WHEN AI GETS AN A: INTRODUCTORY PHYSICS EDUCATION WHEN ANSWERS ARE CHEAP

Marina Babayeva[1], Ralf Widenhorn[2] and Gerd Kortemeyer[3] – DOI: https://doi.org/10.1051/epn/2026210
[1] Department of Physics Education, Faculty of Mathematics and Physics, Charles University, 180 00 Prague, Czech Republic
[2] Department of Physics, Portland State University, Portland, Oregon, USA
[3] Rectorate and AI Center, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland; Michigan State University, East Lansing, MI 48824, USA

Artificial intelligence is entering physics classrooms as a practical tool that can solve physics problems at the level of high-performing students. We argue that to remain relevant, physics education must refocus on scientific practices. Here we share experiences from two complementary settings: laboratory work and written assessment, and conclude with practical recommendations for educators.

Over just a few years, AI systems have progressed from struggling with introductory physics problems [1] to performing at the level of top-scoring students [2]. With today's multimodal "reasoning" models (*e.g.*, GPT-5, Gemini-3), a student can photograph a typical problem, ask for a solution, and often receive a polished, correct answer within seconds. If physics education is reduced to producing the right answer, many students who take physics as a service course may reasonably ask: why bother?

Yet expert physicists know that just the correct answer is rarely the point. "Thinking like a physicist" involves a complex epistemology: how we decide what to trust, how we justify claims, and how we connect models to evidence [3–5]. These are universal competencies that transfer well beyond an introductory "service" course and are worth learning (and persevering) even when answers are readily available. Traditionally, physicists do this through derivations from first principles and through laboratory work.

That is where physics education now needs to move: toward assessing and supporting *ways-of-knowing* rather than answer production — ideally with personalized human guidance, but at scale often with AI assistance. AI can serve in dual roles here: both as a partner that prompts reasoning and as a tool that helps instructors provide feedback at scale. Even though this comes with the risk, as skeptics might put it, of putting the fox in charge of the henhouse. In this work, we offer two perspectives: how an interactive GPT-based assistant can support laboratory learning, and how multimodal AI tools can extend this assistance to homework and exams through structured grading workflows that retain human oversight.

AI as a Virtual Assistant in Physics Laboratories

Imagine a crowded physics lab: thirty students, one teaching assistant, and a lot of waiting. When students get stuck on a question or a calculation, they often lose valuable time or, most importantly, focus on the task before help arrives. To address this, we implemented an AI-based virtual lab assistant designed to provide immediate, low-stakes formative feedback during laboratory activities. The goal of this project was to understand how students use such an assistant and to evaluate whether an AI system can offer meaningful educational support without undermining established pedagogical goals.

Developed at Portland State University, the custom web-based platform integrated a GPT-based chatbot directly into an introductory physics laboratory course (figure 1). The assistant was available first to students working in person and later to online students using experimental tools at home, who typically have even less contact with a teaching assistant. In both contexts, the AI was designed to respond to individual student queries in

real time, offering guidance while avoiding direct solution disclosure. This setup allowed the assistant to act as an always-available first point of support, particularly valuable in remote lab settings where synchronous TA interaction is limited.

Throughout the deployment, the AI assistant demonstrated clear strengths and persistent limitations. The assistant answered more than 85% of student queries correctly and beneficially for students' understanding of the task [6]. It handled theoretical and conceptual questions well, often requiring only minimal contextual information, and provided clear explanations that helped students interpret questions or verify answers. Experimental, numerical, and measurement-based questions, however, were more sensitive to prompt phrasing and context, occasionally resulting in vague, inconsistent, or incorrect responses. Over time, the integration of a feature that could automatically run code and perform calculations significantly improved the system's handling of formula-based and numerical tasks, reducing, but not fully eliminating, errors in calculation and value interpretation.

