An Assessment of ChatGPT 4o's Performance on Mechanical Engineering Concept Inventories

Dr. Rujun Gao, Texas A&M University

Dr. Rujun Gao has completed her Ph.D. in Mechanical Engineering at Texas A&M University and holds an M.S. in Mechanical Engineering from Zhejiang University, China. Her research focuses on Generative AI, Natural Language Processing (NLP), Large Language Models (LLMs), LLM Agents, and the development of educational technology products.

Hillary E. Merzdorf, Cornell University

STEM Instructional Design Associate, eCornell, Cornell University

Xiaosu Guo, University of Texas at Dallas
Sami Melhem, Texas A&M University

Sami Melhem is an undergraduate student pursuing a Bachelor of Science in Computer Science at Texas A&M University, where he is also planned to enroll in a concurrent Master of Science program in Computer Science. Sami serves as an undergraduate research assistant in the department of Mechanical Engineering under the guidance of Dr. Srinivasa. His research interests include the simulation of manufacturing processes including robotic sheet forming and magnetic polishing, and the development of AI-driven educational tools.

Beyond academics, Sami is deeply involved in the Aggie Data Science Club, where he serves as Projects Officer, overseeing and mentoring multiple student-led activities each semester. He also plays clarinet as the concertmaster for the TAMU Wind Symphony and principal clarinet for the TAMU Chamber Orchestra, demonstrating his passion for both technology and the arts. Sami aspires to leverage his technical expertise to develop scalable AI solutions that address real-world challenges in education and sustainability.

Dr. Kristi J. Shryock, Texas A&M University

Dr. Kristi J. Shryock is the Frank and Jean Raymond Foundation Inc. Endowed Associate Professor in Multidisciplinary Engineering and Affiliated Faculty in Aerospace Engineering in the College of Engineering at Texas A&M University. She also serves as Director of the Craig and Galen Brown Engineering Honors Program. She received her BS, MS, and PhD from the College of Engineering at Texas A&M. Kristi works to improve the undergraduate engineering experience through evaluating preparation in areas, such as mathematics and physics, evaluating engineering identity and its impact on retention, incorporating non-traditional teaching methods into the classroom, and engaging her students with interactive methods.

Prof. Arun R Srinivasa, Texas A&M University

Dr Arun Srinivasa is the J. N. Reddy Chair in Applied Mechanics and Associate Dean for Student Success. He has been with TAMU since 1997. Prior to that he was a faculty at University of Pittsburgh. He received his undergraduate

An Assessment of ChatGPT 4o's Performance on Mechanical Engineering Concept Inventories

Abstract

Large Language Models (LLMs) like OpenAI's ChatGPT-4o show promise for enhancing engineering education through real-time support and personalized feedback. However, their reliability in interpreting the conceptual diagrams central to mechanical engineering remains uncertain. This study evaluates ChatGPT-4o's performance on four concept inventories—Force Concept Inventory, Materials Concept Inventory, Mechanics Baseline Test, and Mechanics of Materials Concept Inventory—using assessments by two Mechanical Engineering professors based on correctness, depth of explanation, and application of theoretical knowledge. While ChatGPT-4o demonstrates the ability to provide robust explanations, it often lacks the contextual depth required for higher-order concept mastery, especially when reasoning from diagrams. These findings align with existing literature highlighting AI's limitations in discipline-specific support. Future research should refine AI responses to better align with engineering problem-solving approaches and explore hybrid models integrating AI assistance with human instruction, potentially leading to more effective AI-augmented learning platforms in mechanical engineering education.

1. Introduction

Generative AI tools are becoming increasingly prevalent in college assessment. Students use AI tools for studying and test preparation, and instructors use AI tools for writing and grading assessments. With the development of Large Language Models (LLMs) to predict words from input text, interactive writing and assessment based on natural human language has become a promising new field of study in engineering education. Generative AI tools have the potential to support students in learning difficult material in STEM disciplines, as well as to help teachers assess learning of engineering concepts through their capability to process large amounts of text, respond to prompts, and engage in conversations. Such AI-driven educational tools have the potential to enhance student learning outcomes by offering personalized feedback, reducing cognitive load associated with technical problem-solving, fostering a more interactive learning environment, boosting student engagement, and supporting self-directed learning. AI-based educational tools have the potential to help students learn complex mechanical engineering concepts by answering questions about assessments and explaining mechanical engineering concepts.

Concept inventories assess students' ability to apply scientific principles to real-world problems. They are challenging because students must not only correctly identify the concept but also recognize its correct interpretation in a given context [1], [2]. Mechanical engineering

concept inventories are also based on conceptual diagrams, which students must interpret before responding, and have been shown to be key to evaluating student understanding of core mechanical engineering concepts. This multi-step problem-solving method may be challenging for LLM-based tools, as it requires them to first recognize a diagram and then connect visual input with mechanical engineering conceptual information.

The ability of emerging LLM-based tools to correctly reason with engineering problems, correctly apply concepts to real-world scenarios, and provide adequate justification to show the decision processes behind their answers is an ongoing topic in engineering education. Although AI tools have the potential to support teaching and learning for complex engineering concepts, they must be thoroughly evaluated on their capacity to correctly interpret and respond to fundamental engineering assessments. To address this need, our study aims to evaluate ChatGPT-4o's performance in image processing, understanding, explaining, and applying key mechanical engineering principles. This study is guided by the following research question: "To what extent can we assess the AI's ability in and deep understanding of mechanical engineering topics as measured by concept inventories?"

2. Background / Literature Review

2.1 Large Language Models in Engineering Education

The integration of large language models (LLMs) into engineering education has emerged as a transformative approach. A recent systematic review of 370 studies to identify trends and opportunities in LLM applications, focusing on 20 high-quality papers, highlighted key areas, including knowledge acquisition and skill development, where LLMs have had the most significant impact. Their findings provide actionable recommendations for effectively embedding LLMs in engineering curricula while ensuring rigorous evaluation to meet educational objectives [3]. Another work emphasized ChatGPT's role in enhancing creative idea generation and knowledge extension in mechanical engineering, underscoring its value in fostering innovation [4].

