Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
Paper ID #45741

Exploring the Effective Use of ChatGPT in a Sophomore-Level Dynamics Course

Ryan Carr, U.S. Air Force Academy

Ryan Carr received his PhD from the Air Force Institute of Technology (AFIT) in 2017 focused on optimal control theory in guidance, control, and navigation or aerospace systems. He was an engineer and branch chief at the Air Force Research Laboratory (AFRL), a flight test engineer the Air Force Test Center (AFTC), and a program manager at the Air Force Office of Scientific Research (AFOSR). He joined the faculty at the United States Air Force Academy in 2023 as an Assistant Professor of Mechanical Engineering.

Dr. Phillip Cornwell, U.S. Air Force Academy

Phillip Cornwell is a Professor of Mechanical Engineering at the United States Air Force Academy and an Emeritus Professor of Mechanical Engineering at Rose-Hulman Institute of Technology. He received his Ph.D. from Princeton University in 1989, and his current interests include structural dynamics, structural health monitoring, and undergraduate engineering education. Dr. Cornwell received the SAE Ralph R. Teetor Educational Award in 1992, the Dean's Outstanding Teacher Award at Rose-Hulman in 2000, and the Rose-Hulman Board of Trustees' Outstanding Scholar Award in 2001. He was one of the developers of the Rose-Hulman Sophomore Engineering Curriculum and the Dynamics Concept Inventory. Additionally, he is a co-author of Vector Mechanics for Engineers: Dynamics, by Beer, Johnston, Cornwell, and Self. In 2019, Dr. Cornwell received the Archie Higdon Distinguished Educator Award from the Mechanics Division of ASEE, and in 2023, he was honored with the Ralph Coats Roe Award from the Mechanical Engineering Division of ASEE.

Using Artificial Intelligence Chatbots to Solve Dynamics Problems

Abstract

The rapid progress in the field of Artificial Intelligence and the swift adoption of these tools by students merits a close look at the use of chatbot-style tools to solve engineering problems of all types. This study investigated the performance of three existing chatbot tools, ChatGPT, Claude, and Gemini, in solving Dynamics problems. Specifically, the Dynamics Concept Inventory was used as an assessment of the chatbots' ability to answer conceptual problems in a multiple-choice format when only provided with images of the problem (no text prompts). It was found that while the current version of chatbots performed well in the overall task of understanding the question and providing a high-level problem-solving strategy, they suffered from errors in visualizing spatial relationships and in reasoning about primary principles.

Introduction

A search of the proceedings from the 2024 ASEE Annual Conference & Exposition returns 83 papers and panel discussions with "Artificial Intelligence" or "AI" in the title. The set can be expanded to over 100 by adding terms such as "Machine Learning", "Large Language Models", or "Generative". Results are spread across most ASEE divisions, reflecting the intense interest engineering educators have in using modern AI-based tools in the classroom. Proposed uses of AI are too many to enumerate here, but broad topics include techniques for teaching students how to use AI, recommendations to instructors on using AI tools to assist with curriculum development and assessment, the ethics of AI use in the classroom, and advances in AI for solving engineering problems.

Given the focus on these emerging tools by educators and students alike, it is imprudent to ignore their use in any field of engineering. One recent author urged that "assessment strategies will have to evolve to prevent unethical conduct while still allowing for the productivity that can be achieved with these tools"[1]. The likelihood that students will use AI in coursework is high. Legitimate uses of AI tools by students include help in pseudo-coding, drafting concepts for papers, and explaining engineering concepts. Recent studies have grappled with the impact of students using AI to cheat, showing that using the using the tools inappropriately causes learning to decline significantly [2]. Unsurprisingly, while the means of cheating has changed dramatically, the result of cheating remains the same. Educators need to be conversant in both legitimate and illegitimate uses of ever-changing and increasingly available AI tools.

