Analyzing AI and student responses through the lens of sensemaking and mechanistic reasoning

Dean Zollman, Amogh Sirnoorkar, and James T. Laverty
Department of Physics, Kansas State University, Manhattan, Kansas - 66502

Physics education research (PER) shares a rich tradition of designing learning environments that promote valued epistemic practices such as sensemaking and mechanistic reasoning. Recent technological advancements, particularly artificial intelligence has caught significant traction in the PER community due to its human-like, sophisticated responses to physics tasks. In this study, we contribute to the ongoing efforts by comparing AI (ChatGPT) and student responses to a physics task through the cognitive frameworks of sensemaking and mechanistic reasoning. Findings highlight that by virtue of its training data set, ChatGPT's response provide evidence of mechanistic reasoning and mimics the vocabulary of experts in its responses. On the other hand, half of students' responses evidenced sensemaking and reflected an effective amalgamation of diagram-based and mathematical reasoning, showcasing a comprehensive problem-solving approach. In other words, while AI responses reflected *how physics is talked about*, a part of students' responses reflected *how physics is practiced* during problem solving. We discuss the implications of this finding with an emphasis on epistemology of AI responses and designing next-generation assessments in physics.

2023 PERC Proceedings edited by Jones, Ryan, and Pawl; Peer-reviewed, doi.org/10.1119/perc.2023.pr.Zollman

I. INTRODUCTION

Physics education research shares a rich history of developing learning materials that promote valued epistemic practices. Sensemaking, the process of addressing perceived gaps in understanding, is one such practice [1]. Sensemaking represents one of the primary processes through which disciplinary experts generate new knowledge [2–4]. Sensemaking assists students in better content understanding [5] and facilitates in developing "sophisticated epistemology" [6]. Given this significance, there has been an uptick in investigations concerning the nature of discourse [7] and task features [8–10] that promote sensemaking in physics.

Artificial intelligence (AI) interfaces such as ChatGPT [11] have also gained increased traction in PER. These interfaces are software applications designed to generate human-like responses through Natural Language Processing. The "chatbots" interact with users by generating probabilistic responses to prompts based on large training data sets including physics texts [12]. Researchers have observed these interfaces to "pass" physics standardized assessments such as the Force Concept Inventory [13–15]. Noting this grammatical and technical sophistication, studies have called for exploring the ways in which AI can be leveraged in promoting valued epistemic practices in learning environments [16].

We contribute to these efforts by shifting the focus from analyzing conceptual correctness of AI responses to analyzing them within the frameworks of cognitive processes which are fundamental to scientific inquiry [1, 17]. A better understanding of the AI responses in light of cognitive features can guide the effective integration of emerging technologies in our classrooms and laboratories. We compare the characteristics of responses generated by ChatGPT and introductory students to a physics problem specifically designed to foster the integration of physics principles with everyday ideas. We analyze the responses through the theoretical perspectives of sensemaking and mechanistic reasoning, along with noting the use of representations in making the requisite conclusions.

Findings highlight that by virtue of its training data, ChatGPT mimics the vocabulary of experts in its responses. On the other hand, students' responses reflect an amalgamation of diagram-based and mathematical reasoning, showcasing a comprehensive problem-solving approach. We discuss the implications of these observations for future explorations especially along exploring the epistemology of AI responses and designing assessments. In doing so, we address the following research question: *How does ChatGPT's responses to a physics task compare to students' responses in terms of the characteristics of sensemaking and mechanistic reasoning?*

In the next section, we present the theory of sensemaking and mechanistic reasoning. Thereafter, in Section III, we discuss the physics problem, along with our methods of data collection and analysis. We then present the results of the relative comparison between students' and AI responses on various cognitive features in Section IV before concluding with implications and future work in Section V.

II. THEORY

A. Sensemaking

In the rest of this paper, we adopt the following account of sensemaking from Odden and Russ [1]:

> "*a dynamic process of building or revising an explanation in order to 'figure something out' - to ascertain the mechanism underlying a phenomenon in order to resolve a gap or inconsistency in one's understanding. One builds this explanation out of a mix of everyday knowledge and formal knowledge by iteratively proposing and connecting up different ideas on the subject. One also simultaneously checks that those connections and ideas are coherent, both with one another and with other ideas in one's knowledge system.*"

Based on the above definition, we identify the following markers (henceforth referred to as "sensemaking elements") which together evidence the sensemaking process.

