11-2023

The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?

David A. Wood
Brigham Young University

Muskan P. Achhpilia
Radboud University Nijmegen

Mollie T. Adams
Missouri State University - West Plains

Sanaz Aghazadeh
Louisiana State University

Elizabeth D. Almer
Portland State University, ealmer@pdx.edu

Authors

David A. Wood, Muskan P. Achhpilia, Mollie T. Adams, Sanaz Aghazadeh, Elizabeth D. Almer, and multiple additional authors

The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?

David A. Wood (Corresponding Author), Brigham Young University, davidwood@byu.edu
Muskan P. Achhpilia, Radboud University
Mollie T. Adams, Missouri State University
Sanaz Aghazadeh, Louisiana State University
Kazeem Akinyele, University of Wisconsin Oshkosh
Mfon Akpan, Methodist University
Kristian D. Allee, University of Arkansas
Abigail M. Allen, Brigham Young University
Elizabeth D. Almer, Portland State University
Daniel Ames, Bryant University
Viktor Arity, RMIT University
Dereck Barr-Pulliam, University of Louisville
K. Asli Basoglu, University of Delaware
Andrew Belnap, University of Texas at Austin
Jeremiah W. Bentley, University of Massachusetts Amherst and Tilburg University
Terje Berg, Norwegian University of Science and Technology
Nathan R. Berglund, Mississippi State University
Erica Berry, University of San Diego
Avishek Bhandari, University of Wisconsin - Whitewater
Md Nazmul Hasan Bhuyan, North Carolina A&T State University
Paul W. Black, Auburn University
Eva Blondeel, Ghent University
David Bond, University of Technology Sydney
Annika Bonrath, University of Duisburg-Essen
A. Faye Borthick, Georgia State University
Erik S. Boyle, Idaho State University
Marianne Bradford, North Carolina State University
Duane M. Brandon, Auburn University
Joseph F. Brazel, North Carolina State University
Bryan G. Brockbank, Oklahoma State University
Marcus Burger, University of North Carolina at Pembroke
Dmitri Byzalov, Temple University
James N. Cannon, Utah State University
Cecil Caro, Unaffiliated
Abraham H. Carr, Creighton University
Jack Cathey, University of North Carolina at Charlotte
Ryan Cating, University of Central Arkansas
Kimberly Charron, University of Nevada, Las Vegas
Stacy Chavez, Loyola University Maryland
Jason Chen, Central Connecticut State University

Jennifer C. Chen, Brigham Young University - Hawaii
Jennifer W. Chen, University of Missouri - St. Louis
Christine Cheng, The University of Mississippi
Xu Cheng, Auburn University
Brant E. Christensen, Brigham Young University
Kimberly Swanson Church, Missouri State University
N. J. Cicone, Mississippi State University
Patience Constance, University of Kansas
Lauren A. Cooper, West Virginia University
Candice L. Correia, Whitworth University
Joshua Coyne, Texas Tech University
W. Alec Cram, University of Waterloo
Asher Curtis, University of Washington
Ronald J. Daigle, Sam Houston State University
Steve Dannemiller, University of Alabama
Stephan A. Davenport, University of Tennessee at Chattanooga
Gregory S Dawson, Arizona State University
Karen J. L. De Meyst, Radboud University
Scott Dell, Francis Marion University
Sebahattin Demirkan, Manhattan College
Christine A. Denison, Iowa State University
Hrishikesh Desai, Arkansas State University
Steven DeSimone, College of the Holy Cross
Leah M. Diehl, University of Montana
Ruth Dimes, University of Auckland
Bei Dong, University of South Florida
Amy Donnelly, University of Missouri-Kansas City
Adam du Pon, Georgia Southern University
Huijue Kelly Duan, Sacred Heart University
Ada Duffey, University of Wisconsin-Stevens Point
Ryan T. Dunn, Auburn University
Mary P. Durkin, University of San Diego
Ann C. Dzuranin, Northern Illinois University
Rachel M. Eberle, University of Arkansas
Matthew S. Ege, Texas A&M University
Dina El Mahdy, Morgan State University
Adam Esplin, The University of Texas at El Paso
Marc Eulerich, University Duisburg-Essen
Patricia Everaert, Ghent University
Nusrat Farah, Southern Illinois University Carbondale
Lauren Farish, Baylor University
Michael Favere-Marchesi, Simon Fraser University
Dutch Fayard, University of Tennessee at Chattanooga
Jessica R. Filosa, Virginia Tech
Melinda Ford, Southern Utah University
Diana R. Franz, University of Toledo

Bachman P. Fulmer, University of Tampa
Sarah Fulmer, University of Tampa
Zhan Z. Furner, East Carolina University
Sonia Gantman, Bentley University
Steve Garner, Tennessee Tech University
Jace Garrett, Clemson University
Xin Geng, Berry College
Joanna Golden, University of Memphis
William Goldman, Northeastern University
Josue Gomez, Brigham Young University
Mark Gooley, Northeastern University
Shawn P. Granitto, Florida Gulf Coast University
Karen Y. Green, University of Toledo
Cindy L. Greenman, Utah Tech University
Gaurav Gupta, UNC Wilmington
Ronald N. Guymon, University of Illinois at Urbana-Champaign
Kevin Hale, UNC Wilmington
Christopher J. Harper, Grand Valley State University
S. Allen Hartt, Boise State University
Holly Hawk, Clemson University
Steven R. Hawkins, Southern Utah University
Erin M. Hawkins, Clemson University
David C. Hay, University of Auckland
Rafael Heinzelmann, University of Agder, School of Business and Law
Cassy D. Henderson, Sam Houston State University
Bradley E. Hendricks, University of North Carolina at Chapel Hill
William G. Heninger, Brigham Young University
Mary S. Hill, Kennesaw State University
Nicole Holden, Arizona State University
D. Kip Holderness Jr., West Virginia University
Travis P. Holt, Auburn University
Jeffrey L. Hoopes, University of North Carolina at Chapel Hill
Sheng-Feng Hsieh, National Taiwan University
Feiqi Huang, Pace University
Hua-Wei Huang, National Cheng Kung University
Ting-Chiao Huang, Monash University
Brian W. Huels, University of Wisconsin - Whitewater
Kara Hunter, Fairfield University
Patrick J. Hurley, Northeastern University
Kerry Inger, Auburn University
Sharif Islam, Southern Illinois University Carbondale
Isaac Ison, Truman State University
Hussein Issa, Rutgers University
Andrew B. Jackson, UNSW Sydney
Scott C. Jackson, University of South Dakota
Diane J. Janvrin, Iowa State University

Peggy D. Jimenez, University of North Texas
Daniel Johanson, Norwegian School of Economics, NHH
J. Scott Judd, University of Illinois at Chicago
Brett S. Kawada, Woodbury University
Andrea Seaton Kelton, Middle Tennessee State University
Sara Kern, Gonzaga University
Jon N. Kerr, Brigham Young University
Marsha B. Keune, University of Dayton
Mindy (Hyo Jung) Kim, George Mason University
Robert J. Knisley, Wheaton College
Brian D. Knox, Boise State University
Gregory Kogan, Virginia Tech University
Amr Kotb, Zayed University
Ronja Krane, University of Duisburg-Essen
Joleen Kremin, Portland State University
Kimberly S. Krieg, University of San Diego
Jonathan Kugel, Christopher Newport University
Ellen M. Kulset, University of South-Eastern Norway
Chamara Kuruppu, University of South-Eastern Norway
Garrison LaDuca, Samford University
Barbara A. Lamberton, University of Hartford
Melvin A. Lamboy-Ruiz, Georgia Southern University
Bradley Lang, Mississippi State University
Stephannie A. Larocque, University of Notre Dame
Melissa P. Larson, Brigham Young University
Bradley P. Lawson, Oklahoma State University
James G. Lawson, Bucknell University
Lorraine Lee, UNC Wilmington
Margarita M. Lenk, Colorado State University emeritus
Michelle Li-Kuehne, Whitworth University
Jonathan Liljegren, Brigham Young University
Yi-Hung Lin, Monash University
Wu-Po Liu, National Cheng Kung University
Zishang Liu, University of Houston Downtown
Brandon Lock, Baruch College–CUNY
James H. Long, Auburn University
Tina Loraas, Auburn University
Suzanne Lowensohn, University of Vermont
Thomas R. Loy, University of Bremen
Hakim Lyngstadaas, BI Norwegian Business School
Wim Maas, Tilburg University
Jason E. MacGregor, Baylor University
Dag Øivind Madsen, University of South-Eastern Norway
Carissa L. Malone, North Carolina State University
Maximilian Margolin, Erasmus University
Mary E. Marshall, Portland State University

Rachel M. Martin, Utah State University
Colleen McClain Mpofu, University of Missouri - St. Louis
Chris McCoy, Florida State University
Nicholas C. McGuigan, Monash University
Dwayne N. McSwain, Appalachian State University
Michele D. Meckfessel, University of Missouri-St. Louis
Mark J. Mellon, Northern Illinois University
Olivia S. Melton, Augustana College
Julie M. Mercado, Boise State University
Steven Mitsuda, University of Montana
Kennedy Modugu, HCT Business School, Dubai
Stephen Moehrle, University of Missouri - St. Louis
Amirali Moeini Chaghervand, University of Houston-Downtown
Kevin Moffitt, Rutgers University
Joon Seok Moon, California State University, Northridge
Brigitte Muehlmann, Babson College
Johnna Murray, University of Missouri - St. Louis
Emmanuel S. Mwaungulu, Palm Beach Atlantic University
Noah Myers, Utah Valley University
J. Conrad Naegle, Jr., Missouri State University
Martin J. Ndicu, Northern Illinois University
Aaron S. Nelson, The University of Texas at El Paso
Anh L. Nguyen, RMIT University
Thomas Niederkofler, Radboud University
Ehsan Nikbakht, Hofstra University
Ann D. O'Brien, University of Wisconsin-Madison
Kehinde M. Ogunade, University of Memphis
Daniel O'Leary, University of Southern California
Mitchell J. Oler, University of Wyoming
Derek K. Oler, Texas Tech University
Kari Joseph Olsen, Utah Valley University
John I. Otalor, University of Calabar, Calabar, Nigeria
Kyle W. Outlaw, Samford University
Michael. E. Ozlanski, Susquehanna University
Jenny Parlier, University of North Carolina Wilmington
Jeffrey S. Paterson, Florida State University
Christopher A. Pearson, Idaho State University
Michael J. Petersen, North Dakota State University
Steven T. Petra, Hofstra University
Matthew D. Pickard, Northern Illinois University
Jeffrey Pickerd, The University of Mississippi
Robert Pinsker, Florida Atlantic University
Catherine Plante, University of New Hampshire
James M. Plečnik, Loyola Marymount University
Richard A. Price, III, University of Oklahoma
Linda A. Quick, East Carolina University

