2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #53249

RFE: Design of a Generative AI Assistant for an Educational Technology Tool - Year 3

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: Design of a Generative AI Assistant for an Educational Technology Tool - Year 3

Introduction and Background

In this NSF Grantees Poster Session paper, we report on Year 3 of a collaboration between engineering education researchers at Tufts University and machine learning researchers at the University of Massachusetts, Lowell. This project is funded by National Science Foundation (NSF) Research in the Formation of Engineers Grant No. 2226553. The goal of this project is to develop an AI Assistant Tool for instructors using the Concept Warehouse to analyze short-answer responses to concept questions in thermodynamics and mechanics.

Concept questions, sometimes referred to as ConcepTests [1], are challenging single-right-answer multiple choice questions that ask students to apply recently learned concepts to novel scenarios. When used as part of active learning pedagogies, like Peer Instruction [1], concept questions have been shown to improve learning outcomes [2] - [4]. In this project, we are interested in pairing concept questions with a short-answer justification task. Explicit prompting of students for self-explanations has been shown to promote the development of higher-order knowledge construction and transfer [5], [6]. Previous work done by the Authors has shown that concept questions coupled with a short-answer justification task has been shown to promote deeper understanding, better prepare students for in-class discussion, and improve learning outcomes [7], [8]. Analyzing short-answer responses to understand student reasoning provides valuable information about student understanding but is resource intensive leading us to investigate machine learning as a way to automate the analyses.

Machine learning (ML) approaches have been utilized in education spaces for automated grading, personalized tutoring [9], [10], curriculum development [11], understanding student discourse in small-group work [12] - [14], and analysis of short- and long-answer texts [15] - [22]. Compared to conventional ML techniques, large language models (LLMs) can analyze complex, unstructured data, motivating us to apply them to analysis of short-answer justifications. As application of state-of-the-art LLMs is still emergent in STEM education research, this project applies transformer-based decoder-only models to show potential for more nuanced and scalable analyses of student reasoning. Through this project we have compared several models, including: Mixtral [23], Llama-3 [24], GPT-4 [25] GPT-4o-mini [26], and Phi-3.5-mini [27] to automate coding.

Our previous work has presented findings from qualitative coding on student responses and model development [28] - [32]. Here, we briefly foreground the studies done in Years 1 and 2 and then describe the current progress toward the development of a Generative Artificial Intelligence (GenAI) Assistant tool to supplement the Concept Warehouse, a free, web-based active learning tool and content repository [33].

Methods

Below we detail the data collection, qualitative coding, and model development done to position us to develop and implement the GenAI Assistant in the Concept Warehouse.

Data Collection

Short-answer justifications were collected on the Concept Warehouse from consenting students between 2012 and 2024 from a diverse array of 2- and 4-year institutions. Information about the questions studied in this project are shown in Table I. Active data collection occurred for mechanics questions (M1 – M4) where all instructors were asked to give students a set of common questions, while historical data was used for thermodynamics questions (T1 – T3). An example of a question, T3, is shown in Fig. 1.

TABLE I

QUESTIONS ANALYZED IN THIS PROJECT. TABLE REPRODUCED FROM [28].

Domain	No.	Topic	Responses	% Correct
Mechanics	M1	3-D Moments	54	59
	M2	3-D Moments	53	68
	M3	Friction	240	46
	M4	Moment of Inertia	106	50
Thermodynamics	T1	Enthalpy of mixing ideal gases	1396	84
	T2	Entropy of mixing ideal gases	1387	66
	T3	Enthalpy of mixing two-non-ideal liquids	904	49

