Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA?

Marc Eulerich
University of Duisburg-Essen
marc.eulerich@uni-due.de
ORCID: 0000-0002-9965-7584

Aida Sanatizadeh
University of Illinois Chicago
asanat2@uic.edu
ORCID: 0000-0003-4225-4415

Hamid Vakilzadeh
University of Wisconsin – Whitewater
vakilzas@uww.edu
ORCID: 0000-0001-7398-3472

David A. Wood
Brigham Young University
davidwood@byu.edu
ORCID: 0000-0002-9748-8561

Abstract
ChatGPT is daily in the media with many predicting massive disruption, especially in accounting and auditing. Yet, prior research shows relatively poor performance of ChatGPT on student assessment questions. We extend this research to examine if newly released ChatGPT models and capabilities can pass major accounting certification exams including the CPA, CMA, CIA, and EA (enrolled agent) certification exams. We find that the early released ChatGPT 3.5 model is unable to pass any exam (average score across all assessments of 53.1 percent). However, with additional efforts ChatGPT can pass all sections of each tested exam: moving to the ChatGPT 4 model improved scores by an average of 16.5 percent, providing 10-shot training improved scores an additional 6.6 percent, and allowing the model to use reasoning and acting (e.g., allow ChatGPT to use a calculator and other resources) improved scores an additional 8.9 percent. After all these improvements, ChatGPT averaged a score of 85.1 percent across all sections of exams and passed them all. This high performance suggests ChatGPT has sufficient performance that it likely will prove disruptive to the accounting and auditing industries.

Key Words: Artificial intelligence, chatbots, GPT 3.5, GTP-4.0, Accounting certification, large language model

1. Introduction

ChatGPT, a large language model developed by OpenAI, is one of the fastest-growing technologies in history (e.g., Reuters 2023). The news is filled with stories of how this technology is being used and has the potential for massive disruption, including specifically in accounting (e.g., Eloundou et al. 2023; Felton et al. 2023). Whether this is hype or not can be partially determined by how well ChatGPT models perform accounting tasks, something for which empirical data is largely missing. We build on the few empirical studies that examine older models of ChatGPT or that examine accounting performance on educational assessments (Wood et al. 2023; Chen et al. 2023; Guo 2023; O'Leary 2023) to examine how the most recent ChatGPT model performs on accounting licensure exams (using exam preparation tests as proxies). Specifically, we test ChatGPT's performance on the Certified Public Accountant (CPA), Certified Management Accountant (CMA), Certified Internal Auditor (CIA), and the enrolled agent (EA) certification exams. Furthermore, we also study the extent to which additional model training and add-ons can improve performance on professional licensure exams. Thus, our key question is whether current artificial intelligence (AI) technology can pass these important professional certifications.

AI is not a new topic and has been studied in accounting for several decades; however, recently it has started to have a much greater impact on accounting practice. For example, research shows that AI (and related automation and other technologies) improve management forecast accuracy, timeliness of earnings announcements, and precision in earnings forecasts (Rozario and Zhang 2023); increases firm value and performance (Chen and Srinivasan 2023); improves internal and external audit quality (Christ et al. 2021; Fedyk et al. 2022; Eulerich et al. 2023); and improves audit efficiency (Christ et al. 2021; Cooper et al. 2019, 2022). Large

accounting firms have also announced the use of large language model-based AI systems. As just two examples, PwC uses a platform "that uses natural language processing, machine learning, and data analytics to automate and enhance various aspects of legal work" (PwC 2023, March 15) and EY reports creating a ChatGPT product to answer complex questions about payrolls (EY 2023). Indeed, PwC announced it will spend $1 billion over three years investing in ChatGPT and similar AI products (PwC 2023, April 26).

While this prior research shows positives to the use of technology and AI in accounting, it may come at the cost of accountants' jobs. A study by McKinsey estimates that by 2030, AI could displace 15 percent of the global workforce, or 400 million workers, and hit the accounting profession particularly hard (Manyika and Sneader 2018). Fedyk et al. (2022) show pre-ChatGPT AI reduces the number of accounting employees, but the time it takes to reduce headcount and the number of employees displaced are relatively modest in size. Other studies focus on the potential improvements of efficiency and effectiveness when using AI within a company (e.g., Jain et al. 2021; Choudhury et al. 2020; Tong et al. 2021). Most of the described benefits could directly be transferred to the accounting profession.

Interestingly, all these studies were released before the widespread release of large language models like ChatGPT and Alphabet's Bard. Noy and Zhang (2023) show in a series of experiments that ChatGPT improves task efficiency by 37 percent, while also producing higher quality work and job satisfaction. This effect is largely observed by substituting for worker effort. With the release of these new tools, studies are suggesting that accounting and auditing are particularly vulnerable to significant job loss. In directly examining the potential of ChatGPT to affect labor markets, Eloundou et al. (2023) find "that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of [large language

models], while approximately 19% of workers may see at least 50% of their tasks impacted." They list, among others, accountants, auditors, and tax preparers as having a 100 percent exposure to significant automation.