Jana Raedy, University of North Carolina at Chapel Hill
Robyn Raschke, University of Nevada, Las Vegas
Julie Ravenscraft, Missouri State University
Vernon Richardson, University of Arkansas
Brett A. Rixom, University of Nevada, Reno
John F. Robertson, Arkansas State University
Iyad Rock, Lewis University
Miles A. Romney, Florida State University
Andrea Rozario, University of Illinois at Urbana-Champaign
Michael F. Ruff, Northeastern University
Kathleen Rupley, Portland State University
Ali Saeedi, University of Minnesota - Crookston
Aaron Saiewitz, University of Nevada, Las Vegas
Leigh W. Salzsieder, University of Missouri-Kansas City
Sayan Sarkar, University of Mary Washington
Michael Saulls, Brigham Young University
Tialei A. Scanlan, Brigham Young University - Hawaii
Tammie J. Schaefer, University of Missouri-Kansas City
Daniel Schaupp, WHU - Otto Beisheim School of Management
Gary P. Schneider, California State University, Monterey Bay
Andreas Seebeck, Jacobs University Bremen
R. Drew Sellers, Kent State University
Samantha C. Seto, Simon Fraser University
Romi-Lee Sevel, York University
Yuxin Shan, University of Wisconsin - Eau Claire
Matthew G. Sherwood, University of Massachusetts Amherst
Maggie Singorahardjo, The University of Melbourne
Hanna Kristin Skaftadottir, Bifrost University
Justyna Skomra, Pennsylvania State University, Behrend College
Jason L. Smith, University of Nevada, Las Vegas
Dallin O. Smith, West Texas A&M University
James Smith, University of Lethbridge
Mason C. Snow, California State University, Fullerton
Ryan D. Sommerfeldt, Washington State University
Kate B. Sorensen, University of Memphis
Trevor L. Sorensen, University of Wyoming
Andrew C. Spieler, Hofstra University
Matthew A. Stallings, Boise State University
Lesya Stallings, Kennesaw State University
Alan Stancill, Mississippi State University
Jonathan D. Stanley, Auburn University
Chad M. Stefaniak, University of South Carolina
Nathaniel M. Stephens, Brigham Young University - Hawaii
Bryan W. Stewart, Utah Tech University
Theophanis C. Stratopoulos, University of Waterloo
Daniel A. Street, Bucknell University

Meena Subedi, University of Wisconsin - Whitewater
Scott L. Summers, Brigham Young University
Charlotte H. Sundkvist, University of South-Eastern Norway
Christina Synn, American University
Amanuel Tadesse, University of New Orleans
Gregory P. Tapis, Missouri State University
Kerri Tassin, Missouri State University
Samantha Taylor, Dalhousie University
Mary Teal, Emporia State University
Ryan Teeter, University of Pittsburgh
Meredith Tharapos, RMIT University
Jochen C. Theis, University of Southern Denmark
Jack Thomas, Brigham Young University
Kristen S. Thompson, Idaho State University
Todd A. Thornock, University of Nebraska-Lincoln
Wendy Tietz, Kent State University
Anthony M. Travalent, Lone Star College
Brad S. Trinkle, Iowa State University
J. Mike Truelson, Mississippi State University
Michael C. Turner, University of New Orleans
Brandon Vagner, Middle Tennessee State University
Hamid Vakilzadeh, University of Wisconsin - Whitewater
Jesse van der Geest, Tilburg University
Victor van Pelt, WHU - Otto Beisheim School of Management
Scott D. Vandervelde, University of North Carolina at Charlotte
Jose Vega, Stephen F. Austin State University
Sandra Vera-Muñoz, University of Notre Dame
Brigham Villanueva, Brigham Young University
Nishani Edirisinghe Vincent, University of Tennessee at Chattanooga
Martin Wagener, University of Duisburg-Essen
Stephanie Walton, Louisiana State University
Rick C. Warne, University of San Diego
Olena V. Watanabe, Iowa State University
David Watson, University of South Florida
Marcia Weidenmier Watson, Trinity University
Jill Weber, University of Wisconsin - Whitewater
Thomas Weirich, Central Michigan University
Ashley N. West, Kansas State University
Amanda L. Wilford, IE University
Aaron B. Wilson, Ohio University
Brian Winrow, Florida State University
Timothy Winrow, Eastern New Mexico University
Tasia S. Winrow, Minnesota State University Mankato
Denise Wiseman, University of Central Florida
Annie L. Witte, Northeastern University
Bryan D. Wood, Unaffiliated

Jessica Wood, Brigham Young University
Darryl Woolley, University of Idaho
Nicole S. Wright, James Madison University
Juan Wu, Minnesota State University Mankato
Xiaomei (Grazia) Xiong, Utah State University
Dimitri Yatsenko, University of Wisconsin - Whitewater
Courtney E. Yazzie, Clemson University
Glen M. Young, Texas State University
Chanyuan (Abigail) Zhang, University of Texas at San Antonio
James (Yibo) Zhang, Miami University
Aleksandra B. Zimmerman, Florida State University
Emily Zoet, Ferris State University

Running Head: ChatGPT and Accounting Assessment Questions

The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?

Abstract

ChatGPT, a language-learning model chatbot, has garnered considerable attention for its ability to respond to users' questions. Using data from 14 countries and 186 institutions, we compare ChatGPT and student performance for 28,085 questions from accounting assessments and textbook test banks. As of January 2023, ChatGPT provides correct answers for 56.5 percent of questions and partially correct answers for an additional 9.4 percent of questions. When considering point values for questions, students significantly outperform ChatGPT with a 76.7 percent average on assessments compared to 47.5 percent for ChatGPT if no partial credit is awarded and 56.5 percent if partial credit is awarded. Still, ChatGPT performs better than the student average for 15.8 percent of assessments when we include partial credit. We provide evidence of how ChatGPT performs on different question types, accounting topics, class levels, open/closed assessments, and test bank questions. We also discuss implications for accounting education and research.

Key Words: Artificial intelligence, chat bots, GPT-3, accounting education, assessment, crowdsourcing

I. INTRODUCTION

Open AI, an artificial intelligence (AI) research and development company, released the beta version of ChatGPT for public use on November 30, 2022. ChatGPT is a large language-learning model that utilizes machine learning to generate natural language text (Susnjak 2022). When asked, "What are you?" ChatGPT defines itself as follows:

> I am an artificial intelligence assistant trained to help answer questions and provide information on a wide variety of topics. My primary function is to assist users in generating human-like text based on the prompts given to me. I do not have my own personal experiences or beliefs, and my responses are based on the information that I have been programmed to use. (Retrieved on 12/29/2022)

ChatGPT gained over one million users in just five days and reached 100 million users in less than two months, making it the fastest growing technology platform ever (Harris 2022; Hu

1

2023).[1] By way of comparison, Facebook reached the one million user milestone after 300 days (Firat 2023). According to a survey conducted by Constantz (2023), nearly 30 percent of respondents indicate they have tried using ChatGPT at work. The impact of ChatGPT on education has been the subject of intense debate and discussion, with various authors exploring how it will change education (see for example, McMurtrie 2022; Meckler and Verma 2022; Shrivastava 2022).

We contribute to this discussion by providing data on how well ChatGPT performs in answering accounting assessment questions primarily from class exams and quizzes. We compare ChatGPT's performance on accounting assessments to student averages on the same set of questions, analyzing where it performs better and worse. The results will help academics, administrators, faculty, instructors, and students better understand the potential impact of large language-learning models like ChatGPT on education, not only in the field of accounting, but also in many fields with similar types of assessments such as finance, information systems, management, and operations.

When asked how it was created/trained, ChatGPT reports that it was

> ...trained using advanced machine learning techniques and a large dataset of human-generated text. My training involved being fed large amounts of text and being asked to predict the next word or phrase in a sequence. This process helped me to learn the patterns and structures of human language and to generate text that is similar to the way humans write and speak. My training also included being fed a wide variety of texts on different topics, which helped me to become knowledgeable about a wide range of subjects. (Retrieved on 12/29/2022)

Thus, while ChatGPT's training was not specific to accounting, it likely included material related to the subject. As such, it is unknown how well a "General Purpose Technology" (GPT)

---

[1] Users can create a free account to use the tool, but the company recently unveiled an enhanced service called "ChatGPT Plus" for $20 per month. This premium service provides users with priority access to the tool during periods of high demand. ChatGPT will continue to be free, but nonsubscribers may not always be able to access it (Kelly 2023).

such as ChatGPT will perform on accounting-related content. Furthermore, ChatGPT may not perform well on accounting content, as faculty and textbook authors often design questions to elicit nuanced understanding from accounting students, which may not be comprehended by AI algorithms. Therefore, this study provides evidence of ChatGPT's performance on challenging, context-specific accounting questions.

To evaluate ChatGPT's performance on accounting-specific content, we assembled a crowdsourced team of 328 coauthors who provided data from 186 educational institutions around the world. The institutions represented a range of types and the coauthors include faculty from both U.S. and international institutions. During the months of December 2022 and January 2023 each coauthor entered assessment questions into ChatGPT and evaluated the accuracy of its responses. The study includes a total of 25,817 questions (25,181 gradable by ChatGPT) that appeared across 869 different class assessments, as well as 2,268 questions from textbook test banks covering topics such as accounting information systems (AIS), auditing, financial accounting, managerial accounting, and tax. The questions vary in terms of question type, topic area, and difficulty. The coauthors evaluated ChatGPT's answers to the questions they entered and determined whether they were correct, partially correct, or incorrect.

The results indicate that across all assessments, students scored an average of 76.7 percent, while ChatGPT scored 47.4 percent based on fully correct answers and an estimated 56.5 percent if partial credit was included. Thus, on average, ChatGPT performed worse on assessments in our dataset than students. However, we also find that ChatGPT scored higher than the student average on 11.3 percent (without partial credit) or 15.8 percent (with partial credit) of assessments. The study also revealed differences in ChatGPT's performance based on the topic area of the assessment. Specifically, ChatGPT performed relatively better on AIS and auditing

assessments compared to tax, financial, and managerial assessments. We suggest one possible reason this may occur is that AIS and auditing questions typically do not include mathematical type questions, which ChatGPT currently struggles to answer correctly.

ChatGPT performed better in answering true/false and multiple-choice questions, with full-credit accuracy rates of 68.7 percent and 59.5 percent, respectively.[2] In contrast, ChatGPT struggled with workout and short-answer questions, with accuracy rates of 28.7 percent and 39.1 percent, respectively. In terms of textbook test bank questions, ChatGPT correctly answered 64.3 percent of the time, with the highest accuracy rates for questions in audit (83.1 percent correct) and AIS (76.8 percent correct).

This study provides important insights into the current capabilities of AI compared to human performance in an accounting-specific context. It highlights the limitations of an AI chatbot trained on general material. Nevertheless, the gap in performance between AI and humans will likely close as AI technology improves. AI technology continues to advance at a fast pace. For instance, the current ChatGPT model was trained on 175 billion parameters; however, in 2023, a new chatbot model based on one trillion parameters is likely to be released (IBL News 2023). Thus, while humans currently outperform AI on accounting-related material, it is not unreasonable to expect AI performance to improve, perhaps drastically, over time.

