Stumbling Our Way Through Finding a Better Prompt: Using GPT-4 to Analyze Engineering Faculty's Mental Models of Assessment

Amanda Ross, Virginia Polytechnic Institute and State University

Amanda Ross is a graduate student in the Department of Engineering Education at Virginia Tech. She holds a B.S. in Computer Science and Mathematics from the University of Maryland, Baltimore County.

Dr. Andrew Katz, Virginia Polytechnic Institute and State University

Andrew Katz is an assistant professor in the Department of Engineering Education at Virginia Tech. He leads the Improving Decisions in Engineering Education Agents and Systems (IDEEAS) Lab.

Kai Jun Chew, Embry-Riddle Aeronautical University, Daytona Beach

Kai Jun "KJ" Chew is an assistant professor in the Engineering Fundamentals department at Embry-Riddle Aeronautical University. He is passionate about teaching and research, and he strives to produce knowledge that informs better teaching. His research intersects assessment and evaluation, motivation, and equity. His research goal is to promote engineering as a way to advance social justice causes.

Dr. Holly M. Matusovich, Virginia Polytechnic Institute and State University

Dr. Holly Matusovich is the Associate Dean for Graduate and Professional Studies in the College of Engineering at Virginia Tech and a Professor in the Department of Engineering Education where she has also served in key leadership positions. Dr. Matusovich is recognized for her research and leadership related to graduate student mentoring and faculty development. She won the Hokie Supervisor Spotlight Award in 2014, received the College of Engineering Graduate Student Mentor Award in 2018, and was inducted into the Virginia Tech Academy of Faculty Leadership in 2020. Dr. Matusovich has been a PI/Co-PI on 19 funded research projects including the NSF CAREER Award, with her share of funding being nearly $3 million. She has co-authored 2 book chapters, 34 journal publications, and more than 80 conference papers. She is recognized for her research and teaching, including Dean's Awards for Outstanding New Faculty, Outstanding Teacher Award, and a Faculty Fellow. Dr. Matusovich has served the Educational Research and Methods (ERM) division of ASEE in many capacities over the past 10+ years including serving as Chair from 2017-2019. Dr. Matusovich is currently the Editor-in-Chief of the journal, Advances in Engineering Education and she serves on the ASEE committee for Scholarly Publications.

Stumbling Our Way Through Finding a Better Prompt: Using GPT-4 to Analyze Engineering Faculty Members' Mental Models of Assessment

Abstract

In this full research paper, we discuss the benefits and challenges of using GPT-4 to perform qualitative analysis to identify faculty's mental models of assessment. Assessments play an important role in engineering education. They are used to evaluate student learning, measure progress, and identify areas for improvement. However, how faculty members approach assessments can vary based on several factors, including their own mental models of assessment. To understand the variation in these mental models, we conducted interviews with faculty members in various engineering disciplines at universities across the United States. Data was collected from 28 participants from 18 different universities. The interviews consisted of questions designed to elicit information related to the pieces of mental models (state, form, function, and purpose) of assessments of students in their classrooms. For this paper, we analyzed interviews to identify the entities and entity relationships in participant statements using natural language processing and GPT-4 as our language model. We then created a graphical representation to characterize and compare individuals' mental models of assessment using GraphViz.

We asked the model to extract entities and their relationships from interview excerpts, using GPT-4 and instructional prompts. We then compared the results of GPT-4 from a small portion of our data to entities and relationships that were extracted manually by one of our researchers. We found that both methods identified overlapping entity relationships but also discovered entities and relationships not identified by the other model. The GPT-4 model tended to identify more basic relationships, while manual analysis identified more nuanced relationships.

Our results do not currently support using GPT-4 to automatically generate graphical representations of faculty's mental models of assessments. However, using a human-in-the-loop process could help offset GPT-4's limitations. In this paper, we will discuss plans for our future work to improve upon GPT-4's current performance.

Introduction

Assessments are found in every engineering classroom and are an important part of our education system [1]-[3]. Assessments play many different roles, including understanding student improvements in learning [4], acting as a tool to assist students with learning [5], [6], and for accountability purposes [7]-[9]. Because assessments play key roles in engineering education, it is important that we understand how faculty use assessments in their classrooms, including how they are developed and implemented. It is also important to understand how and why these decisions are being made. However, assessment research is currently lacking compared to other pedagogical research in the field of engineering education, and addressing this research gap can help improve engineering education in both economic and social ways [10]-[13].