Student feedback reflected a generally positive but cautious reception. More than 60% of students who used the assistant in their at-home practice agreed that it helped them better understand the lab material. Many students found the assistant particularly useful for checking answers, clarifying concepts, and gaining confidence before submitting work. For instance, one of the students noted that *"Whenever I got stuck, it would help point me in the right direction"*.

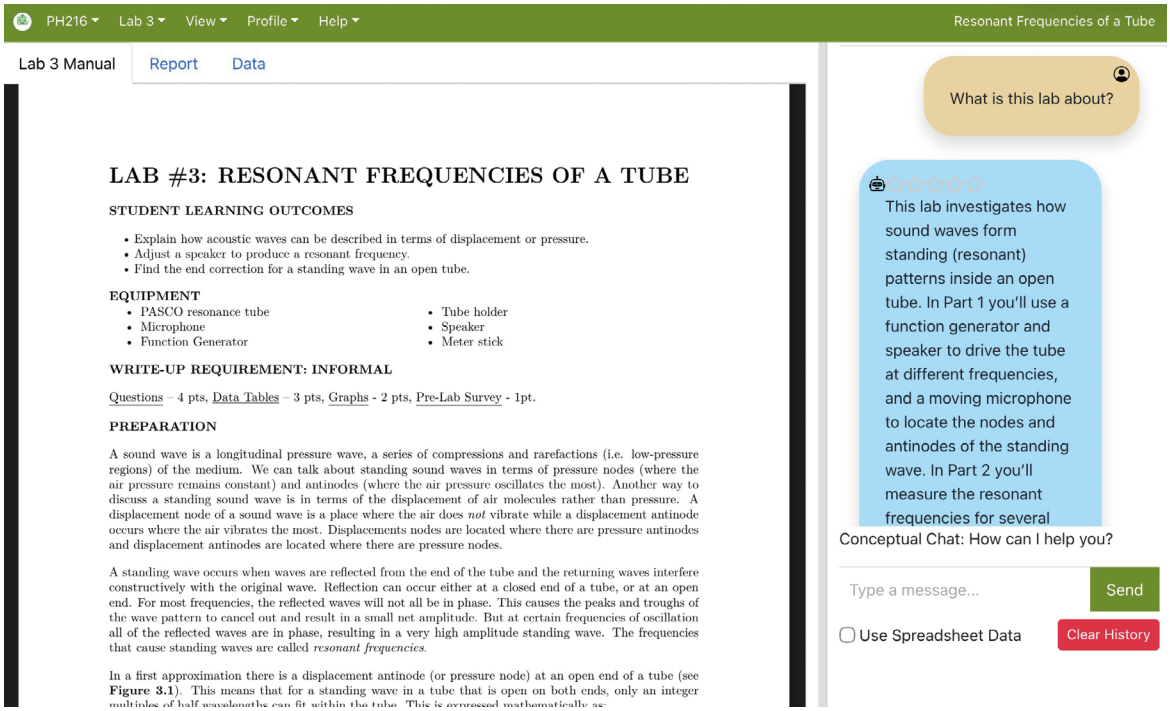
At the same time, students expressed frustration when the AI produced repetitive explanations, overly verbose responses, or numerical errors.

For example, one student mentioned, *"Sometimes it gives overly complicated responses to something simple"*. In several cases, students reported reduced trust after receiving incorrect guidance. During the in-person labs, students who were accustomed to TA interaction continued to prefer human support, while remote students more strongly valued AI's constant availability.

Overall, our experience suggests that an AI-based lab assistant can meaningfully enhance laboratory learning as a complement, rather than a replacement, for human instruction. The assistant proved particularly effective as the first line of support for low-stakes formative assessment, reducing wait times and reinforcing conceptual reasoning for both in-person and at-home students. Recent advances have resolved many early limitations, and the AI assistant now produces detailed explanations, sometimes even more extensive feedback than human teaching assistants due to greater time availability. They do, however, make occasional, unexpected errors. LLM-based tools remain probabilistic systems, underscoring both the significant opportunities and the pedagogical risks of integrating AI into physics laboratory education.

