

Can students spot AI's physics errors?
An experience with GPT in the classroom

CSAAPT Spring 2024
Semi-Virtual Meeting

March 16, 2024

Delaware State University

Generative AI and Physics Education

Since the release of ChatGPT in November 2022, there has been a huge amount of research done on the capabilities of generative AI systems and their potential uses (and misuses) in pedagogical contexts.

Some example specifically concerning Physics:

Liang, Y., Zou, D., Xie, H. et al. Exploring the potential of using ChatGPT in physics education. *Smart Learn. Environ.* 10, 52 (2023)

Forero, M. G. and Herrera-Suárez H. J. ChatGPT in the Classroom: Boon or Bane for Physics Students' Academic Performance? *arXiv:2312.02422 [physics.ed-ph]* (2023)

Ding, L., Li, T., Jiang, S. et al. Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *Int J Educ Technol High Educ* 20, 63 (2023).

PHYS 101 at Harford Community College

PHYS 101 is a standard algebra-based introductory physics course.

First in a 2-course sequence.

Topics include 1d and 2d kinematics, Newton's laws, energy, momentum, rotational motion, gravitation, thermodynamics and fluids.

Three exams, in addition to homework problems, lab activities, and conceptual in-class quizzes

Fall 2023 PHYS 101-02 section was unusually small (9 students).

Used as testing ground for Generative AI-based assignments within coursework.

Three review activities, one before each exam.

Formative assessments, graded but with very small weight.

ChatGPT-based review activities

In the review activities, students were presented with responses generated by ChatGPT to physics questions and problems. (3 or 4 per activity.)

Generated by the instructor through trial-and-error prompting on both ChatGPT 3.5 and ChatGPT 4, the responses contained errors in reasoning comparable to errors commonly exhibited by students of introductory physics.

In some cases, students were told an answer contained an error, and tasked to locate it. In others, they were offered two contrasting answers (usually by versions 3.5 and 4) and the task was to identify the correct answer.

ChatGPT based review activities

In the review activities, students were presented with responses generated by ChatGPT to physics questions and problems. (3 or 4 per activity.)

Generated by the instructor through trial and error prompting on both ChatGPT 3.5 and ChatGPT 4, the responses contained errors in reasoning comparable to errors commonly exhibited by students of introductory physics.

In some cases, students were told an answer contained an error, and tasked to locate it. In others, they were offered two contrasting answers (usually by versions 3.5 and 4) and the task was to identify the correct answer.

After giving answers individually, students discussed again the questions in groups, and resubmitted answers based on group discussion. Each student's grade is the average of their individual and group-based submissions.

(Same grading modality is used for conceptual quizzes in the course.)

Example 1: 1d kinematics

- A stone is thrown downwards from an initial height 2.5 m above the ground giving it an initial speed of magnitude 0.7 m/s. What is its speed when it hits the ground?

To find the speed of the stone when it hits the ground, you can use the following kinematic equation for motion in one dimension: $v^2 = v_0^2 + 2a \Delta y$

Where:

- v is the final velocity (speed) of the stone when it hits the ground.
- v_0 is the initial velocity of the stone.
- a is the acceleration due to gravity (approximately -9.81 m/s^2 , negative because it acts downward).
- Δy is the distance traveled (the height the stone falls, which is 2.5 m in this case).

Let's plug in the values:

$$v_0 = 0.7 \text{ m/s (initial speed)}$$

$$a = -9.81 \text{ m/s}^2 \text{ (acceleration due to gravity, negative because it's downward)}$$

$$\Delta y = 2.5 \text{ m (initial height)}$$

Example 1: 1d kinematics (cont.)

After some correct math, ChatGPT 3.5 reaches the conclusion:

Finally, take the square root to find v : $v = \sqrt{-48.56}$ m/s.

Since the square root of a negative number is an imaginary number, it means that the stone never reaches the ground under these conditions.

Activity then asks students:

The conclusion is clearly wrong. The error is not in the numerical calculation, but happens before. At which point did the reasoning go wrong?

- a) The wrong equation was used (an equation relating v and t should be used instead).
- b) The wrong equation was used (an equation relating Δy and t should be used instead).
- c) The value replaced for v_0 is mistaken.
- d) The value replaced for “ a ” (the acceleration) is mistaken.
- e) The value replaced for Δy is mistaken.

Example 2: rotational dynamics

- A rod of length 1 m and negligible mass has a 1 kg mass attached to it 0.1 m away from one endpoint. The rod rotates around the opposite endpoint. What is the moment of inertia?