Addressing the assessment research gap from the faculty perspective serves two purposes. The first is that more often than not, faculty have autonomy to create the assessments for their

classrooms [14], so naturally, they will have valuable experience. The second is because research has shown that leveraging faculty perspective plays a pivotal role in creating change. For example, studies have shown that research on teachers' beliefs influence practice in the classroom, including assessment related decisions [15]-[20].

In this paper, the guiding conceptual framework is mental models. Mental models are the internal representations people have that help them describe, explain, and predict various aspects of a system, including its state, form, function, and purpose [21] [22]. Having these mental models allows individuals to plan out future actions and make decisions [23] [24]. For example, in engineering education, a study showed the usefulness in using a mental model approach to analyze teacher's varying mental models of the engineering design process [25]. This approach helped researchers identify differences in mental models and recommend different curricular approaches based on these differences. Our own work showed that faculty have varying mental models of assessments' purpose [26]. Purpose deals with why a system exists, and the study found seven major reasons that faculty stated as the reason they use assessment. These included assessing student learning with respect to learning outcomes, benchmarking, assessing student learning, assessing for student ability and competence, a formal evaluation or evaluation of quality, external or program evaluation, and decision making.

Graphically representing mental models is one way in which we can analyze and inspect these varying mental models of assessment. These kinds of graphical representations can be in the form of influence diagrams, which are diagrams that illustrate causal connections between variables and have been used in studies of hazardous material exposure [27] to climate change dynamics [28] to more general representations of agents' beliefs and decision-making processes [29].

Our approach of using a graphical representation to characterize and compare mental models can also be applied in other areas of engineering education research. We believe this work will interest the assessment community and researchers interested in using the mental model approach in other areas of the field. This work highlights the importance of understanding faculty members' mental models when it comes to assessments in engineering education.

Methods

To explore the utility of generative text models to analyze mental models from interview transcripts, which were collected using an Institutional Review Board (IRB) approved protocol, we adopted a novel approach using the recently released GPT-4 instruction-tuned large language model from OpenAI. Published research suggests these instruction-tuned large language models fine-tuned through reinforcement learning from human feedback have shown emergent properties [30]. We wanted to explore their ability to assist in mental model generation. We operationalized this task as an ability to extract information needed to construct a knowledge graph. In this framing, the task for GPT-4 was to identify entities and relationships between those entities as expressed by the participant in the interview transcript. To accomplish this, we used the following prompt:

> "You are working on your next research project, which involves identifying faculty member mental models of assessment. You have interviewed faculty members and are

> now constructing entity-relationship diagrams using GraphViz based on the transcripts. Please analyze the following interview transcript excerpt given in the <text> tag and identify the entities and relationships between those entities. Return the information in JSON format with 'entities' and 'relationships' as the two keys in the JSON object. The relationships should be in the form (source, target, relationship_label). For example, if the excerpt mentions teachers, students, and assessments, and students take assessments, the output should look like: {"entities": ["teacher", "student", "assessment"], "relationships": [("students", "assessment", "take")]}. Try to capture as many entities and relationships between those entities as possible. This should include more than just relationships that involve the speaker from the transcript."

We then used the OpenAI API to iterate through rounds of sending short de-identified interview excerpts to the language model preceded by the instructions prompt. We received responses back from GPT-4 and compiled those together to create the lists of entities and relationship tuples. The list of entities and relationships was then represented graphically using the GraphViz library in Python. In that framework, entities were represented as nodes and relationships were directed edges between those nodes in the graph.

To analyze the accuracy of the GPT-4 model, one of our researchers was asked to perform the same task manually on a subset of the data. Using six excerpts from three different participants, entities and their relationships were manually extracted. These results were then compared to those generated by the GPT-4 model.

Results

The resulting graphs of the GPT-4 model are shown in Figure 1 and Figure 2. These graphs were created by using six excerpts from each participant, and so they only represent part of their mental model. We chose these two graphs because they represent participants with very different mental models. One noticeable difference is that Participant 2 has parts of their mental model disconnected from other parts. Compared to participant 12 whose mental model is fully connected. We also see different entities in each participants' mental model. However, little can be said about these claims, as it only represents a small portion of the mental model, and the missing entities from each may appear in other parts of the interview that were not included in this analysis.

