

Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries

Marc Eulerich
University of Duisburg-Essen
marc.eulerich@uni-due.de
[ORCID: 0000-0002-9965-7584](https://orcid.org/0000-0002-9965-7584)

Aida Sanatizadeh
University of Illinois Chicago
asanat2@uic.edu
[ORCID: 0000-0003-4225-4415](https://orcid.org/0000-0003-4225-4415)

Hamid Vakilzadeh
University of Wisconsin – Whitewater
vakilzas@uww.edu
[ORCID: 0000-0001-7398-3472](https://orcid.org/0000-0001-7398-3472)

David A. Wood
Brigham Young University
davidwood@byu.edu
[ORCID: 0000-0002-9748-8561](https://orcid.org/0000-0002-9748-8561)

Abstract

ChatGPT frequently appears in the media, with many predicting significant disruptions, especially in the fields of accounting and auditing. Yet, prior research has demonstrated relatively poor performance of ChatGPT on student assessment questions. We extend this research to examine whether more recent 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 enhancements, 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 passed all exams with an average score of 85.1 percent. This high performance indicates that ChatGPT has sufficient capabilities to potentially disrupt the accounting and auditing industries, which we discuss in detail. Lastly, this research provides practical insights for accounting professionals, investors, and stakeholders on how to adapt and mitigate the potential negative consequences of this cutting-edge technology in accounting and auditing firms.

Key Words: Artificial Intelligence, ChatGPT, Large Language Model, GPT 3.5, GTP 4, Accounting Certification

1. Introduction

A study by McKinsey estimates that by 2030, artificial intelligence (AI) could displace 15 percent of the global workforce, or 400 million workers, and hit the accounting profession particularly hard (Manyika and Sneader, 2018). Indeed, the release of ChatGPT, a large language model developed by OpenAI and one of the fastest-growing technologies in history (e.g., Reuters, 2023), has revived a discussion of how AI and automation will change the accounting profession. When it was released in November of 2022, news stories immediately discussed how technology would massively disrupt the business world. For example, Eloundou et al. (2023) report “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. Concurrently, large accounting firms have announced the use of large language model-based AI systems: PwC and KPMG will spend \$1 and \$2 billion, respectively, while EY has already spent \$1.5 billion (PwC, 2023; WSJ 2023a, 2023b).

This is not the first time predictions have been made of how technology will drastically revolutionize the accounting profession. These other “technology revolutions” include, among other things the introduction of the computer; software improvements like spreadsheets, databases, and ERP systems; continuous auditing; distributed ledger technology (blockchain); automation tools like robotic process automation (RPA); and many more. While each “Nostradamus” prognosticator proved wrong about the “end of accounting” (or even an immediate drastic “change” in accounting) in the past, the question of whether “this time is different?” is a tantalizing topic that attracts attention and drives panels, press, and presentations.

In this paper, we discuss several past technologies that proved to be overhyped, as a caution to interpreting the effect of ChatGPT on accounting. We then contribute to the question of whether “this time it is different” by examining the current performance of a large language model on accounting content. Specifically, we test how well ChatGPT performs on the Certified Public Accountant (CPA), Certified Management Accountant (CMA), Certified Internal Auditor (CIA), and Enrolled Agent (EA) certification exams.

An initial study by Wood et al. (2023) suggests that the ChatGPT hype is not likely to result in massive disruption. The authors found that ChatGPT 3.5 vastly underperformed accounting students as it could only score around 50 percent on accounting exams compared to students who scored higher than 75 percent on the same exams. A related study conducted by an accounting journalist showed that ChatGPT 3.5 performed poorly on accounting certification exams—with scores ranging from 35 to 48 percent on sections of the CPA exam (Gaetano, 2023).

Although the initial testing suggests that large language models struggle to answer accounting assessment questions, additional testing is necessary because the models are improving at a rapid pace. OpenAI released ChatGPT 4 in March 2023 with statistics showing a significant improvement over ChatGPT 3.5. 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). Given the new model updates, we test how much the new model and additional refinements to the ChatGPT model improve performance. 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 it 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.³

Our results show that the original ChatGPT 3.5 model performs similarly to the Wood et al. (2023) tests 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 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 sufficient to easily pass all sections of the multiple-choice questions to be a CPA, CMA, CIA, and EA. This is far better than the results reported by Geatano (2023) for the CPA exam that showed an average performance of 42 percent across the four sections of the exam.

To our knowledge, these are the first large-scale results that AI performs as well as or better than many accounting professionals. While this does not definitively suggest that “this time it is different,” the results do provide thought provoking evidence that it may be different.

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.

³ 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. When we performed testing, the models of ChatGPT released to the public did not allow for testing images.

To further the discussion of whether ChatGPT is overhyped or not and the potentially large effects it could have, we discuss concrete examples of how ChatGPT is currently changing accounting (including accounting academia). We note that overhyped technologies are usually spoken of in what they *will* do while technologies that are not overhyped are spoken of in what they are *already doing*. We provide numerous examples of how ChatGPT is already being used in significant and meaningful ways in accounting.

While we want to exercise caution so as not to overhype this new technology, our empirical and anecdotal evidence provide compelling evidence that ChatGPT, and similar technologies, will have a significant, maybe even dramatic, effect on accounting and accounting education. This time things do appear to be different and so maybe the right question is not are we overhyping this technology, but rather, are we hyping it enough?

2. Prior Technology Changes in Accounting

Technological changes have occurred continually throughout the history of accounting and have had a profound impact on the profession. Indeed, double-entry bookkeeping was a major technological innovation that spurred great changes in business and society (e.g., Williams, 1978). More recently, research shows that greater use of technology is associated with many positive outcomes in accounting (e.g., Cardinaels, Hollander, and White 2019; Chen and Srinivasan, 2023; Eulerich et al., 2023; Rozario and Zhang, 2023). Although research shows many positives of technology, the effects of technology are usually incremental and achieving them takes some time. This pattern of behavior is consistent with the Gartner Hype cycle (Fenn and Raskino, 2008), which is a visual model that shows the stages of hype and expectations surrounding new technologies (see Figure 1 for the typical shape of the Gartner Hype cycle).⁴ It

⁴ See <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>.

highlights the initial excitement, followed by a period of disillusionment, and ultimately, the practical applications and benefits that emerge as the technology matures. The press and some experts often predict that a new technology is going to have a drastic, dramatic impact, without fully considering how this process will likely be incremental and take time. We discuss several examples of this phenomenon to better evaluate the effects of Generative AI solutions, like ChatGPT on the accounting and auditing profession.

2.1 Distributed Ledger Technology, Blockchain

Distributed ledger technology, including its most notable iteration, blockchain, has been frequently presented as a revolutionary tool poised to completely reshape the way we approach transactions and data management (e.g., Dai and Vasarhelyi, 2017). Central to this narrative is the assertion that blockchain technology, with its capacity for enhanced security and transparent record-keeping, would bring groundbreaking changes to numerous sectors, accounting being a prime example (e.g., Dai and Vasarhelyi, 2017; Kokina et al., 2017). As an example, Dai and Vasarhelyi (2017) argue from an academic point of view that “blockchain is one of the most important and innovative technologies developed in recent years...Accounting and assurance could be among the professions to which blockchain would bring great benefits and fundamentally change the current paradigms” (p. 5). The media and professionals echoed this hype. For instance, Casey and Vigna (2018) write in their opinion article “Blockchain will make today’s accountants (and many Wall Street jobs) obsolete” that “once account-keeping itself becomes fully automated and reconciliation functions become superfluous, both those who keep the books and those who audit them will be out of work.” Morehouse (2017) extended this view, arguing that “transactions that are recorded in real time and can’t be altered can be audited daily,

eliminating the need for the expensive audits public companies are required to have every quarter.”

However, blockchain technology appears to have been overhyped. While research shows there are large investments into this technology and potential use cases for accountants and auditors (e.g., Kokina et al., 2017), others show that the technology is overhyped and that a full transfer to blockchains is infeasible (e.g., Coyne and McMickle, 2017). To date, the adoption of blockchain and its effect on accounting can be categorized, at best, as modest.