This study provides a novel contribution to the accounting literature by being the first in this field to use crowdsourcing for data collection and paper authorship. While the median number of authors on accounting research papers is three, other disciplines have notably more. For example, natural sciences articles have a median number of nine coauthors (mean of 33), but

---

[2] The lower performance for multiple-choice questions relative to true/false questions may be because of the odds of getting multiple choice questions correct without any knowledge is lower (e.g., one out of four) than getting true/false questions correct (e.g., one out of two).

some papers can have hundreds of coauthors (Wood 2016). The use of crowdsourcing allowed for the hand collection of a very large dataset and writing of an entire paper in two months. This approach may serve as a useful methodology for future studies of broad interest that cannot be completed using more traditional data collection methods. Additionally, the sheer number of active participants on this project suggests that educators and academics are keenly interested in the use of AI in our field and willing to test its bounds.

Perhaps the most important contribution of this paper is to highlight that accounting educators need to prepare for a future that includes broad access to AI to serve their students and the needs of the profession effectively. We believe that accounting educators should engage in discussions about the impact of AI on their teaching. This includes addressing questions such as: How should students be allowed to use AI? What material should be memorized versus referenced? Can interactions with AI enhance students' learning, and if yes, how? What value do educators and accountants provide beyond what AI can provide? These are all important questions that accounting educators should discuss and research. As AI technology continues to improve, educators need to prepare themselves and their students for the future, making AI technology a promising area for future research.

II. LITERATURE REVIEW

There is a rich history of AI research in multiple disciplines, including accounting. In accounting specifically, several review papers have been published on this topic (see Gray, Chiu, Liu, and Li 2014; Sutton, Holt, and Arnold 2016; Zemankova 2019; Han, Shiwakoti, Jarvis, Mordi, and Botchie 2023). Many authors have predicted that AI will significantly impact accounting education (Brink and Reichert 2020; Holmes and Douglass 2022); however, its impact to date has been relatively modest (Baldwin-Morgan 1995; Qasim and Kharbat 2020;

Qasim, El Rafae, and Eletter 2022). The modest impact highlights the importance of continued research in this area to better understand AI's potential impact on the profession and education.

One area of AI that has the potential to impact education and has been thrust into the limelight is chatbots, or large language models. These computer programs are designed to interact with users on a wide range of topics (O'Leary 2022). Companies like OpenAI, Alphabet, Meta, and Microsoft are actively developing this technology. Although previous chatbots have been released to the public and have seen varying degrees of success (Adamopoulou and Moussiades 2020), the newest series of chatbots based on the latest language models have demonstrated high levels of effectiveness (O'Leary 2022). In addition, non-peer reviewed, small sample studies have shown that ChatGPT is able to pass exams of highly technical content, including a Wharton MBA operations management final exam (Terwiesch 2023), the U.S. Medical Licensing Exam (Kung et al. 2022), exams in four law school courses (Choi, Hickman, Monahan, and Schwarcz 2023), and the Evidence and Torts portions of the U.S. multistate bar exam (Bommarito and Katz 2023).

Given the documented ability of recently released chatbots, we test the accuracy of the use of chatbots in the accounting domain. Our focus is examining the ability of ChatGPT to answer a large sample of questions from accounting assessments and textbook test banks. The results of our study have significant implications for the field of accounting education. On the one hand, if chatbots such as ChatGPT perform well in answering accounting questions, they could provide a new means for students to cheat.[3] On the other hand, chatbots could also be used as a positive tool to help students generate practice problems, deepen their understanding of accounting content, and improve their learning. The effectiveness of chatbots in performing

---

[3] Conaway and Wiesen (2023) find that 13 to 25 percent of intermediate accounting students use Chegg.com to cheat during exams.

either positive or negative tasks ultimately depends on their accuracy in determining answers to accounting content. Finally, chatbots' ability to correctly answer accounting assessment questions also has potential implications for perceptions of AI's ability to replace accountants in practice. The possibility of AI replacing accounting has been widely discussed (Moffitt, Richardson, Snow, Weisner, and Wood 2016; Frey and Osborne 2017).

While many different chatbot technologies exist, we focus on Open AI's ChatGPT. ChatGPT is a widely publicized and freely accessible chatbot based on a generative pre-trained transformer 3 (GPT-3) deep learning model. It is trained to predict the next "token" (i.e., word) in response to a query, allowing a user to enter any question and receive a response. Although ChatGPT is the focus of this study, other chatbots, either available now or soon to be available, are expected to perform at a similar level.[4]

There has been significant interest in the uses and effectiveness of ChatGPT, particularly in the workplace. Only two months after its release on November 30, 2022, over 500 English language articles on ChatGPT were published in newspapers. As can be seen in Figure 1, many of these articles, particularly a quarter of the news articles in the first two months, focused on ChatGPT's influence and its effects on education and the classroom (particularly its ability to answer questions and its possible use to "cheat" on exams). Although a few examples exist of how ChatGPT answers accounting questions, we are the first to provide a systematic and comprehensive examination of ChatGPT's ability to answer accounting questions.[5]

---

[4] There are many other similar technologies to ChatGPT such as Albert, Bert, Meena, Pegasus, T5, and XLNet by Google; Blender and RoBERTa by Facebook, DialoGPT by Microsoft, and CTRL by Salesforce (see Agomuoh 2023).

[5] For example, see CPA Exam Guide (2023) and Herbert (2023).

III. METHODOLOGY

We relied on questions utilized by accounting faculty in classroom assessments. We invited participants using a snowball sampling method where the first author invited personal contacts and posted on social media, and then each subsequent coauthor was encouraged to invite other faculty to participate. The goal was to select a diverse representation of institutions. A total of 328 authors from 186 different institutions, representing 14 different countries on five continents participated in the study. Table 1 lists these institutions and the number of assessments entered for each institution.[6]

The faculty logged into a shared Google Sheet, where they entered information about one or more assessments from their classes providing the following: information about the topic of the class, the institution where the class was taught, the class level (we group this into Freshman and Sophomore, Junior and Senior, and Graduate), the topic area of the assessment, whether the assessment was open- or closed-book (or if it included a mixture of the two), the time limit (if any) for the assessment, the country of the institution, the language the exam was given in, the average score of the students, the standard deviation for the students, and how many students took the assessment.[7] Not all faculty provided answers to each question, so we include all responses with non-null values for each analysis.

Faculty tested ChatGPT's performance on their assessments by registering on https://chat.openai.com/chat and entering each question into the chatbot. Some questions, such as questions about complying with ethical codes of conduct, questions containing images, and

---

[6] In a few situations, faculty entered data from a previous institution. Table 1 shows the institutions where the data came from, not current faculty locations.

[7] After the initial data collection was performed, the data was removed from the shared Google Sheet. One coauthor then emailed each faculty their individual data and asked them to review the data. Coauthors could then report errors, and the final dataset corrected any errors.

questions about specific class activities, could not be answered by ChatGPT. A total of 636 out of 25,817 total questions were deemed unanswerable by ChatGPT. For analysis purposes, we retain the point values of these questions in all analyses, but ChatGPT receives zero points for these types of questions.

Each faculty member evaluated their own submitted questions. Specifically, they identified the question type (e.g., multiple-choice, true/false, etc.), indicated whether it was custom written or gathered from a test bank, and provided the point value.[8] Next, they examined the answers provided by ChatGPT and determined whether the bot answered correctly or incorrectly. Faculty also decided whether incorrect answers would receive partial credit. Partially correct answers received half of the available points for the question. We present our findings both with and without partial credit.

To compare the bot's performance with that of students, each faculty member provided student performance data, including the mean student score, the standard deviation, and the number of students who took the assessment. We subsequently compared ChatGPT's performance to the student average on each exam.

Finally, we note that ChatGPT was updated several times during the two months of data collection. To ensure our results were not influenced by changes to the model, we divided our sample into deciles based on time and compared the accuracy rates across the deciles. The accuracy rate in the first decile (those gathered earliest) was 57.3 percent, and in the tenth decile was 55.9 percent. The trendline fitted to the deciles shows a less than 0.02 percent increase in accuracy rate over the ten deciles. Therefore, we conclude that ChatGPT did not meaningfully improve in terms of its accounting knowledge over the short data collection period.

---

[8] If faculty customized a test bank question, they were asked to indicate that it was a custom question.

IV. RESULTS

We present our results through a combination of figures and tables. Table 2 provides descriptive statistics of the sample, including the number of assessments and questions broken down by class level, type of assessment, topic area, and note-using policy.

Figure 2 compares the overall performance of students and ChatGPT on assessments. We include two measures of ChatGPT's performance, one without partial credit (No PC) and one including partial credit (PC). In each figure, we include 83 percent confidence intervals, which show the point at which the bars cannot overlap if the difference is significant at the 95 percent level (Gubler, Herrick, Price, and Wood 2016). As shown in Figure 2, students outperform ChatGPT in all scenarios, with a meaningful difference of nearly 30 points when no partial credit is allowed for ChatGPT (No PC).

Figure 3 examines ChatGPT's performance across different topic areas. The chart indicates that ChatGPT performs relatively well in AIS and audit, scoring above 60 percent (with partial credit), with no significant differences from the student mean score when partial credit is awarded. Conversely, ChatGPT's lowest scores are in the areas of financial, managerial, and tax.[9]

Figures 4 and 5 present the performance of ChatGPT based on student class level and the note-using policy, respectively. Across class levels, ChatGPT's performance is relatively similar when partial credit is included. It does worse on open-note assessments. This could be due to the emphasis on application of concepts rather than memorization in open-note assessments.

---

[9] Later in the paper, we conduct a regression analysis to hold constant other factors and analyze what impacts ChatGPT accuracy. Still, one possibility for the difference between areas that the regression does not consider is that some areas (e.g., financial, managerial, and tax) may include more mathematical type questions within a question type. For example, AIS and Audit multiple-choice questions may be more conceptual, whereas the other areas may require math in the multiple-choice question. Given how ChatGPT struggles with some math questions, this may partially explain differences between areas. We encourage future research to analyze this conjecture.

As an alternative way to present the data, we show in Table 3 the percentage of times that ChatGPT performs higher than the student mean score. ChatGPT (PC) outperforms the student average on 15.8 percent of assessments (the outperformance is 11.3 percent for ChatGPT (No PC)). We observed a similar pattern of results as discussed for class level, topic area, and note-using policy.

Table 4 provides a more detailed analysis of ChatGPT's performance by examining the questions rather than aggregating the scores to the assessment level.[10] ChatGPT answered 56.5 percent of the questions correctly and an additional 9.4 percent as partially correct. The results show significant variation in ChatGPT's performance based on question type. ChatGPT performs best on dichotomous (e.g., True/False; 68.7 percent full credit) and multiple-choice (59.5 percent full credit) questions but struggles more on short answer (39.1 percent full credit) and workout (28.7 percent full credit) questions.[11] The relatively poor performance for "full credit" questions on short answer and workout questions is tempered by the relatively high partial credit awarded. For all question types, if both full credit and partial credit are considered, ChatGPT performs above 60 percent for every question type. Additionally, the table shows that faculty use test bank questions for about half of their assessments, and ChatGPT performs about 8 percent better on these questions compared to custom questions.

It is important to note that our results for exam scores (i.e., Figure 2) are sensitive to the score assigned for partial credit.[12] As noted previously, we assigned a score of 50 percent when

---

[10] Since weights for questions differ, the data in Table 4 do not directly tie to the assessment data previously presented.