1. Noticing of inconsistency in understanding.
2. Blending everyday and formal knowledge.
3. Generating and connecting diverse ideas (e.g., conceptual, procedural, and intuitive).
4. Seeking coherence between the generated ideas.
5. Unpacking the mechanism of the phenomenon.

We analyze AI and students' responses to our physics problem (Section III) through the lens of the above-mentioned elements. The fifth sensemaking element – unpacking the underlying mechanism of a phenomenon – however, is a complex process in itself. In the next sub-section we briefly detail what this process is, along with describing its markers.

B. Mechanistic reasoning

The process of unpacking the underlying mechanism of a phenomenon (also known as "mechanistic reasoning") is a form of causal reasoning that entails description of the events and factors responsible for the occurrence of the phenomenon. Mechanistic reasoning entails generating explanations by moving from the observable features of the phenomenon to the underlying entities or processes, often at a micro level. The process of ascertaining the mechanism may also involve transitioning back from the micro to the macro-level features, as well as testing the validity of the explanation by altering the spatial or temporal organization of the involved entities or processes.

You are asked to design a Gravitron for the county fair, an amusement park ride where the rider enters a hollow cylinder, radius of 4.6 m, the rider leans against the wall and the room spins until it reaches angular velocity, at which point the floor lowers. The coefficient of static friction is 0.2. You need this ride to sustain mass between 25-160 kg to be able to ride safely and not slide off the wall. If the minimum ω is 3 rad/s, will anyone slide down and off the wall at these masses? Explain your reasoning using diagrams, equations and words.

FIG. 1. Statement of the Gravitron problem

As an example, consider the phenomenon of the apparent shift in the position of a coin placed in a glass of water as viewed from above. A mechanistic account underlying this observation can be provided using the principle of refraction of light. When a ray of light passes from a denser medium (water) to a rarer medium (air), it deviates away from the normal drawn at the interface between the two media. This deviation results in the formation of an apparent image of the coin, making it appear closer to the water's surface.

Krist *et al.* [18] describe the process of generating mechanistic explanations in terms of the three patterned strategic ways of knowledge-building called "epistemic heuristics":

1. Thinking across scalar levels. Describing the *actors* identified and characterized at the scalar level below the observed phenomenon. In our example, transitioning to the level of *light rays* marks this heuristic.

2. Identifying and unpacking relevant factors. The description of the *activities* engaged by the identified actors at the lower scalar level. This can correspond to the *bending* of light rays relative to the normal and the formation of the apparent image of the coin.

3. Linking to coordinate relationships over time and space. Validating the generated explanation by varying the temporal and spatial organization of the involved *actors* and their *activities*. In our example, reversing the media and concluding that an object in the air medium viewed from the water would be perceived as farther away than its actual position.

We analyze mechanistic reasoning in students' and AI's responses through the lens of the above-mentioned heuristics.

III. METHODS

The objective of the current study is to compare students' and ChatGPT's responses to a physics problem which we refer to as the "Gravitron task" (Refer to Fig 1 for the problem statement). The task was designed using the Three-Dimensional Learning Assessment Protocol (3D-LAP) [19] to elicit the scientific practice of "Developing and Using Models" [20]. It involves a rotating cylindrical amusement ride in which a rider leans against the wall. Given the various parameters of the ride, the task asks to make a claim/prediction whether the rider would slide off the walls. Due to space constraints, we refrain from providing a detailed solution of the task and refer readers to our earlier work [21].

The students' responses to the Gravitron task are derived from think-aloud interviews conducted in Spring 2018. The interviews involved ten introductory students individually solving a total of ten physics problems (with the Gravitation task being seventh on the list) while simultaneously articulating their thoughts aloud. The interview protocol involved asking the participants to treat the problem-solving exercise as an untimed exam. During moments of prolonged silence, the interviewer interjected with questions such as "*What are you thinking?*" to make students articulate their thoughts out loud. Of the ten, two students' responses contained audio/video issues and thus are not part of this study. On the other hand, we recorded ChatGPT's responses to the Gravitron task by using the same problem statement provided to students in the chatbot's interface. In order to maintain symmetry in terms of the number of student responses, ChatGPT was prompted eight separate times by initiating distinct chat sessions, and the responses were recorded accordingly.

As part of the analysis, the students' responses to the Gravitron task were first transcribed by taking into account their verbal arguments and written solutions. The second author then analyzed the transcripts through the lens of the five sensemaking elements described in Section II. The first element, noticing inconsistencies, was identified in the responses by noting puzzling questions [7], intermittent pauses, incomplete arguments [21, 22], etc. The second sensemaking element, blending everyday and formal knowledge, was identified when the responses reflected amalgamation of formal physics principles with the Gravitron's physical system. Arguments such as "the rider being pulled down by the gravity" or "friction preventing the rider from slipping down" were coded as the second element. The third and fourth elements were captured based on the generation and validation of intuitive, conceptual, or procedural ideas.