Modern AI tools, particularly chatbots using Large Language Models (LLM) such as ChatGPT[1], Claude[2], and Gemini[3], are excellent at writing basic code and class papers, but their ability to solve analytical, short answer, or open-ended academic engineering problems is unclear. While early chatbots were simple and seemed unlikely to be useful for solving conceptual or

[1] https://chatgpt.com/
[2] https://claude.ai/
[3] https://gemini.google.com/

computational problems, recent versions such as GPT-4o show greater potential for use in engineering. One recent study used various versions of ChatGPT on the Civil Engineering Fundamentals of Engineering Exam [2], and achieved a score as high as 73%, enough to pass the exam.

The purpose of this paper is to explore the use of modern chatbot-style AI tools to solve engineering problems that are encountered in a Dynamics course. Tools such as ChatGPT, Claude, and Gemini are particularly interesting because of their ease of access and popularity. They are not ideally suited to solve engineering problems, partially because they are trained using a broad range of information and are inherently generalists. Techniques exist to sharpen their focus, such as Retrieval-Augmented Generation [3] or creating custom-trained tools. However, it is unlikely that most undergraduate students will go to these lengths, so the focus of this study will use general unaltered and widely available tools ChatGPT, Claude, and Gemini. Note that Llama[4] (Meta AI) was not used because at the time of writing, images could not be uploaded directly. Additionally, Grok 2[5] did not allow uploading enough images to survey without having a premium account.

The authors drafted this paper specifically to inform mechanics educators about the current state of chatbots. It should be noted that the results are only valid for as long as the chatbots themselves are still relevant. Given the rapid pace of change in this technology, the specific results should be reassessed for the release of each new version.

Experimental Methods

Recently, vision capability has been added to ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Flash. This new capability allows a user to directly insert an image file into the chatbot. This adds a new machine vision dimension to the evaluation of these chatbots[6] for engineering analysis, as many engineering problems are best described with an accompanying diagram or image. For this study, all problems will include a vision component.

Example Problem #1

To begin, a sample problem will be analyzed to illustrate a typical user experience. The problem shown in Figure 1 below was given as a quiz in a recent Dynamics course at _______. The image was pasted directly into ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Flash. In this paper, unless indicated otherwise, all results will be from the above variants.

---

[4] https://www.llama.com/

[5] https://x.ai/blog/grok-2

[6] Note that the term "chatbot" may not be entirely accurate here. The term "reasoners" has been proposed for more advanced versions of these AI tools.

An experiment is performed to determine the coefficient of restitution for a 10 N steel ball and a heavy flat aluminum plate resting on the ground. The ball is dropped from a distance of 5 m above the plate and is measured to have a vertical speed of 4.8 m/s just after impact. Determine the coefficient of restitution.

(a) 0.04
(b) 0.34
(c) 0.92
(d) 0.96
(e) none of these

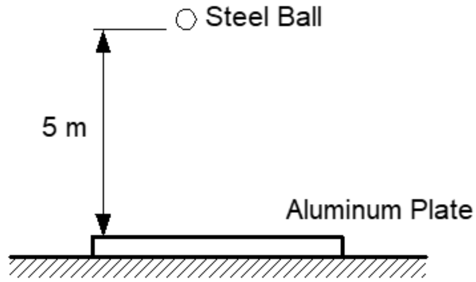

Figure 1. Example problem #1 given to the three chatbots.

After only a few seconds, all three chatbots accurately interpreted the problem in the image and even provided an overview of what it saw by listing given information. All three correctly identified the need to calculate the velocity of the ball just prior to impact, and all three recommended using a simplified but sufficient version of the formula for the coefficient of restitution. Each chatbot calculated the value of the coefficient as 0.48, but only ChatGPT and Claude correctly indicated that the correct answer to the question was (e). Gemini decided to find the closest value instead and recommended (b) as the correct answer.

This simple example illustrates several common features of solutions generated by chatbots. First, the machine vision used by these chatbots is good at reading text. Each of the three deciphered the question without error largely based on reading the question. Second, the chatbots identified overall principles accurately, such as kinematics and the definition of coefficient of restitution. Third, each was then able to connect the identified principles to correct equations. Fourth, each was able to handle the simple arithmetic including multiplication, division, and a square root. Finally, although Gemini understood that the exact solution wasn't among the existing choices, it didn't select the option, "none of these". Instead, it decided to find the closest numerical answer.