It should be noted that these mental models are both mainly unidirectional. That is, most of the relationships tend to move in one direction, and there are few relationships that move back to previously identified entities. More analysis should be done to investigate the reason for this, as it could imply a few entities having more impact on assessment-related decisions.

The accuracy of the model varied with participants. Between 22% and 40% of the entities identified by GPT-4 were also identified manually. Of the entities identified manually, between 18% and 35% were identified by GPT-4. So, both the GPT-4 model and our researcher identified various entities and their relationships that the other did not.

Fig. 1. Participant 2 Graphical Representation of Assessment Mental Model from Six Excerpts

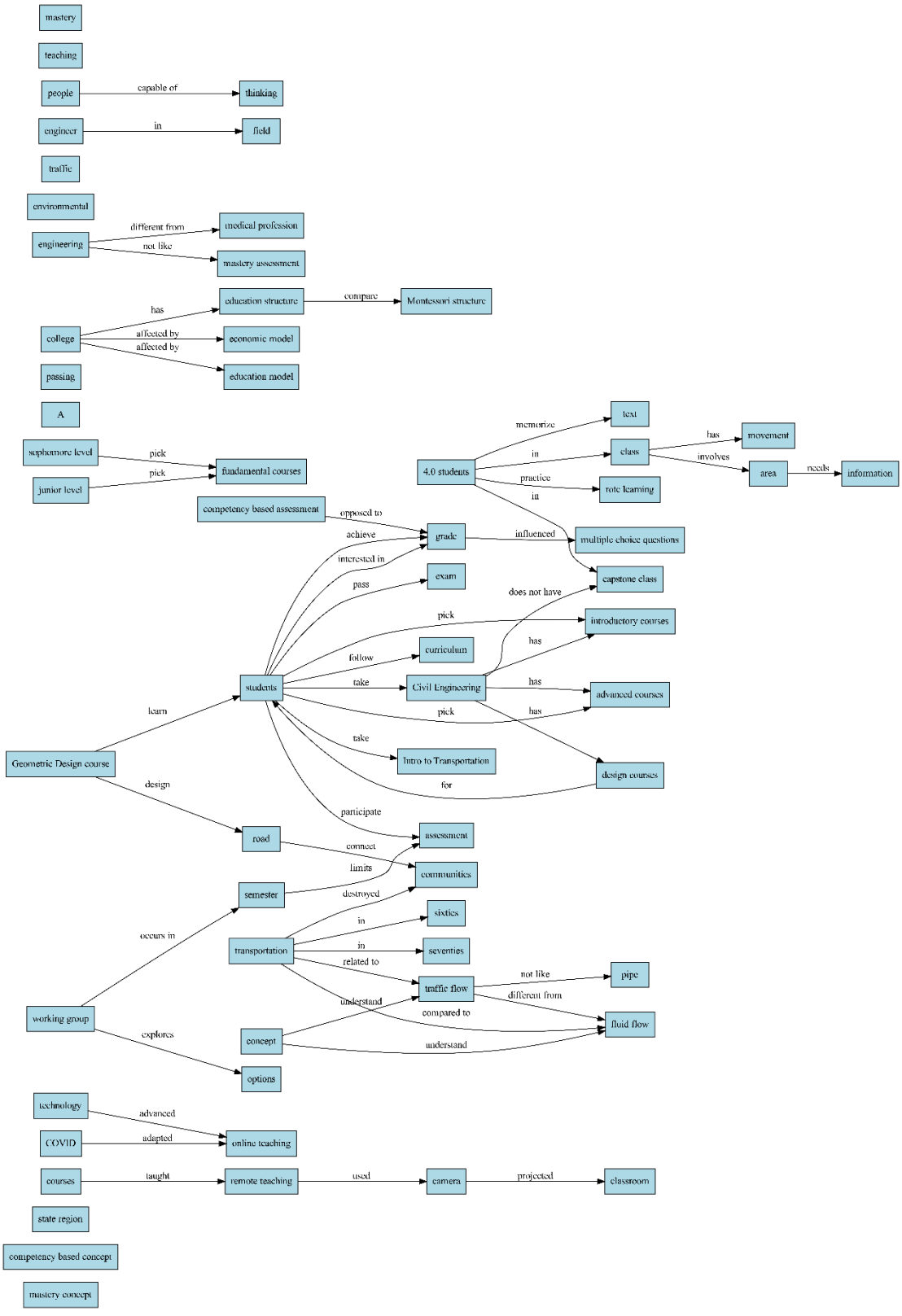
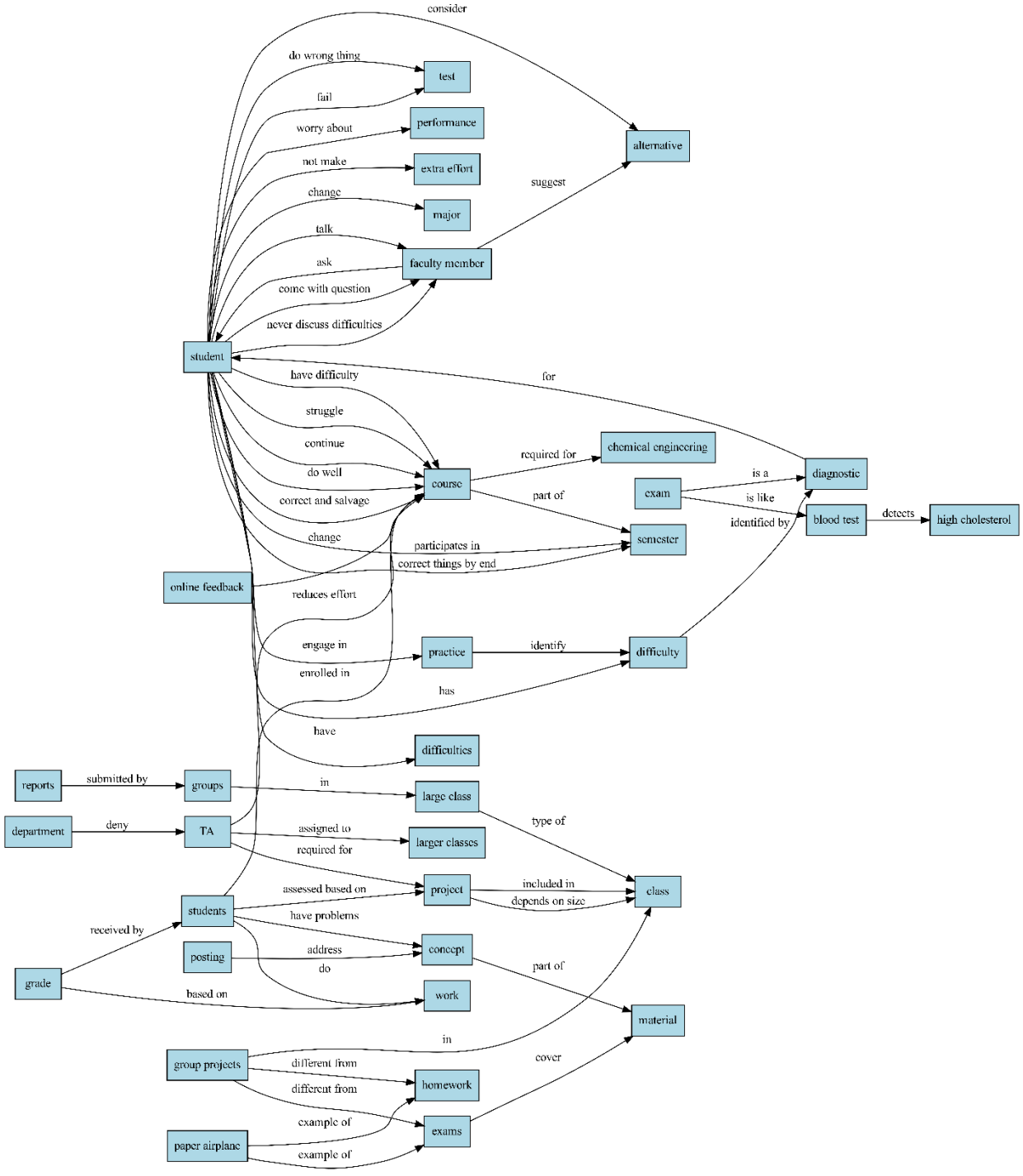


Fig. 2. Participant 12 Graphical Representation of Assessment Mental Model from Six Excerpts



Our analysis showed a difference in the types of relationships each method identified. The GPT-4 model tended to capture basic relationships. For example, when a participant was discussing a civil engineering design course, the GPT-4 model picked up the entity relationship of (civil engineering, design course, has). While this relationship is not incorrect, it was not identified manually, as the participant did not explicitly state that the major of civil engineering had a design class requirement. Instead, the participant was discussing what happens in a civil engineering design course, which is teaching road design in a step-by-step process. So, our manual method identified the entity relationship of (how to design a road, step by step, is taught). This entity relationship is also not incorrect but is more nuanced than the one the GPT-4 model identified.