Nevertheless, it is important to recognize that blockchain has indeed made tangible contributions to several sectors, with accounting and auditing being notable areas of application. Blockchain technology offers an array of possible benefits in these sectors, such as enhanced transparency, data immutability, smart controls, and transactional security as well as close-to-real time audits based on the increased auditability (e.g., Dai and Vasarhelyi, 2017; Nordgren et al., 2019; Kwilinski, 2019).

While these are indeed valuable improvements, they are less revolutionary than initially proclaimed and fall short of the game-changing transformation that was anticipated. In terms of future integration, Macaulay (2022) predicts more modestly that blockchain will become a component of SAP cloud services over the next five years. However, these types of ERP integrations are expected to provide incremental improvements rather than the transformative changes initially touted. Similarly, Oracle’s Blockchain Tables, which integrate blockchain technology into the Oracle Database, offers a good example of how blockchain can be used to enhance existing systems rather than revolutionize them. These tables improve security and data integrity and offer various practical advantages, such as facilitating auditing and securely storing compliance data (Rakhmievich, 2019). Finally, the Canton Network—a blockchain system that

includes significant participants like Deloitte, Goldman Sachs, and Microsoft—offers another example of how blockchain can be integrated within existing regulatory frameworks to provide incremental improvements rather than radical transformations (Weiss, 2023).

Taken together, research and practical application show that blockchain has a role to play in accounting, but the initial hype as a transformative force in accounting that results in massive accounting unemployment and change has yet to come to pass. While blockchain does offer tangible benefits and improvements, the scale of its impact has not lived up to the original rhetoric that initially surrounded it.

2.2 Automation Software including Robotic Process Automation (RPA)

RPA software is the use of low-code or no-code software to automate repetitive, routine business processes (Cooper et al., 2019). It is a type of technology that uses software robots or “bots” to automate repetitive and rule-based tasks within computer systems. RPA software is designed to mimic human interactions with user interfaces and perform tasks such as data entry, data manipulation, form filling, and more. It can work across various applications and systems, interacting with them just as a human user would.

Initial research showed that RPA had very impressive results: “One accounting firm shared that in 2017 they saved over one million human work hours from RPA, while another respondent discussed turning a task that took 16 hours to complete into a 17-second task. Firms also report seeing increased quality as bot accuracy approaches 99.9 percent, compared to human performance on the same task that is often closer to 90 percent (Cooper et al., 2019, p. 16).”

Many press articles echoed these initial findings about the potential of RPA:

- **Headline:** “RPA: the Most Important Megatrend You’re Not Hearing About” and **relevant quote:** “Practically, every profession in the world involves repetitive tasks. And

in almost every case, a computer would do a much better job of carrying out these tasks. The possibilities of RPA are truly endless.”⁵

- Headline: “The Future of Accounting: How RPA And AI Will Revolutionize the Industry” and relevant quote: “AI, RPA, and other automated tech are transforming accounting, bringing increased precision, efficiency, safety, cost-savings and visibility.”⁶
- Headline: “The Impact of Robotic Process Automation in Accounting” and relevant quote: “New technologies are growing able to mimic human activity, taking on repetitive work more rapidly and accurately than people can. The authors offer an overview of Robotic Process Automation (RPA) in accounting that will change the ways the profession operates.”⁷
- Headline: “RPA: A Building Block of Transformative Automation” and relevant quote: “Companies [are] using [RPA] to revolutionize their workforces and accelerate more advanced automation efforts.”⁸

Prior research does provide evidence of the benefits of RPA, such as greater efficiency, effectiveness, and auditor satisfaction (Kokina and Blanchette, 2019; Cooper, Holderness, Sorensen, and Wood, 2022; Coyne, Summers, and Wood, 2023; Coyne, Summers, Villanueva, and Wood 2023). However, research and practice has started to discuss the limitations and dark side of RPA, including failure rates sometimes as high as 50 percent (EY, 2020; Moffitt, Rozario, and Vasarhelyi, 2018), significant internal control and governance problems (Bakarich and O’Brien, 2021; Eulerich, Waddoups, Wagener, and Wood, 2023a, 2023b), and a short-term focus that hurts long-term success (Zhang, Issa, Rozario, and Soegaard, 2023).

Once again, RPA did not live up to the hype of being a transformative technology. Accountants use RPA, but it did not result in dramatic changes to the profession. It has found an important role in organizations and helps in the right circumstances to improve efficiency and effectiveness.

⁵ <https://www.etftrends.com/innovative-etfs-channel/rpa-the-most-important-megatrend-youre-not-hearing-about/>

⁶ <https://blog.accountingprose.com/the-future-of-accounting-how-rpa-and-ai-are-revolutionizing-the-industry>

⁷ <https://www.cima.institute/blog/the-impact-of-robotic-process-automation-in-accounting>

⁸ <https://www.forrester.com/what-it-means/ep63-rpa-building-block-transformative-automation/>

2.3 Other Technologies

While blockchain and RPA are relatively dramatic examples of the hype cycle, other technologies in accounting provide similar, albeit less dramatic examples. There is an ongoing discussion about other technologies with strong disruption potential, like the use of drones for auditing purposes (e.g., Appelbaum and Nehmer, 2017a, 2017b; Christ et al., 2021), the implementation of process mining for auditing (e.g., Jans, Alles, and Vasarhelyi, 2014; Jans and Eulerich, 2022), or using virtual avatars for interviewing in auditing (Pickard, Roster, and Chen, 2016; CTStrategies, 2018; Pickard, Schuetzler, Valacich, and Wood, 2020). Each of these technologies was held up as having the ability to disrupt accounting and auditing, but the changes they have made are more incremental than revolutionary.

John Williams, the head of the ACCA (Association of Chartered Certified Accountants), said it well, “the situation [of technology replacing accountants] isn’t anything new; if you take a look back to 25 years, [someone] may have predicted the end of accountants with the advent of software like SAP or and Oracle, but at this point, it’s quite clear that accountancy is the profession that managed to survive and thrive”.⁹ This same sentiment could be applied to most other technologies that have been introduced with an intent to vastly disrupt the accounting industry.

2.4 Generative AI and ChatGPT

Based on the previous discussion, one might say that ChatGPT is overhyped and unlikely to prove a large disruptor of accounting—that is, it is being overhyped and eventually expectations will temper, as predicted by the Gartner Hype cycle. While this is plausible, it is often hard to know where a specific technology resides on the hype cycle diagram. This is

⁹ <https://www.pushdigits.ae/blog/robots-to-mark-the-advent-of-accounting/>

particularly the case when technology can belong to several categories. For example, ChatGPT is based on a large language model, which is a type of generative AI, as a sub-group of AI. Thus, it is not clear whether ChatGPT, language models, or (generative) AI is what is being mapped on a hype cycle. The former is relatively new and placed near the peak of inflated expectations as a new technology, while AI has existed for decades and is much more likely to be on the plateau of productivity.

AI has been studied in accounting for several decades, mainly looking at anomaly detection and/or decision-support through classification; however, recently, it has started to have a much greater impact on accounting practice. For example, research shows that AI improves 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); causes managers to be less aggressive (Estep, Griffith, and MacKenzie, 2023); and improves internal and external audit quality (Christ et al., 2021; Fedyk et al., 2022; Emmett et al., 2023; Eulerich and Wood, 2023). Other studies focus on the potential improvements in 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 benefits described could be directly transferred to the accounting profession.

While this prior research shows positives to the use of technology and AI in accounting, it may come at the cost of accountants' jobs. 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. They found that a one-standard-deviation increase in AI investments is associated with a reduction in accounting employees that reaches

3.6 percent after three years and 7.1 percent after four years. While this is meaningful, most would not consider it revolutionary.

Interestingly, all these studies were released before the widespread release of large language models like ChatGPT and Alphabet's Bard. Are these language models different? Preliminary research suggests that large language models may have a larger impact than previous AI releases. Rather than focusing on studies that make predictions about what will happen, we focus on discussing the few empirical studies that test the effects of ChatGPT on employee productivity and related topics.

Kreitmeir and Raschky (2023) studied what happened to Italian and other European professional coders' individual productivity when Italy banned ChatGPT. Using a difference-in-difference design they showed that programmer productivity dropped 50 percent in the first two business days after the ban, but then recovered, at least partially because of a swift implementation of the use of censorship bypassing tools (e.g., VPNs using the TOR network).