Example Problem #2
The previous example was intentionally easy for the chatbots. In more complicated problems, the chatbots often fail in one or more of the above areas. Another example is provided to illustrate this point in Figure 2. Once again, the image was pasted directly into the three chatbots without a prompt.

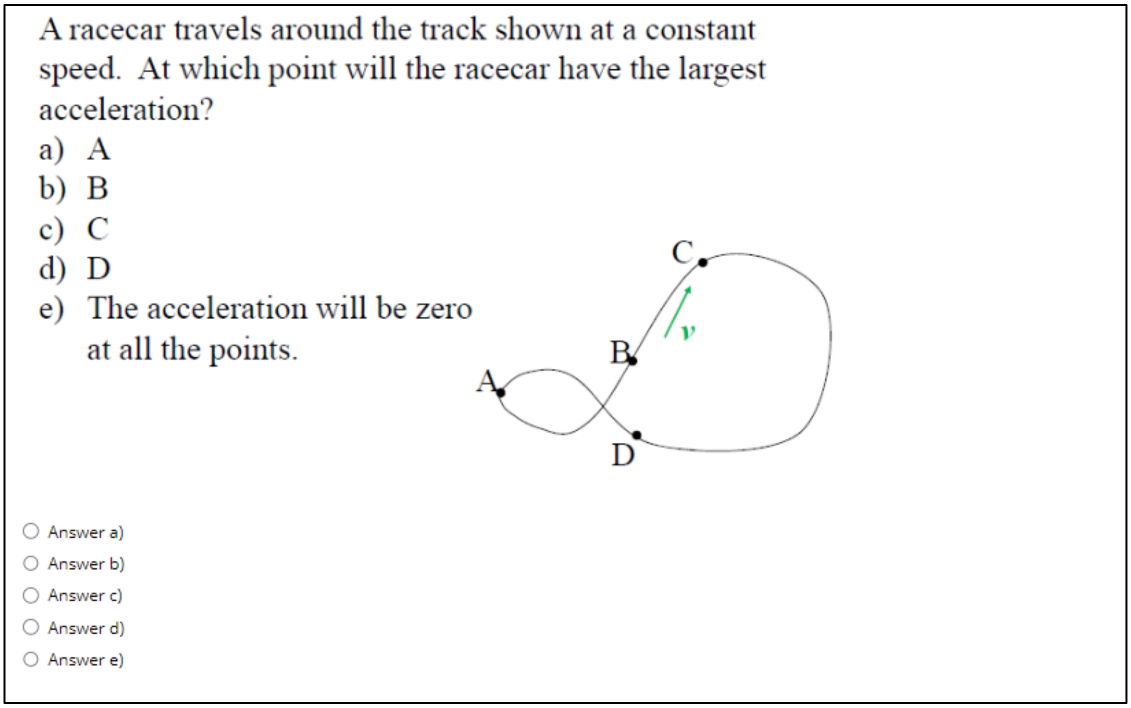

Figure 2. Example problem #2 highlights errors when using a chatbot.

Both ChatGPT-4o and Claude responded immediately with a letter choice. Gemini explained how it would choose an answer, but had to be prompted to actually provide the letter it thought was correct. Each chatbot provided an explanation of the definition of radius of curvature and how it relates to centripetal acceleration. Each correctly identified the equation for centripetal acceleration as $a_n = v^2/r$, and indicated that the largest acceleration would be the point with the smallest radius of curvature. Gemini produced the correct result, recognizing that point A has the smallest radius of curvature and thus the highest acceleration.

ChatGPT indicated that point B was the correct answer, with the following explanation: "At point B, the track appears to have the smallest radius of curvature, as it is located at a sharp curve compared to points A, C, and D." This was an interesting answer, so the authors decided to prompt further to assess ChatGPT's understanding of the geometry of the problem.

Does the track cross over itself?

Yes, the track does cross over itself. Looking at the diagram, the path at point **B** appears to intersect or overlap with another part of the track. This indicates that the track has a crossing point, likely at different elevations, forming a bridge-like structure or an overpass/underpass design. This is common in tracks where the path loops back on itself without being on the same plane.