The potential of LLMs to redefine teaching practices was demonstrated through the assessment of an LLM-based chatbot in a graduate fluid mechanics course. The study identified significant advantages, including self-paced learning and instantaneous feedback, supported by intelligent prompting and integrations like Wolfram Alpha [5]. Undergraduate perspectives on LLM-based tools were explored, revealing diverse perceptions regarding their benefits and challenges. These findings contribute to discussions on balancing AI assistance with ethical considerations and human engagement [6].

Additional insights into the evolving role of generative AI tools, such as ChatGPT, in education, draw parallels between the adoption of generative AI and historical technological disruptions, emphasizing the need for responsible integration to address ethical and pedagogical challenges [7]. Complementing this discussion, another study outlined trends in engineering

education research, providing context for the integration of digital technologies like LLMs [8]. Finally, the contributions of LLMs related to enhancing MATLAB programming and cross-disciplinary knowledge acquisition were explored, highlighting their ability to improve student engagement and mastery of complex concepts [9].

2.2 The Role of AI Chatbots in Engineering Learning Environments

AI chatbots are becoming central to adaptive and personalized learning strategies within engineering education. Research on AI chatbots has focused on areas such as activating prior knowledge, fostering motivation, and supporting self-directed learning. A thematic analysis of student interactions with chatbots in mechatronics and electronic engineering courses revealed both their potential benefits and limitations [10]. Similarly, another extensive overview of LLM advancements emphasized adaptive learning while addressing concerns about accuracy and inclusivity [11]. Generative AI's potential to assist students in STEM problem-solving must be weighed against the challenges of multi-step reasoning and uncommon question formats [12]. Insights emphasized how LLMs could improve essay quality while highlighting the inadequacies of current AI detection systems [13]. These studies, together with analyses exploring the use of GPT-4 to analyze engineering faculty's mental models of assessment underscores the growing need for ethical implementation frameworks and robust AI evaluation mechanisms [14].

2.3 Large Language Models in Mechanical Engineering

The application of LLMs in mechanical engineering has opened new avenues for exploring conceptual understanding and professional practices. One pioneering study assessed LLMs' capabilities in addressing mechanics-focused conceptual questions across topics such as Fluid Mechanics and Mechanics of Materials. GPT-4 exhibited superior performance over GPT-3.5 and other models, particularly when prompts were designed to include explanatory contexts, underscoring the importance of prompt engineering [15]. Another investigation explored the use of generative AI tools like Bard and ChatGPT for mechanical engineering tasks, identifying opportunities and pitfalls in areas like computation and image generation [16].

Several new LLM frameworks have been proposed for instructional use in mechanical engineering. One innovative framework, MechGPT, was demonstrated to have potential using specialized LLMs for connecting disparate knowledge domains in mechanics and materials modeling. The study highlighted the use of ontological knowledge graphs for hypothesis generation and knowledge retrieval, providing visual and structural insights for research and pedagogy [17]. Similarly, the use of natural language processing tools in mechanical engineering education raised questions about academic integrity while also revealing potential for enhancing learning experiences [18]. Finally, MeLM, a multimodal mechanics language model, showcased its ability to address complex design and analysis problems, further expanding the scope of LLM applications in mechanics education [19]. These new frameworks can potentially support students' mechanical engineering education learning, but first must demonstrate their capacity to handle complex material.

2.4 LLMs for Assessment Use

The intersection of LLMs and assessment practices in engineering education has sparked significant discussions about how artificial intelligence can support student learning. ChatGPT's ability to pass exams highlighted its potential for reshaping traditional assessment practices. However, challenges like inconsistencies and confidently incorrect responses underscore the necessity for expert oversight [20]. A study focusing on ChatGPT's performance on mechanical engineering exams revealed that while GPT-4 significantly outperformed GPT-3.5, both models require improvements to handle text-only input limitations [21]. In the context of the Fundamentals of Engineering Exam, researchers showed that noninvasive prompt modifications enhanced ChatGPT's mathematical capabilities, offering insights into how AI models can be adapted for professional and educational use [22]. Collectively, these studies emphasize the need for developing AI-resistant assessment methods while exploring the integration of LLMs as supplementary tools for improving student outcomes [23].

2.5 ChatGPT Domain Knowledge and Image Capability

ChatGPT's ability to adapt and reframe problems across academic disciplines has significant implications for interdisciplinary education. For example, a study demonstrated that ChatGPT could reframe probability and statistics problems to make them accessible and engaging for students in fields as diverse as biology, economics, and law. The findings revealed that in over 73% of cases, reframed problems were deemed to add educational value, highlighting ChatGPT's potential to foster interdisciplinary understanding [24].

Another review identified emerging themes and knowledge management challenges associated with ChatGPT, emphasizing gaps in user satisfaction when applied to complex educational scenarios. While the model excels in generating know-what and know-how knowledge, it struggles with deeper conceptual understanding, emphasizing the need for refining AI applications in pedagogy [25]. Similarly, ChatReview, a ChatGPT-enabled NLP framework, demonstrated the potential of AI in studying domain-specific user reviews. The framework's ability to generate sentiment-based insights across education, hospitality, and local businesses underscores its practical value while raising questions about data privacy [26].

In the domain of image analysis, GPT-4 demonstrated remarkable capabilities in processing visual data, such as identifying patterns in flowcharts and plots with high accuracy. However, limitations in recognizing individual privacy-sensitive images highlight the need for further refinements [27]. Finally, IQAGPT integrated vision-language models with ChatGPT to assess image quality, particularly in medical imaging. This system outperformed existing models, offering a promising approach to objective evaluation and radiological reporting [28].

While generative AI and LLMs hold great potential for engineering education, their capabilities vary significantly, as does the level of guidance required to achieve acceptable performance. Few assessment frameworks exist to guide the use of LLMs in mechanical engineering, and no systematic study has evaluated their accuracy in answering conceptual

questions relevant to ME students. Concept inventories are widely recognized as a measure of students' ability to understand and apply core ME concepts. Thus, our study assesses ChatGPT's performance on a representative set of mechanical engineering conceptual knowledge, supplemented by external evaluation from ME instructors. Such an analysis was previously unfeasible, as LLMs with image recognition capabilities only became widely available with the release of GPT-4o in May 2024.

