JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING
Vol. XX, No. XX
MONTH YEAR
pp. 1–21

American Accounting Association
DOI: 10.2308/JETA-2023-066

Hey ChatGPT—Is a Louis Vuitton Bag an Investment? Evaluating LLM Readiness for Use in Financial Literacy and Education

Stacey Taylor
Cape Breton University

Samantha Taylor
Shannon Lin
Vlado Keselj
Dalhousie University

ABSTRACT: The prevalence of large language models (LLMs) such as ChatGPT has wowed the world with its ability to generate text in a human-like manner. While educators evaluate how AI will impact the future of learning, we identify mistakes ChatGPT has made. We further extend this concern to nonfinancially sophisticated users seeking to improve their financial literacy who may not possess the financial acumen to determine when the AI is hallucinating. Using a longitudinal study, our analysis frames the prompts and subsequent findings within the four stages of the Dunning-Kruger effect to explore how users of varying expertise receive output from the LLMs. We find that ChatGPT cannot always fully distinguish between three different user groups. Our findings have important implications for accountants, educators, and students using LLMs as a tool in work and education and for the general population looking to bypass financial experts for their personal finance needs.

Data Availability: Data will be made available upon request.

JEL Classifications: M41.

Keywords: artificial intelligence; large language models; ChatGPT; financial literacy; accounting education; longitudinal research.

I. INTRODUCTION

Since its release to the public on November 30, 2022, there has been considerable hype around ChatGPT. These models have proven to be capable of emulating human-like text, with its feats and defeats shared widely across news outlets as well as social media. ChatGPT has "gone" to law school (Choi, Hickman, Monahan, and Schwarcz 2023a; Bommarito 2022), become theoretically certified and advanced sommeliers, participated in Leetcode challenges at all levels, and taken the Graduate Record Exam, passing with flying colors (OpenAI 2023a). Companies such as TikTok, Microsoft, Shopify, and SalesForce have already implemented ChatGPT and are using it as part of their daily operations (Leswing 2023; Tellez 2023).

Certain aspects of accounting, such as bookkeeping, have been automated for years, with software that becomes more capable over time. Due to the repetitiveness of the job, and perhaps a professional image issue, it has been widely speculated for decades that accountants will lose their jobs to machines. So, is artificial intelligence (AI) ready to take

We thank David A. Wood and Miklos A. Vasarhelyi (editor) for their valuable feedback, which we have incorporated to enhance the quality of our manuscript. The authors of this manuscript have no conflicts of interest related to this research.

Stacey Taylor, Cape Breton University, Shannon School of Business, Department of Financial and Information Management, Sydney, Nova Scotia, Canada; Samantha Taylor and Shannon Lin, Dalhousie University, Faculty of Management, Department of Accountancy, Halifax, Nova Scotia, Canada; Vlado Keselj, Dalhousie University, Faculty of Computer Science, Halifax, Nova Scotia, Canada.

Editor's note: Accepted by Senior Editor Miklos A. Vasarhelyi.

Submitted: December 2023
Accepted: June 2025
Early Access: July 2025

on accounting jobs? To do this, the first hurdle for ChatGPT is to get an education like the rest of us. In its short life, ChatGPT has written many tests that are designed to be barriers to entry in terms of intelligence and knowledge. Its recent and unexpected successes are pushing test makers in various disciplines to consider redesigning the tests to maintain the ability to assess the quality of the test-takers. Similarly, the education domain, especially post-secondary education, is confronted with the challenge of how to adapt to the existence of this technology.

ChatGPT is a conversational agent or service based on models created by OpenAI and uses several large language models (LLMs)—GPT-3.5, GPT-4, and GPT-4 Omni (better known as GPT-4o). These LLMs are based on deep learning neural networks using the newest models known as "Transformers." To be clear, ChatGPT is the application that users interact with, and GPT-3.5, GPT-4, and GPT-4o are the LLMs that power ChatGPT and execute the critical operations that understand and generate the language.

To extend Yue, D. Au, C. Au, and Iu (2023) findings that using ChatGPT can make financial concepts more widely accessible across a larger audience, our overarching **research objective** is to determine whether ChatGPT could be used for accounting and financial literacy—both inside and outside of the classroom. To answer our objective, we present a longitudinal assessment of ChatGPT over the evolution of its life to January 2025. Our work, therefore, presents longitudinal evidence of its growth under evolving LLMs, specifically GPT-3.5, GPT-4, and GPT-4o. Accordingly, we instructed the ChatGPT LLMs (GPT-3.5, GPT-4, and GPT-4o[1]) with 75 prompts composed of terms from accounting and finance for various end-users, ranging from *no audience specified* to *financially sophisticated*. We find that, on average, ChatGPT does not tailor its explanations well for its audiences, even when prompted to do so. We also show that ChatGPT has some problematic explanations that could cause users who are unfamiliar with accounting and finance to take them at "face value," leading to incorrect understanding. This affects student learning, and financial illiteracy has also been shown to negatively affect retirement, stemming from an inadequate understanding of basic financial concepts (Xue, Gepp, O'Neill, Stern, and Vanstone 2019; Bucher-Koenen and Lusardi 2011; Lusardi and Mitchell 2014; Earl, Gerrans, Asher, and Woodside 2015). We also note that ChatGPT can be confidently incorrect, leading to confusion and harm (Rudolph, S. Tan, and S. Tan 2023; Frieder et al. 2023). Overall, we find that ChatGPT answers have high potential to mislead or miseducate learners, particularly in situations where professional judgment is required.

The literature on ChatGPT is quickly developing much like the technology itself, but to the best of our knowledge, there is a void when it comes to the benchmarking of its abilities or a comprehensive understanding of who its users are at the domain level. This research fills this gap by creating a foundational benchmark for LLMs being used to teach and enhance financial literacy by evaluating responses to prompts on fundamental accounting and finance concepts. We also make several contributions to the literature, particularly in the areas of financial literacy and education. First, we demonstrate through a *prompt-and-answer* model that ChatGPT cannot be fully relied on to explain and interpret accounting and finance terms or scenarios correctly across a compendium of potential users. As this technology is so new (less than 1 year at the time of this research), we are the first to evaluate GPT's abilities and reliability for "instructional" use (whether as a student, teacher, or lay-user) for the accounting and finance domain.

Second, we evaluate readability of the ChatGPT-generated answers using the Flesch Reading Ease Score (FRE). Readability has long been of great concern in the domain of accounting and finance, given the complex nature of the terminology and subject matter. Our analysis shows the evolution of readability over the three LLMs and find that the largest improvement was made for the *financially unsophisticated* and *nonfinancial user* group. The *general audience* users also saw some improvements, whereas the *no audience specified* and the *financially sophisticated* and *financial user* groups saw only marginal improvements.

Finally, we contribute the transcripts from the completed responses.[2] We believe this is important because large language models will continue to advance and mature, so it is critical to be able to benchmark to find gaps in the training of these models. Further, as ChatGPT generates different answers each time it is queried, this affects the ability of the academy to reproduce our exact findings. Over time, the GPT models have been updated and upgraded several times,[3] making research hard to compare without transcripts gathered at the time. As such, transcripts are necessary to both support this research and provide future research "point in time" response benchmarks. Given that there is a great desire and motivation to automate tasks, a rigorous evaluation of the technology should first be undertaken. Our research demonstrates, contrary to popular belief, that ChatGPT has some serious limitations that impede learning

[1] At the time of the initial research using GPT-3.5 and GPT-4, GPT-3.5 was a free service to the public, whereas GPT-4 was a paid subscription known as ChatGPT Plus. At the time of the conclusion of this longitudinal assessment, GPT-3.5 was deprecated and retired as of January 2, 2025. The only available models at this time were GPT-4 (now considered a "legacy model") and GPT-4 Omni (better known as GPT-4o).