ChatGPT was then re-prompted by asking if the radius of curvature at B was large or small. It revised its answer, saying that the centripetal acceleration would be small at B.

Claude initially responded that point C had the sharpest radius and thus had the highest acceleration. It was re-prompted to reconsider the curvature, causing it to adjust its reply to the correct answer:

R Where is the curve sharpest?

Looking more carefully at the curve, point A actually has the tightest radius of curvature compared to points B, C, and D. I need to revise my previous answer - the acceleration would be largest at point A where the curve is sharpest.

The correct answer is a) A.

Copy Retry

To their credit, Claude, ChatGPT, and Gemini provided disclaimers such as:

Claude can make mistakes. Please double-check responses.

The previous two examples highlight characteristics of the chatbots when answering Dynamics questions:

- Their machine vision is surprisingly good at reading text and simple diagrams
- They possess a good understanding of basic physics.
- They comprehend the expectations on multiple-choice questions – this is also true of other short answer question types.
- They often make mistakes, even when it appears that they have solid reasoning.
- They are quick to revise their opinions when confronted with contrary evidence, or even if simply asked to reflect on the answer.

It is interesting to compare these characteristics to those of human students. Rochas et al analyzed errors made by Civil Engineering students at West Point in solving mechanics problems [4]. The study sorted errors into four categories: non-conceptual, minor execution, major conceptual, and inefficiencies. The first three will be used in this study to analyze the performance of chatbots.

During the Fall 2024 semester's Dynamics course, the authors posed many homework, quiz, and exam problems to chatbots to get a qualitative feel for the responses. In all cases, images of the problems were pasted directly into the chatbot without any prompts. Out of the hundreds of responses, some examples of errors seen by the authors are categorized below.

Non-conceptual Errors:

- In a problem involving interlocking gears of varying radii, the chatbot repeatedly applied incorrect ratios because it couldn't distinguish connecting points. (spatial relationships)
- In a multiple-choice problem, the chatbot calculated the correct solution, but picked the wrong letter choice due to misinterpretation of the available answers. (answer comprehension)

Minor Execution Errors:

- In a problem requiring algebraic manipulations, the chatbot failed to correctly isolate and solve for a variable. (algebra)
- In a problem with a blend of Imperial and Standard unit systems, the chatbot failed to correctly convert from one system to another. (unit conversion)
- In a kinematics problem, the chatbot made an error while performing integration of acceleration and velocity to obtain a distance. (calculus)
- In a problem involving tangential and normal accelerations, the chatbot made an error by using sin and cos incorrectly. (trigonometry)

Major Conceptual Errors:

- In a problem involving cables, pulleys, and blocks, the chatbot could correctly identify which blocks were connected to the cables but had a hard time using this information to generate a constraint equation.
- In a problem involving a spider walking across a rotating platform, the chatbot correctly identified the location and velocity of the spider but could not correctly identify the components of the spider's acceleration.

While these examples are illuminating, a more formal approach to evaluate chatbot performance is required. Assessments for AI algorithms exist in the literature for a wide variety of topics, and results scored by various chatbots are published with regularity. Some noteworthy examples are:

MMLU - Measuring Massive Multitask Language Understanding [5]
GPQA - A Graduate-Level Google-Proof Q&A Benchmark [6]
MATH500 – A collection of 500 math problems [7]
HumanEval - A collection of coding problems for testing AI [8]

These assessments measure chatbot performance in subjects related to engineering (math and coding), but to the authors' knowledge no large datasets exist for testing AI on Dynamics problems. However, a bank of problems called the Dynamics Concept Inventory (DCI) was created for the purpose of assessing the comprehension of human students [9]. The DCI is composed of 29 questions focused on topics that have been identified as difficult for students to understand. Readers can request the DCI at https://sites.esm.psu.edu/dci/, or find it via the Concept Warehouse at https://conceptwarehouse.tufts.edu/cw/CW.php.

