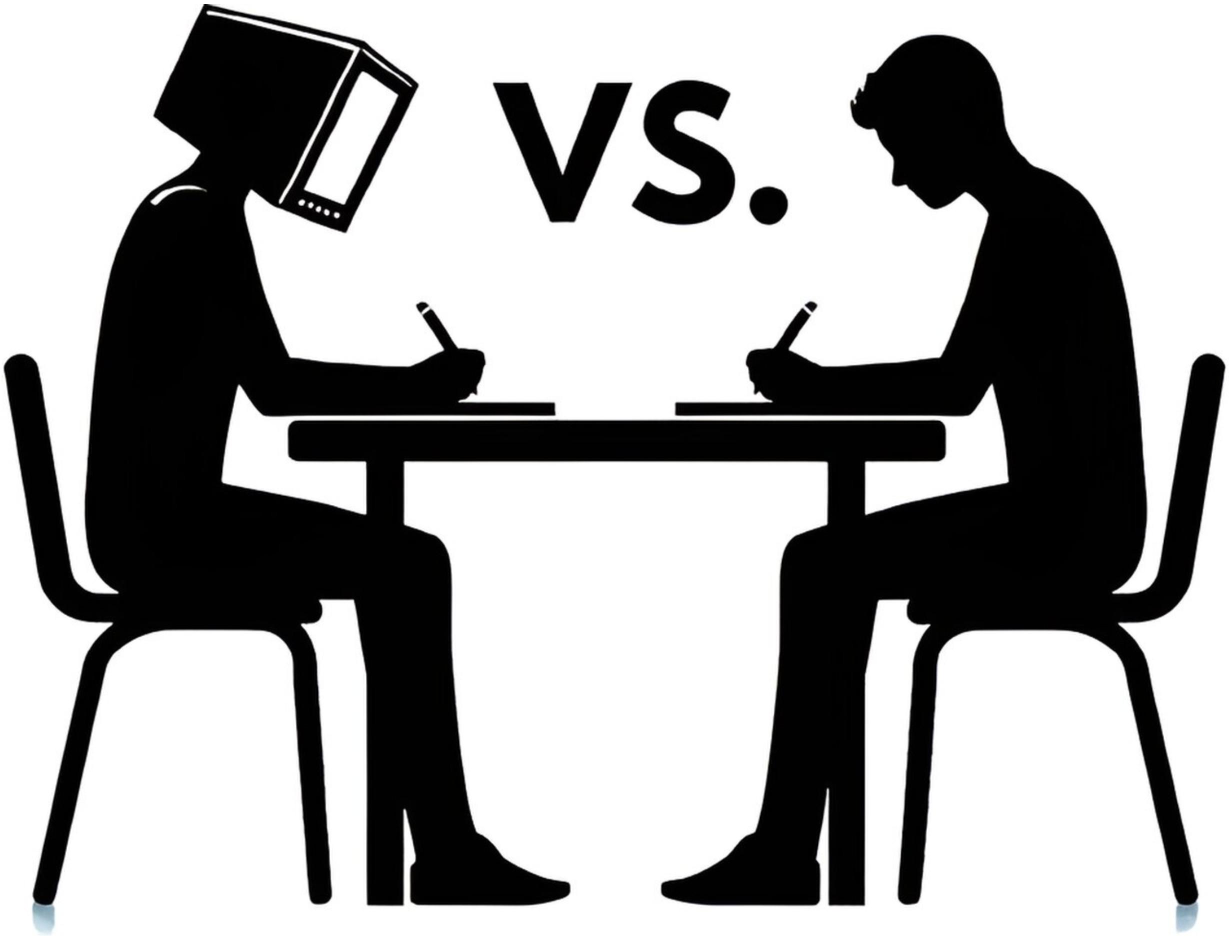
PHYS.ORG

Statistical analysis can detect when ChatGPT is used to cheat on multiple-choice chemistry exams

August 14 2024, by McKenzie Harris



Credit: *Journal of Chemical Education* (2024). DOI: 10.1021/acs.jchemed.4c00165

As the use of generative artificial intelligence continues to extend into all reaches of education, much of the concern related to its impact on cheating has focused on essays, essay exam questions and other narrative assignments. Use of AI tools such as ChatGPT to cheat on multiple-choice exams has largely gone ignored.