We captured the fifth sensemaking element (mechanistic reasoning) by identifying the three epistemic heuristics described in Section II. Students' arguments concerning forces (and other quantities) were coded as the first epistemic heuristic. However, students who engaged only in plug-and-chug of equations, even though the equations entailed force terms were not coded for this heuristic. Identifying the relevant forces (and other quantities) and their interaction on the Gravitron's riders were coded as the second heuristic of mechanistic reasoning. Lastly, transitioning to the macroscopic Gravitron scenario to make the requisite claim marked the final heuristic in the students' and chatbot's responses. In addition, since the problem statement explicitly asked for relevant representations used in making the claims, and the contemporary literature notes the role of representations in mechanistic reasoning [23], we also note diagrams and equations employed

in the responses. In summary, the analysis of students' and ChatGPT's responses involved (i) identifying the first four sensemaking elements, (ii) the three epistemic heuristics of mechanistic reasoning (the fifth sensemaking element), and (iii) representations such as diagrams and equations.

We wish to highlight a few potential limitations in our methodology. First, the nature of responses across the two sets are inherently not identical. And two, since the authors were aware of the source of the solutions, we acknowledge there could be a likely bias while analyzing the responses.

IV. RESULTS

We now discuss how the sensemaking elements manifest in students' and AI's responses to the Gravitron task. These results have also been summarized in Table I.

TABLE I. Summary of the sensemaking and mechanistic reasoning features in students' and ChatGPT's responses to the Gravitron task.

Criteria	Students (Out of 8)	ChatGPT (Out of 8)
Sensemaking	4	-
Mechanistic reasoning	4	8
Sensemaking elements		
Noticing gaps in understanding.	4	-
Blending everyday and formal knowledge.	6	8
Generating and connecting ideas.	5	8
Seeking coherence between the ideas.	4	8
Epistemic heuristics of mechanistic reasoning		
Thinking across scalar levels.	7	8
Identifying and unpacking relevant factors.	4	8
Linking to coordinate relationships over time and space.	6	8
Representations		
Diagrams	7	3
Equations	8	8

1. Noticing inconsistency in understanding

The meta-cognitive activity of noticing gaps in one's knowledge system forms a central feature of the sensemaking process. Four of the eight students' responses reflected evidence of noticing this gap (two of which have been detailed in References [21, 22]). None of the ChatGPT responses reflected this sensemaking element. We believe the implication of this result needs to be reemphasized as it is easy to implicitly model ChatGPT as a "person responding from the other end". As Kortemeyer [13] notes

> "*It is irritatingly hard not to anthropomorphize ChatGPT. As a physics teacher, one invariably finds oneself rooting for the students and thus by extension also for ChatGPT, celebrating its successes and being frustrated about its occasionally inexplicable failures.*"

By virtue of its design, the chatbot generates probabilistic responses from its training data set and at least currently does not possess the meta-cognitive ability to address gaps in its "knowledge system".

2. Use of everyday and formal knowledge

As a task based on real-world scenario, a key feature of the Gravitron task is its potential to facilitate the amalgamation of the physics principles with everyday ideas. Six of the eight responses from students entailed arguments which reflected how various forces (and other quantities) interacted in holding up the Gravitron's rider. On the other end, all of the eight ChatGPT's responses elicited this mode of reasoning (albeit some were conceptually incorrect).

3. Generating and connecting ideas

The Gravitron task requires leveraging conceptual and procedural ideas from both physics and mathematics. Five of the students' responses and all eight of the chatbot's responses reflected this feature. Both sets of responses included arguments involving diverse sets of physical quantities such as forces, accelerations, and momenta along with their directionality. However, students' responses differed from ChatGPT in two ways. They reflected intuitive arguments without any conceptual or procedural justification. Students' responses also included incomplete arguments which were either changed or terminated abruptly. Such iterative construction and/or revision of arguments has been noted as a characteristic feature of students' knowledge building process [1].