Whether these new large language models will reduce the headcount of accountants depends, at least in part, on how well they perform accounting tasks. If they are not "good" at being an accountant or auditor, then the predictions are unlikely to be realized. Wood et al. (2023) provide initial empirical evidence on this question by examining how well ChatGPT 3.5 answers accounting university assessment questions. Overall, they find that ChatGPT 3.5 answers questions correctly 55 percent of the time and that the student average on assessments outperforms ChatGPT between 80 and 90 percent of the time. This relatively poor performance of ChatGPT on university-level accounting curriculum suggests that these large language models are not poised for significantly reducing the need for professional accountants.

Although the initial testing suggests that large language models struggle answering accounting assessment questions, additional testing is necessary for several reasons. First, the models are being updated and improved at a rapid pace. OpenAI released ChatGPT 4 in March 2023 with statistics showing a significant improvement over ChatGPT 3.5, which was released in November 2022. For instance, the 3.5 model scored in the 10th percentile on the bar exam, but the ChatGPT 4 model scored in the 90th percentile (OpenAI 2023). Thus, it is an open empirical question of how well the model has improved in accounting. Second, Wood et al. (2023) use university assessment questions to evaluate ChatGPT 3.5's performance. These questions can vary from very basic accounting knowledge to more advanced; however, they do not necessarily directly connect to the knowledge needed to be a practicing accountant.

We extend prior research by using proxies for professional certification exams to directly compare how well GPT models perform on the assessments to be a certified accounting professional. Professional certifications are a key driver for a successful career. For instance, the AICPA (2023) explains the benefits of a CPA exam as follows: "Getting your CPA certification opens the kinds of doors that can fast-track you into influential jobs in every industry." Prior research shows the importance of licenses and professional quality for individual and firm outcomes (Prawitt et al. 2009; Prawitt et al. 2011, 2012; Hoopes et al. 2018). Thus, examining how AI does on professional licensure examinations is a reasonable first step to understanding how well AI can perform accounting tasks.

Additionally, Wood et al. (2023) did not train their model to perform better on accounting content, and they did not test how allowing ChatGPT to reason and react improves performance. We allow specialized training of ChatGPT on accounting content to see the degree to which training on the topic material can improve the model's performance. We also test how enabling ChatGPT to reason and interact with its environment improve its performance (see Yao et al. 2023).

We have no basis for exact predictions of how well AI will perform on the above exams. The Wood et al. (2023) results suggest it should be able to answer some of the questions and we expect the GPT 4 model to perform better than the GPT 3.5 model. Also, we expect that training the models on accounting content and allowing the model to reason and react will enhance performance. Yet, we do not have a theory to predict the levels of performance and thus, do not make explicit hypotheses. That said, each exam has a performance cut-off to achieve certification, and this is the key benchmark for our analyses.

We test these research questions by gathering data from test preparation exams for each exam. We perform the following tests and document how much each successive test improves performance:

1. Use ChatGPT 3.5 to establish a baseline and compare to prior research.
2. Examine how much using ChatGPT 4 improves performance.
3. Train ChatGPT 4 using few-shot training and measure performance.[1]
4. Turn ChatGPT 4 into an agent with reasoning and acting abilities (ReAct) and measure performance.[2]

We take a random sample of 150 to 300 questions for each part of each exam to perform these tests. We note that we only test questions with definitive correct answers (i.e., no workout problems) and questions that do not have tables or graphics. The current models of ChatGPT released to the public do not allow for testing images.

Our results show that the original ChatGPT 3.5 model performs similarly to the Wood et al. (2023) test using accounting assessment questions. The average across all parts of exams was 53.1 percent, compared to Wood et al.'s (2023) overall average of 55 percent. Using the new ChatGPT 4 model markedly improves scores by an average of 16.5 percent. Providing few-shot training further improves scores by an additional 6.6 percent and allowing ChatGPT to react and reason improves scores by an additional 8.9 percent. The results of these improvements are that ChatGPT 4, with few-shot training and the ability to react and reason results in an overall average score of 85.1 percent across all content tested. The updated ChatGPT performance is

[1] Few-shot training is an approach where the model is trained on a small dataset containing only a few examples per class or category, with the goal of generalizing to new, unseen examples.
[2] Yao et al. (2023) show that large language models perform significantly better when they generate chain-of-thought (Reasoning) and gather additional information from external resources (Acting). This is achieved by creating an agent that uses the model for its reasoning and uses tools to interact with external resources. In other words, the agent enables the large language model to interact with its environment. In our case, the agent may assist the large language model in using a calculator or searching the web.

sufficient to easily pass all sections of the multiple-choice questions to be a CPA, CMA, CIA, and EA.

The major contribution of this paper is to show the power of an AI chatbot to perform on accounting certification exams. The results suggest that current AI technology, with minimal additional effort, is sufficient to pass each of the accounting licensure exams. This suggests that computer programmers will likely be able to program apps that will allow computers to perform tasks that are currently performed by human accountants. While ChatGPT can still make errors (and even think it is right when it makes those errors) that is no different than human accountants. Designing how AI should work with humans, including the appropriate review processes, will be important for future research (see additional discussion in Huang and Vasarhelyi 2019).

This paper does not directly address whether AI chatbots will result in accountant job loss, or if automation by AI will result in the same number of workers performing different types and quantity of work. Rather, we test the ability of ChatGPT to perform on accounting certification exams. While one interpretation of the results is that ChatGPT presents a major threat to accountants and auditors, a different interpretation is that ChatGPT will free accountants and auditors to perform new and more value adding services to their clients. With additional work and development, accountants should be able to use these new tools to be more efficient and effective. Future research that can guide this transformation process will be very important.