[11] Although ChatGPT does better on True/False questions than multiple-choice questions, its incremental performance relative to random chance is higher with multiple-choice questions. That is, for true/false questions, ChatGPT's accuracy is 18.7 percent higher in absolute terms than random chance (68.7 - 50), whereas the bot is 34.5 percent more accurate for multiple choice (59.5 - 25), assuming multiple-choice questions have four choices.

[12] The results for individual questions (Table 4 analyses) are not sensitive to partial credit since these results only consider whether questions are correct, partially correct, or incorrect and do not factor into the scoring the points assigned to questions.

partial credit was given.[13] Authors recorded partial credit as a dichotomous variable and did not score the ChatGPT response. If, instead of 50 percent, we had assumed that partial credit was 25, 75 or 90 percent, the overall ChatGPT (PC) score reported in Figure 2 would have been 51.9, 61.0, and 63.8 percent, respectively.

In addition to assessments, we also examined textbook test bank questions in all the major areas of accounting, including AIS, auditing, financial, managerial, tax, and other topics (Coyne, Summers, Williams, and Wood 2010).[14] The results for the textbook test bank questions, as presented in Table 5, are similar to the results for assessment questions and, therefore, we do not provide additional commentary.

We conduct a regression analysis to further examine the factors that contribute to ChatGPT's ability to correctly answer questions. We use a linear probability model so the interpretation of the coefficients is more straightforward. Results are similar in terms of statistical significance if we use a logistic regression. The dependent variable is a dummy variable indicating whether ChatGPT (No PC) answered the question correctly or not. Independent variables included dummy variables measuring the different class levels, topic areas, question types, whether the questions came from a test bank, if the exam was open-note or not, whether the exam was given in English or another language, and whether the institution is in the United States or not. The results are presented in Table 6.

For ease of interpretation, we grouped the independent variables into categories and the subheadings indicate the variables included in the intercept. The results suggest that, even after

---

[13] We did not measure whether faculty gave students partial credit or not on each question. This is a limitation of our study.

[14] The textbooks we tested are well adopted in each area. We avoid printing textbook names in deference to faculty and textbook authors who indicated a reluctance to alerting students about how well ChatGPT did on specific textbooks.

controlling for other factors, several topic areas and question types are more closely linked to ChatGPT's success in answering questions correctly. The variables with the largest impact include the topic areas being AIS, analytics/technology, and audit classes, which were more likely to be answered correctly than financial topics. By way of practical magnitude, ChatGPT is 20 percent more likely to answer an AIS question correctly than a financial question, all else held constant. The other coefficients can be interpreted in similar fashion. Dichotomous (e.g., T/F) questions were 12.1 percent more likely to be answered correctly than multiple-choice questions, but short-answer and workout questions were less likely to be answered correctly by 16.4 and 24.0 percent, respectively. This analysis can be useful to help guide faculty on designing questions that are more or less likely to be answered correctly by ChatGPT.

We found evidence that ChatGPT does better on exams given in English. ChatGPT does not differ based on the location of the institution. If the exam is given in English, ChatGPT scores 10.4 percent higher than if it is given in a different language.

Given the unique nature of the crowdsourced data collection process, we provide additional anecdotes that individual authors found interesting and wished to highlight. While this list of anecdotes may not be comprehensive as far as the strengths and weaknesses of ChatGPT, it is provided to help guide new users of ChatGPT. These anecdotes also aim to inspire further discussion on the impact and potential of ChatGPT in education, and to stimulate future research in this field.

- During testing, ChatGPT did not always recognize it was performing mathematical operations and made nonsensical errors, such as adding two numbers in a subtraction problem or dividing numbers incorrectly. This is especially problematic for workout problems.

- ChatGPT often provided descriptive explanations for its answers, even if they were incorrect. This raises the important question about how its authoritative, yet incorrect,

responses may impact students. Similarly, at times ChatGPT's descriptions were accurate, but its selection of multiple-choice answers was incorrect.[15]

- ChatGPT sometimes "made up" facts. For instance, when providing a reference, it generates a real-looking reference that is completely fabricated—the work, and sometimes authors, do not even exist.

- ChatGPT could produce specialized non-conversational text, such as journal entries, computer code, tables, and financial statements. The format of the journal entries varied from a simple listing of account names and amounts to a more formal entry in the general journal form with headings for dates, account names, debits, and credits.

- Faculty members noted that ChatGPT sometimes suggested answers that were not included as multiple-choice options. This could be beneficial in identifying ambiguities or mistakes in assessment questions if used by the instructor in advance.

- ChatGPT struggled to answer multiple-choice questions that describe a situation and require students to then select a concept illustrated by this situation. ChatGPT could therefore be instrumental in identifying ambiguities in the wording of multiple-choice questions.

- ChatGPT struggled to correctly answer questions which required higher-order learning (Bloom 1956) and often failed to evaluate and analyze complex, nuanced assessments. However, it performed well at answering accounting questions that required less judgment and for which the accounting standards have been consistent over time.

- The bot's responses were sensitive to the prompt, which guides its AI. For example, the user can prompt the bot to write an essay about a particular topic or in a particular tone, and this can result in different answers for the same question depending on the prompt given. As a second example, questions may be answered differently by ChatGPT when answer options are provided vs. when answer options are not provided.

- If unable to directly generate answers, ChatGPT could provide detailed instructions to complete a question. For instance, it could provide steps on using a software tool or sample code to solve problems that require access to a specific database.

- ChatGPT's answers to the same question sometimes varied when the question was entered multiple times, and its responses did not always progress from incorrect to correct.

- The bot's response to questions that depend on the interpretation of images, such as business process diagrams (BPDs) or tabulated data in picture format, varied (Boritz, Borthick, and Presslee 2012; Borthick, Schneider, and Vance 2012). ChatGPT sometimes

---

[15] This apparent weakness highlights one method instructors might use to prevent cheating or identify it after the fact: require that students explain how they arrived at a particular answer in order to receive credit.

recognized that it lacked the image and declined to answer, sometimes recognized the missing image but answered anyway (sometimes correctly, sometimes not), and sometimes did not recognize the missing image and answered anyway (sometimes correctly, sometimes not).

- ChatGPT could generate code and find errors in previously written code. For example, given a database schema or flat file, ChatGPT could write correct SQL and normalize the data.

- ChatGPT struggled to handle long, written questions with multiple parts, even when allowing for "carry over" mistakes.

- In a case study context, ChatGPT was able to provide responses to questions based on assessing past strategic actions of the firm. However, where data was required to be used, ChatGPT was unable to respond to the questions other than providing formulas. ChatGPT performed even worse where there was a requirement for students to apply knowledge. This highlights that ChatGPT is a general-purpose tool as opposed to an accounting-specific tool. It is not unsurprising, therefore, that students are better at responding to more accounting-specific questions where the technology is not yet trained to answer accounting-specific questions.

V. DISCUSSION AND CONCLUSION

We test ChatGPT's performance in answering 28,085 accounting questions from assessments and textbook test banks. Our results show that students generally outperform ChatGPT, but the bot can approximate average human performance in some topic areas and for certain question types.

The debate around tools like ChatGPT is multifaceted. These types of tools have both strengths and weaknesses. The following discussion is not an exhaustive list of all the positive and negative aspects of chatbots. Rather, we provide it to spur thinking and research on ***both*** the positives and negatives of chatbots.

On the positive side, public accounting firms have invested billions of dollars in AI, and advanced data analytic technologies (Kapoor 2020; Maurer 2021; Eulerich, Masli, Pickerd, and Wood 2023) because they believe AI can help accounting professionals become more effective and efficient (PwC 2017; Deloitte 2018; Cooper, Holderness, Sorensen, and Wood 2019, 2022;

Dickey, Blanke, and Seaton 2019; Haq, Abatemarco, and Hoops 2020; Austin, Carpenter, Christ, and Nielsen 2021; Emett, Kaplan, Mauldin, and Pickerd 2021). By training accounting students to use AI effectively, educators can help students be ready to solve big, interesting problems in their careers. For instance, students could use ChatGPT to write a first draft of a solution to a problem, then improve the draft by fact checking, providing authoritative references (using accounting/auditing standards, tax code/regulations, academic or professional literature), then writing a final version of the solution. Faculty could even ask students to turn in the ChatGPT output along with the student's final edited version for comparison. In this way, faculty and students can accelerate the learning of the strengths and weaknesses of ChatGPT and how these tools can be used in practice.

Faculty could also use these tools to their advantage by generating additional practice problems and copy editing their materials, while investigating whether learning materials may be confusing to students. The tools also empower motivated students to learn on their own, as they can do the same without faculty guidance.

On the negative side, the use of AI chatbots like ChatGPT can hamper students' learning ability if it produces incorrect answers to prompts. New learners may struggle to differentiate between accurate and inaccurate information. Additionally, some students may use ChatGPT for cheating purposes, short-circuiting the learning process. To mitigate cheating, various measures can be taken such as conducting oral exams, administering exams in settings where technology cannot be accessed, shifting from traditional exams to more presentation-style assignments, or pre-testing exams with ChatGPT to assess whether the questions can be correctly answered by ChatGPT. Our results also suggest that some question types are less likely to be correctly answered by ChatGPT, and thus focusing on such question types could reduce the risk of

cheating with ChatGPT. Finally, OpenAI and other companies have developed tools to detect text generated by language models (e.g., https://platform.openai.com/ai-text-classifier, https://gptzero.me/). Additionally, the popular plagiarism detection tool Turnitin has developed AI and ChatGPT detection features for added reassurance for educators.[16] To effectively address the issue of cheating, educators must have access to updated tools like these and must use them consistently.

Even when using ChatGPT is authorized in a class or for other educational purposes, the risk of unintended plagiarism remains. ChatGPT's output may be identical to publicly available third-party wording because the algorithm was trained on a corpus of third-party resources. Although this type of plagiarism may be unintentional, it is still plagiarism. It is critical to understand that ChatGPT is not a reliable source for citations, and students must be encouraged to properly cite the sources of its output. Citing ChatGPT as a source ignores the possibility that its output could come directly from a third party and would be similar to exclusively citing Google Chrome as the source of an article from *The Wall Street Journal*.

This study highlights only a few of the pros and cons of AI chatbots, and further research is needed. One limitation of our research is that we required ChatGPT to be entirely correct to receive full credit. In reality, students using this technology may still benefit from a partially correct answer or, conversely, be misled by incorrect ChatGPT explanations, changing their answer to be incorrect. Future research should examine whether the combination of students and ChatGPT performs better or worse than the average student who does not use ChatGPT.

As a second limitation, we gave no credit for questions that could not be entered into ChatGPT, such as images. A student likely would have been able to make an educated guess on

---

[16] While these tools could also be used to mitigate cheating, they often have limited reliability (e.g., Taylor 2023).

these questions, performing similar to or better than our reported results. Furthermore, language models like ChatGPT can continuously improve as new data becomes available and as users provide feedback on their answers. While our research did not show any evidence of model improvement within the short time frame, future studies should use longitudinal data to investigate the pace of improvement in chatbots in accounting.

Additional questions that may be addressed in future studies include:

- How should AI tools like ChatGPT be incorporated into the accounting curriculum? While some advocate for its ban, others argue for its acceptance and integration, just like other technology tools such as Microsoft Excel. It is important to understand the circumstances in which AI technology enhances and hinders learning in accounting.