AI as a Virtual Assistant for Physics-Problem Solution Feedback and Grading

Even in a digitally saturated classroom, a large fraction of what we value in introductory physics problem solving is still most naturally expressed as handwritten work. Derivations are not just sequences of equations; they are visual objects that encode structure and intent. Mathematical typesetting is cumbersome under time pressure, and many of the marks that matter for reasoning—underbraces and term annotations, arrows ● ● ●



◂ **FIG 1:** Student view of the AI-based lab interface.

• • • indicating substitutions, side notes about limiting cases, quick sketches and free-body diagrams, small tables of units or sign conventions—are frictionless on paper but awkward to represent in plain text. In practice, the "image" of the derivation carries information that instructors use to diagnose understanding [7].

Historically, that same "image-first" reality has been a bottleneck for scalable feedback. In large courses, it is hard to provide rapid, individualized commentary on handwritten reasoning, so we often defaulted to answer checking, delayed feedback, or sparse rubric marks. Modern multimodal AI systems change this constraint: they can ingest a scan of a student's work and produce an interpretable response—imperfectly, but well enough to enable new workflows where the derivation (the way) stays central.

Below are two complementary cases: AI-supported homework feedback (formative) and AI-assisted exam grading (summative), both designed around the same principle: automate only what we can justify, and keep humans accountable for validity and fairness.

Case 1: Homework feedback

The Ethel project describes a practical pathway for giving students feedback on handwritten homework in large-enrollment courses. Students submit scanned PDFs; the system converts the handwriting into a structured representation [8]. The key design choice is that the AI is not asked to "solve the problem from scratch" in a vacuum. Instead, the workflow injects the problem text and an instructor-provided sample solution so the feedback is anchored in the course's notation, definitions, and expected reasoning [8].

A second design lesson is tonal but important: early iterations prompted the system to address students directly and be "encouraging," but the prompts were revised toward more impersonal, task-focused feedback. The authors report that anthropomorphizing quickly wore off and sometimes led to patronizing phrasing or unsolicited study advice — an unhelpful distraction from the goal of clear, actionable commentary.

What did students think? In the reported deployment, students rated the feedback as helpful and correct in about three-quarters of cases. The dominant weakness was not conceptual physics reasoning but handwriting recognition: the system tended to *underestimate* correctness when the OCR/interpretation step misread what students wrote, and students rated recognition accuracy as only about half in those cases.

For our purposes, the takeaways are straightforward and actionable:

- Formative feedback on handwritten work is now feasible at scale, especially when the AI is anchored to course-specific reference materials and sample solutions rather than operating "free-form."
- The practical failure modes often sit "upstream" (recognition and parsing), so any deployment must be designed with transparency and graceful failure: uncertain cases should be flagged, not forced.

Case 2: Exam grading

The summative context is different: exams are high-stakes, and the tolerance for unfairness is low. We treat AI grading as a human-in-the-loop process where the central question is not "can AI grade?" but "when can we trust it, and when must a human intervene?" [9]

The study introduces a practical reliability dial: rubric-level grades are generated by the model, but acceptance is governed by threshold parameters in an independent test-theoretical analysis (a correctness threshold and an uncertainty threshold using Item Response Theory), so that only sufficiently reliable judgments are auto-accepted [9]: *accept only high-confidence "correct" judgments*, while routing "incorrect/uncertain" judgments to humans, albeit at the cost of increasing the instructor's involvement.

We find configurations around $R^2 \approx 0.91$ between TA- and AI-assigned scores when auto-grading roughly half of the grading decisions, and $R^2 \approx 0.96$ when auto-grading about one fifth — illustrating how instructors can choose their balance between workload reduction and conservatism.

There's no magic, though. Attempts to iteratively improve grading rules via detection of problem parts that seemed to have unusual scores (apparently too easy, too hard, *etc.*) produced only minimal gains. Human rubric construction, proofreading, and judgment remain essential—particularly because students can produce diverse solution paths, and handwritten work can include ambiguity that the model interprets inconsistently. In other words, psychometrics helps quantify grading validity and manage risk, but it does not eliminate the need for careful assessment design.