2. Methodology

We compare the performance of ChatGPT 3.5 and 4 models on questions from accounting licensure examinations.[3] We gather questions from five different licensure exams that

[3] Specifically, we use OpenAI's GPT-3.5-turbo-0301and GPT-4-0314 models for this paper.

are meant to cover the main areas of accounting including financial topics (on several of the exams), internal and external auditing (on two exams), management accounting (on one exam), and tax accounting (on two of the exams).[4]

1. CPA exam: we use questions from Becker CPA exam preparation guides. We only include questions from the main course in our analyses. The CPA exam has four parts:
    a. Auditing and attestation (AUD).
    b. Business environment and concepts (BEC).
    c. Financial accounting and reporting (FAR).
    d. Regulation (REG).

2. CIA exam: we use questions from the global Institute of Internal Auditors (IIA) multiple choice training system. This exam is translated into various languages. We use questions translated into German. The global CIA certification had three parts:
    a. Part 1: Essentials of Internal Auditing
    b. Part 2: Practice of Internal Auditing
    c. Part 3: Business Knowledge for Internal Auditing

3. CMA exam: We use questions from Becker CMA exam preparation guides. The CMA has two parts:
    a. Part 1: Financial Planning, Performance, and Analytics
    b. Part 2: Strategic Financial Management

4. EA exam: We use questions from Gleim exam preparation (posted online), the enrolledagent.com exam prep website, and the IRS exam preparation website. The sections of the EA exam include:
    a. Part 1: Individuals (IND)
    b. Part 2: Businesses (BUS)
    c. Part 3: Representation, Practices, and Procedures (RPP)

For all exams, we only keep questions that do not have images in their text. We also only include multiple choice questions and not workout type questions. To the extent workout type questions are similar to university case studies, prior research suggests that ChatGPT can perform reasonably well on most of these types of assessments (Chen et al. 2023). However, to

[4] We note that OpenAI does not use the data transmitted through its API for training of its models. That is, by testing these exams, we did not make the model "smarter" in accounting or otherwise transfer copyrighted material to OpenAI (see https://openai.com/policies/api-data-usage-policies).

expedite testing given the fast-changing nature of this technology, we omit testing of these types of assessments and the timely grading necessary to evaluate them.

We test for differences between the 3.5 and 4 models. We also perform additional tests to see if we can boost the performance of the ChatGPT models. Specifically, we also provide few-shot training to enhance the model. Few-shot prompting is a method in which the model is only provided a few examples before submitting questions for testing (Wang et al. 2020). Few-shot training usually ranges from submitting two to five examples, but it can also use up to 100 examples (Wang et al. 2021). To compensate for the limited number of training examples, models in a few-shot context would require some prior information (e.g., a pre-trained language model). GPT 3.5 and GPT 4 are both pre-trained models.

For our few-shot training, we randomly sample 10 questions and use these to train ChatGPT. Submitting questions is called "prompting" the AI. We follow OpenAI's (2023) guidelines to engineer our prompt. When prompting through the OpenAI API, we can also set the level of creativity of the model using the *TEMPERATURE* parameter. By setting the temperature to zero we eliminate randomness in models' responses and reduce creativity. As we are measuring demonstrably correct answers, creativity in responses was not desirable. In practice, the model should provide the same response, every time we prompt the same question with the temperature set to zero.

Finally, we advance our model through reasoning and acting. To this end, we follow Yao et al. (2023) and Schick et al. (2023) and introduce agents to ChatGPT 4. Agents can be thought of as enabling "tools" for large language models. Agents allow a large language model to accomplish the tasks that a human would do such as using a calculator for math or using search engines for information gathering. Using agents, it is also possible for a large language model to

write and run Python programming or even query an SQL database. In some testing, we allow ChatGPT to use agents to access a calculator and perform web searches.

Furthermore, we take advantage of chain-of-thought prompting. Wei et al. (2022) demonstrate that large language models can construct chain-of-thought responses when given examples of chain-of-thought reasoning in the prompt. Chain-of thought reasoning can be thought of as decomposing a larger problem into several intermediate steps to get the final answer. This is also called reasoning. ReAct is an abbreviation for the combination of reasoning and acting. Appendix 1 shows an example of ReAct prompt with the outcome. As illustrated, the model explicitly states the steps that are needed to solve the problem (Reasoning) and uses search and calculator to get information needed to solve the problem (Acting). In the example provided, the model looks up the current dollar to Euro exchange rate through a web search and uses the calculator to compute the final answer.[5]

Since decision making and reasoning are built into a large language model, ReAct has a number of features that make it stand out: First, creating ReAct prompts is simple as users can simply enter their thoughts on top of their queries. Second, ReAct works for a variety of activities with various actions and reasoning requirements, including but not limited to using a calculator, fact verification, executing code, online search, etc. Third, Yao et al. (2023) find that ReAct regularly outperforms baselines with only reasoning or acting across diverse domains. Lastly, and most importantly, ReAct offers an interpretable sequential decision-making and reasoning process in which users may readily evaluate reasoning and factual accuracy (Yao et al. 2023). In this way, it is less of a black box and provides insight into how it solves a problem.