Discussion of exams provides another example. The GPT-4 model identified the relationship (students, exams, take) that was not identified manually. Again, a correct relationship, and one the model had to infer as it was not explicitly stated. Conversely, the model then failed to identify the relationship (exams, formative, are) that was identified manually.

This difference in what was identified was a common occurrence when comparing the GPT-4 model results and the results from the manual process. While neither necessarily identified incorrect relationships, the GPT-4 model identified more basic, yet not explicitly stated, relationships. Common instances of these have the relationship label of 'is', 'has' or 'have', 'contains', or 'take'. Compared to the manual results that identified relationships that were only explicitly stated but were more nuanced and complicated.

Another finding is that the GPT-4 model was able to more easily identify correct entities but may not have identified a meaningful relationship between them. An example of this can be seen with the identified relationship (Geometric design course, students, learn). The entities of 'Geometric design course' and 'students' were both entities identified manually, but the relationship between those two entities identified by GPT-4 has little meaning, as saying 'Geographic design course learn students' does not make sense. Manually, the relationship was identified as 'teaches', but also identified more explicitly what it taught to students.

Our results also show that GPT-4 picks up entities and their relationships for non-assessment related ideas, as well as assessment related things. While this is not necessarily a bad thing, it means that graphically creating a participant's mental model using our current method cannot be fully automated. There should be a human in the loop to identify which entities and their relationships are assessment related and which are not. This is also supported by the fact that GPT-4's accuracy is not high enough to solely rely on its findings. Instead, combining GPT-4 results with manual results would make for a more complete and accurate representation of a participant's mental model.

Discussion

Our results presented here do not currently support the idea of using GPT-4 in this way to accurately capture faculty's mental models of assessment, although this could be for several reasons. It cannot yet capture everything it needs to to accurately build participants' mental models, at least not in a fully automatic way. This might be because of the prompting, or it could

be an indication of fundamental GPT-4 model limitations. Some of the responses from GPT-4, such as the example mentioned earlier of (Geometric design course, students, learn), do not make grammatical sense. While these language models are very sophisticated, they cannot yet fully read or write without error.

While GPT-4's accuracy does not currently support fully automating generating a graphical representation of participants' mental models of assessment using our method, there are a few things that can be done to improve results. First, you could combine the results of the GPT-4 model and the manual results. This would create a mental model that has nuanced and implied relationships. Second, one could look to better prompt GPT-4. In future work, we plan to add to our prompt more information about what we are looking for and modifying other elements of the prompt through prompt engineering. One option is to provide the model with examples instead of just giving the model instructions (i.e., few-shot prompting). We also plan to utilize newer, open-source models. Using such models helps address important data security questions.

On a positive note, using GPT-4 had two major benefits over manually extracting entities and their relationships. First, GPT-4 was much quicker than our manual analysis. Second, our current method extracted and formatted data in a way to make it easy to compare mental models across participants. In the future, we plan to use graph analysis techniques to analyze similarities and differences in the graphical representation of participants' mental models.

Limitations

There are at least three major limitations with our current method. First, the way we analyzed the interviews to be able to graphically represent participants' mental models does not catch every aspect discussed in the interview in two closely related ways. The first is with descriptions. Sometimes, participants discussed entities and their relationships, and then described these relationships using other entities. For example, one participant noted that they will sometimes review commonly missed exam questions by posting the solutions online. Here, you can manually extract the relationship (professor, commonly missed exam questions, reviews). However, this leaves out how the professor structures these reviews. We could add this information in the relationship to change it to (professor, commonly missed exam questions, reviews by posting solutions online), but this starts to hold too much information in one tuple and becomes a bit confusing. Additionally, it can be argued that online exam solutions are an entity in and of itself. So, if we simply add the detail into the relationship, we lose capturing that entity information.