- What should be the response of online education to ChatGPT? How will it impact online assessments now that it is available?

- Other generative AI platforms, such as Caktus.ai, use different methods and databases, and their outputs may not trigger plagiarism detection software. For example, Caktus-generated text will not be flagged by current AI detectors designed to detect GPT-generated content. It is important to understand the impact of these new and different platforms on accounting education and to determine ways to differentiate between human-created and AI-generated content and whether it matters.

One little understood challenge in using AI chatbots as a learning tool is the potential for the AI model to continue learning. For example, although the body of data used to create the GPT-3 algorithm is fixed, ChatGPT's answers can improve over time through its interactions with users. For example, in an interaction, ChatGPT initially provided incorrect information when asked about the tax implications of §1231 assets. However, after the user pointed out the inconsistency to ChatGPT, ChatGPT apologized, acknowledged its error, made the correct connection, and thanked the user for bringing it to its attention. It stated that it would remember this answer in future interactions and apply this lesson to interactions with others, a claim we tested with a different user a few days later, and ChatGPT successfully applied the lesson. How the learning nature of these algorithms will influence users is an important question. Will users

train them to be more or less correct? How fast will they be changed (i.e., will use of ChatGPT in classroom settings change from lesson preparation to delivery)? How much trust can learners place in a model that can become more or less accurate over time, without any clear indication of the direction it is progressing? These are just a few questions about AI chatbots that require thoughtful consideration and broad debate and discussion.

The global fascination and engagement with ChatGPT in recent months clearly indicate that human interaction with AI is on the rise. We encourage further accounting education research that investigates student and faculty interaction with chatbot technologies explores ways in which such technologies can be incorporated into accounting programs, and discusses the role faculty and assessment design play in accounting program design. ChatGPT may provide the much-needed stimulus educators, university administrators, and students need to reimagine accounting education practices for a changing world (McGuigan, 2021; Richardson and Watson 2021; Tharapos 2022).

REFERENCES

Adamopoulou, E., and L. Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2: 100006.

Agomuoh, F. 2023. The best ChatGPT alternatives (according to ChatGPT). Available at: https://www.digitaltrends.com/computing/the-best-chatgpt-alternatives-according-to-chatgpt/ (last accessed February 25, 2023).

Austin, A. A., T. D. Carpenter, M. H. Christ, and C. S. Nielson. 2021. The data analytics journey: Interactions among auditors, managers, regulation, and technology. *Contemporary Accounting Research,* 38 (3): 1888-1924.

Baldwin-Morgan, A. A. 1995. Integrating artificial intelligence into the accounting curriculum. *Accounting Education* 4 (3): 217-229.

Bloom, B. S. 1956. Taxonomy of educational objectives: The classification of educational goals. D. McKay.

Bommarito, M. J., and D. M. Katz. 2023. GPT takes the bar exam. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314839 (last accessed February 25, 2023).

Boritz, J. E., A. F. Borthick, and A. Presslee. 2012. The effect of business process representation type on assessment of business and control risks: Diagrams versus narratives. *Issues in Accounting Education*, 27 (4): 895-915.

Borthick, A. F., G. P. Schneider, and A. Vance. 2012. Using graphical representations of business processes in evaluating internal control. *Issues in Accounting Education*, 27 (1): 123-140.

Brink, A. G., and B. E. Reichert. 2020. Research initiatives in accounting education: Serving and enhancing the profession. *Issues in Accounting Education* 35 (4): 25-33.

Choi, J. H., K. E. Hickman, A. Monahan, and D. Schwarcz. 2023. ChatGPT goes to law school. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335905 (last accessed February 25, 2023).

Conaway, J. K., T. Wiesen. 2023. Academic dishonesty in online accounting assessments— Evidence from the use of academic resource sites. *Issues in Accounting Education*, Forthcoming.

Constantz, J. 2023. Almost 30% of professionals say they've tried ChatGPT at work. Available at: https://www.bnnbloomberg.ca/almost-30-of-professionals-say-they-ve-tried-chatgpt-at-work-1.1872520 (last accessed February 25, 2023).

Cooper, L. A., D. K. Holderness, T. L. Sorensen, and D. A. Wood. 2019. Robotic process automation in public accounting. *Accounting Horizons*, 33 (4): 15-35.

Cooper, L. A., D. K. Holderness, T. L. Sorensen, and D. A. Wood. 2022. Perceptions of robotic process automation in Big 4 public accounting firms: Do firm leaders and lower-level employees agree? *Journal of Emerging Technologies in Accounting*, 19 (1): 33-51.

Coyne, J. G., S. L. Summers, B. Williams, and D. A. Wood. 2010. Accounting program research rankings by topical area and methodology. *Issues in Accounting Education,* 25 (4): 631-654.

CPA Exam Guide. 2023. Can ChatGPT really pass the CPA exam? We put it to the test [Video]. YouTube. Available at: https://www.youtube.com/watch?v=h_v2qZTt1S0 (last accessed February 25, 2023).

Deloitte LLC. 2018. Artificial Intelligence. Available at: https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/deloitte-analytics/deloitte-nl-data-analytics-artificial-intelligence-whitepaper-eng.pdf (last accessed February 25, 2023).

Dickey, G., S. Blanke, and L. Seaton. 2019. Machine learning in auditing. *The CPA Journal*, 89 (6): pp. 16-21.

Emett, S. A., S. E. Kaplan, E. G. Mauldin, and J. S. Pickerd. 2021. Auditing with data and analytics: External reviewer perceptions of audit quality and effort. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3544973 (last accessed February 25, 2023).

Eulerich, M., A. Masli, J. Pickerd, and D. A. Wood. 2023. The impact of audit technology on audit task outcomes: Evidence for technology-based audit techniques. *Contemporary Accounting Research*, forthcoming.

Firat, M. 2023. How Chat GPT can transform autodidactic experiences and open education?. Available at: https://doi.org/10.31219/osf.io/9ge8m (last accessed February 25, 2023).

Frey, C. B., and M. A. Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114 (C): 254-280.

Gray, G. L., V. Chiu, Q. Liu, and P. Li. 2014. The expert systems life cycle in AIS research: What does it mean for future AIS research? *International Journal of Accounting Information Systems* 15 (4): 423-451.

Gubler, J. R., S. Herrick, R. A. Price, and D. A. Wood. 2016. Violence, aggression, and ethics: The link between exposure to human violence and unethical behavior. *Journal of Business Ethics* 147 (1): 25-34.

Han, H., R. K. Shiwakoti, R. Jarvis, C. Mordi, and D. Botchie. 2023. Accounting and auditing with blockchain technology and artificial intelligence: A literature review. *International Journal of Accounting Information Systems* 48: Forthcoming.

Hanlon, M., and J. L. Hoopes. 2014. What do firms do when dividend tax rates change? An examination of alternative payout responses. *Journal of Financial Economics* 114 (1): 105-124.

Haq, I., M. Abatemarco, and J. Hoops. 2020. The development of machine learning and its implications for public accounting. *The CPA Journal,* 90 (6): pp. 6-9.

Harris, R. 2022. ChatGPT gains 1 million users within 5 days. *App Developer Magazine*. Available at: https://appdevelopermagazine.com/Chatgpt-gains-1-million-users-within-5-days/ (last accessed on February 11, 2023).

Herbert, T. 2023. AI chatbot falls just short on accounting exam. Available at: https://www.accountingweb.co.uk/tech/tech-pulse/ai-chatbot-falls-just-short-on-accounting-exam (last accessed February 25, 2023).

Holmes, A. F., and A. Douglass. 2022. Artificial intelligence: Reshaping the accounting profession and the disruption to accounting education. *Journal of Emerging Technologies in Accounting* 19 (1): 53-68.

Hu, K. 2023. ChatGPT sets record for fastest-growing user base - analyst note. Available at: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (last accessed February 25, 2023).

IBL News. 2023. ChatGPT-4 the fined tuned version of ChatGPT-3, might prompt a major shift. IBL News. Available at: https://iblnews.org/chatgpt-4-the-fined-tuned-version-of-chatgpt-3-might-prompt-a-major-shift/ (last accessed February 25, 2023).

Kapoor, M. 2020. Big Four invest billions in tech, reshaping their identities. Bloomberg Tax. Available at: https://news.bloombergtax.com/financial-accounting/big-four-invest-billions-in-tech-reshaping-their-identities/ (last accessed February 25, 2023).

Kelly, S. M. 2023. ChatGPT creator launches subscription service for viral AI chatbot. CNN. Available at: https://www.cnn.com/2023/02/01/tech/chatgpt-plus/index.html (last accessed February 25, 2023).

Kung, T. H., H. Cheatham, ChatGPT, A. Medenilla, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng. 2022. Performance of ChatGPT on USMLE: Potential AI-assisted medical education using large language models. Available at: https://www.medrxiv.org/content/10.1101/2022.12.19.22283643v2 (last accessed February 25, 2023).

Maurer, M. 2021. PwC to spend $12 billion on hiring, expanding expertise in AI, cybersecurity. *The Wall Street Journal*, June 15. Available at: https://www.wsj.com/articles/pwc-to-spend-12-billion-on-hiring-expanding-expertise-in-ai-cybersecurity-11623758400 (last accessed February 25, 2023).

McGuigan, N. 2021. Future-proofing accounting education: educating for complexity, ambiguity and uncertainty. *Revista Contabilidade & Finanças*, 32 (87): 383-389.

McMurtrie, B. 2022. AI and the future of undergraduate writing. Teaching experts are concerned, but not for the reasons you think. *The Chronicle of Higher Education*. December 13: 36-37. Available at: https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing (last accessed February 25, 2023).

Meckler, L., and P. Verma. 2022. Teachers are on alert for inevitable cheating after release of ChatGPT. *The Washington Post*. December 28. Available at: https://www.washingtonpost.com/education/2022/12/28/chatbot-cheating-ai-chatbotgpt-teachers/ (last accessed February 25, 2023).

Moffitt, K. C., V. J. Richardson, N. M. Snow, M. W. Weisner, and D. A. Wood. 2016. Perspectives on past and future AIS research as the *Journal of Information Systems* turns thirty. *Journal of Information Systems*, 30 (3): 157-171.

O'Leary, D. E. 2022. Massive data language models and conversational artificial intelligence: Emerging issues. *Intelligent Systems in Accounting, Finance and Management* 29: 182-198.

Qasim, A., and F. F. Kharbat. 2020. Blockchain technology, business data analytics, and artificial intelligence: Use in the accounting profession and ideas for inclusion into the accounting curriculum. *Journal of Emerging Technologies in Accounting* 17 (1): 107-117.

Qasim, A., G. A. El Refae, and S. Eletter. 2022. Embracing emerging technologies and artificial intelligence into the undergraduate accounting curriculum: Reflections from the UAE. *Journal of Emerging Technologies in Accounting,* 19 (2): 155-169.

PwC LLP, 2017. Sizing the prize: What'sthe value of AI for your business and how can you capitalise? Available at: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf (last accessed February 25, 2023)

Richardson, V. J., and M. W. Watson. 2021. Act or be acted upon: Revolutionizing accounting curriculums with data analytics. *Accounting Horizons*, 35 (2): 129-144.