4. Seeking coherence between ideas

The generated ideas were then brought together in making the required claim about the riders' status inside the Gravitron. Half of the students and all eight responses from ChatGPT connected and/or justified the solution back in terms of the earlier invoked ideas. However, we observed two key differences between the students' and AI's concluding statements. Interestingly, all of the ChatGPT's responses made the incorrect conclusion that riders would not fall off the Gravitron's walls (as compared to four of the students' conclusions). This observation assumes significance as the chatbot does not "solve" the given problem, but rather produces the most probable result from its training data set. This result indicates that despite all its sophistication in the previous two sensemaking elements, the conclusions from AI should not be taken at its face value. Secondly, three of the ChatGPT conclusions accompanied detailed assumptions under which the conclusions hold true.

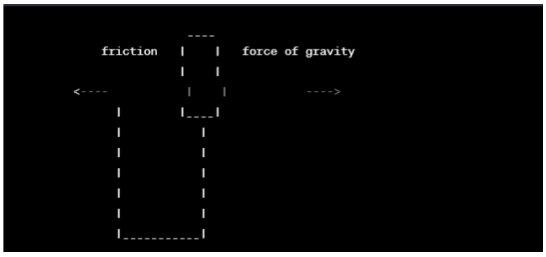

FIG. 2. One of the three diagrams generated by ChatGPT in response to the Gravitron task.

5. Generating mechanistic explanations

5.1 Epistemic heuristics

We see symmetry in both sets of responses on this aspect. While only half of the students' solutions reflected the three epistemic heuristics, all of ChatGPT's responses reflected the "mechanistic sophistication" [17].

5.2 Use of representations

One of the contrasting differences between the two data sets is in terms of diagram-based reasoning. While seven of the eight students' written responses entailed diagrams, only three were noted in ChatGPT's responses. Figure 2 represents one of the three diagrams generated by ChatGPT. The AI-generated diagrams differed from student ones on two features. The AI representations (unlike the student-generated diagrams) did not represent the rider-Gravitron system by highlighting the interaction of relevant forces. And secondly, the responses did not leverage diagram-based arguments into their conclusions. Rather all the three diagrams were generated after making the requisite conclusion. Along with diagrams, equations were the other form of the focused representations. On this end, all eight responses from students and ChatGPT contained equations.

V. DISCUSSION AND CONCLUSION

We analyzed students' and ChatGPT's responses to the Gravitron task (Figure 1) in terms of (i) sensemaking, (ii) mechanistic reasoning, and (iii) the use of diagrams/equations. The findings (summarized in Table I) highlight two key contrasting features between the two sets of responses. Most of the students' reasoning highlight richness in the practices of the problem-solving exercise, especially in blending diagrams with mathematical arguments. That is, students' responses reflected *the practices involved in engaging with physics*. On the other hand, since ChatGPT's responses are drawn from its training data set, they mimic the lexical patterns or vocabulary of an expert's reasoning in physics (despite providing incorrect answers in all of its solutions). In summary, while students' responses reflect *how physics is practiced*, the AI's responses reflect sophistication in *how physics is talked about* during problem solving.

Of the eight students, four evidenced sensemaking and mechanistic reasoning in their responses. On the other hand, all eight AI's responses evidenced the elements of sensemaking (including mechanistic reasoning) with an exception of the meta-cognitive feature of noticing gaps in understanding. This should not come as a surprise as the AI interface is not capable (at least at the moment of writing this paper) of engaging in any meta-cognitive activity. Despite the strong evidence in characteristics of sensemaking and mechanistic reasoning, none of the AI's solutions provided a correct prediction or conclusion. This observation calls for attention on two aspects. One, by virtue of its training data, the AI mimics the vocabulary of an expert on *how physics is talked about* during problem solving. Two, an incorrect answer put forth through sophisticated argument calls for diligence against conflating the conceptual merit of AI's argument with its semantic sophistication. This claim is also in agreement with contemporary studies calling for careful evaluation of AI responses, particularly in terms of scientific accuracy [24].

Yet another contrasting feature between the two sets of responses is difference in the employment of representations during problem solving. While seven of the eight students employed diagrams, only three of the ChatGPT's responses reflected the same. Furthermore, unlike the student-generated diagrams, the AI diagrams did not portray the appropriate Gravitron-rider system with relevant forces and were not integrated in the solutions. Thus students' responses to the Gravitron task reflected the practice of blending diagram-based reasoning with mathematical arguments thereby reflecting the *key practices* of engaging with physics.

For instructors, our observations provide insights on designing assessments using AI. The assessments can focus on asking students to critique and validate AI-generated solutions. Such exercises would serve in promoting the sophisticated vocabulary of *talking about physics* while simultaneously guarding students against conflating the lexical sophistication with conceptual correctness. For researchers, our observations call for investigating the epistemological messages [25] about learning physics conveyed by AI responses to students. That is, unpacking the implicit messages conveyed by AI to students on what counts as "knowing" or "doing" science. Exploring AI responses to well-structured and ill-structured problems also present a potential avenue.