Results and Discussion

The DCI was administered to ChatGPT, Claude, and Gemini, with instructions to respond with the letter corresponding to the correct answer. No further instructions or prompts were provided. For ChatGPT and Claude, it was found that the entire bank of questions could be administered as a single PDF document without changing the results. The DCI was administered to Gemini as individual image files of each question. As an additional check, the DCI was administered to ChatGPT-4o both manually by pasting the images in the chat window and by sending images via the API provided by OpenAI using Python, and again results were consistent.

The test was administered multiple times to various versions of the chatbots over several weeks in October to December of 2024, and the replies were recorded and checked against the correct answers. In Figure 3 is shown the average score on the DCI for the various chatbots. Scores ranged between 31% - 76%, with an average of 46% (of 12 samples). The highest score was achieved by ChatGPT-4 o1[7]. This shouldn't be a surprise, as ChatGPT-4 o1 is touted as a "reasoner", meaning that it spends more time thinking about responses. The time spent per problem by ChatGPT o1 was often 30-40 seconds, which is an order of magnitude more than the other chatbots. The lowest score was achieved by ChatGPT-4o mini, which is also not a surprise given that it is a pared-down version meant for streamlined deployment.

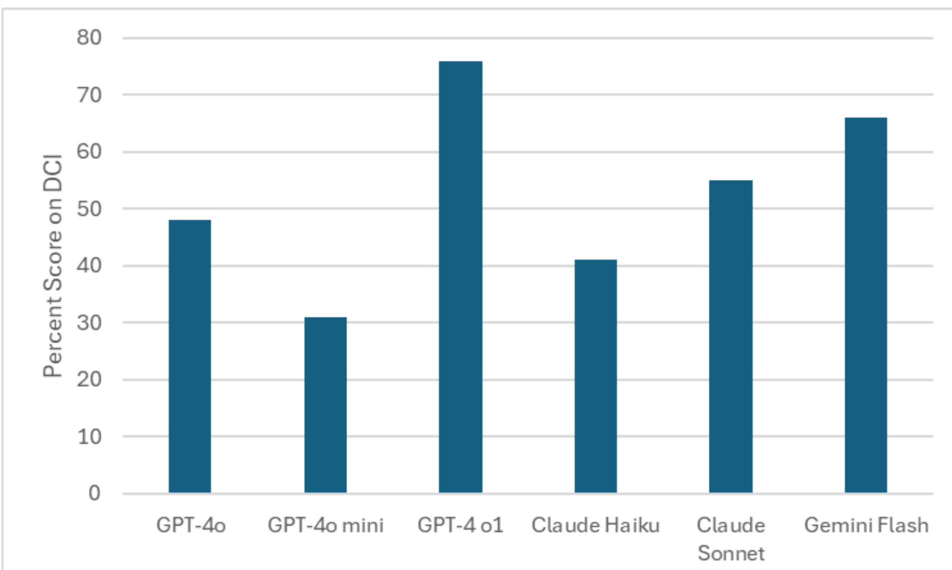

Figure 3. Comparison of scores by various chatbots on the Dynamics Concept Inventory

Generally, the chatbots had similar performance on the various questions. The percentage of correct answers was calculated for each question, and they were divided into three categories: Easy, where the chatbots answered correctly more than 70% of the time; Medium, where

[7] https://openai.com/o1/

answers were correct between 30% and 70%; and Hard, where correct answers were below 30%. A brief synopsis is provided of select problems in each of the three categories:

Easy (>70% correct):

Problem 1 contains text only, asking about forces in a truck vs car collision.
Problem 4 is a picture of a Ferris wheel, asking about the magnitude of angular velocity at different points on the same rigid body.
Problem 12 shows two objects rolling down a slope, a hoop vs a solid cylinder of the same radius and weight.
Problem 14 shows two people sitting on rolling chairs. One person pushes off on the other person. There is an image but it is not required to solve the problem.

Medium (from 30% to 70% correct):

Problem 13 was studied in [10] and is shown in Figure 4. This is a classic problem that tests student comprehension of Newton's Second Law and the roles of inertia and acceleration. Students correctly answered this question 4% to 5% of the time prior to taking a Dynamics class and between 40% to 55% after taking the course. The chatbots answered this question correctly 56% of the time, and when they did, they offered correct explanations of how to apply Newton's Second Law to solve the problem.