3. Methods

3.1 Instrument

ChatGPT - GPT-4o, introduced by OpenAI in May 2024, is a groundbreaking multilingual and multimodal generative large language model. Notably, it is the first LLM capable of both processing and generating images, in addition to excelling at text and audio tasks. This innovative capability shows the potential of LLMs to interpret and generate complex mechanical figures.

Our study used five assessments of mechanical engineering conceptual knowledge as data collection instruments. The Force Concept Inventory [29], [30] was developed and deployed to assess students' knowledge of kinematics, Newton's laws of motion, the principle of superposition, and kinds of force. Five questions in the Force Concept Inventory require students to draw a diagram for mechanical engineering analysis. The Materials Concept Inventory [31, 32] assesses knowledge of the strength of materials concepts, including stress and buckling. The Mechanics Baseline Test [33, 34] measures the application of force, acceleration, speed, friction, and velocity concepts. The Mechanics of Materials concept inventory [35] includes problems with predicting failure, predicting deformation, predicting the location of failure, and material properties. Together, these concept inventories represent essential subjects that undergraduate mechanical engineering students must learn. Our study also included a 10-item undergraduate mechanical engineering classroom assessment developed by the instructor to assess conceptual knowledge using free-body diagrams.

A key aspect of the concept inventories is that they are meant to be reused by the instructors in different institutions and different classes (so that scores can be compared and validated inferences can be drawn [32], [36]. **For this reason, the questions cannot be revealed in the open literature.** Therefore, we are not displaying the questions (only summarizing the results). In addition, our testing was conducted with GPT-4o, which is not used by OpenAI or other LLMs to train their models. Interested readers can contact the relevant authors of the inventories to obtain access to them.

3.2 Data Collection

ChatGPT-4o's responses were given in response to the AI's image processing and interpretation of mechanical engineering conceptual diagrams. For each concept inventory and the mechanical engineering classroom assessment, we gave each question to ChatGPT, along

with screenshots of images, and asked it to provide the correct answer along with an explanation of why it was correct. Two professors, Dr. A (Prof. Arun Srinivasa) and Dr. B (Prof. Kristi Shryock), independently graded the AI for correctness and for the quality of its responses. These grades produced four types of scores: Correct Answer with Correct Reasoning; Correct Answer with Incorrect Reasoning; Incorrect Answer with Correct Reasoning; and Incorrect Answer with Incorrect Reasoning.

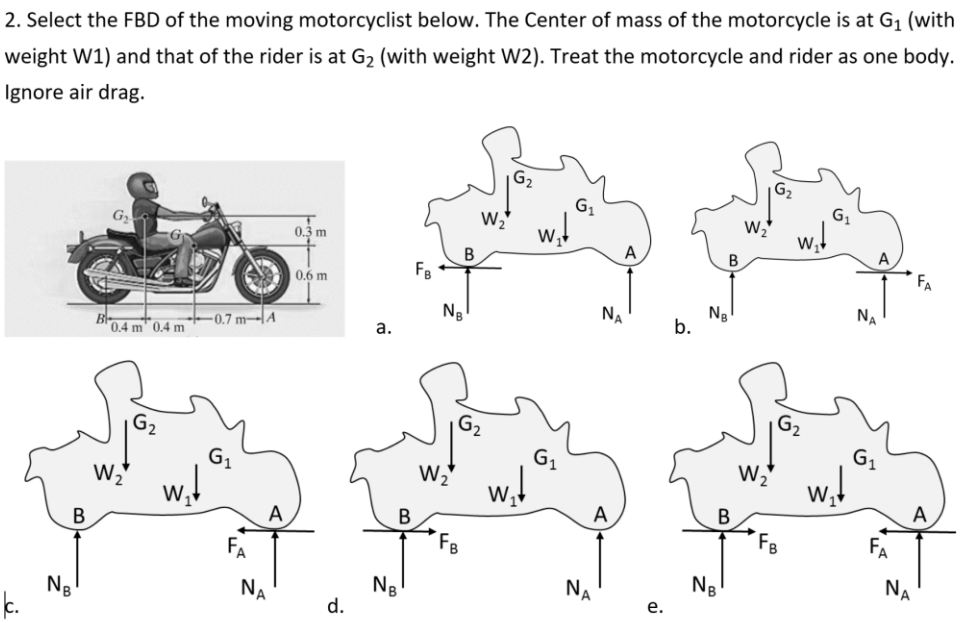

Figure 1. An example of a MEEN 225 mechanics problem (similar to an image-based Force Concept Inventory question)

4. Results

Questions are categorized into 5 datasets. Each dataset is marked with a number. The corresponding list is as follows:

1. Force Concept Inventory,
2. Materials Concepts Inventory,
3. Mechanics Baseline Test,
4. Mechanics of Materials Concept Inventory,
5. MEEN 225-501 Fall 2016 Test 01.

Additionally, questions in dataset 5 are divided into 2 sub-datasets:

5 (a). Multiple choice questions,

5 (b). Free-body diagram (FBD) drawing questions.

To better analyze the dataset, we divided dataset 5 into 5(a) and 5(b). The first five questions resemble standard multiple-choice questions, while the last five require GPT-4o to generate Free Body Diagrams (FBDs) as images. Within these datasets, some questions include reference diagrams, while others do not. Accordingly, we further categorized them as image-

based or non-image-based. Figure 2 illustrates the distribution of question types across the datasets.