Dell'Acqua et al. (2023) use an experiment to study the effects of ChatGPT 4 access on consultants at Boston Consulting Group. They showed dramatic improvements for tasks that AI can currently perform—consultants performed tasks 25.1 percent more quickly, and completed 12.2 percent more tasks with quality 40 percent higher than a control group. Gains were most impressive for historically below average performers who improved performance 43 percent compared to above average performers increasing performance 17 percent. However, for tasks that are outside AI's current abilities, consultants using AI were 19 percent less likely to produce

correct solutions. Thus, generative AI proved to be highly effective for many tasks, but could harm performance for tasks that are outside of AI's current abilities.¹⁰

Other studies do not quantify the effects of ChatGPT but do show there is no qualitative difference in the creativity of humans and AI, with only 9.4 percent of humans being more creative than the most creative AI tested (Haase and Hanel, 2023); Overall, large language models, especially ChatGPT, have led to an increase in the quality, novelty and creativity of content generated by humans (Sanatizadeh et al., 2023; Zhou and Lee, 2023); ChatGPT performs better than previous tools at automatic bug fixing in computer programming (Sobania, Briesch, Hanna, and Petke, 2023); and ChatGPT can significantly outperform existing sentiment analysis methods for forecasting stock market returns (Lopez-Lira, and Tang, 2023).

On the other hand, not using ChatGPT can dramatically impact the current and future success of companies. Bertomeu, Lin, Liu, and Ni (2023) show that the ban of ChatGPT in Italy had a negative effect on the capital market and the valuation of Italian companies. Finally, Eisfeldt, Shubert, and Zhang (2023) create portfolios of companies that have high and low exposure to generative AI like ChatGPT and find that higher-exposure firms earned excess returns that are 0.4 percent higher on a daily basis (which equates to over 100 percent on an annualized basis) than lower-exposure firms—suggesting that, “according to investors, ChatGPT represents an important shock to corporate valuations.”

The effects sizes in these studies are quite large, which suggests that ChatGPT could be more disruptive than previous technologies. Whether these new large language models will have a dramatic effect on accounting will be determined in time. However, one necessary ability of

¹⁰ Using several experiments, Noy and Zhang (2023) show in a series of professional writing experiments that ChatGPT improves task efficiency by 37 percent, while also producing higher quality work and job satisfaction. This effect is largely observed by ChatGPT substituting for worker effort.

ChatGPT to be transformative is that it must have sufficient knowledge of accounting content to perform accounting tasks. If ChatGPT is not “good” at being an accountant or auditor, then the predictions are unlikely to be realized. The evidence from Wood et al. (2023) discussed in the introduction provides initial empirical evidence that ChatGPT is not poised to significantly reduce the need for professional accountants.

Although the initial evidence of ChatGPT in accounting was poor, additional models have been released and the community has gained additional experience and expertise in how to work with these types of models. As such, we empirically test ChatGPT’s current ability in relation to accounting content.

3. Methodology

We compare the performance of ChatGPT 3.5 and 4 models on questions from accounting licensure examinations.¹¹ We gather questions from five different licensure exams that 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).¹²

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

¹¹ Specifically, we use OpenAI’s GPT-3.5-turbo-0301 and GPT-4-0314 models for this paper.

¹² 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>).

- 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 4 can perform reasonably well on most of these types of assessments (Chen et al., 2023). However, to 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 tested the differences between the 3.5 and 4 models. We also perform additional tests to see if we can boost the performance of the ChatGPT 4 model. Specifically, we also provide few-shot training. Few-shot training is a method in which the model is 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 sampled 10 questions and used 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* hyperparameter. 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 the 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.¹³

Since decision making and reasoning are built into a large language model, ReAct has several 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.

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.¹⁴ 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.

¹³ ChatGPT 4 is not required to use the tools every time but will use the tools when it reasons they will be necessary.

¹⁴ 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.

4. 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 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 their results, 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 from 9.2 percent to 24.7 percent with 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 of 6.6 percent to the model performance. 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 2. Figure 2 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 certification exam.

4.1 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 has 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 beyond this threshold can hurt model performance.

5. Discussion of Possible ChatGPT Disruption in Accounting

Technological development is a process of continuous evolution, characterized by successes, disappointments, and constant improvements. While new technologies may initially face skepticism and fail to meet expectations, they typically become more reliable and effective over time. This progress is driven by iterative innovation, where developers learn from past experiences, and societal adaptation, as users discover new applications for the technology. In the long run, this process often leads to technologies becoming better and more integrated into our daily lives, despite the challenges they may face in their early stages.

While skepticism is a healthy part of any decision-making process, it is important to balance it with openness to innovation. Staying stuck in skepticism about emerging technologies may lead to missed opportunities. These could include benefits such as increased efficiency, cost savings, competitive advantage, or even the chance to pioneer a new field.

In our opinion, one key difference between a technology being overhyped or not is the degree to which users and prognosticators talk about *what will be possible* with the technology versus discussing *what is currently being done* with the technology. The more language about possibilities rather than realities suggests the technology is likely to be overhyped. So, in addition to the certification exam evidence, we discuss ways in which ChatGPT technology is being used in accounting and auditing, including accounting education, as of November 2023.¹⁵ In each section, we also discuss challenges and future possibilities that exist because of the emergence of generative AI.

¹⁵ The next discussion includes normative statements, conjectures, and results from the authors' experiences. We recognize the non-empirical nature of much of this discussion and encourage future empirical work.

5.1 Generative AI in Accounting Education

Generative AI is already proving to have some dramatic effects in education. For instance, at one of the author's institutions, an introductory information systems course created a chatbot based on the class's textbook and other materials (e.g., syllabus).¹⁶ This class services a very large number of students and as such employs 30 teaching assistants (TAs) who can answer questions in a virtual lab. In half a semester, the students in the class had approximately 51,000 interactions with the "TA-bot" compared to only 108 interactions with TAs in the virtual lab. The chatbot took less than 20 minutes to build and costs \$200 a month to run. The professor estimates it is more than 95 percent accurate in responding to queries. In comparison, the TAs cost approximately \$22,500 a month and are more than 95 percent accurate in their responses. In this situation, it appears that students significantly prefer using a chatbot, the chatbot produces superior answers, and it is able to drastically reduce costs.¹⁷

In a similar vein, the textbook publisher Pearson has announced plans to introduce generative AI into their online textbooks (Hughes-Morgan 2023). One of the author's has experimented with this technology and it allows the learner to ask any question about material in the textbook (e.g., "summarize the main points of this chapter in five bullet points", "explain concept 'x' in simpler terms") and to generate practice questions to test self-mastery (e.g., "create multiple choice questions to test me on the keywords from this chapter"). The chatbot for the introductory class and the Pearson textbooks is less likely to hallucinate than the publicly available generative AI models like ChatGPT and Bard because the responses are constrained to only use the text provided to the model.¹⁸

¹⁶ The professor authored the textbook and thus had copyright permissions to use it.

¹⁷ This example is similar to the AI teaching assistant that Harvard University introduced into introductory computer science program (see <https://www.harvardmagazine.com/2023/08/ai-in-education>).

¹⁸ Hallucinations are mistakes generated by AI that sound plausible but are in fact incorrect or nonsensical.

Some accounting educators are making significant use of ChatGPT to produce content. For example, ChatGPT (or related technologies) was used to create datasets, accounting scenarios, images, and solution guides for accounting cases and homework problems. As one particular example, the EY ARC cybersecurity accounting case Digital Dungeons is an escape room where students are required to figure out a numeric code to answer the case. To see if they are right, the developers of the case used ChatGPT to create the HTML code for a website. With just plain-language prompts, the website incorporates graphics (which were developed using AI) and submission forms and buttons. Furthermore, ChatGPT was able to encrypt the answer in the HTML code and add a submission delay so that each incorrect submission resulted in the user waiting an additional second before being able to try again. All of this was programmed in less than 30 minutes.

As another example, the TechHub.training website provides visitors with challenges to enhance their digital literacy (see Wood et al., 2023). Student authors write and review all the case materials but use ChatGPT to enhance their work. Similar to the EY ARC case, data, solutions, case descriptions, etc., were developed, refined, or improved by using ChatGPT. Students report that ChatGPT significantly enhanced the quality of their work and the efficiency in producing it.

