Artificial Intelligence Makes the Grade

Language models such as ChatGPT could help university educators provide more consistent and transparent grades for introductory-level physics exams.

By **Susan Curtis**

Much has been discussed about students using artificial-intelligence (AI)-powered chatbots to help write assignments. But physics educators believe that more positive learning outcomes could be achieved by using these powerful language models to improve the assessment of students' work. Some studies have already shown that AI-based language tools can provide fast and accurate grading solutions, particularly for short answers to single-component questions. Now Zhongzhou Chen and Tong Wan at the University of Central Florida have shown that ChatGPT can help university instructors assess more complex answers to introductory-level physics problems, resulting in more consistent grading and more personalized feedback to students [**1**].

Using tools such as ChatGPT to assist the grading process could enable educators to introduce more meaningful assessments of students' work.
Credit: andreaobzerova/stock.adobe.com

Models that can process and generate natural language are particularly suited to grading tasks. These models can generate an outcome from a text-based prompt that describes the problem that has been set, the grading criteria and requirements, and the student's response. Several proof-of-concept studies have applied these models to different assessment scenarios, ranging from simple yes–no answers to more complex, multistep responses, and have shown that machine-generated grades can be as accurate as those provided by human instructors. Common strategies to improve outcomes have been instructing the model to "think through" the reasoning steps before producing the grade, providing some examples of responses and the grades they scored, and selecting the most frequent result from multiple grading attempts.

In their new study, Chen and Wan tested several of these strategies for grading multistep problems that were set in two exams for an introductory-level university course on Newtonian mechanics. Students were asked to explain the reasoning that led to their final solution by providing a written response that included text-based expressions of scientific formulas.

While previous studies have scored such long-format answers on a continuous scale, Chen and Wan sought to capture more detail by defining multiple grading criteria that each assessed a specific component of the answer. Each of these criteria were awarded either 0 or 1 points, which were added up to produce

the grade for that problem. The researchers also decided not to provide the model with any reference examples, which some earlier work has suggested can reduce the grading accuracy for more complicated responses. They ran the grading process five times to obtain the most common outcome, which they found delivered a clear improvement in results.

Chen and Wan compared the machine-generated grades from almost 100 student responses to those produced by two experienced instructors. Initial results suggested that the grading criteria were not specific enough for the model to recognize the variability in the answers, such as different ways of writing mathematical expressions. Once the grading descriptors had been updated to reflect this variability, some 70%–80% of the grades generated by the model agreed with those provided by the two instructors—similar to the level of agreement between the two human graders.

The variance in the grades produced across the five runs of the model was also used to generate a confidence index. Machine-generated grades with a low confidence rating, which accounted for around 10%–15% of the total, were reviewed by expert instructors. While most grades returned by the model were accurate, Chen and Wan found that this checking process identified around 40% of those that were potentially incorrect. With human instructors typically taking 2 or 3 hours to grade 100 student responses, this approach would reduce the hands-on effort to about 15 or 20 minutes.

Finally, the large-language model was tasked with providing feedback on each student response, explaining how the answer addressed each element of the grading scheme. Expert instructors then rated the quality of the feedback messages, which in more than 87% of cases were good enough to provide directly to students with only minor modifications. “It would be impossible for a human grader to provide such targeted feedback to each student, but we were surprised at how easy it was for the AI tool to provide personalized messages that improve the transparency of the grading process,” Chen says.

With a cost of around $5 to grade and provide feedback for 100 student responses, Chen and Wan conclude that AI-assisted grading could save both time and money while maintaining the same grading quality. Gerd Kortemeyer, an expert in the use of AI in education at the Swiss Federal Institute of Technology (ETH) Zurich, agrees that the study “offers further proof that large-language models can be used to assist human graders with giving points and feedback to solutions of open-ended physics problems.” In the longer term, the aim would be to exploit the efficiencies enabled by AI-assisted grading to introduce different types of questions or tasks that could improve learning outcomes for physics students. “Work like this has great promise to provide meaningful assessment at scale,” Kortemeyer says.

Susan Curtis is a freelance science writer based in Bristol, UK.

REFERENCES

1. Z. Chen and T. Wan, “Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy,” **Phys. Rev. Phys. Educ. Res. 21, 010126 (2025).**