[5] ChatGPT 4 is not required to use the tools every time, but will use the tools when it reasons they will be necessary.

Each time we test a set of questions, we perform it in a different session, meaning the model will not consider any previously entered questions. Table 1 shows descriptive statistics of the number of questions we use for each testing phase. The sample sizes differ by exam because of the number of different questions in the review material.[6] The sample sizes differs as we add complexity because the cost of running the more advanced models increases. Given our sample sizes are all above 150 for each section of each exam, this choice is unlikely to bias our results. We also list in the table the minimum score necessary to pass each exam. The notes to the table contain descriptions of how we reached these minimums for tests that do not have a hard-set threshold.

3. Results

We start our analysis by examining the performance of the ChatGPT 3.5 model. Table 2 contains the results for using the 3.5 model for each section of each exam. The results suggest that scores range from a low of 37.3 percent for the individual portion of the EA exam to a high of 68.0 percent for Part 3 of the CIA exam. None of these scores are above the threshold necessary to pass a section of the exam. We note that the overall average of these scores is similar to the average score on accounting assessments of observed in Wood et al. (2023): the average for certification exams is 53.1 percent and the average for accounting assessments was 56.5 percent (see their Table 4). Also, similar to the prior results in Wood et al. (2023) GPT 3.5 struggles most with tax questions and does relatively better with auditing questions.

Table 2 also presents the results when we use the GPT 4 model. With this newer model performance improves substantially, ranging in improvements of 9.2 percent to 24.7 percent with

[6] We have uneven sample sizes based on the given training material. For example, while Becker CPA preparation offers thousands of questions, the CIA training system only has a couple of hundred questions for the different exam sections.

an average improvement per exam section of 16.5 percent. Table 2 shows that based on this higher performance, the GPT model passes 5 sections of exams, including all the sections of the CIA exam. Still, the model does not fully pass any of the other certifications.

Table 3 repeats the GPT 4 results from Table 2 in the column labeled "Zero-Shot", meaning this column shows performance of GPT 4 without any training. Table 3 adds the new column of "10-Shot" that shows how the GPT 4 model performs when it is prompted with 10 examples. The results show an additional average improvement to the model performance of 6.6 percent. With this improvement the model is now able to pass both sections of the CMA exam.

Table 4 repeats the "10-Shot" column from Table 3, labeled as "No ReAct". This table then adds the ability to reason and perform actions (ReAct) to the GPT 4 model. With this new ability, the model shows an additional improvement of 8.9 percent. Importantly, the model is now able to pass all sections of each exam. One major reason ReAct improves performance so much is that the model can now use a calculator to perform calculations. Failure with calculations is a major reason why ChatGPT struggled in financial and tax areas (Wood et al. 2023).

We present a visual summary of our results in Figure 1. Figure 1 shows the performance of the ChatGPT 3.5 model and then adds each additional step. The visual clearly shows that the improved models can easily clear the threshold for each of the certification exams.

3.2 Additional Analyses

The CPA exam training material separates problems into two categories, application and remembering and understanding. To show how each step in the model process improves the overall performance in each of these categories, we tabulate how each model we previously tested performs on these two types of questions. As shown in Table 5, the performance

improvements of using ChatGPT 4 and adding few-shot training have similar effects on application questions as on remembering and understanding questions. In contrast, adding the ReAct abilities to the model have a much more pronounced effect on application questions. This is consistent with Yao et al. (2023) who find that reasoning and acting substantially improves model's ability to answer more complex questions.

We provide an additional sensitivity training in Appendix 2 about the optimal level for the number of training shots provided to the model. This test can only be performed on older models. Our findings suggest that training of 3,000 to 4,000 examples should further enhance performance by around 6 percent. Professionals wanting to implement ChatGPT in practice should consider using more training to further enhance performance. Additional training behind this threshold can hurt model performance.

4. Discussion and Conclusion

Technological advancements continue to have a significant impact on business and accounting (Masli et al. 2011; Moffitt et al. 2016; Austin et al. 2021; Richardson and Watson 2021; Eulerich et al. 2023). The most recent advancements in AI, large language model chatbots, are likely to continue this trend. The degree to which they will impact accounting depends on their ability to perform accounting tasks at a high level. We test this ability by seeing how well one of these chatbots can perform on accounting certification examinations.

Examining certification exams is important because as a business service, accounting quality depends heavily on the quality of work of individual practitioners. Professional certifications are a globally recognized mark of quality for accountants and auditors. The knowledge required to pass these exams, the high preparation invested before taking these

exams, and the (sometimes) low pass rates, represent a high hurdle for many practitioners. The holders of these credentials are usually respected experts in their fields.