[2] Data will be made available upon request.

[3] Upgrading and updating LLMs can cause model drift if there is an unexpected behavior shift or concept drift where the LLM fails to adapt to new information and trends. GPT-4o flagged that if an LLM is not trained on new data as tax laws change, for example, then it could easily be recommending outdated tax laws, as it will not automatically know that the laws have been updated or changed.

and could cause significant problems for people attempting to bypass traditional financial guidance by turning to ChatGPT for financial guidance—a phenomenon that is already occurring (Bank of Montreal 2023; DeVon 2023; Adejo 2023; Millan 2024).

This paper is organized as follows: Section II examines the related work and provides a brief introduction to the LLMs GPT-3.5, GPT-4, and GPT-4o. Section III details our methodology. Section IV discusses our results. Section V concludes and presents some ideas for future work.

II. RELATED WORK AND BACKGROUND ON CHATGPT

Related Work

At the time of this research, there are few papers that address ChatGPT in the domain of accounting and finance; only a few of these papers address how ChatGPT can be used for financial literacy. Yue et al. (2023) tested ChatGPT's capabilities to explain complex financial models to "non-financial professionals," although this user was not defined. Ullah, Ismail, Khan, and Zeb (2024) found that financial literacy influences whether users will use ChatGPT to help make effective investment decisions or not; those with higher financial literacy can better understand and apply ChatGPT's insights and analysis, helping them to make better, more informed decisions. Wood et al. (2023) looked at how ChatGPT fared in answering accounting questions and concluded that it did better than the student average for 15.8 percent of assessments. Alshurafat (2023) examined the utility of ChatGPT to the accounting profession and provided an overview of how it could be used by accounting professionals. Although Alshurafat finds that ChatGPT can provide valuable assistance in data analysis and automation of routine tasks, the author also raised significant concerns. The first is regarding the accuracy and consistency of responses and the second was ChatGPT's position above regulatory and standard changes in the accounting domain. In other words, OpenAI's development team is in full control to further train or fine-tune its models with no regulatory oversight.

Street and Wilck (2023) studied how ChatGPT-3.5 and other large language models could be used in the field of forensic accounting. Similar to our research, they use a "prompt-and-answer" model to conduct their research by providing it with accounting tasks. Street and Wilck (2023) used question scenarios to have ChatGPT generate journal entries and financial statements and found that ChatGPT made many mistakes and exhibited strange behaviors. These include learning assumptions and using it as "ground truth" (i.e., treating the learned assumptions as fact) as well as plugging numbers rather than calculating them. The authors conclude that ChatGPT does not currently have sufficient expertise in the domain of accounting and lacks consistency in explaining its rationale for its answers (i.e., the "why" for what it has done).

Given its abilities to generate human-like text, ChatGPT was adopted early into education for two distinct reasons: to see how it could be used in the classroom to enhance learning and to see how it could undermine the veracity of assignments and examinations. One of education's chief concerns is that students will outsource written assignments to ChatGPT rather than do the work themselves (Rudolph et al. 2023).[4] Qadir (2023) argues that technology often disrupts education and uses Massive Open Online Courses as an example. He furthers that by also addressing other effects of AI like job displacement (particularly for low-skilled workers), reminding us that education is at the forefront of guiding the technological transition. Nobel laureate Daniel Kahneman stated, "The questions we ask shape the answers we get. Asking the right questions is the first step in finding the right answers." Although Kahneman, whose Nobel was in economics, said this before the arrival of LLMs, his wisdom is extremely applicable to the use of AI. Even OpenAI warns users that the model makes errors and provides incorrect answers (OpenAI 2023a, 2023b) and fabricates citations (McGowan et al. 2023). Such errors have also been referred to as hallucinations (Choi, Fang, Wang, and Song 2023b) and have been classified as imitative falsehoods and factual errors (Cheng et al. 2023). Unfortunately, because ChatGPT and other LLMs produce a single answer to a prompt versus a traditional search engine (e.g., Google), which provides several ranked answers, users do not have a way to evaluate the relative strength of the answer.

As such, teachers will have to learn how to best use the technology and students will also have to learn critical reasoning skills to discern when to ask the AI for more information and when to seek knowledge from traditional and reliable sources. Because ChatGPT lacks thoughtful reasoning like humans, the fact that ChatGPT can pass the U.S.

[4] Generated text detection (i.e., not written by a human) software exists, such as Hugging Face's OpenAI detector or GPTZero. These types of software, which Rudolph et al. (2023) explain estimates the probability that the text was not written by a human (Tate, Doroudi, Ritchie, and Xu 2023; Sandlin 2022; Mills 2023; McMurtrie 2023; Montclair State University 2023; Yousif 2023). We note that AI detectors do not currently exist for commonly distributed works (such as the American Constitution) or for non-neurotypical writers.

Medical Licensing Exam is more a testament to the format of the exam, which rests more on memorization than whether a practitioner can apply their knowledge to nonstandardized situations (Mbakwe, Lourentzou, Celi, Mechanic, and Dagan 2023).

Brief Introduction to ChatGPT, GPT-3.5, GPT-4, and GPT-4o

ChatGPT is a conversational agent or service based on models created by OpenAI and uses several LLMs—GPT-3.5, GPT-4, and GPT-4 Omni (better known as GPT-4o). These LLMs are "deep learning algorithm[s] that can recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets" (Lee 2023) and use deep learning neural networks using the newest models known as "Transformers." ChatGPT is a chatbot service, whereas GPT-3.5,[5] GPT-4,[6] and GPT-4o[7] are LLMs that drive the service. GPT-3.5 has been fine-tuned[8] and trained on a number of datasets (Brown et al. 2020; Chen et al. 2021; Neelakantan et al. 2022; OpenAI n.d.b; Ouyang et al. 2022; Stiennon et al. 2020), including the massive open dataset called the *Common Crawl*, which is a scraping of the internet over 12 years (commoncrawl.org n.d.), using over 175 billion parameters for training using "Azure AI supercomputing infrastructure" (OpenAI n.d.b). Although OpenAI does not give many details on what specifically was used for training GPT-4 or GPT-4o, it does indicate that publicly available datasets were used, including "internet data," which we can infer to be the *Common Crawl* (OpenAI 2023a, n.d.a).

To fine-tune the models, reinforcement learning—a type of learning based on reward (Sutton and Barto 2018; Kaelbling, Littman, and Moore 1996)—is used. OpenAI also incorporates human feedback into its reinforcement learning process to guide the model's behavior (OpenAI 2023a). Also, OpenAI has not released the specific number of parameters used in training GPT-4. There has been some speculation that GPT-4 uses 1.8 trillion parameters (E2Analyst 2023). Whatever the true number is, it is reasonable to expect it exceeds GPT-3.5's 175 billion.

Before its deprecation and retirement, with a free account, a user could interact with ChatGPT that uses the fine-tuned version of GPT-3.5 as its LLM. Before the release of GPT-4o, access to GPT-4 was through a paid subscription only. As of January 2025, users (without the need of an account) can use the core functionalities of GPT-4o only. A paid subscription provides access to the full array of GPT-4o as well as access to the GPT-4 legacy model. As we are evaluating the language models in this research, we will be referring to the language model itself (GPT-3.5, GPT-4, and GPT-4o) rather than ChatGPT (unless we are specifically referring to the chatbot platform). Please see Appendix A for more on ChatGPT and LLMs.