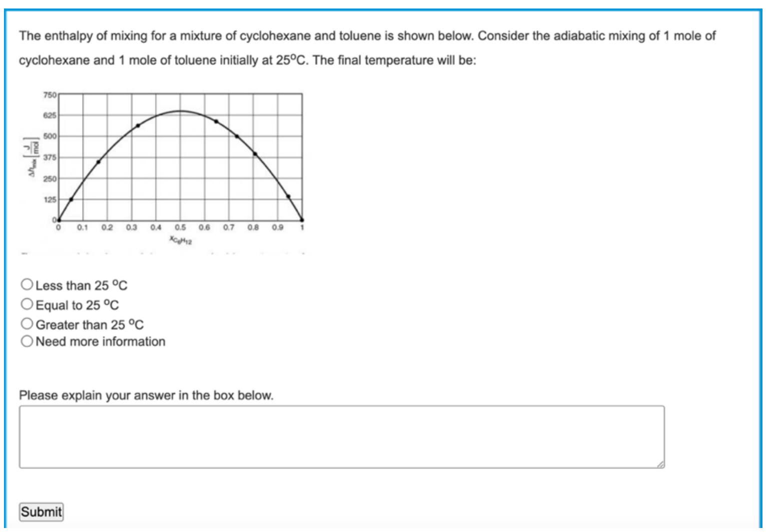

Fig. 1. Example of a concept question analyzed in this study.

Data Analysis

Fig. 2 shows the entire process for qualitative coding and training/testing of LLMs.

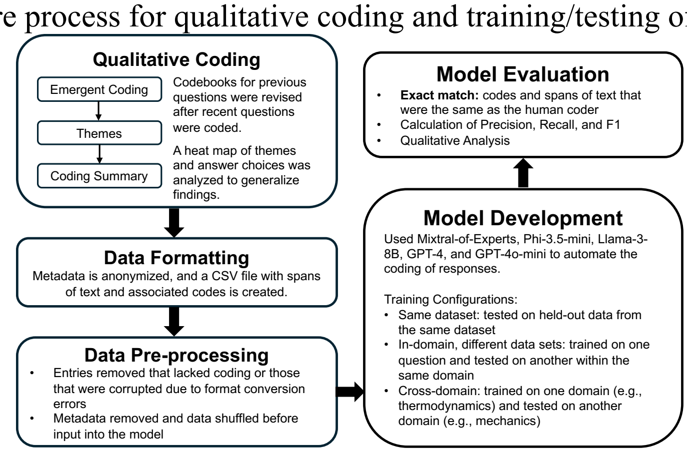

Fig. 2. Overall process describing the qualitative coding and model. Reproduced from [28].

Qualitative Coding
A two-stage constant comparative coding approach [34] was completed from 2021-2024 for 3000+ short-answer justifications across seven questions, shown in Table I. We utilized a resources framing [35], [36] to guide our coding process. Stage 1 coding involved a mixture of *a priori* and emergent coding. Stage 2 involved condensing codes into cognitive resources and associated themes.

Machine Learning
The ML problem was formulated as a sequence labeling task where spans of student justifications were labeled with manual codes. We then trained Mixtral [23], Llama-3 [24], GPT-4 [25], GPT-4o-mini [26], and Phi-3.5-mini [27] to automate coding. Mixtral, Llama-3, and Phi-3.5-mini were all trained using fine-tuning [37] and GPT-4, GPT-4o-mini were trained using in-context learning [38]. Evaluation was done an exact match approach using precision, recall, and F1 scores as metrics of model performance. Precision is the ratio of the correct codes generated by the model relative to the total number of model-generated codes. Recall is the percentage of human codes that the model was able to generate correctly. Finally, F1 is the harmonic mean of precision and recall.

Findings

In this section we detail the results from qualitative coding and training of large language models as a foundation for the GenAI Assistant.
Research Question 1: What ideas do students use to explain their reasoning when writing short-answer responses to conceptually challenging questions?
Students activate a variety of cognitive resources to write short-answer justifications. For example, in question T3 (Fig. 1), we observed that students who used resources related to molecular-scale reasoning were less likely to get the correct answer choice, suggesting that the connective tissue often provided when thinking about molecular reasoning may not be fully developed in this context [28].
Research Question 2: How well do transformer-based machine learning models replicate the human-coded data?
Mixtral and Llama-3 performed best at within the same-dataset, in-domain coding tasks, especially as the training set size increased. Phi-3.5-mini, while effective for mechanics questions, struggled in thermodynamics. In contrast, GPT-4 and GPT-4o-mini stood out for their robust generalization across in- and cross-domain tasks [28]. Selected results are shown in Table II.