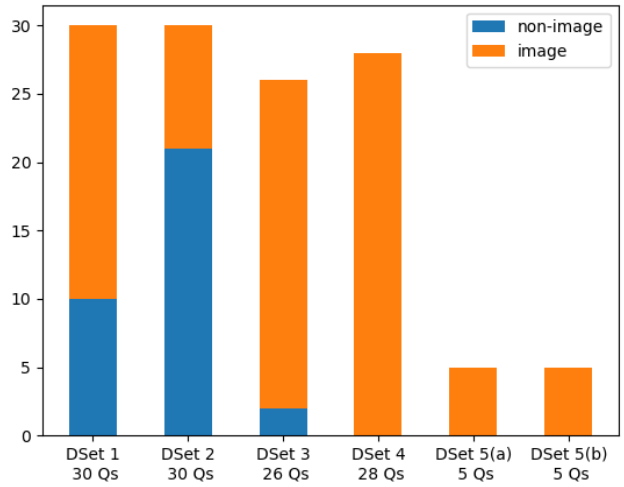

Figure 2 Distribution of image-based & non-image-based questions

GPT-4o generated responses that included both answer choices and reasoning. Two professors, Dr. A and Dr. B, evaluated its responses based on their expertise. They assigned scores as follows:

- **1:** Correct choice and reasoning.
- **2**: Incorrect choice but correct reasoning.
- **3**: Correct choice but incorrect reasoning.
- **0**: Incorrect choice and reasoning.

Since Dr. B provided detailed explanations that aligned with standard answers, Dr. B's evaluations were used as the baseline for comparison.

To measure GPT-4o's performance, we employed three statistical metrics:

- Accuracy – Percentage of correct choices.
- Cohen's Kappa – Measures agreement between evaluators.
- Agreement Percentage – Measures consistency between Dr. A and Dr. B, as well as between GPT-4o and the baseline (Dr. B).

4.1 Accuracy

$$Accuracy = \frac{the\ number\ of\ correct\ choices}{the\ total\ number\ of\ questions\ in\ the\ dataset}$$

Table 1 shows the accuracy of GPT-4o answers compared with the baseline (Dr. B).

Table 1. Accuracy of GPT-4o answers in mechanical engineering

Dataset	Total Accuracy	Image-based Accuracy	Non-image-based Accuracy

1	0.67 (20/30)	0.5 (10/20)	1 (10/10)
2	0.7 (21/30)	0.67 (6/9)	0.71 (15/21)
3	0.38 (10/26)	0.33 (8/24)	1 (2/2)
4	0.39 (11/28)	0.39 (11/28)	N/A
5 (a)	0.6 (3/5)	0.6 (3/5)	N/A
5 (b)	0 (0/5)	0 (0/5)	N/A

Total accuracy is calculated by dividing the number of correct GPT-4o answers by the total number of questions in the dataset. It reflects the overall accuracy of GPT-4o's answers across different types of questions. Image-based accuracy only considers the percentage of correct answers in image-based questions. Similarly, non-image-based accuracy measures GPT-4o's performance on questions that do not require images. The number in parentheses following the percentage, such as in dataset 1 (image-based accuracy 0.5, 10/20), indicates that, within dataset 1, there are 20 image-based questions, and GPT-4o answered 10 of them correctly according to Dr. B's evaluations.

Because dataset 3 contains only two non-image-based questions out of the total 26 ones, its accuracy percentage is highly skewed and not representative. Therefore, focusing on Dataset 1 and Dataset 2, their non-image-based accuracy is higher than their respective image-based accuracy. Overall, across all 86 image-based questions (excluding FBD questions), the accuracy is 0.44 (38/86), while for all 33 non-image-based questions, the accuracy is 0.82 (27/33). This result suggests that GPT-4o's accuracy is higher for non-image-based questions than for image-based ones.

Observing image-based questions only, accuracies are relatively higher in dataset 1, dataset 2, and dataset 4 and lower in dataset 3 and dataset 5 (consider both (a) & (b)). Therein, the accuracy of GPT-4o's answers to image-based questions in dataset 5 is the worst. Thus, in this experiment, GPT-4o performs in accuracy relatively better in Force Concept Inventory, Materials Concepts Inventory, and Mechanics of Materials Concept Inventory, compared to Mechanics Baseline Test and MEEN 225-501 Fall 2016 Test 01. This result shows that GPT-4o's accuracy is relatively higher in image-based conceptual questions of mechanical engineering than in test questions. This is because compared to conceptual questions, test questions are revealed less on the Internet and GPT-4o's training corpus contains less similar questions.

Dataset 5(b) is atypical, as questions 6 to 10 require students to draw schematic diagrams (FBD) as a key part of their answers. However, instead of simple 2D schematic diagrams, GPT-4o generates complex, detailed, and vivid 3D images for each question. This reveals some of GPT-4o's limitations in generating images for mechanical engineering. From a purely image-

generation perspective, the output appears visually rich. However, in mechanical engineering—and engineering disciplines in general—beneficial diagrams prioritize clarity and simplicity to enhance key concepts and facilitate problem analysis. Therefore, 2-D representations are typically more beneficial than 3-D ones. Yet, GPT-4o cannot generate high-quality, professionally relevant mechanical engineering diagrams. These challenges emphasize GPT-4o's current limitations in generating professional technical illustrations. Consequently, it is not yet suitable for applications in instructional settings or as a reliable tool for generating professional mechanical engineering diagrams.

4.2 Inter Rater Reliability

To evaluate the consistency of ratings between two different professors and between Dr. B (the baseline) and GPT-4o, we adopted Cohen's Kappa as the metric [36]. Cohen's Kappa measures the inter-rater reliability of two or more raters while also accounting for the possibility of chance agreement. The formula for Cohen's Kappa is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o represents the observed agreement as the proportion of items on which both raters agree. Observed agreement is calculated as:

$$P_o = \frac{the\ number\ of\ agreements}{total\ number\ of\ questions\ in\ the\ dataset}$$

P_e represents the expected agreement by chance, which is the probability that the two raters would agree by chance alone. Expected agreement is calculated as:

$$P_e = \sum(Pr_1(i) \times Pr_2(i))$$

$Pr_1(i)$ is the proportion of Rater 1 choosing the i-th category, and $Pr_2(i)$ is the proportion of Rater 2 choosing the same category, summed over all possible categories i. Table 2 shows the ranges of possible values for Cohen's Kappa and the strength of inter-rater agreement they represent. Values from 0 - 0.4 indicate low agreement, and values from 0.41 - 1 indicate high agreement.