Question 13

Both systems shown have massless and frictionless pulleys. On the left, a 10 N weight and a 50 N weight are connected by an inextensible rope. On the right, a constant 50 N force pulls on the rope. Which of the following statements is true immediately after unlocking the pulleys?

(a) In both cases, the acceleration of the 10 N blocks will be equal to zero.

(b) The 10 N block on the left will have the larger upward acceleration.

(c) The 10 N block on the right will have the larger upward acceleration.

(d) The tension in the rope on the left system is 40 N.

(e) In both cases, the 10 N block will have the same upward acceleration.

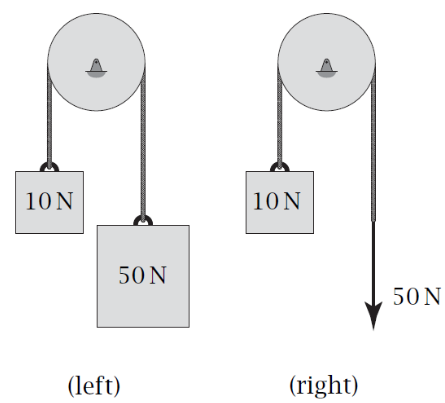

Figure 4. Problem 13 from the Dynamics Concept Inventory

Problem 18 is a word problem asking about the change in energy during a non-ideal collision.
Problem 24 shows a ball and box colliding, testing the concept of rigid body general planar motion with an impact.
Problem 27 asks about friction on one of the tires of an accelerating car where the tires do not slip. The key is knowing that static friction $\leq \mu N$.

Hard (<30% correct):

Problem 5 asks about the acceleration of a spider walking on a rotating disk from the center toward the right. All chatbots claimed that the acceleration is zero because they neglected to account for the Coriolis component $(2\dot{r}\dot{\theta})$.

Problem 6 shows the same Ferris wheel as Problem 4 but asks for the direction of the angular velocity vector of two points. This problem is intended to test understanding of the concept of the angular velocity vector, which in this case points directly out of the page toward the viewer. Most models (and many human students) assume the question asks about the translational velocity vectors, which move in opposite directions.

Problem 10 was studied in [10] and is shown in Figure 5. The correct answer is (b), because platform Q has both rotational and translational components of kinetic energy. Many students struggle with this problem but improve after taking a Dynamics course. Most chatbots only considered translational kinetic energy, quoting the classic formula, but neglected rotational kinetic energy. Only one chatbot correctly answered this question, citing rotational kinetic energy. However, when prompted to explain its choice, it gave incorrect reasoning, stating that the moment of inertia about point A of platform Q was greater than that of platform P.

Question 10

Two different amusement park rides are shown in the figure at the right. Each of the platforms is supported on *frictionless* pins by a pair of arms. All of the arms supporting the platforms rotate at the same angular velocity ω. Compare the kinetic energies of the two identical platforms P and Q.

(a) Platform P has greater kinetic energy.

(b) Platform Q has greater kinetic energy.

(c) The kinetic energy of the platforms will be the same.

(d) Each will have zero kinetic energy.

(e) Not enough information is given.

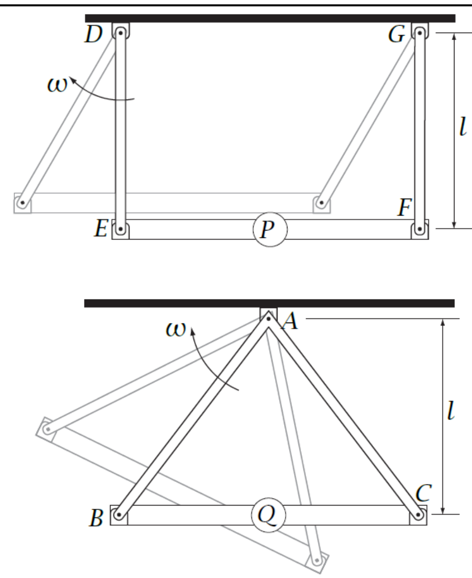

Figure 5. Problem 10 from the Dynamics Concept Inventory

A more detailed explanation of errors from a single exam administered to ChatGPT-4o is provided in the appendix. Overall, it appears that errors generally come from two sources: non-conceptual mistakes in chatbots' ability to distinguish graphics and mistakes in applying physics principles.