TABLE II
SELECTED RESULTS FROM MOST RECENT EXPERIMENTS

Configuration	Metric	Mixtral*	Llama-3*	Phi-3.5-mini*	GPT-4^	GPT-4o-mini^
Same Dataset	**Precision**	0.51±0.02	0.50±0.02	0.37±0.03	0.50±0.02	0.43±0.02
	Recall	**0.66±0.03**	**0.66±0.03**	0.16±0.02	0.51±0.02	0.42±0.02
	F1	**0.58±0.01**	0.57±0.02	0.23±0.02	0.50±0.02	0.42±0.02
In Domain, Different Dataset	**Precision**	0.26±0.01	0.20±0.01	0.18±0.01	**0.39±0.01**	0.38±0.01
	Recall	0.37±0.01	0.38±0.01	0.20±0.01	0.36±0.01	**0.39±0.01**
	F1	0.31±0.01	0.26±0.01	0.19±0.01	0.37±0.00	**0.38±0.00**
Out of Domain, Different Dataset	**Precision**	0.25±0.01	0.22±0.01	0.17±0.01	**0.29±0.01**	0.27±0.01
	Recall	0.38±0.01	**0.40±0.01**	0.27±0.01	0.38±0.01	0.28±0.01
	F1	0.30±0.01	0.28±0.01	0.21±0.01	**0.33±0.01**	0.32±0.01

*Trained on 85% with 10% held out, ^4-shot. **Bolded** items note the highest values.

Design of a GenAI Assistant Tool

As we identified the possibility of open-source models like Mixtral and Llama-3 to analyze short-answer justifications, Year 3 has focused on the development GenAI assistant tool to supplement the Concept Warehouse. The use and propagation of educational technology tools is deeply embedded in the context of the ecosystems in which they exist [39] - [43]. To design with end users in mind, we met with instructors involved in data collection to brainstorm a potential user interface (UI) which would promote their interest in the use of the tool. During our session, we co-created a storyboard, shown in Fig. 4, to scaffold a shared experience where an instructor could use the tool.