In the coming semester, we will use this approach operationally—not as full automation, but as conservative triage. The workflow is deliberately designed to keep learners and educators in control:

1. AI proposes grades only where it meets strict acceptance thresholds.
2. Students are empowered as the first line of accountability: they can veto the AI grading of any problem part they believe was misjudged.
3. Vetoed parts go to a teaching assistant for final adjudication; that judgment is final.

This "student veto" mechanism does two things simultaneously. It makes the system more legitimate (students are not trapped by machine judgment), and it bounds workload by escalating only contested parts. Our current calibration yields <2% false positives on accepted "correct" judgments, which is why we are confident we are

not "giving away the farm." At the same time, we will be explicit with students about an uncomfortable truth: TA review is the gold standard, but it is not infallible—there will be cases where a human second look disagrees with (and may even be harsher than) the AI's initial decision.

Implications and Recommendations

The integration of artificial intelligence into physics education is not a distant prospect - it is already reshaping classroom practices. Our experience with AI-based tools in the physics classroom demonstrates clear benefits: reduced waiting times for assistance, fast feedback, and opportunities for personalized learning. However, these advances do not eliminate the need for human oversight. Generative models remain prone to occasional inaccuracies and inconsistencies, particularly when interpreting experimental data or applying nuanced judgment. At the same time, these developments raise important questions about pedagogy, ethics, and the evolving role of educators.

Based on our experience:

- Effective AI integration favors seamless incorporation into existing workflows without mandating use.
- Human oversight remains essential, especially for summative assessment.
- Thoughtful task design and flexible answer expectations help avoid misleading feedback.
- Students should be supported in developing critical AI literacy.
- Ethical, transparent practices addressing privacy, bias, and equity are necessary to sustain trust.

Looking Ahead

AI will not replace educators, but it will redefine their role. Teachers will increasingly act as designers of learning environments where AI augments instruction rather than dictates it. The challenge, and opportunity, lies in using these tools to enhance engagement and conceptual understanding while safeguarding the integrity of physics education. ■

About the Authors

Marina Babayeva obtained her PhD from the Department of Physics Education at Charles University in Prague, Czech Republic. Her research focuses on technology-enhanced physics learning and connecting theory, practice, and classroom experience.

Ralf Widenhorn is an education researcher in the Department of Physics at Portland State University. His research interests are physics lab instruction, physics for life science students, and the use of technology in education.

Gerd Kortemeyer is a member of the rectorate of ETH Zurich. He is also an Associate Professor Emeritus at Michigan State University. His research focuses on technology-enhanced learning of STEM disciplines.

References

[1] G. Kortemeyer, *Physical Review Physics Education Research* **19(1)**, 010132 (2023)

[2] G. Kortemeyer, *The Physics Teacher* **64(1)**, 8 (2026)

[3] A. Van Heuvelen, *American Journal of physics* **59(10)**, 891 (1991)

[4] C. Wieman & K. Perkins, *Physics today* **58(11)**, 36 (2005)

[5] K. E. Gray, W. K. Adams, C. E. Wieman & K. K. Perkins, *Physical Review Special Topics—Physics Education Research* **4(2)**, 020106 (2008)

[6] T. Kregear, M. Babayeva & R. Widenhorn, *Analysis of student interactions with a large language model in an introductory physics lab setting. International Journal of Artificial Intelligence in Education* (2025). https://doi.org/10.1007/s40593-025-00489-3.

[7] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson & J. Yang, *Physical review physics education research* **12(1)**, 010130 (2016)

[8] G. Kortemeyer, *The Physics Teacher* **62(8)**, 698 (2024)

[9] G. Kortemeyer & J. Nöhl, *Physical Review Physics Education Research* **21(1)**, 010136 (2025)