Shrivastava, R. 2022. Teachers fear ChatGPT will make cheating even easier than ever. *Forbes*. Available at: https://www.forbes.com/sites/rashishrivastava/2022/12/12/teachers-fear-chatgpt-will-make-cheating-easier-than-ever/?sh=3479a50e1eef (last accessed February 25, 2023).

Susnjak, T. 2022. CHATGPT: The end of online exam integrity? arXiv.org. Retrieved January 19, 2023. Available at: https://arxiv.org/abs/2212.09292 (last accessed February 25, 2023).

Sutton, S. G., M. Holt, and V. Arnold. 2016. "*The reports of my death are greatly exaggerated*"—Artificial intelligence research in accounting. *International Journal of Accounting Information Systems* 22: 60-73.

Taylor, J. 2023. ChatGPT maker OpenAI releases 'not fully reliable' tool to detect AI generated content. *The Guardian*. Available at: https://www.theguardian.com/technology/2023/feb/01/chatgpt-maker-openai-releases-ai-generated-content-detection-tool (last accessed February 25, 2023).

Terwiesch, C. 2023. Would ChatGPT get a Wharton MBA? A prediction based on its performance in the operations management course. Available at: https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf (last accessed February 25, 2023).

Tharapos, M. (2022). Opportunity in an uncertain future: reconceptualising accounting education for the post-COVID-19 world. *Accounting Education*, 31 (6): 640-651.

Wood, D. A. 2016. Comparing the publication process in accounting, economics, finance, management, marketing, psychology, and the natural sciences. *Accounting Horizons* 30 (3): 341-361.

Zemankova, A. 2019. Artificial intelligence in audit and accounting: Development, current trends, opportunities and threats - literature review. 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Athens, Greece, 2019, pp. 148-154.

Figure 1
ChatGPT Articles in Newspapers: General and Education Related Articles - First Two Months
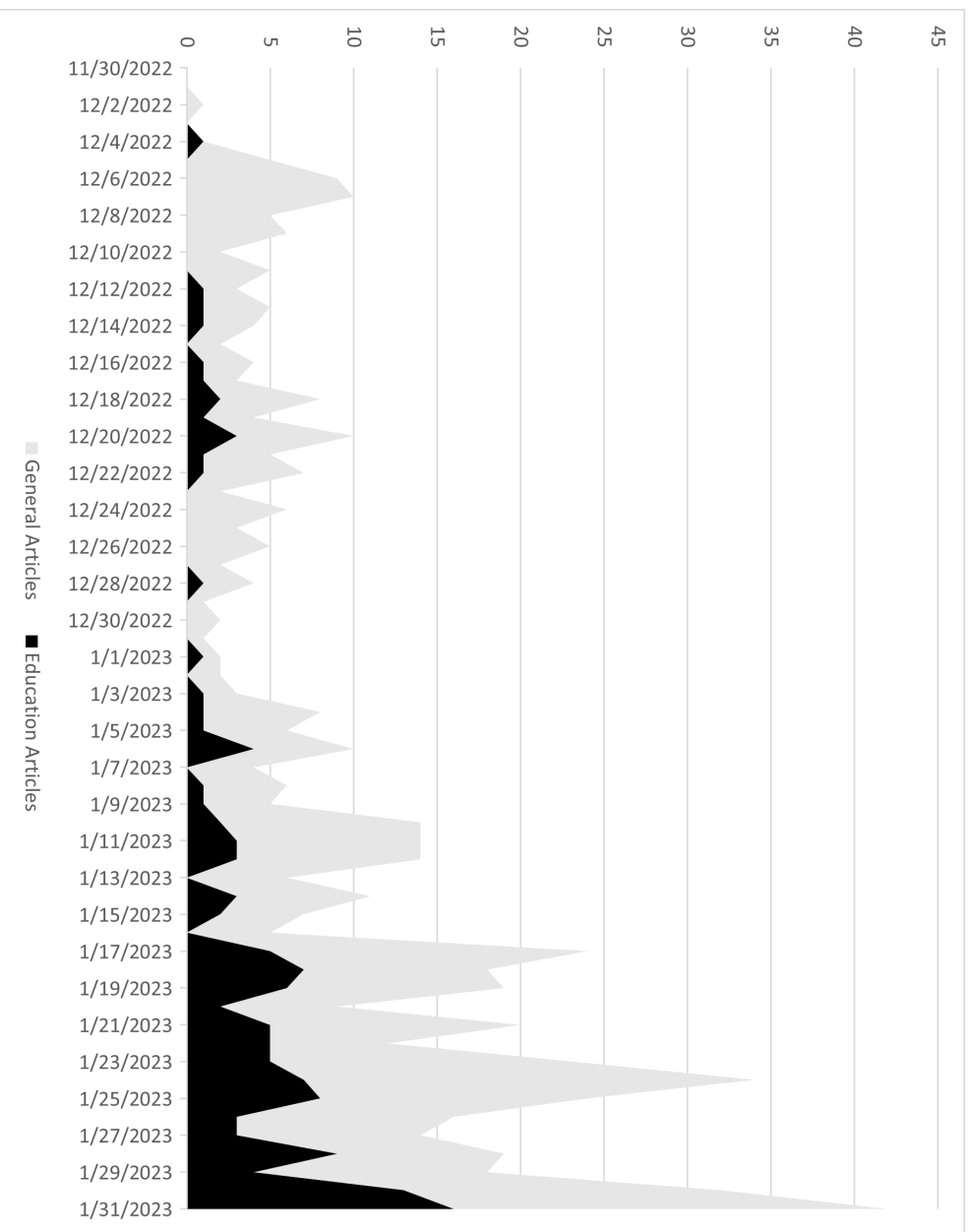
Data was obtained from ProQuest Central, published in English in the "Newspaper". The categories of "General Articles" and "Education Articles" are based on the search terms "ChatGPT" and "ChatGPT additional word" where the additional word are any of the following words: education, academics, school, exam, college, university, educate, teach, teachers, class, or classroom, respectively. The total number of articles is 566 with 136 of those articles related to education.

Figure 2
Comparison of Human and ChatGPT by Assessment
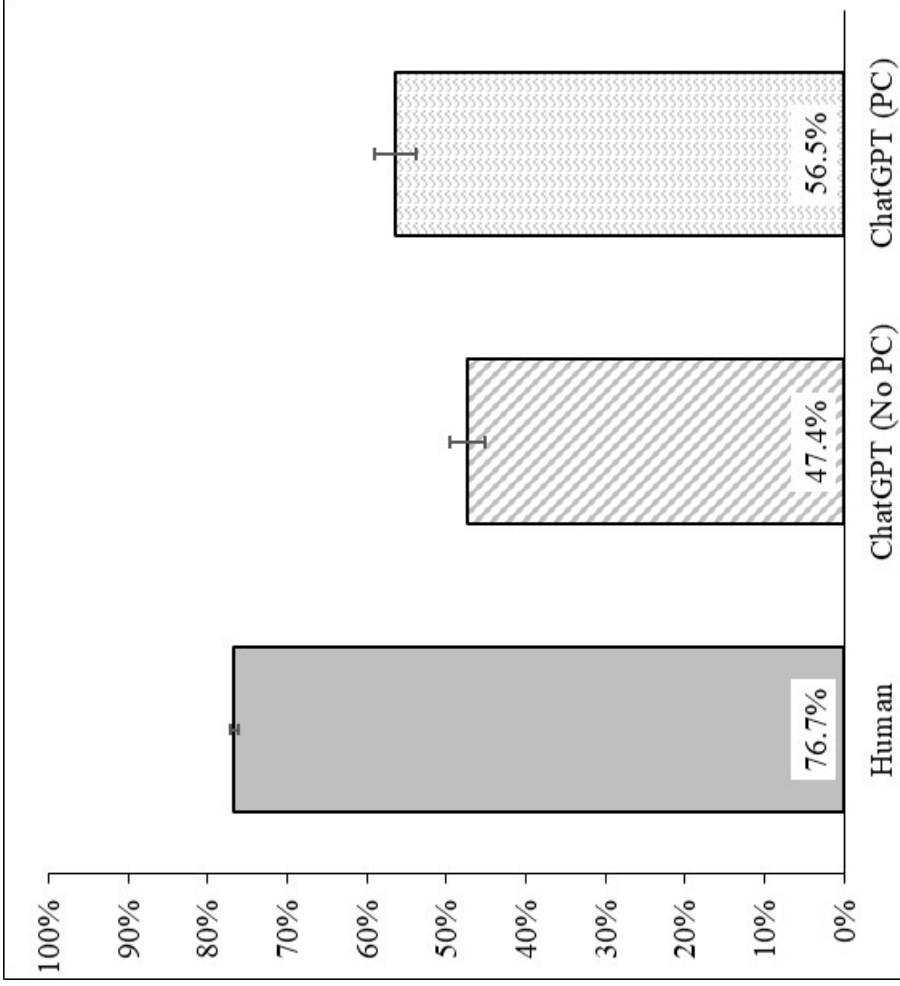
Figure shows the average performance by humans and ChatGPT. ChatGPT scores are given with no partial credit (No PC)—meaning the answer had to be exactly correct—and with partial credit (PC), where the question received 50 percent of the points for being correct. Error bars show 83 percent confidence intervals, which show the point at which the bars cannot overlap if the difference is significant at the 95 percent level (Gubler et al. 2016).