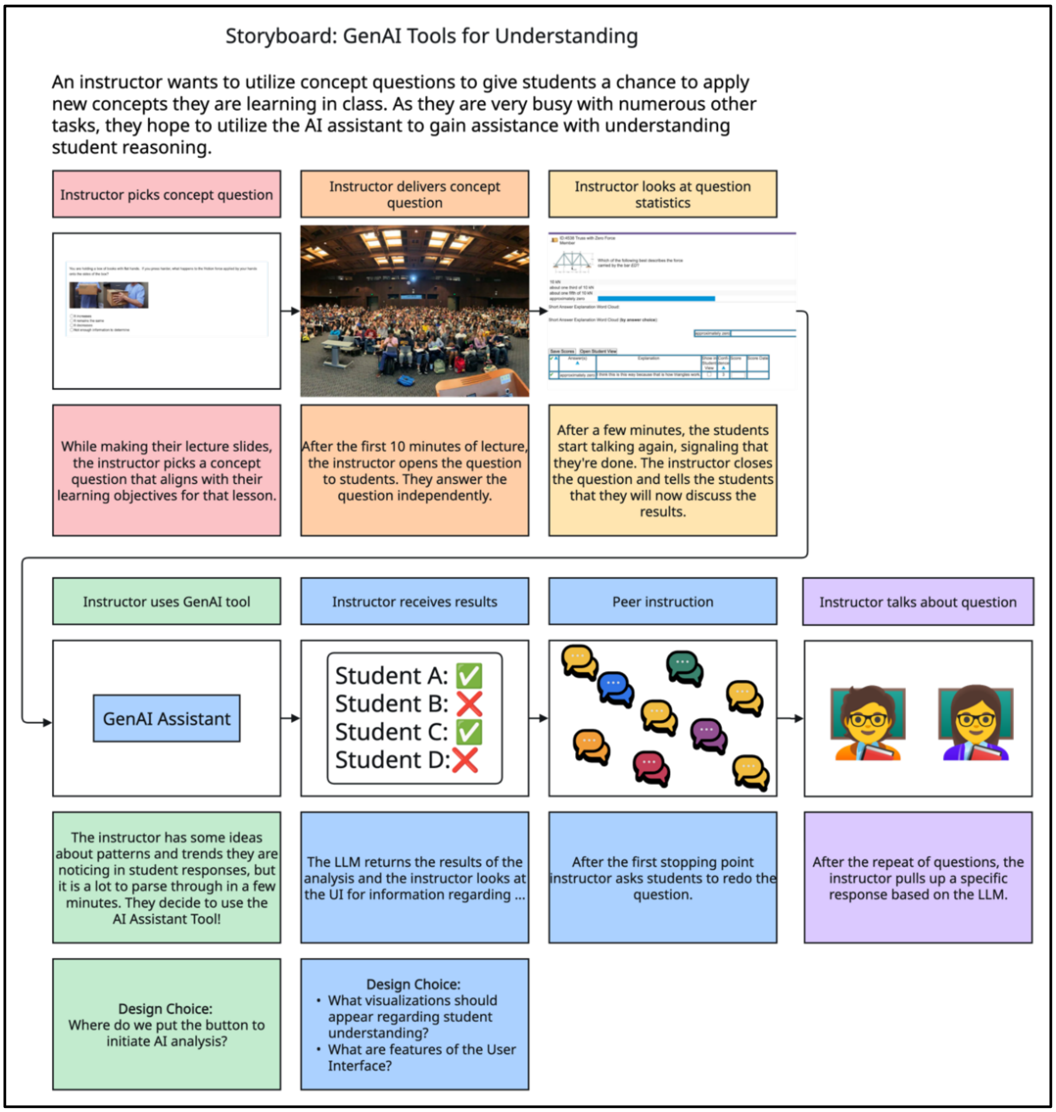

Fig. 4. Storyboard developed with instructors to center our co-design session around a shared experience.

We then asked instructors to brainstorm and draw their ideal UI for the tool. Each instructor was given their own slide on a Zoom whiteboard. Examples of their brainstorming is in Fig. 5, where they discuss key aspects of their ideal UI.

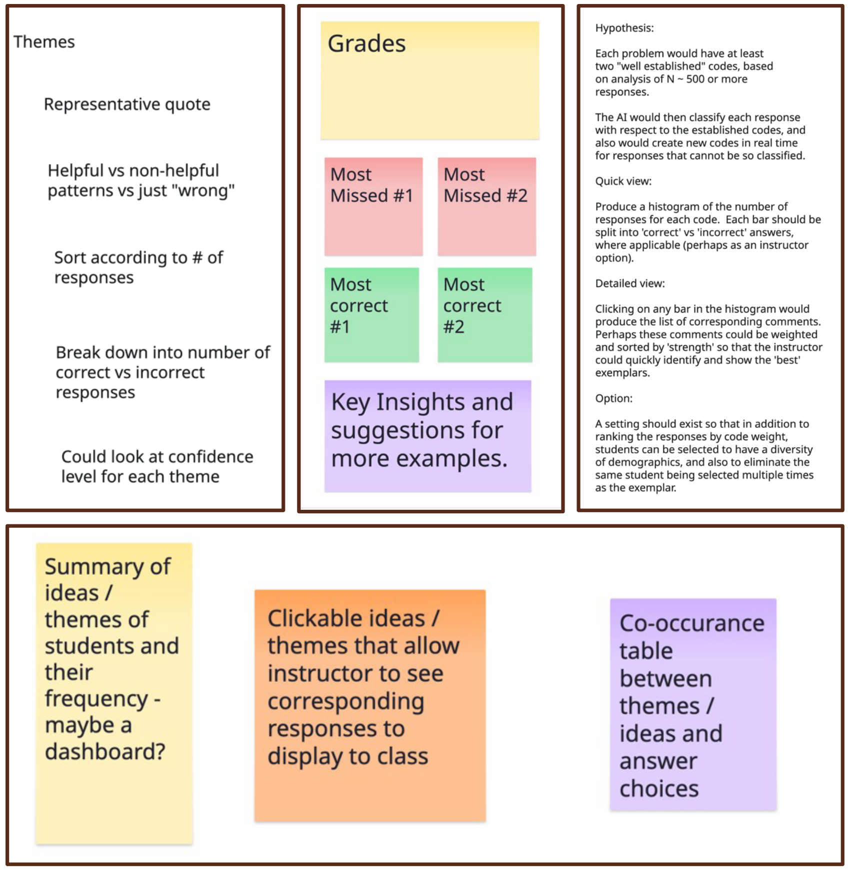

Fig. 5. Samples of instructor ideas for AI Assistant Tool UI.