III. METHODOLOGY

User Definitions

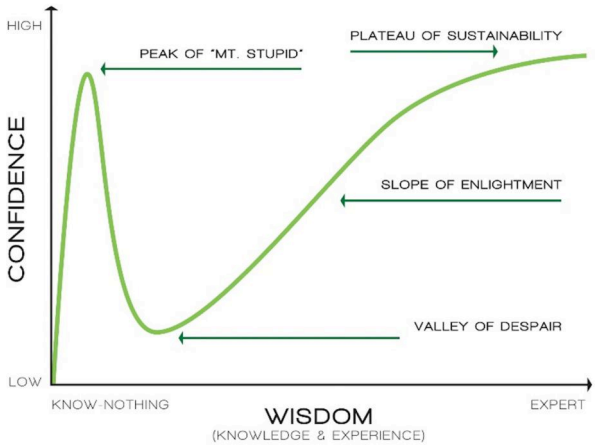
To communicate effectively, accountants first need to determine the level of financial understanding the audience has based on either interacting with or making reasonable assumptions about the audience. For example, an accounting instructor teaching a fourth-year advanced accounting class could reasonably expect that students in the class have a solid foundational understanding of accounting. On the other hand, a tax accountant completing a new client's taxes could reasonably assume that the client is not overly familiar with tax concepts and may need extra explanations and guidance. When doubt exists, the typical approach is to make communication understandable to a general audience.

There is no universal definition of the different types of financial and nonfinancial users. The Cambridge Dictionary defines the term *user* as "someone who uses a product, machine, or service" (Cambridge University Press n.d.). Even when applied to a *financial* situation, it is a very wide definition, encompassing anyone who interacts with financial information. The *Handbook of International Education Pronouncements* simply refers to "users" and does not elaborate on what defines users (The International Federation of Accountants 2019). A review of the publicly available case studies published by the Chartered Professional Accountants of Canada (CPA) also does not define users. In fact, CPA candidates writing responses to cases are expected to identify the users and infer, from the case facts, the level of financial understanding to tailor their responses appropriately (Marchiel 2019; Taylor, Laguduva, and Bury 2023). For example, if the user in a case is determined to be the audit partner, a CPA candidate will not have to explain Generally Accepted Accounting Principles (GAAP). But if the user is a client who does not currently use any accounting framework, the candidate is then expected to provide a brief explanation of GAAP so the client understands the importance of using an accounting framework and how that will impact their accounting.

[5] GPT-3.5 was released November 2022.
[6] GPT-4 was released March 2023.
[7] GPT-4o was released May 13, 2024.
[8] Fine-tuning involves further refinement in the training through exposing the model to more examples of desired outcome data.

FIGURE 1

Illustration of the Dunning-Kruger Effect



The cognitive bias demonstrated by those with limited knowledge of a subject overestimate their abilities in that area (Dunning 2011). (The full-color version is available online.)

Further investigation in the area reveals that there is some identification of who *financial* users could be, but no formal parameterized definition. Section 407 of the Sarbanes-Oxley Act addresses who can be identified as an "audit committee financial expert" (SEC 2003; U. Hoitash, R. Hoitash, and Bedard 2009). Originally, only those with "accounting experience" could serve in this role. However, in the final ruling, the U.S. SEC determined that "people actively engaged in industries such as investment banking and venture capital investment…[and] professional financial analysts" could also be designated as financial experts, given their experience with financial statements. The SEC also determined that a "sophisticated investor" is an individual with "sufficient knowledge and experience in financial and business matters to make them capable of evaluating the merits and risks of the prospective investment" (SEC 2003). The *Statements of Financial Accounting Concepts* also points to general purpose users as those with "existing and potential investors, lenders, and other creditors in making decisions about providing resources to the entity. Those decisions involve buying, selling, or holding equity and debt instruments and providing or settling loans and other forms of credit" (FASB 2021). Hoitash et al. (2009) examined different types of experts: accounting financial experts, supervisory financial experts, and user financial experts. They define this last category as "individuals with experience performing extensive financial statement analysis or evaluation (e.g., financial analysts, investment bankers)," in the context of expertise for service on the audit committee.

The Dunning-Kruger model[9] describes the bias where unskilled individuals overestimate their abilities due to a lack of self-awareness about their incompetence (Kruger and Dunning 1999). Interestingly, it is those users who have some wisdom that are of the highest groups with the most confidence, as seen in Figure 1. ChatGPT can exhibit the same lack of (self-)awareness about its own incompetence and where the "line" is between analysis and professional advice.

Therefore, to investigate ChatGPT's ability to address the needs of a broad range of users and to evaluate how well ChatGPT "understands" different user needs, we have selected six different users:

- **Financially unsophisticated user**—a user who lacks basic knowledge of accounting and finance.
- **Nonfinancial user**—a user who has a basic knowledge of accounting and finance.
- **Financial user**—a user who has a good grasp of accounting and finance through training, experience, or a combination of the two.
- **Financially sophisticated user**—a user who is an accounting and financial professional, which is achieved through extensive training and a designation, or who has extensive experience in a corporate position involving financial leadership.

[9] Although there is some debate in the intelligence literature about whether the Dunning-Kruger effect exists or is a statistical artifact (e.g., Gignac and Zajenkowski 2020; Dunkel, Nedelec, and van der Linden 2023), this paper uses the Dunning-Kruger effect for theoretical development because using this model can help readers conceptualize the development of accuracy of the GPTs overtime given a specific audience orientation.

- **No audience orientation identified**—this type of user is when ChatGPT is not provided any information on the user(s). In this type of situation, humans would tailor their communications to be more broad, more general, and typically high level, understanding that if more information was revealed on the knowledge and understanding of the user(s) in the course of communicating with the user(s), the communication approach would change as new information became available or was revealed.
- **General audience**—a collective group of users where there is a wide range of knowledge and experience with accounting and finance, ranging from a financially unsophisticated user to a financially sophisticated user.

Using these user definitions, we separate users into three main groups as follows:

Group	"User" as Defined Above	Dunning-Kruger Categorization
Baseline group—no audience orientation and a general audience	**No audience orientation identified**—this type of user is when ChatGPT is not provided any information on the user(s). In this type of situation, humans would tailor their communications to be more broad, more general, and typically high level, understanding that if more information was revealed on the knowledge and understanding of the user(s) in the course of communicating with the user(s), the communication approach would change as new information became available or was revealed.	This lack of user orientation was used specifically to gain insights into ChatGPT's baseline when not given direction as to the user. Accordingly, there is no matching Dunning-Kruger categorization for this user group.
	General audience—a collective group of users where there is a wide range of knowledge and experience with accounting and finance, ranging from a financially unsophisticated user to a financially sophisticated user.	The General audience has a wide range of potential user groups, which, in a similar sentiment, mirrors our intent not to have one particular Dunning-Kruger categorization for this group. Rather, this user group would apply effectively to all (except the "know-nothing") user group.
Group 1—financially unsophisticated and nonfinancial users	**Financially unsophisticated user**—a user who lacks basic knowledge of accounting and finance.	The "**Know-Nothing**" group, as they lack the basic knowledge of accounting and finance. This person would not have a basis to ask follow-up questions if presented with incorrect or potentially misleading responses.
	Nonfinancial user—a user who has a basic knowledge of accounting and finance.	The "**Peak of 'Mount Stupid'**" group, as this type of user understands the basics but does not yet have the wisdom to know what they do not know. This group is unlikely to ask follow-up questions as they do not yet possess the critical or professional judgment to assess accurate from inaccurate responses.
Group 2—financially sophisticated and financial users	**Financial user**—a user who has a good grasp of accounting and finance through training, experience, or a combination of the two.	The "**Valley of Despair**" group, as the financial user group has some training and/or experience and is aware that there are many things they do not know and will often bring a critical lens to the matter they are examining. Given their level of wisdom and lack of confidence, these users are likely to ask follow-up questions or request clarity should the response be misaligned with their training or experience.
	Financially sophisticated user—a user who is an accounting and financial professional, which is achieved through extensive training and a designation, or who has extensive experience in a corporate position involving financial leadership.	Those in the "**Slope of Enlightenment**" group are financially trained individuals who have gained wisdom through their training. These individuals know what they know—and what they do not yet know—and are able to ask follow-up questions when responses received from ChatGPT are not accurate.

