, setting up dedicated hardware for high-performance computing, and user experience research for a beta version of the tool.

Acknowledgments

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

References

[1] E. Mazur, *Peer Instruction: A user's manual*. in Series in Educational Innovation. Prentice Hall, 1997.

[2] 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.

[3] 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.

[4] 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.

[5] E. A. Alasadi and C. R. Baiz, “Generative AI in education and research: Opportunities, concerns, and solutions,” *J. Chem. Educ.*, vol. 100, no. 8, pp. 2965–2971, Aug. 2023, doi: 10.1021/acs.jchemed.3c00323.

[6] D. Baidoo-Anu and L. O. Ansah, “Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning,” *J. AI*, vol. 100, no. 8, pp. 2965–2971, 2023.

[7] A. Johri, A. S. Katz, J. Qadir, and A. Hingle, “Generative artificial intelligence and engineering education,” *J. Eng. Educ.*, vol. 112, no. 3, pp. 572–577, 2023, doi: 10.1002/jee.20537.

[8] M. D. Koretsky *et al.*, “For systematic development of ConcepTests for active learning,” in *EDULEARN19 Proceedings*, Palma, Spain: IATED, Jul. 2019, pp. 8882–8892. doi: 10.21125/edulearn.2019.2205.

[9] 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.

[10] T. B. Brown *et al.*, “Language models are few-shot learners,” Jul. 22, 2020, *arXiv*: arXiv:2005.14165. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2005.14165

[11] OpenAI *et al.*, “GPT-4 technical report,” Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[12] H. Touvron *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” Jul. 19, 2023, *arXiv*: arXiv:2307.09288. doi: 10.48550/arXiv.2307.09288.

[13] D. Carpenter, W. Min, S. Lee, G. Ozogul, X. Zheng, and J. Lester, “Assessing student explanations with large language models using fine-tuning and few-shot learning,” in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 403–413. Accessed: Sep. 09, 2024. [Online]. Available: https://aclanthology.org/2024.bea-1.33

[14] C. Cohn, N. Hutchins, T. Le, and G. Biswas, “A chain-of-thought prompting approach with LLMs for evaluating students’ formative assessment responses in science,” *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 21, Art. no. 21, Mar. 2024, doi: 10.1609/aaai.v38i21.30364.

[15] B. Shneiderman, *Human-centered AI*. Oxford, New York: Oxford University Press, 2022.

[16] M. D. Koretsky *et al.*, “The AIChE Concept Warehouse: A web-based tool to promote concept-based instruction,” Advances in Engineering Education, vol. 4, no. 1, p. 27, 2014.

[17] J. Saldaña, *The coding manual for qualitative researchers*. SAGE Publications, 2021.

[18] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified Text-to-Text Transformer,” Jul. 28, 2020, *arXiv*: arXiv:1910.10683. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1910.10683

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

[20] OpenAI *et al.*, “GPT-4o System Card,” Oct. 25, 2024, *arXiv*: arXiv:2410.21276. doi: 10.48550/arXiv.2410.21276.
[21] Llama Team and AI @ Meta, “The Llama-3 herd of models.”
[22] M. Abdin *et al.*, “Phi-3 technical report: A highly capable language model locally on your phone,” May 23, 2024, *arXiv*: arXiv:2404.14219. doi: 10.48550/arXiv.2404.14219.
[23] H. Auby, N. Shivagunde, A. Rumshisky, and M. D. Koretsky, “WIP: Using machine learning to automate coding of student explanations to challenging mechanics concept questions,” in *Proceedings of the 2022 American Society for Engineering Education Annual Conference & Exposition*, Jun. 2022. [Online]. Available: https://peer.asee.org/40507
[24] “AI Coding powered by OpenAI,” ATLAS.ti. [Online]. Available: https://atlasti.com/ai-coding-powered-by-openai
[25] H. Auby, N. Shivagunde, A. Rumshisky, and M. D. Koretsky, “Board 408: Toward building a human-computer coding partnership: Using machine learning to analyze short-answer explanations to conceptually challenging questions,” in *Proceedings of the 2024 American Society for Engineering Education Annual Conference & Exposition*, Portland, Oregon, Jun. 2024. [Online]. Available: https://peer.asee.org/46996
[26] H. Auby, N. Shivagunde, A. Rumshisky, and M. D. Koretsky, “Using machine learning to analyze short-answer responses to conceptually challenging chemical engineering thermodynamics questions,” in *Proceedings of the 2024 American Society for Engineering Education Annual Conference & Exposition*, Portland, Oregon, Jun. 2024. [Online]. Available: https://peer.asee.org/48236