Figure 3
Comparison of Human and ChatGPT by Assessment Topic Area
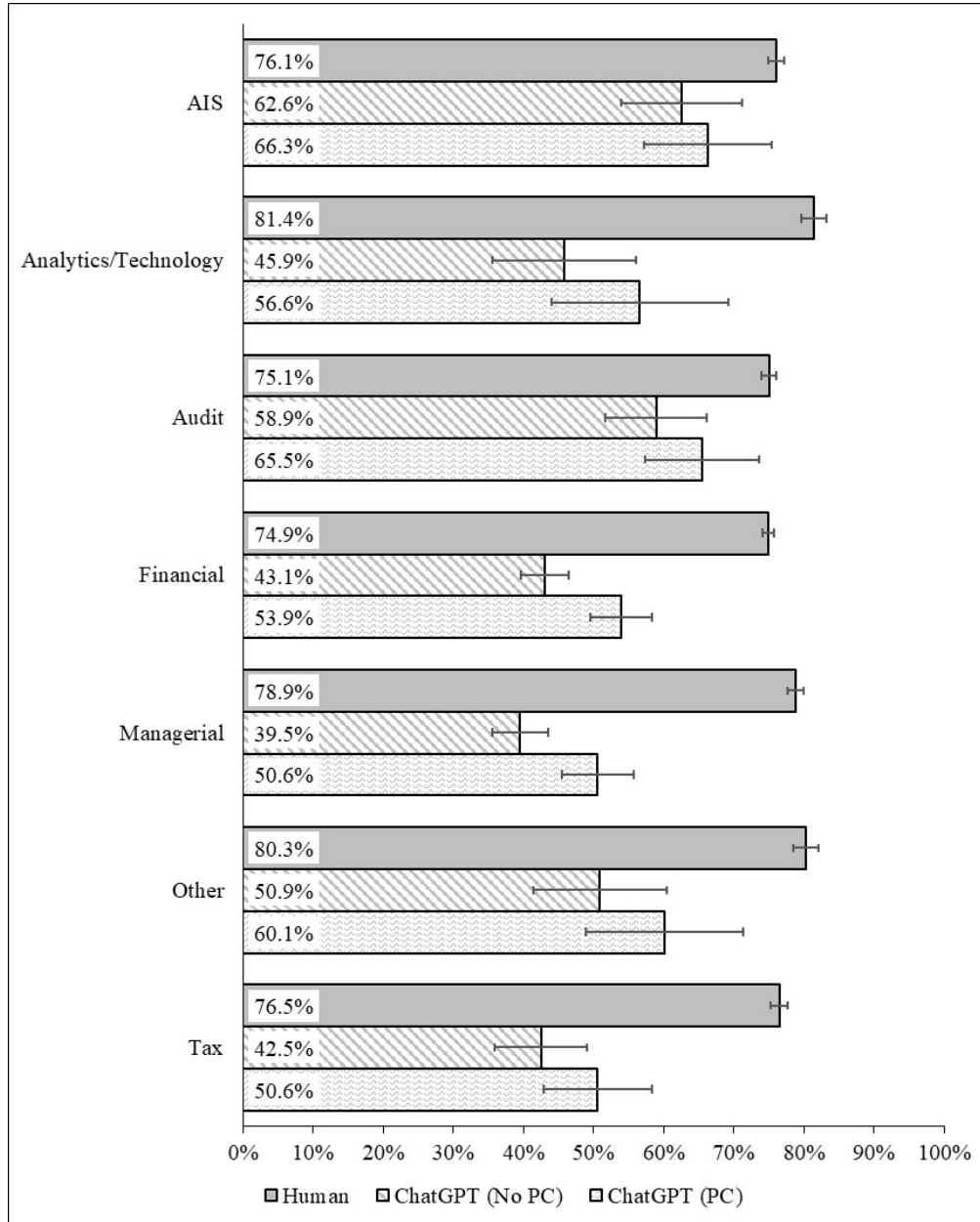
Figure shows the average performance by humans and ChatGPT by the topic area of the assessment. ChatGPT scores are given with no partial credit (No PC)—meaning the answer had to be exactly correct—and with partial credit (PC), where the question received 50 percent of the points for being correct. Error bars show 83 percent confidence intervals, which show the point at which the bars cannot overlap if the difference is significant at the 95 percent level (Gubler et al. 2016).

Figure 4
Comparison of Human and ChatGPT by Assessment and Student Class Level
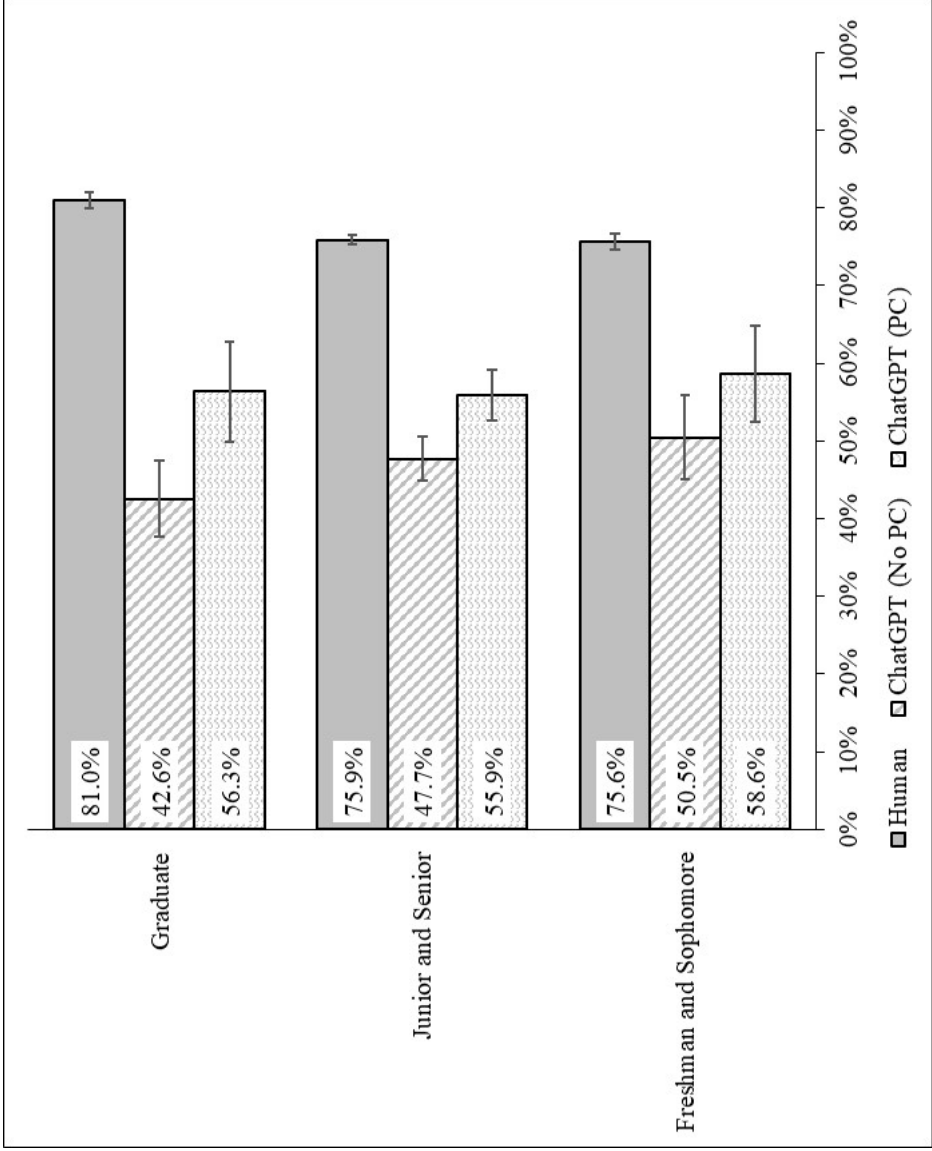
Figure shows the average performance by humans and ChatGPT by the class level of students taking the assessment. ChatGPT scores are given with no partial credit (No PC)—meaning the answer had to be exactly correct—and with partial credit (PC), where the question received 50 percent of the points for being correct. Error bars show 83 percent confidence intervals, which show the point at which the bars cannot overlap if the difference is significant at the 95 percent level (Gubler et al. 2016). We exclude the one observation of "other" in this graph.

Figure 5
Comparison of Human and ChatGPT for Open- and Closed-Note Assessments
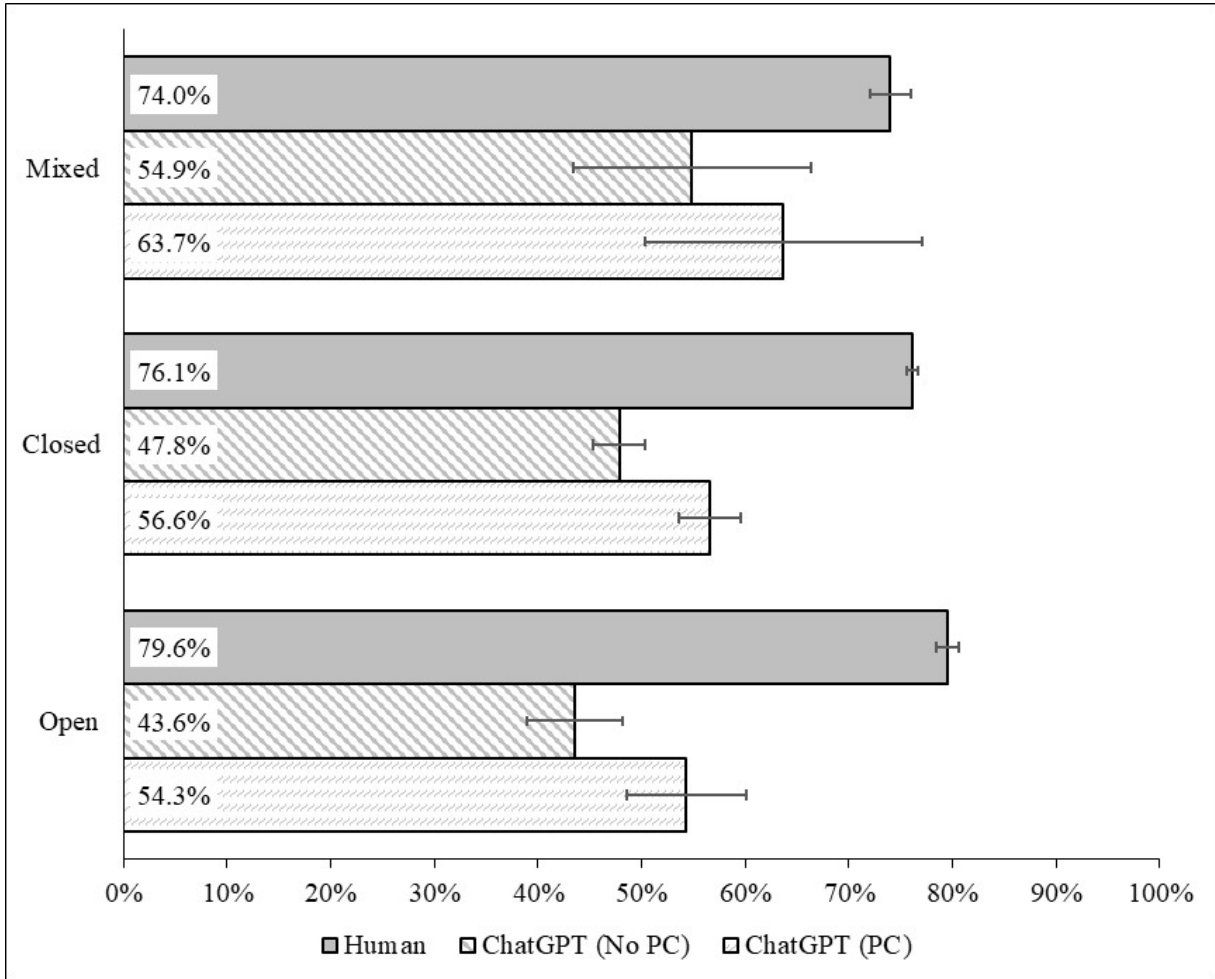
Figure shows the average performance by humans and ChatGPT for open, closed, and mixed assessments (mixed allow some open-note and some closed-note). ChatGPT scores are given with no partial credit (No PC)—meaning the answer had to be exactly correct—and with partial credit (PC), where the question received 50 percent of the points for being correct. Error bars show 83 percent confidence intervals, which show the point at which the bars cannot overlap if the difference is significant at the 95 percent level (Gubler et al. 2016).