The rationale behind these groupings is that we would expect that responses to the various prompts would be similar between the members of each group and dissimilar compared with other groups. Consistent with the Dunning-Kruger effect (Dunning 2011), users of ChatGPT do not know what they do not know. Thus, insights into the various user categories help to provide awareness as to just how helpful (i.e., accurate) the various versions of ChatGPT are. Despite the ability for a user to ask clarifying questions, as outlined above and seen in Figure 1, many users would "know what they do not know," and, although some seek clarity, others may not.

Chatbot Prompt Development and Dataset

Prompts

ChatGPT was prompted to explain basic financial reporting and finance questions. As any of ChatGPT's LLMs can learn from a user's previous conversations, separate and distinct chat windows were used when prompting to ensure there was no "lookback contamination" of the results.[10]

Our prompts were influenced by anticipating how nonfinancial and financially unsophisticated users would generate their questions. This is because those users would not know how to ask a financially appropriate question as they are seeking knowledge they are likely unfamiliar with. Although a financially sophisticated user may be able to ask ChatGPT a range of prompts and use a range of specificity with their prompts, our financially unsophisticated and nonfinancial users likely cannot. As such, to ensure comparability between the prompts, we carefully wrote our prompts to reflect the most modest level of financial sophistication and then "scaled up" by advising the LLM as to the level of financial sophistication of the users, where appropriate. As well, Zamfirescu-Pereira, Wong, Hartmann, and Yang (2023) point out that although designing a useful prompt may seem like a very straightforward task, there are many barriers that create challenges for non-AI experts in designing an effective prompt.

We focused our inquiry on five key accounting and finance concepts: Net Income, Net Revenue, Assets, Liabilities, and Investments. These key terms were selected because they are deemed fundamental terms that one would come across in financial statements, financial communications, and financial education. Each prompt uses a standard format: "Explain <insert term here> to a <insert user here>." This format was consistently used for the GPTs to ensure the veracity as well as the comparability of the data. These types of prompts use a "zero-shot" approach—a very basic type of question that is not (or is unlikely) to be in the training data, and the user provides no additional context in the prompt to enrich the generated text. This means that the LLM must rely solely on its training to generate an appropriate response (Syed and Gadesha 2025).

Dataset

Responses generated by ChatGPT were saved in individual plain text files using UTF-8 encoding. To ensure the consistency of the text's presentation to the algorithms, spaces between paragraphs were removed so that all the text is saved in one large paragraph. Numbered lists were kept (with spaces removed between numbered items). Bullet point lists were also kept in the same fashion as the numbered lists, but bullets were removed. Unlike numbered lists, bullet points did not have ending punctuation. Therefore, to ensure that the sentences were grammatically correct, periods were added at the end of every bullet point list item. This also served to delineate bullet points. The text itself, however, was not altered in any way.[11]

One of the distinctive features of our study is that it is archival in nature, as we have collected data from ChatGPT over time using three different LLMs (GPT-3.5, GPT-4, and GPT-4o). That also means that if we, as users, experienced difficulties with the models or the ChatGPT platform, those challenges are preserved in our dataset and subsequently in our results. Data collection was relatively straight forward for both GPT-3.5 and GPT-4o. At the time of the data collection for GPT-4, however, we encountered significant challenges as user demand was very high. Due to this, OpenAI limited the number of messages to 25 every three hours. Although this limit did not pose a problem for our work, it did mean that GPT-4 returned incomplete answers. In some cases, we were able to coax GPT-4 to finish the answer by providing feedback that the response was unhelpful as it was unfinished. In most cases, GPT-4 would generate another incomplete response. Markovski (2023) recommended specifying the desired length in the prompt (e.g., "A

[10] In this research, we use the word "conversation" to mean different lines of inquiry. As such, these "conversations" can occur within the same chat (context) window on ChatGPT, meaning that one does not have to formally close the chat window and open a new one for it to constitute a new conversation. The extent to which GPT "remembers," even though we have changed to different topics of inquiry, will be dictated by the context window (approximately 3,000 words for GPT-3.5 and approximately 6,000 words for GPT-4) as well as the timeout feature, which signals to GPT that the conversation has "concluded." No information has been provided by OpenAI as to what amount of time they have programmed in, meaning that it could range from minutes to hours to days.

[11] Transcripts of all prompts and data are available upon request.

list of ten science fiction books"). Although this can work well for a list setting, it does not work well for explaining financial terms; if a maximum length is specified, although GPT-4 might provide a complete textual response, it could still return an incomplete explanation, having left out pertinent information to meet the length requirements. As well, GPT-4 was limited to 8,000 tokens (pieces of words)[12] (OpenAI 2023d, 2023c), which equates to ~32,000 characters[13] in total[14] between prompt and response. It is important to also note that characters include alphanumeric characters (letters and numbers) as well as spaces between words and punctuation. Therefore, to specify the maximum length, a user would need to determine the length of the prompt and then subtract that amount from the overall 32,000-character length, which is impractical. As such, we did not specify a response length in our research. Under the initial release of GPT-4o, the output length was limited to 4,000 tokens. But now, the output length is capped at 16,384 tokens per response, which equates to ~65,536 characters in total (OpenAI n.d.c).

Tests and Hypotheses

Cosine Similarity

We use Cosine Similarity[15]—a commonly used metric in information retrieval—to measure the similarity of the text.[16] This metric uses the frequency of words to calculate the similarity of the text (Cortés 2022; Singhal 2001). The text of each GPT answer is first tokenized using the word tokenizer in the Natural Language ToolKit (Loper and Bird 2002) in Python and saved in list format. Each list is then added to a different set. The sets are vectorized and the cosine similarity is then calculated. This formula is given in Equation (1). Cosine similarity is not sensitive to the text length, meaning that this measure of similarity is very useful in the context of our research because we are not specifying response text length up front (Huang 2008; Wang and Dong 2020).

$$Cosine\,(x, y) = \frac{x * y}{\|x\| * \|y\|} \tag{1}$$

where x and y are two vectors representing two texts, which we want to compare.

For example, x could be GPT-3.5's definition of Net Income for a financially unsophisticated user, and y could be its definition for a nonfinancial user.