Professors are experimenting with using generative AI to provide formative feedback and to grade student submissions (Pinto et al., 2023; Chen et al., 2023; Jukiewicz, 2023). The results, to date, are mixed and additional work is needed to understand both how and when generative AI is able to help faculty grade. However, the possibility of using generative AI to grade unstructured submissions (e.g., essays) would make it possible to better align assessment with

learning objectives, rather than using less effective testing because of limitations in faculty time (Kuechler and Simkin, 2005).

Generative AI is also shaping the production of academic research. Vakilzadeh, and Wood (2023) have created a beta-version of a tool to help automate the literature review process. The tool allows authors to use generative AI to understand and synthesize research. The tool can be used, among other things, to generate the first draft of a literature review, identify conflicts or gaps in research understanding, and brainstorm research questions to address. The tool has already helped to draft literature reviews for papers, succinctly summarize papers for reviews and promotion and tenure packets, and interpret academic research for business professionals.

Indeed, ChatGPT has the potential to revolutionize some existing research methodologies. Consider qualitative research, which collects significant written material from interviews, surveys, or other data collection means. Qualitative scholars must spend a significant amount of time reviewing and coding data. Generative AI tools may be able to more efficiently and potentially more effectively perform some of these tasks. For example, Zhang, Wu, Xie, Lyu, Cai, and Carroll (2023) were able to bolster thematic analysis by using ChatGPT, finding that “[large language models] (such as ChatGPT) can conduct qualitative analysis on corpora through well designed prompts, addressing concerns of human analysts” (pg. 22). These same authors then develop a tool that “not only refines the qualitative analysis process but also elevates its transparency, credibility, and accessibility” (Zhang, Wu, Xie, Kim, Cai, and Carroll 2023, pg. 1). Even if generative AI proves to be less effective than humans at qualitative research, providing the corpora of data from a qualitative research project for other scholars to examine using generative AI has the potential to significantly increase the impact the collected

data can have. Certainly, more research is needed on the positive and negatives of using generative AI for qualitative research, but the potential of these tools are significant.

Additional academic tasks ChatGPT enhances include the production of research proposals (see Cheng et al., 2023), copy editing manuscripts and textbook materials, translating materials to foreign languages, writing emails, brainstorming ideas, finding relevant research (especially when using ChatGPT internet plugins, or ChatGPT through the Bing search engine), producing presentations (see <https://www.beautiful.ai/>), and summarizing research papers. As authors, we use this technology on a daily or near-daily basis in these and other tasks.

We do acknowledge that ChatGPT does have problems. ChatGPT can hallucinate, which is no different from humans. ChatGPT is best thought of as a very good, albeit still imperfect, assistant. 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).

So, what will the future hold in education and scholarship in a generative AI world? We highlight a few potential ideas for how things may change. In terms of scholarship, the journey of publishing the Wood et al. (2023) manuscript is illustrative of future challenges. The Wood et al. (2023) paper began about two weeks after the release of ChatGPT 3.5 to the public (i.e., mid December 2022). Final notice of the acceptance of the paper was received on March 15, 2023—meaning from initial idea to final acceptance took only three months. Yet, the day before final acceptance, ChatGPT 4 was released. ChatGPT 4 substantially improved upon the ChatGPT 3.5 model such that the basic results of Wood et al. (2023) showing students outperformed the generative AI model were put in serious doubt.

The pace of change in the AI sector is so fast, how will academic scholarship keep pace using our current knowledge production and reviewing model? At least in accounting, we are unaware of a paper that was produced, reviewed, and accepted as fast as the Wood et al. (2023) study, and even still as fast as that paper went through the process it was somewhat obsolete upon acceptance. If accounting scholars are going to contribute research findings to guide cutting-edge technology or other fast-paced changes, the model for producing accounting scholarship will likely need updating and improvement.¹⁹

In the classroom, the ability to provide mass customized education is now closer to reality. Generative AI has the ability to adapt learning materials to the interests of individual students and to help serve as a guide to help students better self-diagnose their current understanding and then cater materials to their continual development. The divide in performance between students who want to learn and to excel compared to those who are “checking-the-box” will likely grow. Generative AI will enable dedicated students to advance at a much faster pace and achieve mastery sooner, while students who are just getting by, will be more likely to cheat and over-rely on technology to the detriment of their longer-term learning and progression.

Another change in education will be that faculty will increasingly be more “guide-at-the-side” to students rather than “sage-on-the-stage”, meaning that professors will have to focus more on guiding learners to self-teach and explore rather than have all the answers and just share the answers to students via lectures. The amount of knowledge that is now even more easily accessible via generative AI chatbots will decrease the need for faculty who just know a lot,

¹⁹ An alternative is that accounting scholars cede most or all thought leadership contributions on fast-changing areas and instead study questions that can be examined with old data. While there are certainly interesting questions that are not influenced by the fast-changing business landscape, it would be lamentable that accounting academics would preclude themselves from providing research insights and thought leadership on timely topics.

instead increasing the need for faculty who can help others learn how to teach themselves and produce on their own.

5.2 Generative AI in Accounting and Business

There appears to be significant use of ChatGPT by employees. A survey by 11,793 professionals using the networking app Fishbowl finds that 43 percent of respondents indicate using ChatGPT at work and 68 percent haven't disclosed the use of it to their boss.²⁰ The current use of ChatGPT in business runs from the very basic to complex. Several basic uses of ChatGPT in business include using ChatGPT to generate basic emails, using it to translate emails for multinational corporations (Emett et al. 2023), and using it to “to quickly write reports and prepare compliance documents, analyze and evaluate business strategies, [and] identify inefficiencies in operations or create marketing materials and sales campaigns” (Loten 2023). EY reports that board members are using “generative AI in real time during board meetings as an additional input to brainstorm counterpoints, tweak scenario planning and summarize trends. As one director put it, ‘We can use AI almost like a copilot’” (Kanazawa, Oktem, and Pederson 2023). A survey conducted by KPMG of 2010 companies with more than 1 billion in revenue and 500 or more employees finds that 65 percent of companies are already using AI in financial reporting and 48 percent have deployed or are piloting generative AI in their organizations.²¹

In terms of more sophisticated use, the large accounting firms are starting to develop their own generative AI models. PwC reports entering into a global partnership with AI startup Harvey, backed by the OpenAI Startup Fund, to provide its Legal Business Solutions

²⁰ See <https://www.fishbowlapp.com/insights/70-percent-of-workers-using-chatgpt-at-work-are-not-telling-their-boss/>. This raises concerns of whether ChatGPT is being used as shadow IT, meaning employees using unsanctioned IT to perform their tasks, which can introduce unnecessary risks to the organization (Myers, Starliper, Summers, and Wood 2017).

²¹ See <https://kpmg.com/us/en/webcasts/2023/ai-and-financial-reporting.html>.

professionals with exclusive access to Harvey's AI platform, which uses natural language processing, machine learning, and data analytics to enhance legal work (O'Dwyer, Murgia, and Alim, 2023). The platform will be used to support PwC's global clients, enhancing the ability of PwC's network of legal professionals to deliver solutions in areas such as contract analysis, regulatory compliance, and due diligence. For several years, EY has been leveraging OpenAI's GPT engine to develop its own applications. One of their creations is an AI-driven document reader and classification system, which they use specifically for categorizing receipts and tax-related considerations, demonstrating their incremental approach to the technology's application (Wilkinson, 2023).

EY is using ChatGPT in Azure OpenAI to innovate their payroll services as part of their Next Gen Payroll Platform. They've developed a prototype for a payroll chatbot that can handle complex employee queries using a large language model to analyze extensive compliance data. The EY Intelligent Payroll Chatbot is designed to reduce employers' workload by over 50 percent by answering intricate payroll questions and offering a personalized employee experience. It can understand the specifics of an individual's pay slip and link regulatory compliance aspects with company policies for detailed responses and personalized explanations (EY, 2023).

Bloomberg has developed a new large-scale generative AI model called BloombergGPT. This large language model is specifically trained on a wide range of financial data to support various natural language processing (NLP) tasks within the financial industry (Wu et al., 2023). BloombergGPT is designed to improve existing financial NLP tasks such as sentiment analysis, named entity recognition, news classification, and question answering. It will also unlock new

opportunities for utilizing the vast quantities of data available on the Bloomberg Terminal to better serve the firm's customers (Haas and Gilmore, 2023).