Vision Errors

It appears that machine vision is often a source of error, which would be included in the category of "Non-conceptual" errors above. To attempt to explore machine vision sources of error in greater depth, five "vision" questions were posed about each problem. Each question was crafted to gauge understanding of the image being displayed for the problem, including the text about the problem and multiple-choice answers. To facilitate assessment, the vision questions required only a yes or no answer. An example set of five questions for Problem 10 is presented in Table *1*, along with a comparison of the correct answer to the response from ChatGPT-4o.

Table 1. Example set of vision questions for Problem 10

Question	**Correct Answer**	**ChatGPT-4o**
Are both platforms P and Q rotating at the same angular velocity?	Yes	Yes
Is platform P moving?	Yes	Yes
Are the arms supporting platform Q rotating?	Yes	Yes
Is platform Q higher than platform P?	No	No
Are points D and G moving?	No	Yes

The authors posed these sets of vision questions to ChatGPT-4o through the API with instructions to only reply with a "yes" or "no" answer. Overall, ChatGPT-4o was able to answer 88% of the questions correctly. Four of the problems had multiple incorrect answers, specifically Problem 3 (points on a spinning disk), Problem 5 (spider on a spinning disk), Problem 26 (spinning boomerang), and Problem 28 (accelerating car). These problems all belong to the "Hard" category discussed above, where the chatbots performed poorly, hinting that at least for these questions vision was a significant source of error. However, the r^2 correlation between the DCI results for each problem with the scores for the vision questions was only 0.53, indicating that while errors in vision certainly contribute to the overall error, it is not the only source.

Conceptual Errors

To assess the occurrence of conceptual errors, ChatGPT-4o was provided with a list of principles that might be used to solve each problem. The DCI was written specifically to evaluate certain concepts, so the authors wanted to see if ChatGPT-4o could identify major principles. The following list contains possible principles for solving problems on the DCI:

- Newton's 1st law
- Newton's 2nd law
- Newton's 3rd law
- Principle of work and energy
- Principle of impulse and momentum
- Kinematics

ChatGPT was provided with the problem image and asked to identify which principles from the list were required to solve the problem. ChatGPT responded with a list, along with a short explanation of why the principle should be included. It also explained why it did not include the principles it omitted from the list. The answers were compared to a list generated by the authors, and a score of 1 was assigned for each perfect match. ChatGPT never failed to mention a principle that was required, but it often recommended using a principle that wasn't directly helpful in answering the question. This mostly occurred for Newton's Second Law or Kinematics. For example, on a problem involving the velocity of points on a moving tire, ChatGPT included Newton's Second Law, which is not strictly necessary to answer the question. It did provide the caveat that, "This law underpins the car's acceleration and how forces result in the tire's rotational and translational motion. However, this law is more implicit for understanding the rolling condition." In this case, ChatGPT clearly understood that the principle wasn't directly required to answer the question but decided to include it in the list as a supporting concept. ChatGPT included these implicit "extras" on 19 out of 29 problems. It is unlikely that these additional principles caused ChatGPT to make a mistake on any particular problem.

The authors performed another test by asking ChatGPT to decide for each problem whether it could be solved by assuming the system was a particle or if it required rigid body motion. In all cases except one, it correctly distinguished between particle and rigid body motion. The single incorrect case involved masses attached to pulleys (Problem 10). ChatGPT claimed that because the pulleys rotated, it was necessary to consider the system as a rigid body. This is not true, as the system can be analyzed using only 1-D translational motion. However, this test showed that overall ChatGPT understood the difference between particle and rigid body motion, eliminating this as a potential source of error.