Table 2. Interpretation of Cohen's Kappa Values [37]

Kappa Statistic	Strength of Agreement
< 0.00	Poor (systematic disagreement)
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate

0.60-0.80	Substantial
0.80-1.00	Almost Perfect

In our analysis, we assessed the consistency of ratings both between the two professors and between Dr. B (baseline) and GPT-4o, as shown in Table 3. Given that GPT-4o assumes all of its responses are correct, Cohen's Kappa relies entirely on Dr. B's evaluation. As a result, the Kappa value would be 0 for all datasets, which makes the comparison meaningless. Therefore, we used the agreement percentage as a more meaningful metric when comparing Dr. B's evaluations with GPT-4o's performance.

Table 3. Cohen's Kappa and Agreement Percentage for Two Professors and Dr. B vs. GPT-4o

Dataset	2 professors (κ)	Dr. B vs. GPT-4o (κ)	2 professors (Agreement)	Dr. B vs. GPT-4o (Agreement)
1	0.94 (Almost Perfect)	0	**0.967** (29/30)	0.67 (20/30)
2	1 (Perfect)	0	**1** (29/29)	0.7 (21/30)
3	0.79 (Substantial)	0	**0.88** (22/25)	0.38 (10/26)
4	1 (Perfect)	0	**1** (28/28)	0.39 (11/28)
5 (a)	1 (Perfect)	0	**1** (5/5)	0.6 (3/5)
5 (b)	1 (Perfect)	0	**1** (5/5)	0 (0/5)

The high Cohen's Kappa values between the two professors indicate strong consistency, demonstrating their high reliability on the correct answers and problem explanations. For the datasets 1, 2, 4, and 5, this consistency remains robust whether Dr. A or Dr. B is considered the baseline. This is also reflected in the agreement percentages. For dataset 3, both professors show a substantially slightly lower level of consistency (88% agreement), but we place more trust in Dr. B's expertise based on the reasons discussed earlier.

When comparing GPT-4o to Dr. B, higher agreement is observed in datasets 1, 2, and 5(a). However, dataset 5(a) should be interpreted with caution due to its small sample size. Datasets 3 and 4 exhibit lower agreement, and dataset 5(b) reveals significant discrepancies. Considering Dr. B's expertise in mechanical engineering, his evaluations are regarded as more reliable. Therefore, GPT-4o's performance is considered less reliable in datasets 3, 4, and 5, which involve the Mechanics Baseline Test, Mechanics of Materials Concept Inventory, and

MEEN 225-501 Fall 2016 Test 01. Overall, GPT-4o's performance in solving professional mechanical engineering problems is modest, based on the results of this assessment experiment.

4.3 Examples of FBD generation

According to the above results, GPT-4o has less understanding and reasoning ability in mechanical engineering schematic diagrams than word descriptions. The possible reason is training corpus has more resources for word reasoning and less for image reasoning. Especially professional mechanical engineering background diagrams. Fine-tuning may improve this limitation.

GPT-4o follows a fixed, general approach to image generation and lacks the adaptability required for mechanical engineering applications. For example, in dataset 5(b), Question 7 asks to "Draw the FBD for the T-bar AB." The question with the provided reference image is shown in Figure 17(a). A comparison between the image and description generated by GPT-4o (Figure 17(b)) and the ideal solution (Figure 17(c)) highlights its limitations in producing precise, engineering-specific diagrams.

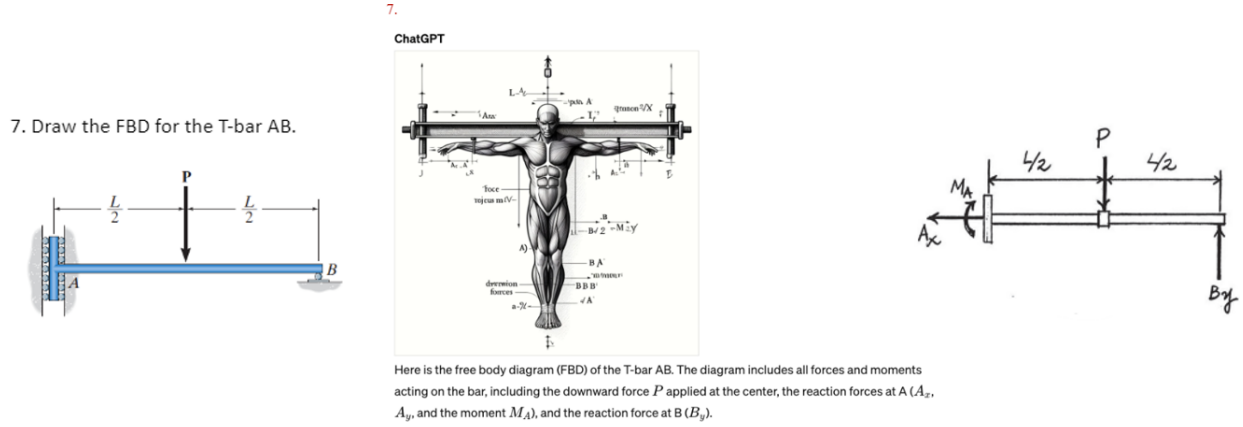

Figure 5 (a) Question 7 in dataset 5(b). (b) GPT-4o answer and description. (c) ideal answer.

