ChatGPT is still no match for humans when it comes to accounting

April 20 2023, by Todd Hollingshead

In a test of ChatGPT's ability to handle accounting assessments, it still couldn't compete with the student's level. Credit: Nate Edwards/BYU

Last month, OpenAI launched its newest AI chatbot product, GPT-4. According to the folks at OpenAI, the bot, which uses machine learning to generate natural language text, passed the bar exam with a score in the 90th percentile, passed 13 of 15 AP exams, and got a nearly perfect

score on the GRE Verbal test.

Inquiring minds at Brigham Young University (BYU) and 186 other universities wanted to know how OpenAI's tech would fare on accounting exams. So, they put the original version, ChatGPT, to the test. Their research is described in *Issues in Accounting Education*.

The researchers say that while it still has work to do in the realm of accounting, it's a game changer that will change the way everyone teaches and learns—for the better.

"When this technology first came out, everyone was worried that students could now use it to cheat," said lead study author David Wood, a BYU professor of accounting. "But opportunities to cheat have always existed. So for us, we're trying to focus on what we can do with this technology now that we couldn't do before to improve the teaching process for faculty and the learning process for students. Testing it out was eye-opening."

Since its debut in November 2022, ChatGPT has become the fastest growing technology platform ever, reaching 100 million users in under two months. In response to intense debate about how models like ChatGPT should factor into education, Wood decided to recruit as many professors as possible to see how the AI fared against actual university accounting students.

His co-author recruiting pitch on social media exploded: 327 co-authors from 186 educational institutions in 14 countries participated in the research, contributing 25,181 classroom accounting exam questions. They also recruited undergrad BYU students (including Wood's daughter, Jessica) to feed another 2,268 textbook test bank questions to ChatGPT. The questions covered accounting information systems (AIS), auditing, financial accounting, managerial accounting and tax, and varied

in difficulty and type (true/false, multiple choice, short answer, etc.).

Although ChatGPT's performance was impressive, the students performed better. Students scored an overall average of 76.7%, compared to ChatGPT's score of 47.4%. On a 11.3% of questions, ChatGPT scored higher than the student average, doing particularly well on AIS and auditing. But the AI bot did worse on tax, financial, and managerial assessments, possibly because ChatGPT struggled with the mathematical processes required for the latter type.

When it came to question type, ChatGPT did better on true/false questions (68.7% correct) and multiple-choice questions (59.5%), but struggled with short-answer questions (between 28.7% and 39.1%). In general, higher-order questions were harder for ChatGPT to answer. In fact, sometimes ChatGPT would provide authoritative written descriptions for incorrect answers, or answer the same question different ways.

"It's not perfect; you're not going to be using it for everything," said Jessica Wood, currently a freshman at BYU. "Trying to learn solely by using ChatGPT is a fool's errand."

The researchers also uncovered some other fascinating trends through the study, including:

- ChatGPT doesn't always recognize when it is doing math and makes nonsensical errors such as adding two numbers in a subtraction problem, or dividing numbers incorrectly.
- ChatGPT often provides explanations for its answers, even if they are incorrect. Other times, ChatGPT's descriptions are accurate, but it will then proceed to select the wrong multiple-choice answer.
- ChatGPT sometimes makes up facts. For example, when

providing a reference, it generates a real-looking reference that is completely fabricated. The work and sometimes the authors do not even exist.

That said, authors fully expect GPT-4 to improve exponentially on the accounting questions posed in their study, and the issues mentioned above. What they find most promising is how the chatbot can help improve teaching and learning, including the ability to design and test assignments, or perhaps be used for drafting portions of a project.

"It's an opportunity to reflect on whether we are teaching value-added information or not," said study co-author and fellow BYU accounting professor Melissa Larson. "This is a disruption, and we need to assess where we go from here. Of course, I'm still going to have TAs, but this is going to force us to use them in different ways."

More information: The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?, *Issues in Accounting Education* (2023). DOI: 10.2308/ISSUES-2023-013

Provided by Brigham Young University