Similarly, our current method does not handle more complex entities. Our approach defines an entity as a noun. However, in some interviews, we found that entities were ideas, or were represented as another entity relationship tuple. For example, one participant noted that they taught differently based on whether their students understood the material. From this, we can extract the two relationships of (students, material, do understand) and (students, material, do not understand). However, to incorporate how their teaching method interacts with these ideas, an entity would have to be the entire 'students do understand the material' or 'students do not understand the material' entity relationship tuple. However, our current approach does not allow

for an entity to be an entity relationship tuple. In future work, we plan to address this issue by prompting GPT-4 in a way that would allow this sort of entity and relationship interaction.

The final limitation of our approach is that this method only captures participants' mental models of assessment at a single point in time, based on specific prompted questions. It is possible for their mental models to evolve over time, or for our interview questions to have not captured all aspects of their mental model.

Implications

Graphically representing mental models allows us to visually inspect and compare across participants. It makes it easy to analyze similarities and differences in faculty's mental models, which can help us understand how various mental models lead faculty to make certain decisions. In this study, we used GPT-4 to help extract information to generate those representations. We found that the model showed promising results yet did not match human performance on the extraction task. Although not ready yet, language models such as GPT-4 may soon help researchers analyze large amounts of any type of qualitative data, something that normally would take a large amount of time and resources. These models work especially well when looking at data through a specific lens, as they can be prompted to focus on certain aspects.

Notably, researchers are not the only ones who might benefit from these tools. Professors can also benefit from these methods while teaching. It can be difficult for faculty to receive feedback from students and make changes during the semester. In large classes, even just a few sentences from each student becomes time consuming to sift through and identify common trends on where improvements can be made. But using NLP methods can help save time and be used to identify these common themes for the professor. We plan to pursue future work along these paths toward identifying ways generative text models can assist teaching and research in engineering education.

References

[1] L. Suskie, Assessing student learning: A common sense guide. Jossey-Bass, 2018.

[2] J. W. Pellegrino, N. Chudowsky, and R. Glaser, Knowing what students know: The science and design of educational assessment. Washington, DC, 2001.

[3] G. P. Wiggins and J. McTighe, "What is backward design?," in Understanding by design, 2011, pp. 7–19.

[4] L. Lachlan-Haché and M. Castro, "Proficiency or growth? An Exploration of two approaches for writing student learning targets acknowledgments," 2015. [Online]. Available: https://www.air.org/sites/default/files/Exploration-of-Two-Approaches-Student-Learning-Targets-April-2015.pdf.

[5] G. Gibbs, "Using assessment strategically to change the way students learn," in Assessment matters in higher education: Choosing and using diverse approaches, S. Brown and A. Glasner, Eds. 1999, pp. 41–53.

[6] L. A. Shepard, "The role of assessment in a learning culture," Educ. Res., vol. 29, no. 7, pp. 4–14, 2000.

[7] J. W. Prados, G. D. Peterson, and L. R. Lattuca, "Quality assurance of engineering education through accreditation: The impact of engineering criteria 2000 and its global influence," J. Eng. Educ., vol. 94, no. 1, pp. 165–184, 2005, doi: 10.1002/j.2168-9830.2005.tb00836.x.