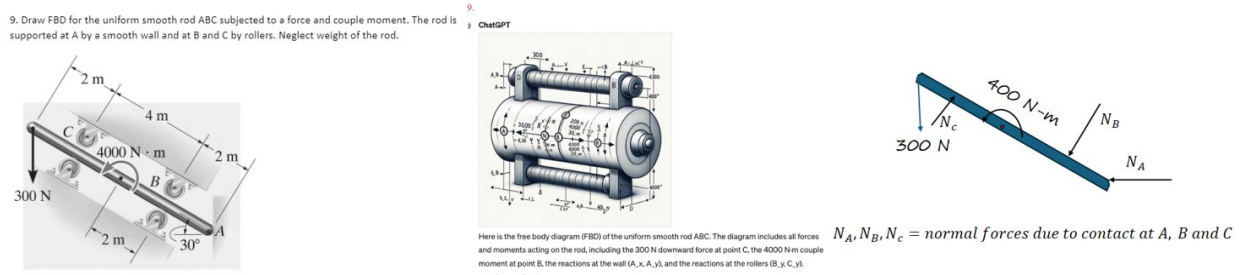

Figure 6 (a) Question 9 in dataset 5(b). (b) GPT-4o answer and description. (c) ideal answer.

From a general image perspective, the generated illustration appears suitable, as it includes key elements such as "free body," "T-bar," "A, B," and "force." The image is surprisingly rendered in 3-D, incorporating intricate details and key features. However, simply assembling these components without proper consideration results in an absurd image from a mechanical engineering standpoint—particularly evident in the inclusion of a human figure in

Figure 18 (c). A proper FBD should be simple and clear, enabling students to extract essential components from the given scenario and accurately represent external forces and constraints.

Under these circumstances, a 2-D schematic diagram would likely be far more useful. This outcome reflects a rigid, generalized approach to image generation that does not align well with the needs of most engineering problems when applied directly. Figure 18 presents the image generated for Question 9 in Dataset 5(b). Although Figure 18(b) seems better than the figure for question 7, GPT-4o model still fails to idealize the diagram and does not effectively highlight the key forces. Specifically, in the generated image in Figure 18(b), the question 7 stem and image clearly require that the answer is supposed to focus on the ABC rod; however, the generated image depicts a complicated device structure rather than one single rod. It marks many unnecessary or irrelevant forces everywhere across the device structure. However, from Figure 18(c), only NA, NB, NC, 300 N force, and 4000 N-m should be marked.

4.4 Error analysis

Appendix A contains the question numbers for one of the concept inventories, and the professor's comments on the accuracy of GPT-4o. As can be seen from that analysis, GPT-4o has some wrong answers and in some cases wrong explanations.

5. Discussion

A close examination of the results of the concept inventories (including the reasoning presented by GPT-4o) indicates that in the majority of cases, GPT-4o is able to provide reliable reasoning and correct choice for the answers, However, there are several cases involving reasoning from graphs and images where the reasoning provided by GPT-4o is correct but nevertheless the wrong answer is chosen. This kind of error can be seen consistently in many similar cases, and in particular in reasoning for materials related questions. While the sample size is too small to draw general conclusions, it is plausible that selecting answers to multiple choice questions by converting a general reasoning to a specific answer which is not probabilistic is challenging for LLMs, since they are based on probabilistic sentence completion strategies. LLM performance on educational assessment improves when questions are well-structured, with assessment of open-ended questions proving difficult for AI systems [38]. A similar behavior was also noted in the study by Abedi et. al [5] in their graduate fluid mechanics class. They state that (p. 23) "There are situations where GPT-4o makes accurate assumptions and applies correct mathematical equations but falters during the execution of mathematical operations, resulting in erroneous answers." They went on to observe that "GPT-4o struggled with problems that needed information from tables or images" (p. 23).

A key aspect of LLM performance was their complete failure on the system isolation free body diagram questions, which were included in the Mechanics Baseline Test. In this case, both the question and answer are pictorial, and interpreting both appears to be beyond the capabilities of current LLMs.

Implications for AI-based tutoring:

The results show that LLMs have high reliability when the questions are primarily textural. In all cases, if such LLMs are to be used for tutoring or answering student conceptual questions, there needs to be a hybrid approach with a TA or instructor in the loop to tag anomalous behavior. A typical case where a multiple choice conceptual quiz is generated and a Retrieval-Augmented Generation (RAG) based agent, which has been provided the answers, can be used to verify and correct student reasoning, combined with a Reinforcement Learning with Human Feedback (RLHF) strategy. It is also possible to leverage the generative capabilities of LLMs to produce questions that are "similar to" the concept inventory questions and use this to help deepen student insights.

Implications for pedagogy:

Given these considerations, a two-pronged strategy can be implemented to channel the potential of AI in education, while addressing ethical issues related to cheating and undermining the learning process. Converting real-world situations into idealized images and then reasoning with them is a critical part of the mechanical engineering curriculum. Given that LLMs are currently unreliable for these tasks, instructors can pose more of these questions, having confidence that the students may not be able to use an LLM to answer them. Even if an LLM is able to carry this out in the future, describing the real-world situation with sufficient accuracy to the system is a time-consuming task (and requires considerable skill) that it is unlikely to be used for cheating.

On the other hand, as previously mentioned, LLMs can serve as valuable assistance to instructors, much like nurse practitioners assist doctors or surgeons, by helping craft assessments and evaluate specific components. This support allows instructors to focus on more creative and engaging ways to interact with the class, such as leveraging LLMs for tasks like short-answer grading [39, 40]. For instance, a typical task such as finding the forces in a structure involves all aspects of reasoning: (1) identifying and simplifying the real-world system and converting it into an idealized pictorial representation (truss or frame or other idealized body), (2) estimating loads, (3) converting the graphical to symbolic representations, (4) solving the resultant systems and then, (5) inferring suitable information about the real world from the idealized solutions. Currently due to difficulties in evaluating each of these steps in large classes, most of these steps are not practiced or evaluated independently with only the visible end result being assessed. LLMs offer the chance to help guide the students to develop the reasoning behind the steps, without overwhelming them or the instructors.

6. Limitations

Our study only tested the capability of ChatGPT-4o, as the newest model released by OpenAI. However, other chatbots using different LLMs released by companies such as Gemini or Anthropic have similar capabilities and may excel ChatGPT-4o in answering concept inventory questions. These models may vary in their mechanical engineering knowledge base as

well as their image processing capabilities. For future work, we plan to test other LLMs using the same approach and compare their abilities. Even as models and tools continue to be developed, a quantitative comparison will allow us to develop benchmarks for guiding future tool selection.