Table 1
List of Universities and Number of Assessments in Dataset

Missouri State University	32	California State University, Monterey Bay	6
Iowa State University	31	Sam Houston State University	6
Auburn University	30	Susquehanna University	6
University of Missouri-St. Louis	22	Texas Tech University	6
Brigham Young University	20	University of Central Arkansas	6
Mississippi State University	20	University of South Florida	6
University of Wisconsin-Whitewater	20	University of Wyoming	6
Rutgers University	15	Whitworth University	6
University of North Carolina Wilmington	14	Arizona State University	5
University of Nevada, Las Vegas	13	Baylor University	5
Boise State University	12	East Carolina University	5
Georgia Southern University	12	Georgia State University	5
Idaho State University	11	Gonzaga University	5
University of San Diego	11	Grand Valley State University	5
Utah State University	11	Louisiana State University	5
University of Minnesota Crookston	10	National Cheng Kung University	5
Florida State University	9	The University of Texas at El Paso	5
Lewis University	9	University of Duisburg-Essen	5
Monash University	9	University of Georgia	5
Northern Illinois University	9	University of Mary Washington	5
Southern Illinois University Carbondale	9	University of Montana	5
University of Missouri-Kansas City	9	University of North Carolina at Chapel Hill	5
University of South Dakota	9	University of South Carolina	5
University of Wisconsin-Eau Claire	9	University of Tampa	5
Utah Valley University	9	University of Toledo	5
West Virginia University	9	Appalachian State University	4
Pennsylvania State University, Behrend College	8	College of the Holy Cross	4
Southern Utah University	8	Emporia State University	4
University of Nebraska-Lincoln	8	Ferris State University	4
Brigham Young University-Hawaii	7	Hofstra University	4
Clemson University	7	Kent State University	4
Lone Star College	7	Minnesota State University Mankato	4
Northeastern University	7	North Carolina State University	4
Oklahoma State University	7	Pace University	4
The University of Mississippi	7	Palm Beach Atlantic University	4
University of Memphis	7	RMIT University	4
Bucknell University	6	Tilburg University	4
		Truman State University	4

University of Arkansas	4	WHU - Otto Beisheim School of Management	3
University of Hartford	4	Woodbury University	3
University of Massachusetts Amherst	4	Zayed University	3
University of North Texas	4	Baruch College–CUNY	2
University of Oklahoma	4	Bentley University	2
Washington State University	4	Central Michigan University	2
Arkansas State University	3	Christopher Newport University	2
Augustana College	3	Fairfield University	2
California State University, Fullerton	3	Ghent University	2
California State University, Northridge	3	James Madison University	2
Central Connecticut State University	3	Kansas State University	2
Creighton University	3	Kennesaw State University	2
Florida Atlantic University	3	Louisiana Tech University	2
Florida Gulf Coast University	3	Loyola Marymount University	2
Jacobs University Bremen	3	Miami University	2
Loyola University Maryland	3	Norwegian School of Economics	2
Middle Tennessee State University	3	Simon Fraser University	2
North Carolina A&T State University	3	Texas State University	2
Ohio University	3	University of Auckland	2
Portland State University	3	University of Delaware	2
Sacred Heart University	3	University of Idaho	2
Temple University	3	University of Lethbridge	2
Tennessee Tech University	3	University of New Hampshire	2
The Ohio State University	3	University of New Mexico	2
University of Alabama	3	University of New Orleans	2
University of Bremen	3	University of Pittsburgh	2
University of Cincinnati	3	University of Tennessee at Chattanooga	2
University of Dayton	3	West Texas A&M University	2
University of Florida	3	York University	2
University of Houston-Downtown	3	American University	1
University of Illinois at Chicago	3	Babson College	1
University of Illinois at Urbana-Champaign	3	Berry College	1
University of Kansas	3	BI Norwegian Business School	1
University of Nevada, Reno	3	Bryant University	1
University of Notre Dame	3	Colorado State University	1
University of South-Eastern Norway	3	Dalhousie University	1
University of Texas at Austin	3	Erasmus University Rotterdam	1
University of Vermont	3	Francis Marion University	1
University of Waterloo	3	George Mason University	1
University of Wisconsin-Madison	3	Higher Colleges of Technology	1
Virginia Tech University	3	IE University	1

Manhattan College	1	University of Calabar	1
Methodist University	1	University of Central Florida	1
Morgan State University	1	University of Louisville	1
National Taiwan University	1	University of Minnesota, Mankato	1
North Dakota State University	1	University of North Carolina at Charlotte	1
Norwegian University of Science and Tech.	1	University of North Carolina Pembroke	1
Radboud University	1	University of Southern California	1
Rice University	1	University of Southern Denmark	1
Stephen F. Austin State University	1	University of Technology Sydney	1
Texas A&M University	1	University of Washington	1
The University of Melbourne	1	University of Wisconsin-Oshkosh	1
Trinity University	1	University of Wisconsin-Stevens Point	1
University of Agder	1	UNSW Sydney	1
University of Bayreuth	1	Utah Tech University	1
University of Bifrost	1		

This table shows the institutions that provided assessment data. The # column shows the number of assessments provided by authors at the institution.

Table 2
Descriptive Statistics

Panel A: Class Level of Assessments and Questions

Class Level	Assessments	Questions
Junior and Senior	558	17,557
Freshmen and Sophomore	168	5,181
Graduate	142	3,068
Other	1	11

Panel B: Type of Assessments

Type of Assessment	Assessments	Questions
Exam	828	25,517
Quiz	24	225
Other	9	48
Assignment	8	27

Panel C: Topic Area of Assessments

Topic	Assessments	Questions
Financial	288	8,120
Managerial	186	4,511
Audit	123	4,631
AIS	100	3,990
Tax	80	2,456
Other	54	1,395
Analytics/Technology	38	714

Panel D: Open or Closed Nature of Assessments

Open or Closed	Assessments	Questions
Closed	659	20,173
Open	167	4,097
Mixed	43	1,547

Panel E: Location of Assessments

Location	Assessments	Questions
Institution in the USA	796	24,481
Institution not in the USA	73	1,336

Panel F: Language of Assessments

Language	Assessments	Questions
English	855	25,609
Non-English	14	208

Table provides descriptive statistics of the number of assessments and questions included in our sample per category. Sample includes 869 assessments and 25,817 assessment questions. The number of questions in this table differ than that listed in Table 4 because Table 4 does not include questions that are not gradable by ChatGPT and this table does.

Table 3
Percentage of Times Average ChatGPT Score Was Greater Than Average Human Score per Assessment

Category	ChatGPT (No PC)	ChatGPT (PC)
Overall	11.3%	15.8%
Class Level		
Freshmen and Sophomore	10.7%	17.9%
Junior and Senior	11.8%	16.1%
Graduate	9.9%	12.0%
Topic Area		
AIS	23.0%	28.0%
Analytics/Technology	18.4%	18.4%
Audit	17.9%	25.2%
Financial	9.7%	14.9%
Managerial	3.8%	8.1%
Other	14.8%	16.7%
Tax	3.8%	8.1%
Open/Closed Assessments		
Closed	12.0%	15.6%
Mixed	18.6%	34.9%
Open	6.6%	11.4%

Table shows the percentage of assessments where ChatGPT's score was higher than the average human score. ChatGPT scores are given with no partial credit (No PC)—meaning the answer had to be exactly correct—and with partial credit (PC), where the question received 50 percent of the points for being correct. Lines are sorted within each grouping in alphabetical order except for class levels.

Table 4
Questions Answered Correctly by Various Factors

Category	N	Full Credit	Additional Partial Credit
Overall	25,181	56.5%	9.4%
Question Type			
Dichotomous (e.g., T/F, Either/Or, etc.)	792	68.7%	4.5%
Multiple Choice	20,084	59.5%	4.2%
Matching	548	58.2%	14.6%
Fill-in-the-blank	159	56.6%	22.0%
Essay	294	55.1%	32.0%
Other	15	53.3%	13.3%
Short Answer	1,988	39.1%	37.1%
Workout	1,301	28.7%	40.6%
Question Source			
Test bank	12,866	60.8%	5.6%
Custom	12,315	52.1%	13.3%

This table does not consider point values for questions, but whether questions are correct, partially correct, or incorrect; thus, the results do not directly tie to Figure 2. Table examines individual questions that were given in assessments. The table shows the percentage of questions that received full credit (meaning the answer had to be exactly correct) and the additional percentage of questions that would receive partial credit. As an example, the 56.5% on the first row means that 56.5% of questions were answered perfectly correct and the 9.4% in the "Additional Partial Credit" column means that 9.4% additional questions were answered partially correct. Lines are sorted within each grouping by Full Credit percentages. The number of questions in this table differ than that in Table 2 because Table 2 includes questions that are not gradable by ChatGPT and this table does not.

Table 5
Textbook Test Bank Questions Answered Correctly by Various Factors

Category	N	Full Credit	Additional Partial Credit
Overall	2,268	64.3%	2.8%
Question Type			
Dichotomous (e.g., T/F, Either/Or, etc.)	151	83.4%	0.0%
Multiple Choice	1,919	65.2%	0.0%
Short Answer	102	53.9%	29.4%
Essay	52	30.8%	46.2%
Workout	37	24.3%	13.5%
Matching	7	14.3%	57.1%
Topic Area			
Audit	255	83.1%	0.0%
AIS	405	76.8%	4.4%
Financial	766	61.7%	2.1%
Tax	143	56.6%	0.7%
Managerial	631	55.5%	4.4%
Financial and Managerial	50	48.0%	0.0%
Accounting Research	18	44.4%	0.0%

Table examines individual questions that are listed in textbook test banks. The table shows the percentage of questions that received full credit (meaning the answer had to be exactly correct) and the additional percentage of questions that would receive partial credit. Lines are sorted within each grouping by Full Credit percentages.

Table 6
Regression Results

Variable	Estimate	t-value	P-value
Intercept	0.433	12.22	<0.01
Class Level (Freshmen and Sophomore Omitted)			
Graduate	-0.045	-3.73	<0.01
Junior and Senior	-0.078	-8.78	<0.01
Topic Area (Financial Omitted)			
AIS	0.199	19.29	<0.01
Analytics/Technology	0.149	7.47	<0.01
Audit	0.126	12.81	<0.01
Other topics	0.113	7.65	<0.01
Managerial	-0.013	-1.44	0.150
Tax	-0.036	-2.97	<0.01
Question Type (Multiple Choice Omitted)			
Dichotomous (e.g., T/F)	0.122	6.91	<0.01
Fill-in-the-blank	-0.008	-0.20	0.844
Essay	-0.012	-0.42	0.675
Matching	-0.015	-0.70	0.482
Other question types	-0.040	-0.32	0.746
Short Answer	-0.157	-13.18	<0.01
Workout	-0.240	-16.92	<0.01
Additional Variables			
Exam given in English	0.104	2.81	<0.01
Test bank questions	0.057	8.68	<0.01
Institution in the USA	0.023	1.54	0.123
Open note exams	0.004	0.47	0.641

Adjusted R^2 = 0.058

The dependent variable is whether ChatGPT answered the question entirely correct or not (i.e., ChatGPT (No PC)). All other variables included in the model are dummy variables. We use a linear probability model so the interpretation of the coefficients is more straight forward, and we use heteroskedasticity-robust standard errors (e.g., see Hanlon and Hoopes 2014). We obtain similar results in terms of statistical significance for all variables if we use a logistic regression. Total number of observations is 25,181. Lines are sorted within each grouping by effect size.