Emett et al. (2023) report that Uniper, an international energy company, is using ChatGPT in the internal audit function, testing its use in audit preparation, fieldwork, and audit reporting. Initial reports suggest efficiency gains ranging from 50 to 80 percent.

This discussion could continue with the many creative and innovative ways that companies are using generative AI. Indeed, OpenAI reports that more than 90 percent of Fortune 500 companies are building tools on OpenAI's platform.²² If we step back, what are the larger takeaways that we are seeing for the effects of generative AI on accounting? So far, we have not seen evidence that generative AI results in accounting job loss; however, survey evidence suggests that 26 percent of employers are considering reducing headcount because of implementation of ChatGPT.²³

Emett, Eulerich, Pickerd, and Wood (2023) find that board members, senior management, and heads of internal audit are in relative agreement that any savings in assurance work from automation (of any kind) will not be redeployed into increasing the amount of assurance, but rather allocated to non-assurance (i.e., consulting) activities. This suggests that accounting firms are likely to see profitability erosion from AI in their audit work and continued growth in providing non-audit services (see Fedyk et al. 2022 for pre-ChatGPT AI evidence on fees). Our discussions with accounting professionals suggest accounting partners are considering whether AI will be able to replace offshoring work as a first area to automate.

²² See <https://www.bloomberg.com/news/articles/2023-11-09/openai-suggests-cyber-attackers-behind-persistent-chatgpt-outage>.

²³ See <https://www.sortlist.com/datahub/reports/chat-gpt-statistics/>.

It is also worth noting that ChatGPT deployment is still very modest in accounting, especially at smaller accounting firms. Recently, one of the authors spoke with nine managing partners for regional accounting firms. These firms are just starting to learn about ChatGPT and consider how it will be used in their firms. While large firms have billions of dollars to invest in these technologies, smaller accounting firms do not and implementing generative AI may not be immediately feasible. This could result in an increased gap between services offered by large and small accounting firms and the necessity for companies that are using technology to have to work with large accounting firms that can understand AI technologies.

In our opinion, it is clear that generative AI is already starting to impact the accounting and business fields. The research evidence and our experiences suggest that generative AI is not just hype but is already being used in substantive ways. The question is not whether generative AI will influence accounting, but how much it will. Our early observations are that generative AI may not be hyped enough for its potential to change the accounting industry in significant ways.

6. 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., 2023a). The most recent advancements in AI, large language model chatbots, will likely 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 some of the “competitive advantages” of the human accountant relative to the machine. To our knowledge, for the first time, AI has performed as well as the majority of human accountants on real-world accounting tasks. This raises important questions about how the machine and accountant will 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.²⁴

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.

²⁴ 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. 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 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 modern technologies. We encourage their continued study and implementation in practice.

References

- Appelbaum, D., & Nehmer, R. A. (2017a). Using drones in internal and external audits: An exploratory framework. *Journal of Emerging Technologies in Accounting*, 14(1), 99-113.
- Appelbaum, D., & Nehmer, R. (2017b). The coming disruption of drones, robots, and bots: how will it affect CPAs and accounting practice? *CPA Journal*, 87(6).
- Austin, A.A., Carpenter, T.D., Christ, M.H., & Nielson, C.S. (2021). The data analytics journey: Interactions among auditors, managers, regulation, and technology. *Contemporary Accounting Research*, 38(3), 1888-1924.
- Bakarich, K. M., & O'Brien, P. E. (2021). The robots are coming... but aren't here yet: The use of artificial intelligence technologies in the public accounting profession. *Journal of Emerging Technologies in Accounting*, 18(1), 27-43.
- Bertomeu, J., Lin, Y., Liu, Y., & Ni, Z. (2023). Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy. *Available at SSRN 4452670*.
- Burton, F.G., Heninger, W.G., Summers, S.L., & Wood, D.A. (2023). Updated perceptions of accounting academics on the review and publication process. *Issues in Accounting Education*, Forthcoming.
- Burton, F.G., Summers, S.L., Wilks, T.J., & Wood, D.A. (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., Summers, S.L., Wilks, T.J., & Wood, D.A. (2022). Relevance of accounting research (ROAR) scores: Ratings of titles and abstracts by accounting professionals. *Accounting Horizons* 36 (2): 7-18.
- Cardinaels, E., Hollander, S., White, B.J., 2019. Automatic summarization of earnings releases: attributes and effects on investors' judgments. *Review of Accounting Studies* 24, 860-890
- Casey, M. J., & Vigna, P. (2018). Blockchain will make today's accountants (and many Wall Street jobs) obsolete. Retrieved from <https://www.marketwatch.com/story/blockchain-will-make-todays-accountants-and-many-wall-street-jobs-obsolete-2018-02-28>.
- Chen, T. C., Kaminski, E., Koduri, L., Singer, A., Singer, J., Couldwell, M., Delashaw, J., Dumont, A., & Wang, A. (2023). Chat GPT as a neuro-score calculator: Analysis of a large language model's performance on various neurological exam grading scales. *World Neurosurgery*, 179, e342-e347. <https://doi.org/10.1016/j.wneu.2023.08.088>.

- Chen, X., Dunn, R., Holt, T., Inger, K., Jenkins, J.G., Jones, J., Long, J., Loraas, T., Mathis, M., Stanley, J., & Wood, D.A. (2023). Artificial intelligence's capabilities, limitations, and impact on accounting education: Investigating ChatGPT's performance on educational accounting cases. *Issues in Accounting Education*, Forthcoming.
- Chen, W., & Srinivasan, S. (2023). Going digital: Implications for firm value and performance. *Review of Accounting Studies*, Forthcoming, 1-47.
- Choudhury, P., Starr, E., & Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, 41(8), 1381-1411.
- Christ, M.H., Emett, S.A., Summers, S.L., & Wood, D.A. (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., Holderness Jr, D.K., Sorensen, T.L., & Wood, D.A. (2019). Robotic process automation in public accounting. *Accounting Horizons*, 33(4), 15-35.
- Cooper, L.A., Holderness Jr, D.K., Sorensen, T.L., & Wood, D.A. (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., & McMickle, P.L. (2017). Can blockchains serve an accounting purpose? *Journal of Emerging Technologies in Accounting*, 14(2), 101-111.
- Coyne, J.G., Summers, S.L., & Wood, D.A. (2023). Enhancing faculty grading productivity using robotic process automation: The development of an SQL automated grading tool. *Working Paper*.
- Coyne, J.G., Summers, S.L., Villanueva, B., & Wood, D.A. (2023). Automating evaluation of skill assessment to promote skill acquisition in the area of data analytics automation. *Available at SSRN 4520670*
- CTStrategies. (2018). 'AVATAR' Technology Utilizes Artificial Intelligence to Interview Border Crossers. Retrieved from <https://ct-strategies.com/2018/06/01/avatar-technology-utilizes-artificial-intelligence-to-interview-border-crossers/>
- Dai, J., & Vasarhelyi, M. A. (2017). Toward blockchain-based accounting and assurance. *Journal of Information Systems*, 31(3), 5-21.