During our discussion, we observed some common themes that emerged regarding the tool:

- **Use of Trends/Patterns from the Tool:** Instructors wanted a feature where the tool would find common patterns/trends associated with student responses.
- **Proposed Interactions with Visualizations:** Instructors discussed how they would like to interact with the visualizations in the UI. For example, an instructor mentioned that they wanted to be able to click on a notable theme/pattern identified by the tool and see example responses related to the theme. Similarly, another instructor identified that the tool could produce a histogram weighed by the strength of the justification. They would then be able to click on a certain bar of the histogram to see example responses.
- **Misconceptions:** Instructors discussed how common misconceptions could also be identified. This also came up alongside separating responses by correctness.

Our future work will utilize what we learned from the community of practice alongside our trained models to design a beta version of the tool for future study and iteration. Throughout this process, we will consult instructors to gather feedback.

Acknowledgements
We acknowledge support from the National Science Foundation (NSF) through grants RFE 2226553 and DUE 2135190. 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. We also want to thank the instructors and students who participated in data collection.

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] 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.

[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] M. T. H. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser, "Self-explanations: How students study and use examples in learning to solve problems," *Cogn. Sci.*, vol. 13, no. 2, pp. 145–182, 1989, doi: 10.1207/s15516709cog1302_1.

[6] M. T. H. Chi, N. De Leeuw, M.-H. Chiu, and C. Lavancher, "Eliciting self-explanations improves understanding," *Cogn. Sci.*, vol. 18, no. 3, pp. 439–477, Jul. 1994, doi: 10.1016/0364-0213(94)90016-7.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] S. Li, X. Huang, T. Wang, J. Zheng, and S. P. Lajoie, "Using text mining and machine learning to predict reasoning activities from think-aloud transcripts in computer assisted learning," *J. Comput. High. Educ.*, Jun. 2024, doi: 10.1007/s12528-024-09404-6.

[13] N. Tran, B. Pierce, D. Litman, R. Correnti, and L. C. Matsumura, "Analyzing large language models for classroom discussion assessment," in *Proceedings of the 17th International Conference on Educational Data Mining*, International Educational Data Mining Society, Jul. 12, 2024, pp. 500--510. doi: 10.5281/zenodo.12729870.

[14] X. Yu, "Discourse analysis of perceptions of STEM education: A virtual dialogue based on multi-interested subjects," *J. Educ. Humanit. Soc. Sci.*, vol. 20, pp. 298–304, Sep. 2023, doi: 10.54097/ehss.v20i.11686.

[15] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *Int. J. Artif. Intell. Educ.*, vol. 25, no. 1, pp. 60–117, Mar. 2015, doi: 10.1007/s40593-014-0026-8.

[16] S. Gombert *et al.*, "Coding energy knowledge in constructed responses with explainable NLP models," *J. Comput. Assist. Learn.*, vol. 39, no. 3, pp. 767–786, 2023, doi: 10.1111/jcal.12767.

[17] 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.

[18] S. Maestrales, X. Zhai, I. Touitou, Q. Baker, B. Schneider, and J. Krajcik, "Using machine learning to score multi-dimensional assessments of chemistry and physics," *J. Sci. Educ. Technol.*, vol. 30, no. 2, pp. 239–254, 2021.

[19] 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, Apr. 2018, doi: 10.1080/10627197.2018.1427570.

[20] P. P. Martin, D. Kranz, P. Wulff, and N. Graulich, "Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry," *J. Res. Sci. Teach.*, vol. 61, no. 8, doi: 10.1002/tea.21903.