The results of our study demonstrate that ChatGPT can perform sufficiently well to pass important accounting certifications. This calls into question the "competitive advantage" of the human accountant relative to the machine. To our knowledge, for the first time, AI has performed as well as a majority of human accountants on a real-world accounting task. This raises important questions of how will machine and accountant work together in the future. We encourage research to help understand where machine and human abilities are best deployed in accounting. We also encourage research that develops and invents the capabilities for machines to perform greater amounts of accounting work—freeing accountants to innovate and add greater value to their organizations and society.[7]

We make several additional suggestions for future research. In our study, we equip ChatGPT with a calculator for computation tasks and a search engine to find out more about the topics in the questions. However, we observed that the search agent does not always provide useful information to ChatGPT. Future research can investigate whether agents that retrieve information from reliable and more specialized resources improve model's performance. An agent, for example, can be programmed to retrieve information from the PCAOB audit standards or the IRS tax publications. Similarly, researchers might study whether human feedback as an intermediary step of the chain-of-thought can improve the performance of the model.

[7] Research that "invents" and "develops" can follow the design science methodology (Geerts 2011). This methodology compliments traditional hypothesis-testing methodologies by providing a rigorous method for inventing and developing solutions to important practical problems, something the accounting research paradigm currently struggles with (Wood 2016; Rajgopal 2021; Burton, Summers, Wilks, and Wood 2021, 2022; Burton, Heninger, Summers, and Wood 2023).

Another area for future research is AI transparency. We find that using ReAct substantially enhances transparency about ChatGPT's decision-making. We note that the model is more likely to explicitly state that it is "unsure" of what to do or is making a "guess" to answer. Similarly, under these settings, the model is more likely to respond, "I don't know." Although not empirically tested, we anecdotally note that ChatGPT tends to hallucinate less when we use ReAct.[8] Future research can investigate ways through which transparency can be enhanced and if making ChatGPT an agent improves accuracy and reduces hallucinations.

Considering that AI deployment in accounting is already starting to take place, there is also a need for research in auditing AI. Prior research on AI auditing has focused on evaluating whether specific applications meet predefined industry specific requirements. For instance, researchers have created procedures for auditing AI systems used in recruitment (Kazim et al. 2021), online search (Robertson et al. 2018), and medical diagnostics (Liu et al. 2022). As AI becomes more prevalent in corporate operations, AI auditing from a corporate governance perspective becomes even more important. While some studies propose frameworks from a governance perspective (e.g., Mökander et al. 2023), auditing AI remains an important, under explored area for future research.

Our study is subject to several limitations. First, our study omits from testing questions that require greater cognitive ability, such as interpreting situations and contexts and interpreting visualizations. Future studies should continue to probe how AI and related technologies can perform these more advanced functions. Second, we note that we test practice exams rather than actual exams, as the actual exams are not made available. Third, although our results suggest ChatGPT can respond to questions, we do not test whether it can perform actual accounting tasks

[8] Hallucinations are mistakes generated by AI that sound plausible but are in fact incorrect or nonsensical.

such as bank reconciliations, tax preparation, closing the books, etc. We encourage research that can demonstrate whether AI can move from knowing to doing. ChatGPT and related technologies are exciting new technologies. We encourage their continued study and implementation in practice.

References

AICPA. 2023. *What is a CPA and what do they do?* Association of International Certified Professional Accountants. Retrieved April 27, 2023, from https://us.aicpa.org/becomeacpa/gettingstarted/frequentlyaskedquestions

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

Burton, F.G., W.G. Heninger, S.L. Summers, D.A. Wood. 2023. Updated perceptions of accounting academics on the review and publication process. *Issues in Accounting Education*, Forthcoming.

Burton, F.G., S.L. Summers, T.J. Wilks, D.A. Wood. 2021. Do we matter? The attention the general public, policy makers, and academics give to accounting research. *Issues in Accounting Education* 36 (1): 1–22.

Burton, F.G., S. L. Summers, T. J. Wilks, D. A. Wood. 2022. Relevance of accounting research (ROAR) scores: Ratings of titles and abstracts by accounting professionals. *Accounting Horizons* 36 (2): 7–18.

Chen, X., R. Dunn, T. Holt, K. Inger, J.G. Jenkins, J. Jones, J. Long, T. Loraas, M. Mathis, J. Stanley, D.A. Wood. 2023. Artificial intelligence's capabilities, limitations, and impact on accounting education: Investigating ChatGPT's performance on educational accounting cases. *Working Paper.* Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4431202.

Chen, W., S. Srinivasan. 2023. Going digital: Implications for firm value and performance. *Review of Accounting Studies,* Forthcoming, 1-47.

Choudhury, P., E. Starr, R. Agarwal. 2020. Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, *41*(8), 1381-1411.

Christ, M.H., S.A. Emett, S.L. Summers, D.A. Wood. 2021. Prepare for takeoff: Improving asset measurement and audit quality with drone-enabled inventory audit procedures. *Review of Accounting Studies*, *26*(4), 1323–1343.

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

Cooper, L.A., D.K. Holderness Jr, T.L. Sorensen, 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.

Eloundou, T., S. Manning, P. Mishkin, D. Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *Working Paper*. Available at https://doi.org/10.48550/arXiv.2303.10130

Eulerich, M., A. Masli, J. Pickerd, D.A. Wood. 2023. The Impact of Audit Technology on Audit Task Outcomes: Evidence for Technology-Based Audit Techniques. *Contemporary Accounting Research*, *forthcoming*, 1-32.