A Florida State University chemist is half of a research partnership whose latest work is changing what we know about this type of cheating, and their findings have revealed how the use of ChatGPT to cheat on general chemistry multiple-choice exams can be detected through specific statistical methods. The work was published in *Journal of Chemical Education*.

"While many educators and researchers try to detect AI assisted cheating in essays and open-ended responses, such as Turnitin AI detection, as far as we know, this is the first time anyone has proposed detecting its use on multiple-choice exams," said Ken Hanson, an associate professor in the FSU Department of Chemistry and Biochemistry. "By evaluating differences in performances between student- and ChatGPT-based multiple-choice chemistry exams, we were able to identify ChatGPT instances across all exams with a false positive rate of almost zero."

Researchers collected previous FSU student responses from five semesters worth of exams, input nearly 1,000 questions into ChatGPT and compared the outcomes. Average score and raw statistics were not enough to identify ChatGPT-like behavior because there are certain questions that ChatGPT always answered correctly or always answered incorrectly resulting in an overall score that was indistinguishable from students.

"That's the thing about ChatGPT—it can generate content, but it doesn't necessarily generate correct content," Hanson said. "It's simply an answer generator. It's trying to look like it knows the answer, and to someone who doesn't understand the material, it probably does look like a correct answer."

By using fit statistics, researchers fixed the ability parameters and refit the outcomes, finding ChatGPT's response pattern was clearly different from that of the students.

On exams, high-performing students frequently answer difficult and easy questions correctly, while average students tend to answer some difficult questions and most easy questions correctly. Low-performing students typically only answer easy questions correctly. But on repeated attempts by ChatGPT to complete an exam, the AI tool sometimes answered every easier question incorrectly and every hard question correctly. Hanson and Sorenson used these behavior differences to detect the use of ChatGPT with almost 100-percent accuracy.

The duo's strategy of employing a technique known as Rasch modeling and fit statistics can be readily applied to any and all generative AI chat bots, which will exhibit their own unique patterns to help educators identify the use of these chat bots in completing multiple-choice exams.

The research is the latest publication in a seven-year collaboration between Hanson and machine learning engineer Ben Sorenson.

Hanson and Sorenson, who first met in third grade, both attended St. Cloud State University in Minnesota for their undergraduate degrees and stayed in touch after moving into their careers. As a faculty member at FSU, Hanson became curious about measuring how much knowledge his students retained from lectures, courses and lab work.

"This was a conversation that I brought to Ben, who's great with statistics, computer science and data processing," said Hanson, who is part of a group of FSU faculty working to improve student success in gateway STEM courses such as general chemistry and college algebra. "He said we could use statistical tools to understand if my exams are good, and in 2017, we started analyzing exams."

The core of this Rasch model is that a student's probability of getting any test question correct is a function of two things: how difficult the question is and the student's ability to answer the question. In this case, a student's ability refers to how much knowledge they have and how many of the necessary components are needed to answer the question they have. Viewing the outcomes of an exam in this way provides powerful insights, researchers said.

"The collaboration between Ken and I, though remote, has been a really seamless, smooth process," Sorenson said. "Our work is a great way to provide supporting evidence when educators might already suspect that cheating may be happening. What we didn't expect was that the patterns of artificial intelligence would be so easy to identify."

More information: Benjamin Sorenson et al, Identifying Generative Artificial Intelligence Chatbot Use on Multiple-Choice, General Chemistry Exams Using Rasch Analysis, *Journal of Chemical Education* (2024). DOI: 10.1021/acs.jchemed.4c00165

Provided by Florida State University

Citation: Statistical analysis can detect when ChatGPT is used to cheat on multiple-choice chemistry exams (2024, August 14) retrieved 14 August 2024 from https://phys.org/news/2024-08-statistical-analysis-chatgpt-multiple-choice.html