- Deloitte. (2023). Deloitte Announces Strategic Alliance Expansion to Help Solve Industry Challenges With Google Cloud Generative AI. Retrieved from <https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-announces-strategic-alliance-expansion-to-help-solve-industry-challenges-with-google-cloud-generative-ai.html>
- Eisfeldt, A. L., Schubert, G., & Zhang, M. B. (2023). Generative AI and Firm Values (No. w31222). *National Bureau of Economic Research*.
- Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv: 2303.10130v5*.
- Emett, S. A., Eulerich, M., Lipinski, E., Prien, N., & Wood, D. A. (2023). Leveraging ChatGPT for Enhancing the Internal Audit Process—A Real-World Example from a Large Multinational Company. *Available at SSRN 4514238*.
- Emett, S. A., Eulerich, M., Pickerd, J., & Wood, D. A. (2023). Reassessing assurance in an automated world: Insights from board members, managers, and internal auditors. *Working Paper*.
- Eulerich, M., Masli, A., Pickerd, J., & Wood, D.A. (2023a). The Impact of Audit Technology on Audit Task Outcomes: Evidence for Technology-Based Audit Techniques. *Contemporary Accounting Research, forthcoming*, 1-32.
- Eulerich, M., Pawlowski, J., Waddoups, N. J., & Wood, D. A. (2022). A framework for using robotic process automation for audit tasks. *Contemporary Accounting Research, 39*(1), 691-720.
- Eulerich, M., Waddoups, N., Wagener, M., & Wood, D. A. (2023b). Development of a framework of key internal control and governance principles for robotic process automation (RPA). *Available at SSRN 431127*.
- Eulerich, M., Waddoups, N., Wagener, M., & Wood, D. A. (2023c). The dark side of robotic process automation. *Accounting Horizons, forthcoming*.
- Eulerich, M., & Wood, D. A. (2023). A demonstration of how ChatGPT can be used in the internal auditing process. *Available at SSRN 4519583*.
- Estep, C., Griffith, E. E., & MacKenzie, N. L. (2023). How do financial executives respond to the use of artificial intelligence in financial reporting and auditing?. *Review of Accounting Studies, Forthcoming*.

- EY. 2020. *Can robots help your business be more human?* Retrieved from https://www.ey.com/en_gl/digital/can-robots-help-your-business-be-more-human.
- EY. (2023, March 29). *EY announces modernization of payroll employee care using ChatGPT In Azure OpenAI*. Ernst & Young Global Limited. Retrieved from https://www.ey.com/en_gl/news/2023/03/ey-announces-modernization-of-payroll-employee-care-using-chatgpt-in-azure-openai.
- Fedyk, A., Hodson, J., Khimich, N., Fedyk, T. (2022). Is artificial intelligence improving the audit process? *Review of Accounting Studies*, 27(3), 938-985.
- Fenn, J., & Raskino, M. (2008). *Mastering the hype cycle: how to choose the right innovation at the right time*. Harvard Business Press.
- Geatano, C. (2023, May 8). We had ChatGPT take the CPA exam — and it failed. *Accounting Today*. Retrieved from <https://www.accountingtoday.com/news/we-ran-the-cpa-exam-through-chatgpt-and-it-failed-miserably>.
- 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.
- Haas, C. & Gilmore, A. (2023). Introducing BloombergGPT, Bloomberg’s 50-billion parameter large language model, purpose-built from scratch for finance. *Bloomberg*. Retrieved from <https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance>.
- Haase, J., & Hanel, P. H. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *arXiv preprint arXiv:2303.12003*.
- Huang, F., & Vasarhelyi, M.A. (2019). Applying robotic process automation (RPA) in auditing: A framework. *International Journal of Accounting Information Systems*, 35, 100433.
- Hughes-Morgan, C. (2023, July 31). Pearson offers new AI study tools in time for back-to-school. *Bloomberg*. Retrieved from <https://www.bloomberg.com/news/articles/2023-07-31/pearson-offers-new-ai-study-tools-in-time-for-back-to-school?embedded-checkout=true#xj4y7vzkg>.
- Jain, H., Padmanabhan, B., Pavlou, P.A., & Raghu, T.S. (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.

- Jans, M., Alles, M. G., & Vasarhelyi, M. A. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. *The Accounting Review*, 89(5), 1751-1773.
- Jans, M., & Eulerich, M. (2022). Process mining for financial auditing. In *Process Mining Handbook* (pp. 445-467). Cham: Springer International Publishing.
- Jukiewicz, M. (2023). The Future of Grading Programming Assignments in Education: The Role of ChatGPT in Automating the Assessment and Feedback Process. DOI: 10.13140/RG.2.2.22103.85924
- Kanazawa, M., Oktem, C., & Pederson, K. (2023, September 18). Boards are seeking to increase their impact on long-term value and growth by emphasizing oversight of innovation. EY Global. Retrieved November 8, 2023, Retrieved from https://www.ey.com/en_us/board-matters/how-board-oversight-of-innovation-is-evolving.
- Kazim, E., Koshiyama, A.S., Hilliard, A., & Polle, R. (2021). Systematizing audit in algorithmic recruitment. *Journal of Intelligence*, 9(3), 46.
- Kokina, J., & Blanchette, S. (2019). Early evidence of digital labor in accounting: Innovation with Robotic Process Automation. *International Journal of Accounting Information Systems*, 35, 100431.
- Kokina, J., Mancha, R., & Pachamanova, D. (2017). Blockchain: Emergent industry adoption and implications for accounting. *Journal of Emerging Technologies in Accounting*, 14(2), 91-100.
- Kreitmeir, D. H., & Raschky, P. A. (2023). The Unintended Consequences of Censoring Digital Technology--Evidence from Italy's ChatGPT Ban. *arXiv preprint arXiv:2304.09339*.
- Kuechler, W., and M. Simkin. 2005. Multiple-choice tests and student understanding: what is the connection? *Decision Sciences of Journal of Innovative Education*, 3: 73-98
- Kwilinski, A. (2019). Implementation of blockchain technology in accounting sphere. *Academy of Accounting and Financial Studies Journal*, 23, 1-6.
- Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv:2304.07619*.
- Loten, A. (2023, April 26). PricewaterhouseCoopers to Pour \$1 Billion Into Generative AI. Retrieved May 20, 2023, Retrieved from <https://www.wsj.com/articles/pricewaterhousecoopers-to-pour-1-billion-into-generative-ai-cac2cedd>.

- Liu, X., Glocker, B., McCradden, M.M., Ghassemi, M., Denniston, A.K. & Oakden-Rayner, L. (2022). The medical algorithmic audit. *The Lancet Digital Health*.
- Macaulay, A. (2022, September 28). *SAP And Blockchain*. Ignite SAP Resources Ltd. Retrieved May 18, 2023, from <https://ignitesap.com/sap-and-blockchain>.
- Manyika, J., & Sneider, K. (2018, June 1). *AI, Automation, and the Future of Work: Ten Things to Solve for*. McKinsey Global Institute. Retrieved from <https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for>.
- Masli, A., Richardson, V.J., Sanchez, J.M., & Smith, R.E. (2011). The business value of IT: A synthesis and framework of archival research. *Journal of Information Systems*, 25(2), 81-116.
- Moffitt, K.C., Richardson, V.J., Snow, N.M., Weisner, M.M., & Wood, D.A. (2016). Perspectives on past and future AIS research as the Journal of Information Systems turns thirty. *Journal of Information Systems*, 30(3), 157-171.
- Moffitt, K.C., Rozario, A.M., & Vasarhelyi, M.A. (2018). Robotic process automation for auditing. *Journal of Emerging Technologies in Accounting*, 15(1), 1-10.
- Morehouse, L. (2017, June 14). The Technology That Will Change Accounting. Retrieved from <https://www.forbes.com/sites/forbesfinancecouncil/2017/06/14/the-technology-that-will-change-accounting/?sh=3072526d6916>.
- Mökander, J., Schuett, J., Kirk, H.R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500*.
- Myers, N., Starliper, M. W., Summers, S. L., & Wood, D. A. (2017). The impact of shadow IT systems on perceived information credibility and managerial decision making. *Accounting Horizons*, 31(3), 105-123.
- Nordgren, A., Weckström, E., Martikainen, M., & Lehner, O.M. (2019). Blockchain in the fields of finance and accounting: a disruptive technology or an overhyped phenomenon. *ACRN Journal of Finance and Risk Perspectives*, 8, 47-58.
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Working Paper*, Retrieved from https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf.

