WHEN THE ANSWER IS FREE

Will Yeadon – Durham University, UK – DOI: https://doi.org/10.1051/epn/2026212

Generative AI can now solve the vast majority of physics problems a typical student will encounter until the end of a master's course. As the cost of producing correct answers approaches zero, physics education must pivot from valuing the "product" to valuing the "process." This shift requires a reimagining of how we teach, assess, and verify understanding.

The End of the Struggle

Current large language models perform extremely well on structured physics questions, including short calculations, standard derivations, and brief written explanations. Earlier versions often made arithmetic mistakes, but these errors are becoming less common as models gain access to integrated calculators and symbolic tools. Figure 1 illustrates this rapid advancement. Reevaluating earlier work on how ChatGPT-3.5 answered physics questions [1] shows 81% at GCSE (for age 16 in the UK), 63% at A-Level (for age 18 in the UK), and 23% on university textbook questions. By contrast, ChatGPT-5.2 reached 99% for GCSE, 96% for A-Level, and 94% for university questions. While invigilated pen-and-paper exams remain immune to AI, they are only one part of a broader assessment landscape.

If a student pastes a homework question into a chatbot, they receive a fluent and usually correct answer. The correct formula appears, the numerical result is calculated, and the explanation is structured. In this scenario, the struggle is removed. However, a struggle in learning physics is not an inefficiency, but arises from the cognitive effort required for learning. It is the moment when a student realizes they do not know which equation applies, or when algebra fails and must be retraced.

When answers are free, that friction is reduced. While AI could theoretically strengthen learning by functioning as an adaptive tutor, the reality is that when students are tasked solely with getting the answer correct, they will use AI as a shortcut. It is therefore unsafe to design assessments on the assumption that AI cannot perform a given calculation. The more stable approach is to assume that while AI can do a task, producing an answer is not the same as understanding it.

Reimagining Assessment at Secondary Education

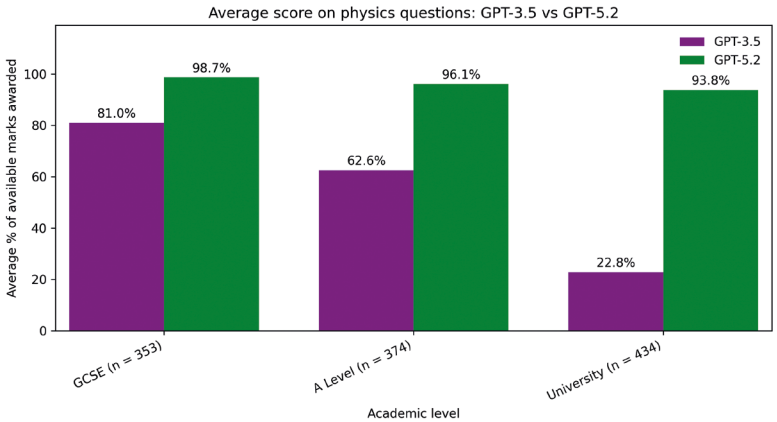
Secondary education has limited flexibility to respond because national curricula and examination boards change slowly. Schools cannot easily replace written homework with oral examinations due to constraints on staff time. However, homework can be redesigned to focus on the "why" rather than the "what."

For GCSE students, tasks can shift toward explaining puzzling phenomena. Puzzles naturally resist over-reliance on AI by making student thinking visible. Instead of a standard numerical answer, students might be asked to record a short reflection on how they arrived at their ideas or what they found confusing about a specific concept. These critical thinking-based tasks not only enhance physics understanding but equip students for the wider world.

A-Level has much the same constraints as earlier schooling. Examinations are high stakes with increased conceptual depth. If homework tasks at this stage can be fully automated, the signal about student understanding weakens. A more sustainable approach combines structured practice with in-class verification. Students could be asked to submit work developed with AI assistance but then be required to "teach back" or explain key steps under light supervision to prove understanding.

Reimagining Assessment at University

At the university level, assessment is more varied, providing flexibility to move away from take-home tasks like problem sheets and scientific coding challenges, which AI can generate in seconds. Not all forms of assessment are equally vulnerable. Examples include laboratory work where students must interpret real-world data; oral examinations where students must defend their ● ● ●



▼ **FIG. 1:** Comparison of GPT-3.5 and GPT-5.2 performance across academic levels. The rapid closing of the gap at the University level highlights the increasing vulnerability of traditional problem sets.

▲ **FIG. 2:** The "Human-in-the-loop" assessment model. A student "Interface" mediates between an AI "Oracle" and a student "Hand," ensuring cognitive engagement through mandatory interpretation.

reasoning in real time; and whiteboard reasoning, which forces the translation of thought into a visible, interactive process. However, all of these methods require significant staff time.

One approach worth considering is to use longer project-based work where assessments are given earlier in modules with class time dedicated to iterative development. This better matches professional research and the wider world of employment. While this doesn't prevent AI use, marking schemes can be designed to assess student learning and their engagement with specific class material or unique datasets that AI has not seen.

Another idea to embrace AI is the "human-in-the-loop" structure, such as pair programming with an AI oracle. In this model, one student has access to an AI "oracle" and acts as an "interface," while the other does not but acts as the "hand" to complete an assessment. Information must pass through human explanation and interpretation before it becomes part of the submission. For this to work practically, there must be physical isolation between the "hand" and the "oracle" so that the "interface" is required. This style of assessment accounts for the collaborative nature of modern work while ensuring individual students remain cognitively engaged.