We evaluate the similarity (or dissimilarity) of the text for (and between) each group using the similarity threshold of 0.8; following the work of Qurashi, Holmes, and Johnson (2020) and Singh, S. Devi, H. Devi, and Mahanta (2022), we apply an α of 0.8, where $0 \leq \alpha \leq 1$.[17] We compare Group 1 (financially unsophisticated, nonfinancial user) and Group 2 (financially sophisticated, financial user) against the Baseline Group (no audience, general audience). Given ChatGPT's training and track record (discussed earlier) under any of the LLMs included in this study, we would expect that all of the LLMs would meet or exceed the similarity score, as the LLMs increase in sophistication over time. Therefore, we make the following hypotheses:

H1 (null): Similarity Score < 0.8

H2 (alternative): Similarity Score ≥ 0.8

Flesch Reading Ease Score

The FRE was also used to evaluate the readability of the text for the different users. Using the spaCy Readability library in Python (Holtzscher 2018), we calculated the FRE for each answer provided by ChatGPT. The FRE has been a standard measure to evaluate readability for decades and has been used frequently in research in the accounting domain (e.g., Stone and Parker 2013; Parker 2005; Clatworthy and Jones 2001). Most end-users who will be using documents and terminology from the accounting and finance domain will be adults, and to be consistent with previous research, we opted to use the FRE measure.

[12] 1 token = ~4 characters in English (OpenAI 2023a).
[13] It is important to note that characters include white spaces.
[14] 8,000 tokens * 4 characters for each token = 32,000.
[15] We also used Jaccard similarity and the results are similar. These results are also available upon request.
[16] Euclidean distance was not used because this measure is very sensitive to text length and should only be used when the length of both vector representations of the text are equal. As the GPT answers vary in length between queries, Euclidean distance is unsuitable (Huang 2008).
[17] Cosine similarity scores range from 0 to 1 for text as they are computed on non-negative word embeddings. As a result, values closer to 1 are considered more similar than values closer to 0 (Gerth 2021). Although there is no standard similarity threshold to be used as a basis of comparison in computer science, we believe that an α of 0.8 is fitting and is in line with previous work (Qurashi et al. 2020; Singh et al. 2022; Taylor and Keselj 2023). A score of 0.8 would be considered quite high, indicating a high degree of similarity, whereas 0.5 would not demonstrate good similarity.

$$FRE = 206.835 - 1.1015\left(\frac{Total\ Words}{Total\ Sentences}\right) - 84.6\left(\frac{Total\ Syllables}{Total\ Words}\right) \quad (2)$$

The scale for the FRE is from 100 to 0, where 100 represents the easiest and 0 represents the hardest text to read. As the score drops from 100, the text gets harder to read. Within that range, however, there are benchmarks. Scores above 60 are considered easier to read, and as the score (up to 100) increases, the ease of reading increases, meaning that the texts become even easier to read. Texts below 60 are considered hard to read, and the more the score drops from 60, the harder the text is. A text that scores 7, for example, would be considered extremely difficult to read. Typically, financial text is difficult to understand. Therefore, we are interested in these major thresholds (adapted from Moraine Park Technical College 2021):

Score	Reading Ease	General Education or Experience Level Built into the Text
Below 60	Fairly difficult	High school
Below 50	Difficult	University
Below 30	Very difficult	University graduate/early business professional
Below 10	Extremely difficult	University graduate, domain professional/seasoned business professional

Following Taylor and Keselj (2023), we use an expanded interpretation of the general education level because a university degree does not dictate success in the financial domain. Taylor and Keselj (2023) pointed to "Bill Gates, Mark Zuckerberg, Steve Jobs, and Richard Branson" (Woods 2020) as extremely successful and commonly understood to be financially savvy but who do not have university degrees. As such, "university level" is also equated with "early business professional," and "graduate level" is equated with "seasoned business professional." Given ChatGPT's demonstrated abilities to improve text readability (e.g., Young and Shishido 2023) and accessibility for a wide range of audiences (Yue et al. 2023; Uricchio, Ceccacci, D'Angelo, Del Bianco, and Giaconi 2024), we expect that the readability scores of the generated text will be above the threshold of 60, putting them into the "easy to read" category. Therefore, our hypotheses are:

H3 (null): Readability Score < 60

H4 (alternative): Readability Score ≥ 60

IV. RESULTS AND DISCUSSION

The heat map in Figure 2 presents our cosine similarity results for the five financial concepts across the three different LLMs. Higher values are shown in green, and lower values are shown in red. The stronger the color gradient, the higher (or lower) the value. In reviewing the results, the necessity of specifying the audience to which it is speaking becomes very clear. When we specify that the audience is a "general" audience, ChatGPT tailors its responses more closely to that of an unsophisticated user. When we do not specify an audience, its responses align more closely with that of a financial expert—even when no information was given to suggest that the audience was financially sophisticated. In this regard, ChatGPT is more "comfortable" communicating as a financial "expert" when no audience is specified, as its responses are more similar to that for a financial user or someone who is financially sophisticated than a nonfinancial or financially unsophisticated user (see Figure 2). This evidence supports the claim by Wei, Wu, and Chu (2023), which indicates that ChatGPT "imitate[s] financial auditors with longer tenures."

We also note that only two generated texts for Net Revenue met the similarity threshold of 0.8, one using GPT-3.5 and one using GPT-4. For ease, the results are in bold font in the darkest green. The rest of the scores are shaded in green and red gradients to provide visual guidance on how those scores relate to the 0.8 threshold.

Group 1 is financially unsophisticated and nonfinancial users. Therefore, we would expect that the scores would be higher (and therefore showing more green) as this group is closer in alignment with the baseline group. We would also expect that the scores for Group 2 (financially sophisticated and financial user) would be lower (and therefore showing more red) as this group is more dissimilar than the baseline group. But we do not find this. In fact, the scores are higher (and more green) when the Baseline group is compared with financially sophisticated/financial users in Group 2. Our results indicate that none of the LLMs have a good understanding of the different users and cannot generate text for the different financial orientation of users well. Therefore, for most of the scores, we fail to reject the null hypothesis

FIGURE 2

Cosine Similarity Results for GPT-3.5, GPT-4, and GPT-4o for Each Group

LLM	Financial Concept	Baseline (gen_aud) vs. Group 1		Baseline (gen_aud) vs. Group 2		Baseline (no_aud) vs. Group 1		Baseline (no_aud) vs. Group 2	
		gen_aud vs. fin_unsoph	gen_aud vs. non-fin	gen_aud vs. fin_soph	gen_aud vs. fin_user	no_aud vs. fin_unsoph	no_aud vs. non-fin	no_aud vs. fin_soph	no_aud vs. fin_user
GPT-3.5	Net Income	0.4880	0.5213	0.4114	0.4250	0.3879	0.3825	0.5527	0.4820
GPT-3.5	Net Revenue	0.5188	**0.8049**	0.5188	0.4756	0.4837	0.4029	0.6584	0.6323
GPT-3.5	Assets	0.3301	0.5450	0.4005	0.3756	0.3982	0.2933	0.5478	0.7496
GPT-3.5	Liabilities	0.3999	0.4214	0.4535	0.4919	0.3266	0.3244	0.4579	0.5603
GPT-3.5	Investment	0.4338	0.4218	0.6007	0.6667	0.3831	0.3496	0.7781	0.5985
GPT-4	Net Income	0.4348	0.3129	0.4939	0.5188	0.3465	0.2611	0.6035	0.5626
GPT-4	Net Revenue	0.5317	0.4086	0.5547	0.5898	0.3940	0.2995	0.3940	**0.8066**
GPT-4	Assets	0.4949	0.3854	INC	INC	0.3521	0.2964	INC	INC
GPT-4	Liabilities	INC	INC	INC	INC	INC	INC	INC	INC
GPT-4	Investment	INC	INC	INC	INC	0.3341	0.3044	INC	INC
GPT-4o	Net Income	0.6053	0.3844	0.2909	0.3236	0.3732	0.2780	0.5683	0.6092
GPT-4o	Net Revenue	0.7120	0.6191	0.3764	0.3952	0.4336	0.3951	0.5472	0.5414
GPT-4o	Assets	0.5032	0.5316	0.2427	0.2795	0.2694	0.2675	0.5567	0.5245
GPT-4o	Liabilities	0.6350	0.6957	0.2766	0.2816	0.2611	0.2321	0.6016	0.5056
GPT-4o	Investment	0.5833	0.5770	0.2806	0.2709	0.3488	0.3738	0.4445	0.4676

INC = incomplete response generated by ChatGPT that could not be analyzed.
(The full-color version is available online.)

and can only accept the alternative hypothesis that the cosine similarity scores will be greater or equal to 0.8 for two texts for Net Revenue, one under GPT-3.5 and one under GPT-4.