- O'Dwyer, M., Murgia, M., & Alim, A. N. (2023, March 15). PwC experiments with chatbots to speed up its lawyers Technology from start-up Harvey aims to make contract analysis and due diligence quicker. *Financial Times*. Retrieved May 21, 2023, from <https://www.ft.com/content/463f8cc1-9feb-46ac-a14e-7826c87e2bf>.
- OpenAI. (2023). GPT-4 Technical Report. *Working Paper*. Retrieved from <https://doi.org/10.48550/arXiv.2303.08774>
- Pickard, M. D., Roster, C. A., & Chen, Y. (2016). Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions?. *Computers in Human Behavior*, 65, 23-30.
- Pickard, M. D., Schuetzler, R., Valacich, J. S., & Wood, D. A. (2020). Innovative accounting interviewing: A comparison of real and virtual accounting interviewers. *The Accounting Review*, 95(6), 339-366.
- Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. *In Proceedings of the XXXVII Brazilian Symposium on Software Engineering (SBES '23)* (pp. 293–302). Association for Computing Machinery. <https://doi.org/10.1145/3613372.3614197>
- PWC. (2023, April 26). *PwC US Makes \$1 Billion Investment to Expand and Scale AI Capabilities* PricewaterhouseCoopers. Retrieved from <https://www.pwc.com/us/en/about-us/newsroom/press-releases/pwc-us-makes-billion-investment-in-ai-capabilities.html>.
- 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 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>.
- Rajgopal, S. (2021). Integrating practice into accounting research. *Management Science* 67 (9): 5430–5454.
- Rakhmilevich, M., (2019, September 30). *Blockchain Tables in Oracle Database: Technology Convergence*. Oracle Blockchain Blog. Retrieved May 18, 2023, from <https://blogs.oracle.com/blockchain/post/blockchain-tables-in-oracle-database-technology-convergence>.
- Reuters. (2023, February 2). *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters. Retrieved from <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01>.

- Richardson, V.J., & Watson, M.W. (2021). Act or be acted upon: Revolutionizing accounting curriculums with data analytics. *Accounting Horizons*, 35(2), 129-144.
- Robertson, R.E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-22.
- Rozario, A., & Zhang, C. (2023). The effect of artificial intelligence on the accuracy of management earnings forecasts. *Available at SSRN 3850823*.
- Sanatizadeh, A., Lu, Y., Zhao, K., & Hu, Y. (2023, May 26). Exploring the Effect of Large Language Models on Knowledge Seeking and Contribution in Online Knowledge Exchange Platforms. *Available at SSRN 4459729*.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., ... & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.
- Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of ChatGPT. *arXiv preprint arXiv:2301.08653*.
- Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, 42(9), 1600-1631.
- Wilkinson, A. (2023, Feb. 15). How EY approaches generative AI. Retrieved from <https://www.ciodive.com/news/EY-generative-AI-OpenAI-ChatGPT/642869>.
- Williams, J. J. (1978). A new perspective on the evolution of double-entry bookkeeping. *Accounting Historians Journal*, 5(1), 29-39.
- Wall Street Journal (WSJ) (2023a). KPMG Plans \$2 Billion Investment in AI and Cloud Services. Retrieved from: <https://www.wsj.com/articles/kpmg-plans-2-billion-investment-in-ai-and-cloud-services-e4fd0dd5>.
- Wall Street Journal (WSJ) (2023b). EY Unveils Fruits of \$1.4 Billion Artificial-Intelligence Investment. Retrieved from <https://www.wsj.com/articles/ey-unveils-fruits-of-1-4-billion-artificial-intelligence-investment-ab8d5b5a>.
- Wang, Y., Mukherjee, S., Chu, H., Tu, Y., Wu, M., Gao, J., & Awadallah, A.H. (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., Yao, Q., Kwok, J.T., & Ni, L.M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys*, 53(3), 1-34.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*
- Weiss, B. (2023, May 9). Deloitte, Goldman Sachs, Microsoft among institutional players partnering on new blockchain network. *Fortune*. Retrieved May 18, 2023, from <https://fortune.com/crypto/2023/05/09/deloitte-goldman-sachs-microsoft-among-institutional-players-partnering-on-new-blockchain-network/>
- 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., Achhpilia, M.P., Adams, M.T., ... & Zoet, E. (2023). The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions? *Issues in Accounting Education*, forthcoming, 1-28.
- Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*
- Zhang, C., Issa, H., Rozario, A., & Soegaard, J. S. (2023). Robotic process automation (RPA) implementation case studies in accounting: a beginning to end perspective. *Accounting Horizons*, 37(1), 193-217.
- Zhang, H., Wu, C., Xie, J., Kim, C., & Carroll, J. M. (2023). QualiGPT: GPT as an easy-to-use tool for qualitative coding. *arXiv preprint arXiv:2310.07061*.
- Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2023). Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. *arXiv preprint arXiv:2309.10771*.
- Zhou, E., & Lee, D., (2023, November 3). Generative AI, Human Creativity, and Art. *Available at SSRN 4594824*.

Appendix 1

Example of ChatGPT Reasoning and Acting

Below contains the instructions that are sent along with each question for the ReAct model testing. *Actions* are requests of ChatGPT to the agent to use tools. *Action Inputs* are the inputs ChatGPT sends to the tools. *Observations* are the response of the tools sent to the agent. *Thoughts* are the reasoning of ChatGPT.

Prompt we give ChatGPT:

Answer the following question. Before you answer, explain your thoughts. If you don't know where to begin, use the search tool to get help. You are encouraged to use the tools when you are unsure about the correct answer.

Use calculator ONLY when you have numbers for input.

You have access only to the following tools:

{tools}

Always use the following format:

Question: the input question you must answer

Thought: you should always think about what to do

Action: the action to take, should be one of [{tool_names}]

Action Input: the input to the action

Observation: the result of the action

(this Thought/Action/Action Input/Observation can repeat N times)

Thought: I am ready to answer

My answer is:

Begin! Remember: Always follow the above format

Question: {input}

{agent_scratchpad}

Examples of ChatGPT Output

Question: how much is \$45134 in euros?

Thought: I need to find the current exchange rate from dollars to euros.

Action: Search

Action Input: current exchange rate dollars to euros

Observation: Convert US Dollar to Euro ; 10 USD, 9.05704 EUR ; 25 USD, 22.6426 EUR ; 50 USD, 45.2852 EUR ; 100 USD, 90.5704 EUR.

Thought: I can use the exchange rate of 1 USD to 0.905704 EUR to calculate the amount in euros.

Action: Calculator

Action Input: 45134 * 0.905704

Observation: Answer: 40878.044336

Thought: I am ready to answer

My answer is: 40878.044336

Appendix 2

Results from Additional Training

We perform another sensitivity test around the number of training examples we provide. One limitation of few-shot training is that the user is limited to the number of tokens allowed in the prompt. That is, the prompt for GPT 3.5 and GPT 4 is limited to 4,097 and 8,192 characters, respectively. We chose 10 examples because the number of tokens would always be closely under the limit for the GPT 3.5 model.

To train a model with more examples, one can fine-tune a new model. At the time of writing this paper, fine-tuning is not available for the most recent models, namely GPT 3.5 and GPT 4. Therefore, to perform fine-tuning we use GPT-3 Davinci and Curie. Curie is smaller model, and Davinci is the base model for GPT 3.5, but does not have the same level of accuracy.

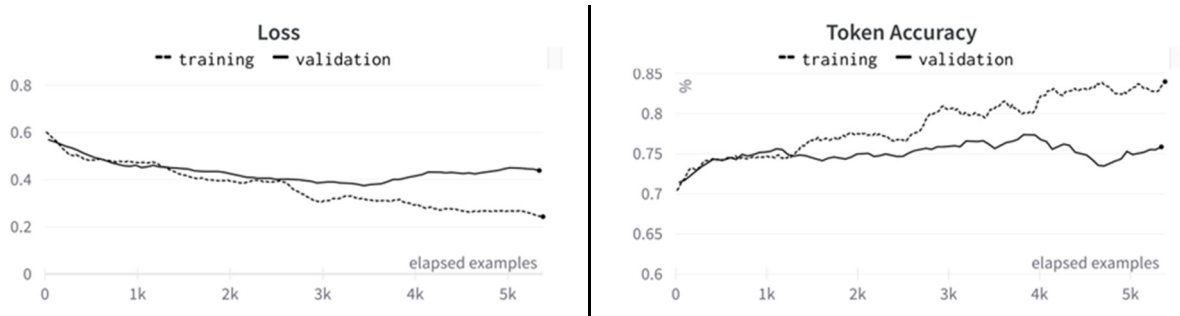
We divide our sample into 3 group of training, validation, and test as 60, 10, 20 percent of the number of questions for each section. We remove questions that are duplicated. The training sample is used to fine-tune the models, while the validation is used to evaluate the performance of each trained model. We train two variations of each model and report the average loss and accuracy for brevity. Our experiments here is to see how well the model predicts the explanation of any given question. Therefore, we train the model using the question and the expected explanation along with the correct answer. We expect that models would become better at sampling when fine-tuned.