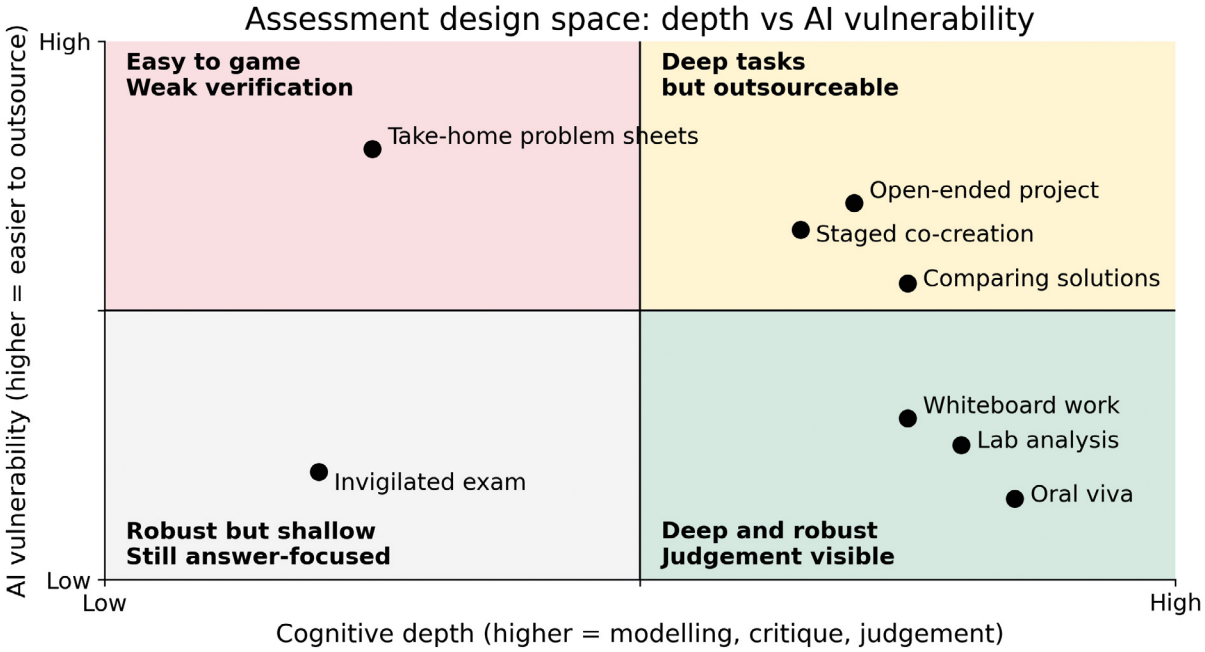
Verification in an AI-Saturated World

If AI can generate solutions on demand, the problem for educators is not generation but verification: how do we know that a student understands the work they submit? Historically, verification was indirect, a student who repeatedly produced correct homework solutions was assumed to understand the material. That assumption no longer holds. As correct output decouples from cognitive effort, the signal provided by homework has become noisy. When a measurement becomes a target, it ceases to be a good measurement, a phenomenon known as Goodhart's Law. This forces educators to articulate more clearly what they value.

Even without generative AI, assessment can be reimagined to include different forms of verification. Short oral follow-ups constitute one example: after submitting a problem sheet, a student might be asked to explain one solution in two or three minutes. Not to repeat memorized text, but to respond to a new variation of the question. Even light-touch questioning can help students internalize their reasoning.

Another approach uses comparative reasoning. Instead of asking for a single answer, students can be given two competing model outputs and asked to evaluate which is more physically plausible and why. This shifts the emphasis from production to judgment. The ability to critique a solution is cognitively different from the ability to copy one. There is also value in staged co-creation with AI, a student might first submit a prediction without AI, then refine it using AI, and finally reflect on

▶ **FIG. 3:** The assessment design space. Effective modern assessment should move toward the lower-right quadrant, prioritizing deep cognitive tasks with robust human verification.



the differences. This makes the interaction with AI explicit rather than covert.

Figure 3 summarizes this assessment landscape. Tasks that are shallow and easily outsourced sit in the upper left, providing weak evidence of understanding. Moving rightward increases cognitive depth, while moving downward strengthens verification. The most durable forms of assessment lie in the lower-right quadrant, where students must model, critique, and justify their thinking in real time. The point is not to eliminate AI, but to shift assessment toward forms where judgment, not answer production, becomes the primary signal of learning.

Ultimately, the goal should not be to simulate a world in which AI does not exist, but to ensure that students develop durable reasoning skills within a world where it does. Physics depends on modelling. Real problems require selecting assumptions, identifying scales, and deciding which effects can be neglected. In some sense, these un-verifiable tasks are a more accurate representation of physics than the neat reproduction of a textbook calculation.

What Do We Want Students to Learn?

It is tempting to ask which tasks AI cannot yet perform, but this framing will age poorly as capabilities change. A more stable question is what we want students to gain from studying physics. Do we value the speed of calculation, or the capacity to model, approximate, and question?

Noticing when a result violates physical intuition requires experience and reflection. Revising a model when assumptions fail requires flexibility. These skills prepare students for research, industry, and public life. If answers become cheap, modelling and critical evaluation move to the foreground. When the answer is free, understanding becomes the scarce resource and the opportunity exists to design learning environments where reasoning cannot be separated from explanation. ■

About the Author

Will Yeadon is an Assistant Professor of Physics at Durham University. His research examines the impact of generative artificial intelligence on physics education and assessment. He holds a PhD from the University of Sheffield, where he developed a novel computational fluid dynamics model for ultra-thin welding processes.

Reference

[1] W. Yeadon and T. Hardy, *Phys. Educ.* **59**, 025010 (2024)