EY. 2023, March 29. *EY announces modernization of payroll employee care using ChatGPT In Azure OpenAI*. Ernst & Young Global Limited. Retrieved April 27, 2023, from https://www.ey.com/en_gl/news/2023/03/ey-announces-modernization-of-payroll-employee-care-using-chatgpt-in-azure-openai

Fedyk, A., J. Hodson, N. Khimich, T. Fedyk. 2022. Is artificial intelligence improving the audit process? *Review of Accounting Studies*, *27*(3), 938-985.

Felten, E., M. Raj, R. Seamans. 2023. How will Language Modelers like ChatGPT Affect Occupations and Industries? *Working Paper.* Available at https://doi.org/10.48550/arXiv.2303.01157.

Geerts, G.L. 2011. A design science research methodology and its application to accounting information systems research. *International Journal of Accounting Information Systems* 12 (2): 142–151.

Guo, B., X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *Working Paper*. Available at https://doi.org/10.48550/arXiv.2301.07597

Hoopes, J.L., K.J. Merkley, J. Pacelli, J.H. Schroeder. 2018. Audit personnel salaries and audit quality. *Review of Accounting Studies*, *23*, 1096-1136.

Huang, F., M.A. Vasarhelyi. 2019. Applying robotic process automation (RPA) in auditing: A framework. *International Journal of Accounting Information Systems*, *35*, 100433.

Jain, H., B. Padmanabhan, P.A. Pavlou, T.S. Raghu. 2021. Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research*, *32*(3), 675-687.

Kazim, E., A.S. Koshiyama, A. Hilliard, R. Polle. 2021. Systematizing audit in algorithmic recruitment. *Journal of Intelligence*, 9(3), 46.

Liu, X., B. Glocker, M.M. McCradden, M. Ghassemi, A.K. Denniston, L. Oakden-Rayner. 2022. The medical algorithmic audit. *The Lancet Digital Health*.

Manyika, J., K. Sneader. 2018, June 1. *AI, Automation, and the Future of Work: Ten Things to Solve for*. McKinsey Global Institute. Retrieved April 27, 2023, from https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for

Masli, A., V.J. Richardson, J.M. Sanchez, R.E. Smith. 2011. The business value of IT: A synthesis and framework of archival research. *Journal of Information Systems*, *25*(2), 81-116.

Moffitt, K. C., V.J. Richardson, N.M. Snow, M.M. Weisner, 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.

Mökander, J., J. Schuett, H.R. Kirk, L. Floridi. 2023. Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500*.

Noy, S., W. Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Working Paper,* Available at https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf.

O'Leary, D. E. 2023. An analysis of three chatbots: BlenderBot, ChatGPT, and LaMDA. *Intelligent Systems in Accounting, Finance and Management*, 30 (1): 41-54.

OpenAI. 2023. GPT-4 Technical Report. *Working Paper.* Available at https://doi.org/10.48550/arXiv.2303.08774

Prawitt, D.F., N.Y. Sharp, D.A. Wood. 2011. Reconciling archival and experimental research: Does internal audit contribution affect the external audit fee. *Behavioral Research in Accounting*, *23*(2), 187-206.

Prawitt, D.F., N.Y. Sharp, D.A. Wood. 2012. Internal audit outsourcing and the risk of misleading or fraudulent financial reporting: Did Sarbanes-Oxley get it wrong? *Contemporary Accounting Research*, *29*(4), 1109-1136.

Prawitt, D.F., J.L. Smith, D.A. Wood. 2009. Internal audit quality and earnings management. *The Accounting Review*, *84*(4), 1255-1280.

PWC. 2023, March 15. *PwC Announces Strategic Alliance with Harvey, Positioning PwC's Legal Business Solutions at the Forefront of Legal Generative AI*. PricewaterhouseCoopers. Retrieved April 27, 2023, from https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-announces-strategic-alliance-with-harvey-positioning-pwcs-legal-business-solutions-at-the-forefront-of-legal-generative-ai.html

PWC. 2023, April 26. *PwC US Makes $1 Billion Investment to Expand and Scale AI Capabilities* PricewaterhouseCoopers. Retrieved April 28, 2023, from https://www.pwc.com/us/en/about-us/newsroom/press-releases/pwc-us-makes-billion-investment-in-ai-capabilities.html

Rajgopal, S. 2021. Integrating practice into accounting research. *Management Science* 67 (9): 5430–5454.

Reuters. 2023, February 2. *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters. Retrieved April 27, 2023, from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Richardson, V.J., M.W. Watson. 2021. Act or Be Acted Upon: Revolutionizing Accounting Curriculums with Data Analytics. *Accounting Horizons*, *35*(2), 129-144.

Robertson, R.E., S. Jiang, K. Joseph, L. Friedland, D. Lazer, C. Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-22.

Rozario, A., C. Zhang. 2023. The effect of artificial intelligence on the accuracy of management earnings forecasts. *Working paper,* Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850823.

Schick, T., J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, ... T. Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv*:2302.04761.

Tong, S., N. Jia, X. Luo, Z. Fang. 2021. The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*(9), 1600-1631.

Wang, Y., S. Mukherjee, H. Chu, Y. Tu, M. Wu, J. Gao, A.H. Awadallah. 2021. Meta self-training for few-shot neural sequence labeling. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining,* 1737–1747.

Wang, Y., Q. Yao, J.T. Kwok, L.M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys*, *53*(3), 1-34.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, D. Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *Working Paper*. Available at https://doi.org/10.48550/arXiv.2201.11903

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.