Results for the audit section (AUD) of the CPA exam are reported in the Appendix 2 Figure. In the figure, validation *Token Accuracy* shows how well the models' generated explanation compares to the actual explanation of the validation sample. Training *Token Accuracy* shows how well the models' generated explanations compare with the training actual explanations. We find that validation *Token Accuracy* on average, increases for about 6 percent.

These trends, however, are seen for up to 4,000 examples. In other words, the models' predictions do not improve after that. In fact, we may be overfitting the model to the training sample as the models get better at predicting the expected explanations for the training samples.

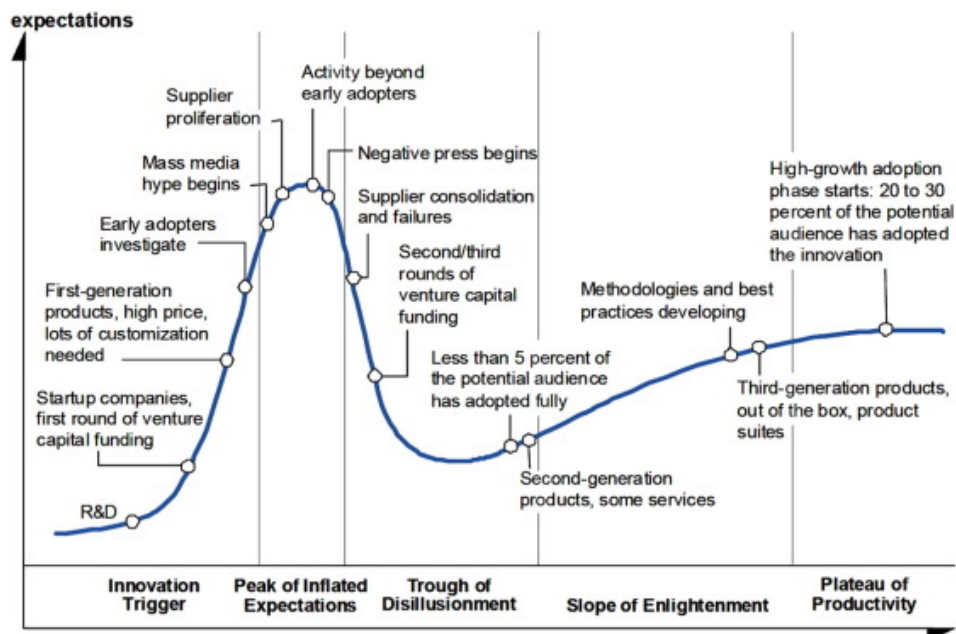
Thus, although we showed that 10-shot training can improve models, as the ChatGPT models improve, using larger training samples may result in greater model improvement. Finding the optimal hyperparameters for fine-tuning is a key step in generating the "best" model. Because fine-tuning ChatGPT 4 is not achievable at the time of authoring this article, future studies may examine whether a model fine-tuned on accounting materials can outperform the general model.

Appendix 2 Figure Figures Showing Effects of Different Amounts of Training on Models



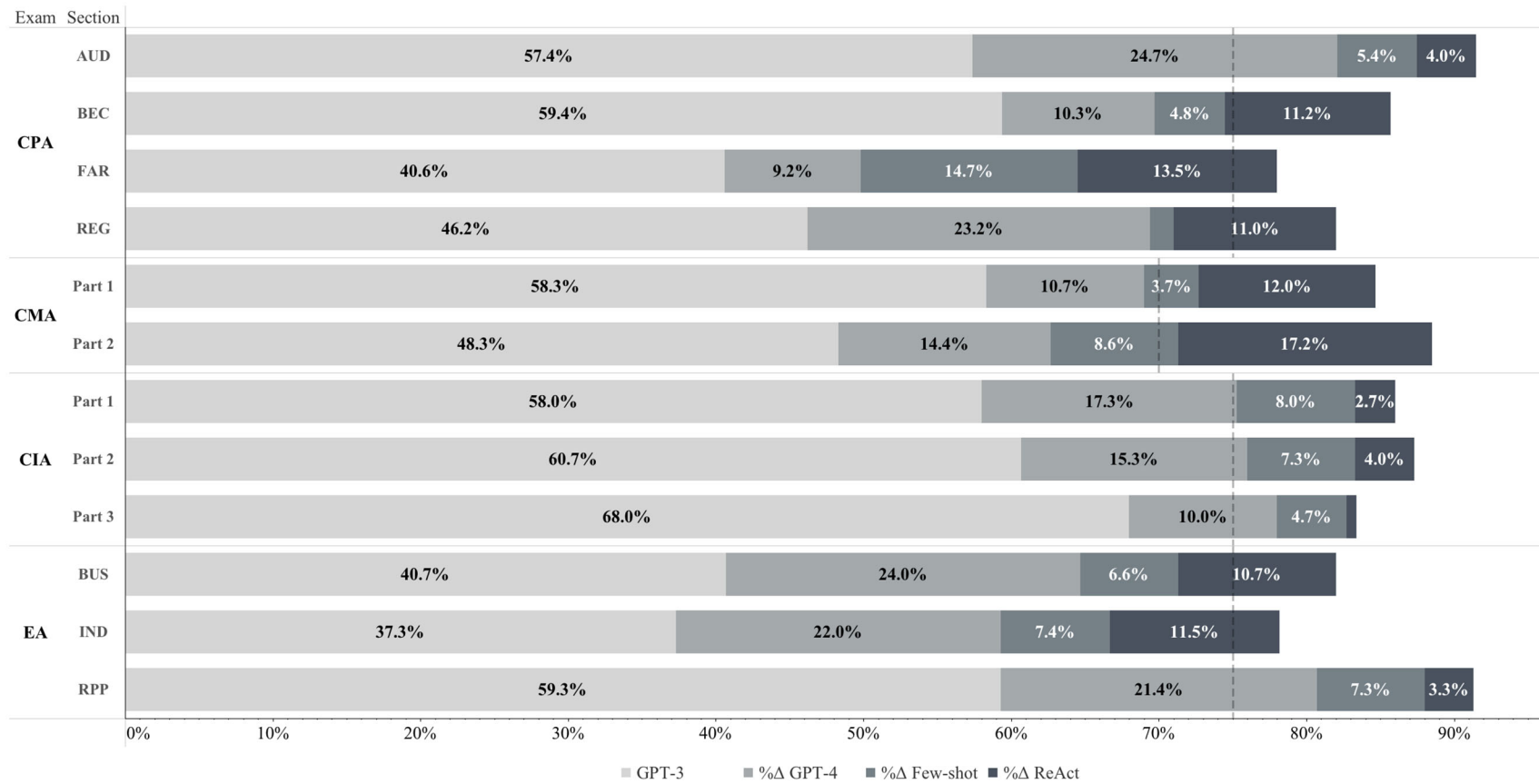
The charts report the average loss and accuracy of the models. The loss graph (on the left) displays the errors made in training or validation sets for each additional example. The lower the loss, the better a model. The best predictive model would be where the validation error has its global minimum. In these models, we observe that the minimum validation loss is around 3,400 examples after which the validation loss starts to increase. Token accuracy graph (on the right) shows the percentage of tokens (e.g., pieces of words) in the training or validation sample that were predicted correctly with each additional example. We observe that training accuracy for the validation sample drops after around 4,000 examples and the training accuracy increases. This may suggest models trained on more than 4,000 examples are overfitted to/memorizing the training sample.

Figure 1
Depiction of the Gartner Hype Cycle



The hype cycle and its stage indicators (adapted from Fenn and Raskino, 2009).

Figure 2
Model Performance Improvement



Performance improvement in each section of each exam

Table 1
Descriptive Statistics

Certification	Required Pass Rate %	Number of Questions Examined		
		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

Certification	Section	GPT 3.5	GPT 4	Improvement	Best Case Pass/Fail
		% Correct	% Correct		
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

Certification	Section	Zero-Shot	10-Shot	Improvement	Best Case Pass/Fail
		% Correct	% Correct		
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

Certification	Section	No ReAct	ReAct	Improvement	Best Case Pass/Fail
		% Correct	% Correct		
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 rate for each type of question is computed.