The moment of inertia of a system of point masses is given by the sum of the products of each mass and the square of its distance from the axis of rotation. In this case, we have a rod of negligible mass and a point mass. Since the rod has negligible mass, we only need to consider the point mass.

For the 1 kg mass located 0.1 m from the endpoint, the moment of inertia is:

$$I = m \cdot r^2 = (1 \text{ kg}) \cdot (0.1 \text{ m})^2 = 0.01 \text{ kg} \cdot \text{m}^2$$

Thus, the moment of inertia of the system is $0.01 \text{ kg} \cdot \text{m}^2$.

Activity then asks students if answer is correct. (This was Chat GPT 4!)

Example 3: thermodynamics

- Question posed to two AI text generators: When a heat engine (such as the one found in a car) does a cycle, is the net change in the internal energy of the gas positive, negative or zero?

The first one's output [ChatGPT-4] was:

The net change in the internal energy of the gas in a heat engine, like the one found in a car, undergoing a complete cycle is zero. This is based on the First Law of Thermodynamics, which states that the change in internal energy of a system is equal to the heat added to the system minus the work done by the system.

In a complete cycle of a heat engine, the system returns to its initial state. Since internal energy is a state function (meaning it depends only on the current state of the system and not on the path taken to reach that state), the internal energy at the end of the cycle is the same as it was at the beginning. Therefore, the net change in internal energy over one complete cycle is zero.

This principle is fundamental to the operation of cyclic heat engines, where the objective is to convert heat energy into work, while the internal energy remains unchanged over each cycle.

Example 3: thermodynamics (cont.)

The second one's output [ChatGPT-3.5] was:

In a heat engine, such as the one found in a car or any other internal combustion engine, the net change in the internal energy of the gas is typically negative over the course of a complete cycle.

[... Cycles] consist of a series of processes, including compression, combustion, and expansion, which collectively convert thermal energy (heat) into mechanical work. During the combustion process, heat is added to the gas, causing an increase in internal energy. However, during the expansion and exhaust processes, the gas does work on the engine's pistons, and some of its internal energy is converted into mechanical work, resulting in a net decrease in internal energy.

The second law of thermodynamics states that not all of the heat energy added to the system can be converted into useful work, and some of it must be rejected as waste heat. This waste heat is transferred to the surroundings, and as a result, the internal energy of the gas decreases, making the net change in internal energy negative. [...]

So, in summary, the net change in the internal energy of the gas in a heat engine cycle is typically negative, as heat is converted into mechanical work, and some heat is rejected to the surroundings.

Students are asked which answer is correct.

Grades

Students did **much worse** in this assignment than would be predicted by their general performance in the course.

Average grade of all students in all three review activities (excluding zeroes due to absences) was **54.1%**.

This includes the group discussion submissions!

Grades

Students did **much worse** in this assignment than would be predicted by their general performance in the course.

Average grade of all students in all three review activities (excluding zeroes due to absences) was **54.1%**.

This includes the group discussion submissions!

By contrast, average grade of all students in conceptual quizzes (which also includes both individual and group submissions) was **75.7%**.

Average grade in exams (which include both conceptual questions and calculational problems, with formula sheet allowed) was **80.4%**.

Clearly something about this assignment “throws off” students when compared to more conventional assessment methods.

Overall assessment of activity

Most conventional physics assessment activities test students' knowledge by asking them to apply it straightforwardly to answer a question, e.g.:

“What would happen in this situation?” (qualitatively or quantitatively)

Call these “Level 1 questions”.

This assignment goes meta (“Level 2”) by asking “Is this answer to a Level 1 question correct?”

Requires not only physics knowledge, but also skills in critical close reading.

Tests for the ability to see through errors in a text which is **by design** generated to be plausible-sounding.

From this point of view, it is not surprising students do worse!

Summary and outlook

- Trial assignment requiring students to spot errors in AI-generated answers to physics questions and problems.
- Assignment done first individually and then through group discussion.
- Grades much worse than in other more traditional assignments in the same course.
 - Interpretation: Students find it easier to approach “Level 1” questions than “Level 2” questions.

Summary and outlook

- Trial assignment requiring students to spot errors in AI-generated answers to physics questions and problems.
- Assignment done first individually and then through group discussion.
- Grades much worse than in other more traditional assignments in the same course.
 - Interpretation: Students find it easier to approach “Level 1” questions than “Level 2” questions.
- Do we think “Level 2” thinking (critical engagement with possibly wrong physics content) is a valuable skill for students to develop?

If so, we need to think specifically about how approach the development of these skills, without assuming it will come directly with “Level 1” knowledge.