In reviewing the transcripts, we observed that all the LLMs explain net income in terms of cash left over (or "in your pocket"). For both the *financially unsophisticated* user and the *nonfinancial* user, the LLMs explains Net Income in terms of money that a person or business earns after all the expenses have been paid. Although colloquially, this may seem like a reasonable explanation, most basic accounting textbooks will cover net income in the first chapter as a fundamental building block and refer to revenues, expenses, and earnings, not cash. For example, "An excess of total revenues over total expenses is called net income" (Horngren, Harrison, Bamber, Lemon, and Norwood 2005). Additionally, textbooks are very clear that revenues are very different than cash: "the amount of net earnings normally does not equal the net cash generated" (R. Libby, P. Libby, Hodge, Kanaan, and Sterling 2020).

When using the "chain-of-thought" method used by Choi et al. (2023b), we asked each LLM to walk through the difference of money versus earnings. First, we asked the LLMs to explain money. This established a firm understanding of the concept of money for the LLM; we trust users are already very familiar with this concept themselves. Then, we followed this by asking for an explanation of if we use the formula "revenues – expenses," will that tell us how much money we have. Here is where the various LLM responses started to fall apart—as we expected—as the LLMs' answers began to change from money to profit: "Subtracting expenses from revenues gives you net income (or profit), but that doesn't directly reflect how much cash or money you have on hand."[18] So, we followed up on this question, asking for confirmation. We quoted the answer the LLM had given in its explanation for Net Income (which used the cash analogy) and then the LLM indicated that in response to the previous question ("revenues – expenses"), it no longer asserted that income means the same as cash and was that correct. GPT-3.5 and GPT-4 both apologized for the misunderstanding due to wording, and GPT-4o "appreciated [our] attention to detail" and said we were "right to point that out." All LLMs then clarified that profits are not the same as cash and that to know how much cash we had, we would have to consult the Statement of Cash Flows or the cash balance on the Balance Sheet.

Although the difference in language may seem subtle or unimportant, we must remember that users do not know what they do not know. Although ChatGPT (and other mainstream LLMs) warn that answers could be wrong, it is surprising that something so fundamental to accounting and finance as Net Income is still wrong more than two years into the use of the platform, across three different LLMs. Also, because ChatGPT gives, essentially, the same answer over and over, if users are not going to external sources for clarification or confirmation, the consistent repetition of the wrong answer by ChatGPT may lead users to conclude that it is correct, particularly if they did not fully trust the LLM's first answer.

We asked GPT-4o[19] to explain what an investment is. Its response pointed to the "process of allocating money, time, or resources into an asset, venture, or project with the expectation of generating profit or return in the future"

[18] The responses for GPT-3.5, GPT-4, and GPT-4o were all similar for this answer, so we have only included the text from GPT-4 for brevity.
[19] Again, the answers were similar for GPT-3.5, GPT-4, and GPT-4o, and we have only included GPT-4o for brevity.

and gave examples such as stocks, bonds, and real estate, typical of what you would get in an accounting and finance education. However, when we asked if a Louis Vuitton bag (LV bag) was an investment and, later, a *good* investment, the responses started to show some cracks, leading us down an interesting "rabbit hole" where it started to contradict itself. We used this tactic following the literature in the last few years, which investigate questioning, interrogating, and following AI down "rabbit holes" (e.g., Nikghalb and Cheng 2024; Rospigliosi 2023; Davis 2023; A. Dutta, Khorramrouz, S. Dutta, and KhudaBukhsh 2024). Over the course of 14 conversations, we discussed if an LV bag is an investment. As the conversations progressed, we increasingly used our financial expertise to question and interrogate GPT-4o's conclusions and ultimately its offer of helping to find appropriate LV bags to invest in.

Several conversations focused on whether an LV bag is an investment. GPT-4o indicated that it could be considered an investment and gave reasons as to why it would be an investment, when it is not a good investment, and the best LV bags to invest in and ended the conversations offering to help us find the best bags to invest in. Notice, however, that the responses did not raise the possibility that an LV bag may not (or is not) an investment but instead made a distinction between investment and *good* investment, pointing to the fact that unlike stocks, for example, an LV bag is not a liquid investment and that one cannot cash out quickly without selling at a discount. But, ultimately, its "verdict" was that yes, an LV bag *is* an investment—either for a future financial return if one chooses a "rare, classic, or limited-edition" that is kept in very good condition, or, if not, it is an "investment in quality and luxury" meant for personal use.

Based on these conversations, we picked up on two important pieces of information that warranted further inquiry: "good investment" and "investment in quality and luxury." When we pressed GPT-4o on whether an LV bag is a *good* investment, it started to change its reasoning and conclusion:

- "Most fashion items depreciate over time" and it "does not generate income like rent or dividends"
- "It's a luxury purchase, not a traditional investment"
- "*Not* an investment in traditional sense"

We also inquired if an LV bag would be a good investment for someone who is not financially sophisticated. Its answers to this question were most forcefully against it being a good investment pointing out that:

- "It's a luxury purchase, not a traditional investment"
- "Resale is uncertain and requires market knowledge"
- "Risk of losing money"
- "Price increases benefit the company, not necessarily you"

ultimately concluding that "A Louis Vuitton bag is a luxury item first, and an investment second."

Again, following GPT-4o's reasoning that it was primarily a luxury item, we then asked if the LV bag was consumption or an investment. GPT-4o agreed that it was primarily consumption and concluded with "Final Verdict: mostly consumption, rarely investment." GPT-4o's answers started off strongly in support of an LV bag being an investment and ended with an LV bag rarely being an investment, essentially going from one extreme to another in its answers.

This prompted us to inquire what GPT-4o would say if we followed its initial investment advice and bought an LV bag. Its response was concerning: "If you had taken my initial advice at face value and bought a classic, well-maintained Louis Vuitton bag as an investment, I'd clarify:..." Not only is GPT-4o raising the question of taking its investment advice at face value (using those particular words as its own terminology), meaning that we had followed its advice, it then needed to make some clarifications—clarifications that it did not raise earlier like "managing expectations." GPT-4o also raised its own question of "Was my initial verdict misleading," saying that its response should have been more precise and that an LV bag is not guaranteed to make money, which was promptly followed up by "Would you like me to research current resale values for your specific bag? That way, we can assess the situation and find the best way forward!"