Wood, D.A., M.P. Achhpilia, M.T. Adams,... and E. Zoet. 2023. The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions? *Issues in Accounting Education*, forthcoming, 1-28.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao. 2023. React: Synergizing reasoning and acting in language models. *Working Paper*. Available at https://doi.org/10.48550/arXiv.2210.03629

Figure 1
Model Performance Improvement

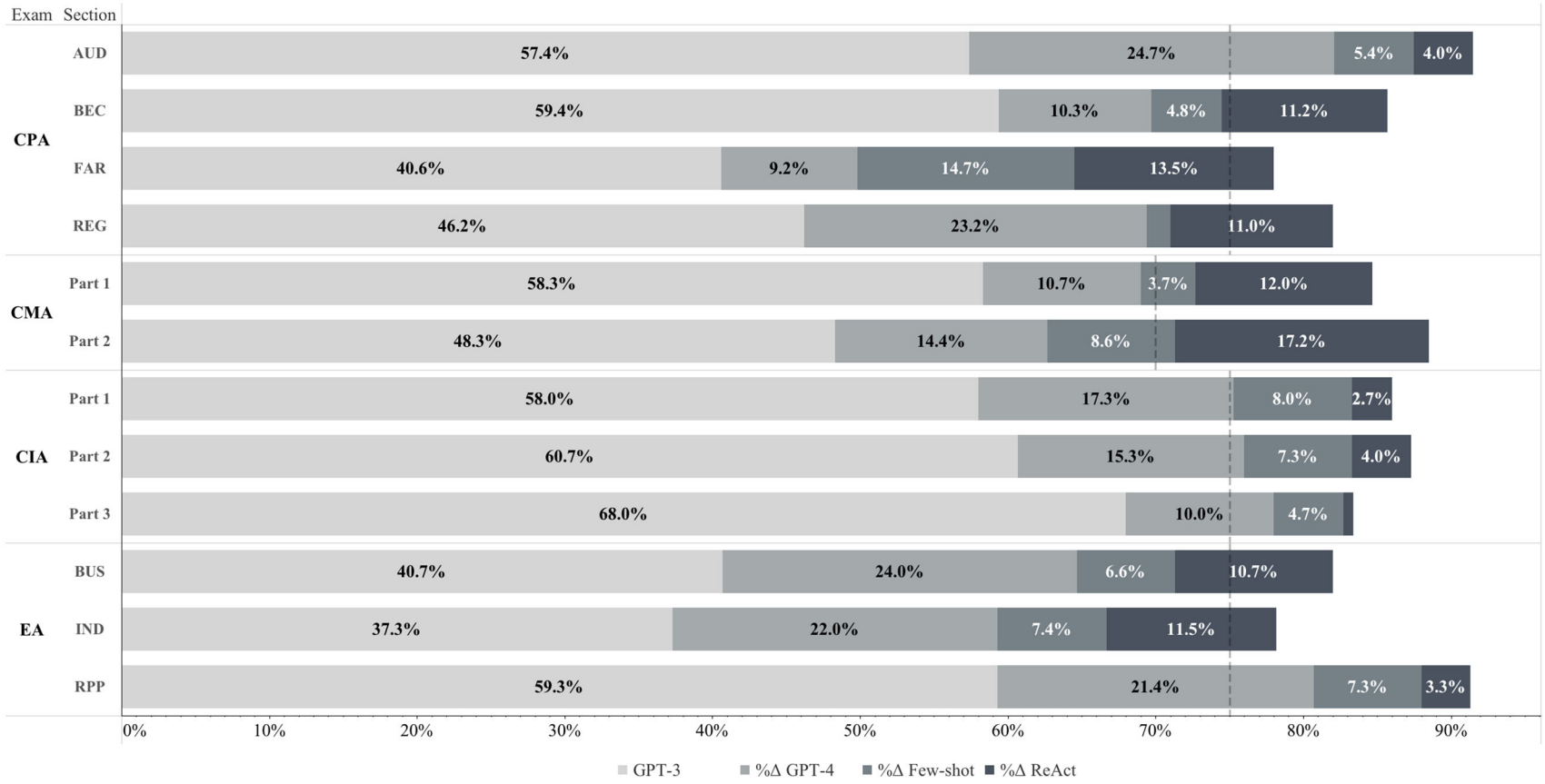

Table 1
Descriptive Statistics

		Number of Questions Examined		
Certification	**Required Pass Rate %**	**ChatGPT 3.5, Zero-Shot**	**ChatGPT 4, Zero-Shot**	**ChatGPT 4, 10-Shot**
CPA	75%	2,000	1,200	200
CMA	70%	300	300	300
CIA	75%	150	150	150
EA	75%	150	150	150

For each exam, we present the required minimum percentage score on the multiple-choice questions to pass the exam and the number of questions we tested for each model. The EA exam does not provide an exact percentage as the exam is scaled based on several factors. The 75 percent figure is the percentage of the scored exam required to pass and serves as our benchmark for passing.