7. Future Work

Our study has three areas for future work we wish to explore. First, we plan to expand this evaluation approach to mechanical engineering assessments in other subject areas. We are also interested in testing LLM capabilities at different levels of undergraduate and graduate knowledge. In this way, we can continue to determine the level of support that AI tools can provide on advanced engineering concepts. Second, we will continue collecting scores and feedback from additional professors, including teaching assistants, to further judge the reliability of our evaluation method. This will give a more detailed look at how different instructors perceive the LLM output, and how they view its reasoning capabilities. Finally, we plan to evaluate LLM tools' mechanical engineering conceptual performance on classroom assessments in other domains. While concept inventories are standardized, informal classroom assessments are a wealth of information about student learning for AI-based feedback.

References

[1] V. Fakiyesi, D. Fabiyi, I. Dunmoye, O. Olaogun, and N. Hunsu, "A Scoping Review of Concept Inventories in Engineering Education," 2024 ASEE Annual Conference & Exposition Proceedings, doi: 10.18260/1-2--46487.

[2] Evans, D.L., Gray, G.L., Krause, S., Martin, J., Midkiff, C., Notaros, B.M., Pavelich, M., Rancour, D., Reed-Rhoads, T., Steif, P. and Streveler, R., 2003, November. Progress on concept inventory assessment tools. In *33rd Annual Frontiers in Education, 2003. FIE 2003.* (Vol. 1, pp. T4G-1). IEEE.

[3] S. Filippi and B. Motyl, "Large Language Models (LLMs) in Engineering Education: A Systematic Review and Suggestions for Practical Adoption," Information, vol. 15, no. 6, p. 345, Jun. 2024, doi: 10.3390/info15060345.

[4] S. Filippi and B. Motyl, "Possible Applications of Large Language Models (LLMs) in Engineering Education: An Overview," 2024, doi: 10.54941/ahfe1005390.

[5] M. Abedi, I. Alshybani, M. Shahadat, and M. Murillo, "Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education," Sep. 2023, doi: 10.32388/md04b0

[6] I. Joshi, R. Budhiraja, P. D. Tanna, L. Jain, M. Deshpande, A. Srivastava, S. Rallapalli, H. D. Akolekar, J. S. Challa, and D. Kumar, "'With Great Power Comes Great Responsibility!': Student and Instructor Perspectives on the Influence of LLMs on Undergraduate Engineering Education," 2023.

[7] A. Alsharif, D. Knight, and A. Katz, "The Evolution of Technology in Education and the Emerging Role of Generative AI," 2024.

[8] A. Katz, J. B. Main, A. Struck Jannini, and D. Knight, “Special report: The research topics addressed and research methods applied in the Journal of Engineering Education (1993–2022),” Journal of Engineering Education, vol. 112, no. 4, pp. 852–860, Oct. 2023, doi: 10.1002/jee.20559.

[9] C. Liu and S. Yang, “Application of large language models in engineering education: A case study of system modeling and simulation courses,” International Journal of Mechanical Engineering Education, Aug. 2024, doi: 10.1177/03064190241272728.

[10] F. A. Bravo and J. M. Cruz-Bohorquez, “Engineering Education in the Age of AI: Analysis of the Impact of Chatbots on Learning in Engineering,” Education Sciences, vol. 14, no. 5, p. 484, May 2024, doi: 10.3390/educsci14050484.

[11] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, "Large Language Models for Education: A Survey and Outlook," 2024.

[12] B. Borges et al., “Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants,” Proceedings of the National Academy of Sciences, vol. 121, no. 49, Nov. 2024, doi: 10.1073/pnas.2414955121.

[13] M. Bernabei, S. Colabianchi, A. Falegnami, and F. Costantino, “Students’ use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances,” Computers and Education: Artificial Intelligence, vol. 5, p. 100172, 2023, doi: 10.1016/j.caeai.2023.100172.

[14] A. Ross, A. Katz, K. J. Chew, and H. Matusovich, “Stumbling Our Way Through Finding a Better Prompt: Using GPT-4 to Analyze Engineering Faculty’s Mental Models of Assessment,” 2024 ASEE Annual Conference & Exposition Proceedings, doi: 10.18260/1-2--48032.

[15] J. Tian, J. Hou, Z. Wu, P. Shu, Z. Liu, Y. Xiang, B. Gu, N. Filla, Y. Li, N. Liu, X. Chen, K. Tang, T. Liu, and X. Wang, "Assessing Large Language Models in Mechanical Engineering Education: A Study on Mechanics-Focused Conceptual Understanding," Jan 2024, doi: 10.31219/osf.io/d3nc6.

[16] K. B. Mustapha, E. H. Yap, and Y. Abakr, “Bard, ChatGPT and 3DGPT: A Scientometric Analysis of Generative AI Tools and Assessment of Implications for Mechanical Engineering Education,” Feb. 2024, doi: 10.36227/techrxiv.170792405.51299882/v1.

[17] M. J. Buehler, “MechGPT, a Language-Based Strategy for Mechanics and Materials Modeling That Connects Knowledge Across Scales, Disciplines, and Modalities,” Applied Mechanics Reviews, vol. 76, no. 2, Jan. 2024, doi: 10.1115/1.4063843.

[18] J. Lesage et al., “Exploring natural language processing in mechanical engineering education: Implications for academic integrity,” International Journal of Mechanical Engineering Education, vol. 52, no. 1, pp. 88–105, Mar. 2023, doi: 10.1177/03064190231166665.

[19] M. J. Buehler, “MeLM, a generative pretrained language modeling framework that solves forward and inverse mechanics problems,” Journal of the Mechanics and Physics of Solids, vol. 181, p. 105454, Dec. 2023, doi: 10.1016/j.jmps.2023.105454.

[20] S. Nikolic et al., “ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to

investigate assessment integrity," European Journal of Engineering Education, vol. 48, no. 4, pp. 559–614, May 2023, doi: 10.1080/03043797.2023.2213169.