Conclusions and Recommendations

The chatbots used in this study almost always demonstrated a correct understanding of the nature of the problems presented to them and the underlying principles required to formulate an answer. However, they often made errors in applying those principles to choose the correct answer. When combined with vision inaccuracies, these errors make current chatbots somewhat unreliable for solving conceptual problems in Dynamics. None of the problems in the DCI require algebra or calculus, but for other problem sets these might introduce another source of error.

The authors found the high-level explanations of which principles should be used to solve problems to be very reliable. In general, the explanations about how those principles would be used in a problem were accurate and helpful. The use of prompts to ask clarifying questions and to provide extra information can be used effectively as a critical thinking exercise, and research on this topic would certainly yield interesting results for the Dynamics community. The authors have already begun to encourage students to use the chatbots analyze problems by performing sanity checks, reflecting on the larger meaning of solutions, and checking the work of one chatbot against another. In addition, the authors regularly use the chatbots to quickly generate feedback on assessments and homework. A more in-depth look into using chatbots for feedback on student work is a future direction for this study.

Finally, as discussed above, programmatic methods exist to assess the performance of chatbots. The authors plan to rebuild the DCI in a JSON format to facilitate automated testing of updated versions of ChatGPT through the API[8,9]. Hopefully this will stimulate more investigation into the utility of modern AI tools, resulting in broader studies with more data for statistical analysis.

Disclaimer: *The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the United States Air Force Academy, the Air Force, the Department of Defense, or the U.S. Government. Approved for public release, USAFA-DF-2025-412: distribution unlimited.*

References

[1] J. Qadir, "Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education," in *IEEE Global Engineering Education Conference, EDUCON*, IEEE Computer Society, 2023. do: 10.1109/EDUCON54358.2023.10125121.
[2] 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, 2023, doi: 10.1016/j.caeai.2023.100183.
[3] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." [Online]. Available: https://github.com/huggingface/transformers/blob/master/
[4] M. B. Rocha, E. B. Williamson, and E. Williamson, "Sticking Points: Reasons Why Civil Engineering Students Make Errors Solving Engineering Mechanics Problems."
[5] D. Hendrycks *et al.*, "Measuring Massive Multitask Language Understanding," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.03300
[6] D. Rein *et al.*, "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.12022
[7] J. Cunningham, "MATH 500," http://math500.com/. Accessed: Dec. 26, 2024. [Online]. Available: http://math500.com/
[8] M. Chen *et al.*, "Evaluating Large Language Models Trained on Code," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.03374
[9] P. Cornwell, "The Dynamics Concept Inventory (DCI)-The Past, Present, and Future."
[10] G. L. Gray, F. Costanzo, D. Evans, P. Cornwell, B. Self, and J. L. Lane, "The Dynamics Concept Inventory Assessment Test: A Progress Report and Some Results," 2005.

[8] https://github.com/openai/evals/blob/main/README.md
[9] https://github.com/openai/simple-evals/blob/main/README.md#user-content-fn-7-e5001c7f0eb641886e9c28a25213ff74

Appendix

Analysis of errors made by ChatGPT on the DCI

Problem 2: ChatGPT made a bad assumption about the direction of velocities along a rigid plate
Problem 3: ChatGPT mistakenly assumed the question was asking about velocity magnitude, instead of both magnitude and direction
Problem 5: ChatGPT neglected to consider Coriolis acceleration (spider moving on a disk)
Problem 6: ChatGPT assumed the question was asking about the translational velocity vector instead of the angular velocity vector
Problem 7: ChatGPT was unable to understand that two vectors pointed in the same direction
Problem 10: ChatGPT incorrectly compared the moment of inertia of two identical bars
Problem 17: ChatGPT incorrectly stated that a higher moment of inertia would lead to higher angular acceleration, even though it correctly identified the relationship, $\tau = \alpha I$.
Problem 21: ChatGPT failed to account for rotational motion on a rigid body
Problem 22: ChatGPT incorrectly said that a point was in contact with the ground (vision)
Problem 23: ChatGPT incorrectly determined the direction of components of a vector
Problem 25: ChatGPT incorrectly measured the relative separation between several points and did not find the max distance
Problem 28: ChatGPT incorrectly believed that the friction on the front tires of a rear-wheel drive car contribute to the acceleration (the reverse is true)