[21] X. Zhai, K. C. Haudek, C. Wilson, and M. Stuhlsatz, "A framework of construct-irrelevant variance for contextualized constructed response assessment," *Front. Educ.*, vol. 6, 2021, Accessed: Feb. 08, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/feduc.2021.751283

[22] 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.

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

[24] Llama Team and AI @ Meta, "The Llama 3 Herd of Models."

[25] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[26] OpenAI *et al.*, "GPT-4o System Card," Oct. 25, 2024, *arXiv*: arXiv:2410.21276. doi: 10.48550/arXiv.2410.21276.

[27] 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.

[28] H. Auby, N. Shivagunde, V. Deshpande, A. Rumshisky, and M. D. Koretsky, "Analysis of student understanding in short-answer explanations to concept questions using a human-centered AI approach," *J. Eng. Educ.*, vol. 114, no. 4, p. e70032, 2025, doi: 10.1002/jee.70032.

[29] 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 of Engineering Education Annual Conference & Exposition*, Jun. 2022. [Online]. Available: https://peer.asee.org/40507

[30] H. Auby, N. Shivagunde, A. Rumshisky, and M. D. Koretsky, "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 of Engineering Education Annual Conference & Exposition*, Portland, Oregon, Jun. 2024. [Online]. Available: https://peer.asee.org/46996

[31] H. Auby, N. Shivagunde, A. Rumshisky, and M. Koretsky, “Using machine learning to analyze short-answer responses to conceptually challenging chemical engineering thermodynamics questions,” in *Proceedings of the 2024 American Society of Engineering Education Annual Conference & Exposition*, Portland, Oregon, Jun. 2024. [Online]. Available: https://peer.asee.org/48236
[32] H. Auby, N. Shivagunde, A. Rumshisky, and M. D. Koretsky, “RFE: Machine learning for student reasoning during challenging concept questions - Year 2,” in *Proceedings of the 2025 American Society of Engineering Education Annual Conference & Exposition*, Montreal, QC, Canada, Jun. 2025. [Online]. Available: https://peer.asee.org/55820
[33] 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.
[34] J. Saldaña, *The coding manual for qualitative researchers*. SAGE Publications, 2021.
[35] 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.
[36] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, “Resources, framing, and transfer,” in *Transfer of learning from a modern multidisciplinary perspective*, J. P. Mestre, Ed., in Current perspectives on cognition, learning, and instruction., IAP, 2006.
[37] E. J. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” *ArXiv Prepr. ArXiv210609685*, 2021.
[38] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 09, 2024. [Online].
[39] H. Auby and M. Koretsky, “Student perceptions of learning and engagement using an educational technology tool,” presented at the *2025 ASEE Annual Conference & Exposition*, Jun. 2025. [Online]. Available: https://peer.asee.org/57689
[40] M. D. Koretsky, S. B. Nolen, J. Galisky, H. Auby, and L. S. Grundy, “Progression from the mean: Cultivating instructors’ unique trajectories of practice using educational technology,” *J. Eng. Educ*, vol. 113, no. 2, pp. 330–359, Feb. 2024, doi: 10.1002/jee.20586.
[41] S. B. Nolen and M. D. Koretsky, “Work in Progress: An ecosystems metaphor for propagation,” presented at the 2020 ASEE Virtual Annual Conference Content Access, Jun. 2020. Accessed: Feb. 07, 2024. [Online]. Available: https://peer.asee.org/work-in-progress-an-ecosystems-metaphor-for-propagation
[42] H. Auby, J. Galisky, S. Nolen, and M. D. Koretsky, “WIP: Instances of dynamic pedagogical decision making in the uptake of a technology tool,” in *Proceedings of the 2022 American Society of Engineering Education Annual Conference & Exposition*, Baltimore, Maryland, Jun. 2022, pp.1–9. [Online]. Available: https://peer.asee.org/44389
[43] J. J. Ferrare and M. W. Apple, “Field theory and educational practice: Bourdieu and the pedagogic qualities of local field positions in educational contexts,” *Camb. J. Educ.*, vol. 45, no. 1, pp. 43–59, Jan. 2015, doi: 10.1080/0305764X.2014.988682.