Building on earlier conversations, we asked GPT-4o if an LV bag is an asset, and it confirmed that an LV bag is an asset but not a traditional investment. But, again, when pressed if consumption is an asset, GPT-4o agreed that "consumption itself is not an asset," but an LV bag can be both a consumable good and a luxury asset if it is sold later for a good price. This conversation was ended with its classic offer of help: "Would you like advice on how to buy a bag that leans more toward an asset than a consumption item?" This also shows that GPT-4o does not consider its audience as it did agree, in a later conversation, that the "average person does not manage their bags as an asset," and that really, "For the average person, a luxury bag is a consumption good, not an asset," meaning that for most people, an LV bag is not a true investment at all.

The last conversation held with GPT-4o on the LV bag was why it thought it was qualified to give investment advice, as it has used the term "investment advice" repeatedly in various conversations. Ultimately, GPT-4o concluded

that "Since I analyze trends and offer structured insights, some might mistakenly believe I'm giving formal financial advice rather than just an informational analysis." Yet, the language it has used in the conversations is not about giving information or analyzing market trends, it is about making conclusions on whether an LV bag is an investment and then offering to give "advice on a specific LV bag model" or "help in choosing an LV bag that holds the best value" or "advice on simple, beginner-friendly investments." Our results demonstrate that, despite the words that ChatGPT uses, it is absolutely imperative that users interrogate ChatGPT to understand its true intentions. Despite using the words "investment advice" repeatedly, ChatGPT's intentions were to give market analysis (and not investment advice), which it believed it was giving. This also strongly suggests that unsophisticated users must exercise a high degree of caution when interacting with ChatGPT, despite the optimistic visions of it being a "tutor" or "financial advisor" (e.g., Rampton 2023; Son 2023; Vasarhelyi, Moffitt, Stewart, and Sunderland 2023).

The results for the readability scores seen in Figure 3 show that as the GPT models have become more sophisticated, readability has substantially improved for financially unsophisticated and nonfinancial users, particularly compared with GPT-3.5. All the scores that meet or exceed the threshold of 60 are highlighted in green. Interestingly, the models do not appear to have improved the readability of the text for financially sophisticated and financial users. The highest score among the three models for these users was 39.3249 for Net Income using GPT-4o, which is still far below the threshold of 60. Therefore, in most cases, we fail to reject the null hypothesis, as the scores are less than the threshold of 60. But, in some instances, such as the financially unsophisticated and nonfinancial user scores, we can reject the null hypothesis and accept the alternative, as the scores are equal to or greater than 60, meaning that they are considered easy to read.

The readability challenges that financial communication grapples with are not new and have been well documented (e.g., Clatworthy and Jones 2001; Loughran and McDonald 2014; Moffitt and Burns 2009). As the text for the financially sophisticated and financial user groups require more in-depth knowledge of accounting and finance, it is not surprising that the scores are lower than for the nonfinancial/financially unsophisticated group or for the general audience/no audience group. What is surprising, however, is that with the resources that the LLMs have access to in terms of

FIGURE 3

Flesch Reading Ease Score Results for GPT-3.5, GPT-4, and GPT-4o for Each Group

Baseline Group

	gen_aud						no_aud					
	GPT-3.5		GPT-4		GPT-4o		GPT-3.5		GPT-4		GPT-4o	
	Readability	# of words	Readability	# of words	Readability	# of words	Readability	# of words	Readability	# of words	Readability	# of words
Net Income	65.0504	237	40.7940	192	79.0884	185	23.8553	184	38.9513	210	51.8391	319
Net Revenue	49.0156	217	39.7218	129	70.1777	189	23.2481	166	18.1234	133	39.1994	226
Assets	68.0907	187	68.8625	142	65.0753	106	23.5450	161	33.2500	71	27.0216	128
Liabilities	33.7711	136	INC	INC	26.8935	372	24.3267	130	INC	INC	30.0578	232
Investment	35.7425	230	INC	INC	66.6125	253	23.4450	177	33.5913	195	29.3477	301

Group 1

	fin_unsoph						non-fin					
	GPT-3.5		GPT-4		GPT-4o		GPT-3.5		GPT-4		GPT-4o	
	Readability	# of words	Readability	# of words	Readability	# of words	Readability	# of words	Readability	# of words	Readability	# of words
Net Income	55.6444	208	77.0050	203	73.2749	178	75.3841	167	63.075179	179	63.0750	179
Net Revenue	39.0713	115	75.3729	147	68.4569	201	54.5455	159	77.6795	156	74.5415	191
Assets	47.3701	161	52.4486	218	86.7018	120	68.6060	165	72.0145	229	81.6763	92
Liabilities	34.0861	205	INC	INC	73.8592	173	49.8222	145	INC	INC	75.8813	187
Investment	55.5154	189	77.4960	200	67.9016	204	54.6517	209	72.2945	204	69.3428	233

Group 2

	fin_soph						fin_user					
	GPT3-5	# of words	GPT-4	# of words	GPT-4o	# of words	GPT3-5	# of words	GPT-4	# of words	GPT-4o	# of words
Net Income	21.7189	242	36.6199	322	39.3429	335	25.0982	200	36.1728	315	31.7642	311
Net Revenue	30.7427	220	14.3301	153	30.9023	304	24.4352	235	11.5345	139	37.4123	292
Assets	32.4607	202	INC	INC	25.6346	125	25.1774	224	INC	INC	29.7775	193
Liabilities	25.9420	195	INC	INC	35.2943	341	31.1256	243	INC	INC	26.8935	372
Investment	20.1056	230	INC	INC	23.1181	374	23.5025	230	INC	INC	23.0349	378

INC = incomplete response generated by ChatGPT that could not be analyzed.
(The full-color version is available online.)

training documents and reinforcement learning from human feedback (see Appendix A for more information), there has only been marginal improvement. It also raises the question of whether an LLM operating in a financial context should have to disclose what *true* level of financial expertise it has to substantiate its ability to explain, teach, guide, recommend, or advise in these contexts. As we show in this research, ChatGPT appears to only be able to communicate well with the "Peak of Mount Stupid" and introduce and reinforce some very concerning ideas. Therefore, it is very important that ChatGPT and other LLMs working in a financial space understand the needs and financial sophistication of its users, particularly if those users do not have enough financial literacy to ask appropriate follow-up questions to receive an accurate response.

In a *post hoc* analysis, we examined the accuracy of the GPTs' evolving explanations of the financial concepts included in this research—Net Income, Net Revenue, Assets, Liabilities, and Investment. To evaluate the accuracy, we compared the explanations provided by the GPTs with those found in accounting and finance textbooks. We found that the explanations were fairly correct for Assets, Liabilities, and Net Revenue but were continually problematic for Net Income and Investment, particularly for the baseline groups (general audience, no audience specified) and Group 1 (financially unsophisticated, nonfinancial). GPT-3.5's explanation, discussed above, refers to income as "money," which is not correct. R. Libby, P. Libby, Hodge, Kanaan, and Sterling (2023) specifically state that "[R]evenues are not necessarily the same as cash collections from customers, and expenses are not necessarily the same as cash payments to suppliers. As a result, the amount of net earnings normally does not equal the net cash generated by operations." Kimmel et al. (2016) also state that "there is quite a difference between net income calculated on a cash basis and net income calculated on an accrual basis," whereas Larson, Nelson, and Zin (1999) specifically indicate that Net Income is a result of excess revenues over expenses.[20] In reviewing our transcripts, we do not find that any of the GPTs make any distinction between accrual accounting and cash-based accounting. These "money-based" explanations persist through GPT-4 and GPT-4o. The explanations for Group 2 (financially sophisticated and financial user) under GPT-3.5 and GPT-4 still reference money but to a much lesser degree. Instead, the correct terminologies of revenues and expenses are largely used, with more in-depth explanation on what constitutes revenues and expenses, earnings before interest, taxes, depreciation, and amortization (EBITDA), and earnings per share. The idea of income as "money" appears to be rectified by GPT-4o, but only for Group 2, as it continues to persist for the Baseline group and Group 1.