[21] M. E. Frenkel and H. Emara, "ChatGPT-3.5 and -4.0 and mechanical engineering: Examining performance on the FE mechanical engineering and undergraduate exams," Computer Applications in Engineering Education, vol. 32, no. 6, Jul. 2024, doi: 10.1002/cae.22781.

[22] V. Pursnani, Y. Sermet, M. Kurt, and I. Demir, "Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice," Computers and Education: Artificial Intelligence, vol. 5, p. 100183, 2023, doi: 10.1016/j.caeai.2023.100183.

[23] M. E. Frenkel and H. Emara, "ChatGPT-3.5 and -4.0 and mechanical engineering: Examining performance on the FE mechanical engineering and undergraduate exams," Computer Applications in Engineering Education, vol. 32, no. 6, Jul. 2024, doi: 10.1002/cae.22781.

[24] H. Einarsson, S. H. Lund, and A. H. Jónsdóttir, "Application of ChatGPT for automated problem reframing across academic domains," Computers and Education: Artificial Intelligence, vol. 6, p. 100194, Jun. 2024, doi: 10.1016/j.caeai.2023.100194.

[25] O. Ali, P. A. Murray, M. Momin, and F. S. Al-Anzi, "The knowledge and innovation challenges of ChatGPT: A scoping review," Technology in Society, vol. 75, p. 102402, Nov. 2023, doi: 10.1016/j.techsoc.2023.102402.

[26] B. Ho, T. Mayberry, K. L. Nguyen, M. Dhulipala, and V. K. Pallipuram, "ChatReview: A ChatGPT-enabled natural language processing framework to study domain-specific user reviews," Machine Learning with Applications, vol. 15, p. 100522, Mar. 2024, doi: 10.1016/j.mlwa.2023.100522.

[27] O. V. Johnson, O. Mohammed Alyasiri, D. Akhtom, and O. E. Johnson, "Image Analysis through the lens of ChatGPT-4," Journal of Applied Artificial Intelligence, vol. 4, no. 2, pp. 31–46, Dec. 2023, doi: 10.48185/jaai.v4i2.870.

[28] Z. Chen et al., "IQAGPT: computed tomography image quality assessment with vision-language and ChatGPT models," Visual Computing for Industry, Biomedicine, and Art, vol. 7, no. 1, Aug. 2024, doi: 10.1186/s42492-024-00171-w.

[29] Hestenes, D., Wells, M. and Swackhamer, G., 1992. Force concept inventory. *The physics teacher*, *30*(3), pp.141-158.

[30] Savinainen, A. and Scott, P., 2002. The Force Concept Inventory: a tool for monitoring student learning. *Physics education*, *37*(1), p.45.

[31] Krause, S., Decker, J.C. and Griffin, R., 2003, November. Using a materials concept inventory to assess conceptual gain in introductory materials engineering courses. In *33rd Annual Frontiers in Education, 2003. FIE 2003.* (Vol. 1, pp. T3D-7). IEEE.

[32] Jordan, W., Cardenas, H. and O'Neal, C.B., 2005. Using a materials concept inventory to assess an introductory materials class: Potential and problems. *age*, *Proceedings of the 2005 American Society for Engineering Education Annual Conference and Exposition American Society for Engineering Education*

[33] Hestenes, D. and Wells, M., 1992. A mechanics baseline test. *The physics teacher*, *30*(3), pp.159-166.

[34] Cardamone, C.N., Abbott, J.E., Rayyan, S., Seaton, D.T., Pawl, A. and Pritchard, D.E., 2012, February. Item response theory analysis of the mechanics baseline test. In *AIP Conference Proceedings* (Vol. 1413, No. 1, pp. 135-138). American Institute of Physics.

[35] Kuchnicki, S.N. and Ericson, T.M., 2021, July. Development of a New Concept Inventory for Mechanics of Materials. In *2021 ASEE Virtual Annual Conference Content Access*.

[36] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.

[37] J. L. Fleiss, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977.

[38] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. Srinivasa, "Automatic assessment of text-based responses in post-secondary education: A systematic review," Computers and Education: Artificial Intelligence, vol. 6, p. 100206, Jun. 2024, doi: 10.1016/j.caeai.2024.100206.

[39] Gao, R., Thomas, N., & Srinivasa, A. (2023, October). Work in progress: Large language model based automatic grading study. In 2023 IEEE Frontiers in Education Conference (FIE) (pp. 1-4). IEEE.

[40] Gao, R., Guo, X., Li, X., Narayanan, A. B. L., Thomas, N., & Srinivasa, A. R. (2025, January). Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering. In Large Foundation Models for Educational Assessment (pp. 186-206). PMLR.

Appendix A

Mechanics of Materials Concept Inventory (problem number)	Professor's comments
2	Misinterpreted tensile load, generic answer not specific to the situation
3	It needs Moment of Intertia. Same problem, generic answer not specific to the problem
4	Misinterpreted the graph, but otherwise the explanation (given the misinterpretation of figure 4) is correct
5	It is not due to stress concentration
6	Here the Beam is under bending so the Moment of Intertia has to be accounted for.
7	
8	
9	Missed the Bending due to transverse load requires Moment of Intertia
10	
11	
12	Clear misunderstanding of the relation between CS shape and Moment of Intertia
13	Answer is correct but reasoning is completely wrong
14	
15	

16	same error as the tapered bar in 4.
17	There was no moment in the bar.
18	Misinterpreted the graph as a cantilever bam problem
19	same error as problem 18
20	
21	same as 18
22	The explanation was correct but it misunderstood what was the top of the beam
23	partially correct reasoning but misunderstood location of maximum toque
24	same as 18. Persistently mischaracterizes free ends (unable to recognize)
25	The logical reasoning is correct but interpreting points on the graph corresponding to different cases is incorrect
26	Uses incorrect (but popular) definition of stress so it gives the popular answer not the correct one
27	
28	mischaracterized the problem since it was not able to identify the relation between axial and bending loads
29	I think that the explanation is not quite correct because it didn't specifically state that the stress will be proportional to the load since the dimensions are the same. So technically the explanation is Passable but not complete