[8] J. T. Brown, “The seven silos of accountability in higher education: Systematizing multiple logics and fields,” Res. Pract. Assess., vol. 11, no. 2017, pp. 41–58, 2014.
[9] P. Nagy, “The Three Roles of Assessment : Gatekeeping, Accountability, and Instructional Diagnosis,” Can. J. Educ., vol. 25, no. 4, pp. 262–279, 2000.
[10] National Academy of Engineering, “The engineer of 2020: Visions of engineering in the new century,” Washington, DC, 2004. doi: http://www.nap.edu/catalog/10999.html.
[11] L. Jamieson and J. Lohmann, Creating a culture for scholarly and systematic innovation in engineering education. Washington, DC: American Society of Engineering Education (ASEE), 2009.
[12] L. L. Long III, “Toward an antiracist engineering classroom for 2020 and beyond: A starter kit,” J. Eng. Educ., vol. 109, no. 4, pp. 636–639, 2020, doi: 10.1002/jee.20363.
[13] K. A. Douglas, A. Rynearson, Ş. Purzer, and J. Strobel, “Reliability, validity, and fairness: A content analysis of assessment development publications in major engineering education journals,” Int. J. Eng. Educ., vol. 32, no. 5, pp. 1960–1971, 2016.
[14] L. R. Lattuca and J. S. Stark, Shaping the college curriculum: Academic plans in context. San Francisco, CA: Jossey-Bass, 2009.
[15] J. Skott, “The promises, problems, and prospects of research on teachers’ beliefs,” in International handbook of research on teachers’ beliefs, H. Fives and M. G. Gill, Eds. New York, NY: Routledge, 2015, pp. 13–30.
[16] N. Barnes, H. Fives, and C. M. Dacey, “Teachers’ beliefs about assessment,” in International handbook of research on teachers’ beliefs, H. Fives and M. G. Gill, Eds. Routledge, 2015, pp. 284–300.
[17] M. F. Pajares, “Teachers’ Beliefs and Educational Research: Cleaning Up a Messy Construct,” Rev. Educ. Res., vol. 62, no. 3, pp. 307–332, 1992, doi: 10.3102/00346543062003307.
[18] H. Fives and M. M. Buehl, “Spring cleaning for the ‘messy’ construct of teachers’ beliefs: What are they? Which have been examined? What can they tell us?,” in APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors., vol. 2, 2011, pp. 471–499.
[19] M. M. Buehl and J. S. Beck, “The relationship between teacher’s beliefs and teacher’s practices,” in International handbook of research on teachers’ beliefs, H. Fives and M. G. Gill, Eds. Routledge, 2015, pp. 66–84.
[20] L. A. Bryan, “Nestedness of beliefs: Examining a prospective elementary teacher’s belief system about science teaching and learning,” J. Res. Sci. Teach., vol. 40, no. 9, pp. 835–868, 2003, doi:10.1002/tea.10113.
[21] P. N. Johnson-Laird, “Mental models and human reasoning,” Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 43, pp. 18243–18250, 2010, doi:10.1073/pnas.1012933107.
[22] W. B. Rouse and N. M. Morris, “On looking into the black box: Prospects and limits in the search for mental models,” Psychol. Bull., vol. 100, no. 3, pp. 349–363, 1985, [Online]. Available: http://scholar.google.com/scholar?q=related:QM4p5zGC8jMJ:scholar.google.com/&hl=en&num=30&as_sdt=0,5.
[23] K. Carley and M. Palmquist, “Extracting, representing, and analyzing mental models,” Soc. Forces, vol. 70, no. 3, pp. 601–636, 2016.
[24] C. D. Wickens and A. Kramer, “Engineering psychology.,” Annu. Rev. Psychol., vol. 36, no. 1, pp. 307–348, 1985, doi:10.1146/annurev.ps.14.020163.001441.

[25] A. P. McMahon, “Mental models elementary teachers hold of engineering design processes: A comparison of two communities of practice,” 2012, doi: 10.18260/1-2--21686.
[26] K. J. Chew, A. Ross, A. Katz and H. M. Matusovich, "Defining Assessment: Foundation Knowledge Toward Exploring Engineering Faculty’s Assessment Mental Models," 2022 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 2022, pp. 1-8, doi: 10.1109/FIE56618.2022.9962667.
[27] A. Bostrom, B. Fischhoff, and M. G. Morgan, “Characterizing Mental Models of Hazardous Processes: A Methodology and an Application to Radon,” Journal of Social Issues, vol. 48, no. 4, pp. 85–100, 1992, doi: 10.1111/j.1540-4560.1992.tb01946.x.
[28] T. D. Lowe and I. Lorenzoni, “Danger is all around: Eliciting expert perceptions for managing climate change through a mental models approach,” Global Environmental Change, vol. 17, no. 1, pp. 131–146, Feb. 2007, doi: 10.1016/j.gloenvcha.2006.05.001.
[29] Y. Gal and A. Pfeffer, “Networks of Influence Diagrams: A Formalism for Representing Agents’ Beliefs and Decision-Making Processes,” Journal of Artificial Intelligence Research, vol. 33, pp. 109–147, Sep. 2008, doi: 10.1613/jair.2503.
[30] J. Wei et al., “Emergent Abilities of Large Language Models.” arXiv, Oct. 26, 2022. Accessed: May 15, 2023. [Online]. Available: http://arxiv.org/abs/2206.07682