Our future work would involve extending similar methodology in analyzing students and AI solutions across other epistemic practices such as modeling and argumentation. Given the evidence of epistemic heuristics of mechanistic reasoning in AI's responses, we also seek to explore the solutions in finer details using alternate frameworks [17] and across enhanced version of ChatGPT and other AI interfaces.

VI. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2013339.

[1] T. O. B. Odden and R. S. Russ, Defining sensemaking: Bringing clarity to a fragmented theoretical construct, Science Education **103**, 187 (2019).

[2] M. J. Ford, A dialogic account of sense-making in scientific argumentation and reasoning, Cognition and Instruction **30**, 207 (2012).

[3] L. K. Berland and B. J. Reiser, Making sense of argumentation and explanation, Science education **93**, 26 (2009).

[4] B. A. Danielak, A. Gupta, and A. Elby, Marginalized identities of sense-makers: Reframing engineering student retention, Journal of Engineering Education **103**, 8 (2014).

[5] M. A. Cannady, P. Vincent-Ruz, J. M. Chung, and C. D. Schunn, Scientific sensemaking supports science content learning across disciplines and instructional contexts, Contemporary Educational Psychology **59**, 101802 (2019).

[6] T. J. Bing and E. F. Redish, Epistemic complexity and the journeyman-expert transition, Physical Review Special Topics-Physics Education Research **8**, 010105 (2012).

[7] T. O. B. Odden and R. S. Russ, Vexing questions that sustain sensemaking, International Journal of Science Education **41**, 1052 (2019).

[8] A. Sirnoorkar and J. T. Laverty, Theoretical exploration of task features that facilitate student sensemaking in physics, arXiv preprint arXiv:2302.11478 (2023).

[9] O. Sand, T. O. Odden, C. Lindstrom, and M. Caballero, How computation can facilitate sensemaking about physics: A case study (2019).

[10] E. Kuo, M. M. Hull, A. Elby, and A. Gupta, Assessing mathematical sensemaking in physics through calculation-concept crossover, Physical Review Physics Education Research **16**, 020109 (2020).

[11] https://openai.com/product/chatgpt.

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, Language models are few-shot learners, Advances in neural information processing systems **33**, 1877 (2020).

[13] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Physical Review Physics Education Research **19**, 010132 (2023).

[14] C. G. West, Ai and the fci: Can chatgpt project an understanding of introductory physics?, arXiv preprint arXiv:2303.01067 (2023).

[15] C. G. West, Advances in apparent conceptual physics reasoning in gpt-4, arXiv e-prints , arXiv (2023).

[16] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, A swot analysis of chatgpt: Implications for educational practice and research, Innovations in Education and Teaching International , 1 (2023).

[17] R. S. Russ, R. E. Scherr, D. Hammer, and J. Mikeska, Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science, Science education **92**, 499 (2008).

[18] C. Krist, C. V. Schwarz, and B. J. Reiser, Identifying essential epistemic heuristics for guiding mechanistic reasoning in science learning, Journal of the Learning Sciences **28**, 160 (2019).

[19] J. T. Laverty, S. M. Underwood, R. L. Matz, L. A. Posey, J. H. Carmel, M. D. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, and M. M. Cooper, Characterizing college science assessments: The three-dimensional learning assessment protocol, PloS one **11**, e0162333 (2016).

[20] N. R. Council *et al.*, *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (National Academies Press, 2012).

[21] A. Sirnoorkar, J. T. Laverty, and P. Bergeron, Sensemaking and scientific modeling: Intertwined processes analyzed in the context of physics problem solving, arXiv preprint arXiv:2207.03939 (2022).

[22] A. Sirnoorkar and J. Laverty, A methodology for identifying task features that facilitate sensemaking, in *Physics Education Research Proceedings* (2021) pp. 396–401.

[23] V. De Andrade, Y. Shwartz, S. Freire, and M. Baptista, Students' mechanistic reasoning in practice: Enabling functions of drawing, gestures and talk, Science Education **106**, 199 (2022).

[24] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, How do physics students evaluate chatgpt responses on comprehension questions? a study on the perceived scientific accuracy and linguistic quality, arXiv preprint arXiv:2304.05906 (2023).

[25] R. S. Russ, Characterizing teacher attention to student thinking: A role for epistemological messages, Journal of Research in Science Teaching **55**, 94 (2018).