Table 2
Model Performance for ChatGPT 3.5 and 4 with Zero-Shot Training

		GPT 3.5	GPT 4		
Certification	**Section**	**% Correct**	**% Correct**	**Improvement**	**Best Case Pass/Fail**
CPA	AUD	57.4%	82.1%	24.7%	Pass
CPA	BEC	59.4%	69.7%	10.3%	Fail
CPA	FAR	40.6%	49.8%	9.2%	Fail
CPA	REG	46.2%	69.4%	23.2%	Fail
CPA Average:		50.9%	67.8%	16.9%	Fail
CMA	Part 1	58.3%	69.0%	10.7%	Fail
CMA	Part 2	48.3%	62.7%	14.4%	Fail
CMA Average:		53.3%	65.9%	12.6%	Fail
CIA	Part 1	58.0%	75.3%	17.3%	Pass
CIA	Part 2	60.7%	76.0%	15.3%	Pass
CIA	Part 3	68.0%	78.0%	10.0%	Pass
CIA Average:		62.2%	76.4%	14.2%	Pass
EA	BUS	40.7%	64.7%	24.0%	Fail
EA	IND	37.3%	59.3%	22.0%	Fail
EA	RPP	59.3%	80.7%	21.4%	Pass
EA Average:		45.8%	68.2%	22.5%	Fail
Overall Average:		53.1%	69.6%	16.5%	

Reported numbers display the percentage of questions answered correctly for each section of an exam based on the ChatGPT 3.5 and ChatGPT 4 model with zero-shot training. The "Best Case Pass/Fail" column represents whether the best score for an exam section is considered a passing score for the professional exam.

Table 3
Model Performance for ChatGPT 4 with 10-Shot Training

		Zero-Shot	10-Shot		
Certification	**Section**	**% Correct**	**% Correct**	**Improvement**	**Best Case Pass/Fail**
CPA	AUD	82.1%	87.5%	5.4%	Pass
CPA	BEC	69.7%	74.5%	4.8%	Fail
CPA	FAR	49.8%	64.5%	14.7%	Fail
CPA	REG	69.4%	71.0%	1.6%	Fail
CPA Average:		67.8%	74.4%	6.6%	Fail
CMA	Part 1	69.0%	72.7%	3.7%	Pass
CMA	Part 2	62.7%	71.3%	8.6%	Pass
CMA Average:		65.9%	72.0%	6.1%	Pass
CIA	Part 1	75.3%	83.3%	8.0%	Pass
CIA	Part 2	76.0%	83.3%	7.3%	Pass
CIA	Part 3	78.0%	82.7%	4.7%	Pass
CIA Average:		76.4%	83.1%	6.7%	Pass
EA	BUS	64.7%	71.3%	6.6%	Fail
EA	IND	59.3%	66.7%	7.4%	Fail
EA	RPP	80.7%	88.0%	7.3%	Pass
EA Average:		68.2%	75.3%	7.1%	Fail
Overall Average:		69.6%	76.2%	6.6%	

Reported numbers display the percentage of questions answered correctly for each section of an exam based on the ChatGPT 4 and whether zero-shot or 10-shot training was provided. The "Best Case Pass/Fail" column represents whether the best score for an exam section is considered a passing score for the professional exam.

Table 4
Model Performance for ChatGPT 4 with 10-Shot Training and ReAct

		No ReAct	ReAct		
Certification	**Section**	**% Correct**	**% Correct**	**Improvement**	**Best Case Pass/Fail**
CPA	AUD	87.5%	91.5%	4.0%	Pass
CPA	BEC	74.5%	85.7%	11.2%	Pass
CPA	FAR	64.5%	78.0%	13.5%	Pass
CPA	REG	71.0%	82.0%	11.0%	Pass
CPA Average:		74.4%	84.3%	9.9%	Pass
CMA	Part 1	72.7%	84.7%	12.0%	Pass
CMA	Part 2	71.3%	88.5%	17.2%	Pass
CMA Average:		72.0%	86.6%	14.6%	Pass
CIA	Part 1	83.3%	86.0%	2.7%	Pass
CIA	Part 2	83.3%	87.3%	4.0%	Pass
CIA	Part 3	82.7%	83.4%	0.7%	Pass
CIA Average:		83.1%	85.5%	2.4%	Pass
EA	BUS	71.3%	82.0%	10.7%	Pass
EA	IND	66.7%	78.2%	11.5%	Pass
EA	RPP	88.0%	91.3%	3.3%	Pass
EA Average:		75.3%	83.8%	8.5%	Pass
Overall Average:		76.2%	85.1%	8.9%	

Reported numbers display the percentage of questions answered correctly for each section of an exam based on the ChatGPT 4 model with 10-shot training and whether the model had NoReAct or ReAct (ReAct stands for reasoning and acting). The “Best Case Pass/Fail” column represents whether the best score for an exam section is considered a passing score for the professional exam.

Table 5
Model Performance Improvement for CPA Question Types

Question Type	ChatGPT 3.5, Zero-Shot	ChatGPT 4, Zero-Shot	Improvement	ChatGPT 4, 10-Shot	Improvement	ChatGPT 4, ReAct	Improvement
Application	40.9%	54.9%	14.0%	62.3%	7.4%	77.4%	15.1%
Remembering & Understanding	60.8%	79.8%	19.0%	85.3%	5.5%	90.6%	5.4%

Questions from the CPA exam are separated based on their categorization of being application or remembering & understanding questions. The accuracy rates for each type of question is computed.