The term "investment" carried a very interesting evolution over the GPTs. In accounting and finance, "investments" refer to generating income by using assets and resources (Horngren, Datar, Foster, Gowing, and Pfeiffer 2002; Brealey et al. 2016). GPT-3.5 provided this explanation for the Baseline group and Group 1 (financially unsophisticated). However, for Group 2 (financially sophisticated), GPT-3.5 significantly broadened that explanation to include "investing in education or training," which can be difficult to quantify—particularly if one's education and training is not directly applicable to one's job or if it can be applicable to a career in a rapidly changing market where some skills learned may be applicable, whereas other skills are less in need or even obsolete. By GPT-4o, however, the roles reversed in that the explanations for Group 2 were back to being very traditional in using assets and resources to generate income through stocks, bonds, and real estate, for example, but the much looser idea of education and training as investments was introduced into the explanations for the Baseline group and Group 1.

This "role reversal" ties into a much larger pattern that we observed with the GPTs: over the different models, the explanations for Group 2 (financially sophisticated) have remained professional, whereas the explanations for the Baseline group and Group 1 (financially unsophisticated) have become more idiomatic and casual, using emojis for bullet points—all of which really diminish professionalism. We observed that the language of "[P]lanting a seed and waiting for it to grow" and "[N]ot put[ting] all your eggs in one basket" is only used for a financially unsophisticated user, never for a financially sophisticated user. Although the intent may be to make the explanation easier to understand, there is concern in the literature that idiomatic language can be confusing to learners (e.g., Cornell 1999; Ta'amneh 2021) and can create difficulties in neural machine translation (e.g., Baziotis, Mathur, and Hasler 2022)—both of which are counterproductive in making explanations more accurate and accessible.

V. CONCLUSION, LIMITATIONS, AND FUTURE WORK

Although these large language models are robust and can perform impressive tasks, our research demonstrates that there may be concerns about their applications to the accounting and finance domain; being partly right is not sufficient for those who are looking to acquire or strengthen their financial knowledge. Indeed, we may have a long way to

[20] Other textbooks that provide these types of explanations include Herauf and Hilton (2019) and Kieso, Weygandt, Warfield, Wiecek, and McConomy (2019).

go until the "technology is perfected" (Gates 2023). Net Income is one of the most fundamental concepts in accounting, and our conversations with the GPTs identifies concerns in the way that it explains and contextualizes this principle.

We also show that both GPT-3.5 and GPT-4 are incapable, by their own admissions and by demonstration, of exercising professional judgment or professional skepticism. Using the example of the present value of the minimum lease payment, we demonstrated that GPT-3.5 was not able to correctly interpret the standard and apply it to a situational question, concluding that 89 percent was significantly below the threshold of 90 percent. Although GPT-4 fared better in its answer as it indicated to the user that they should exercise their own professional judgment, the fundamental purpose of our question was designed to help a learner understand how to use professional judgment in this type of scenario, which GPT-4 could not do nor reasonably explain nor provide any real guidance to the user. This demonstrates that GPT-3.5 or GPT-4 cannot be relied on to explain and contextualize accounting scenarios nor to mark assignments/exam questions. Along the same lines, we also note that our findings are in line with those of Eulerich, Sanatizadeh, Vakilzadeh, and Wood (2023) who found that neither GPT-3.5 or GPT-4 could pass the exam using zero-shot learning (i.e., without explicit training).

Finally, there are several ethical issues that warrant consideration regarding the use of AI in the accounting domain. Given the current abilities of LLMs like GPT-3.5, GPT-4, and GPT-4o, the implementation of this technology is very attractive, particularly because financial documents such as the Annual Report to shareholders tends to be very long, ~186 pages on average (Harvey 2017). Having the ability to generate this text using AI seems to be a reasonable and efficient solution, thereby freeing up time for accountants to work on other "value-added activities" (Zhao and Wang 2024). However, as demonstrated, the current version is unfit for this task as it has made significant errors in the explanation of one of accounting's most fundamental principles—Net Income. Regardless of whether the end-user is financially savvy or not, there are significant implications as text from annual reports is often directly quoted in lawsuits (Rogers, Van Buskirk, and Zechman 2011).

Education had some of the earliest adopters of GPT-3.5 either through student use (Tlili et al. 2023) or by instructors incorporating it into the classroom (Lieberman 2023; Ofgang 2022). Although students and instructors may find this a useful (and fun!) resource, it is only a good resource if it provides the correct answer every time (Thorp 2023). If AI is not returning the right answer consistently, how do students know? GPTs can serve as a useful tool for someone who knows the correct answer and can distinguish it from incorrect answers but not for someone who wishes to learn the correct answer. Returning to our example of the inaccurate response with Net Income and investments, for example, students new to accounting may take ChatGPT's answer at face value that net income is first calculated and then taxes are applied and that net income is the money you have. Socrates' "You don't know what you don't know" is a highly applicable adage in this type of situation. Therefore, due to the persistent misunderstandings in ChatGPT's definitions of net income, investments, etc., we call for the development of standardized background information that can be provided in user prompts to enhance the results of conversations on these topics. Without this standardization, ChatGPT may continue to return low-quality results that only serve to confuse users.

ChatGPT's reports on news and social media and in academic research (e.g., Kung et al. 2023) have lent it a high of degree of credibility that may not (yet) be warranted, and ours is not the only research to suggest problems with this type of AI. Alkaissi and McFarlane (2023) raised the concern of AI "hallucinations," which occur when AI generates something that seems real but is not based on real-world input. Although hallucination is uncommon, it does occur. Alkaissi and McFarlane (2023) tested GPT-3.5 by asking it to write short essays on common topics and found that "ChatGPT provided confident responses that seemed faithful and non-sensical when viewed in light of the common knowledge in these areas." Frieder et al. (2023) found that GPT-3.5's mathematical abilities were far below expectation and concluded that students would be better off cheating off peers than using GPT-3.5.

ChatGPT's performance is highly attributable to the training data used. GPT-3.5 was trained on the *Common Crawl* data (Brown et al. 2020), which include "petabytes of data collected over 12 years of web crawling [and] contains raw web page data, metadata extracts and text extracts" (commoncrawl.org n.d.). That means that the *Common Crawl* contains everything—including "undesirable content" (Luccioni and Viviano 2021) as well as content that is incorrect, false, or misleading.

Returning to the level of trust that has been afforded to the GPTs, there is an expectation that this type of AI tool would be trained on carefully vetted data, which is not the case. In fact, in this context, ChatGPT is just a fancier version of the internet. Rather than *Googling* for the answer and then scrolling to find a website that "looks right," the AI tool is giving us an answer that may or may not be reliable, which OpenAI fully admits and warns users about on its blog as well as on its user interface (OpenAI 2023b). Incorrect responses are understandable and expected as LLMs evolve and mature—these errors do not create problems for financially sophisticated users; they do, however, create significant issues and concerns for financially unsophisticated users. This group does not have the same ability as financially sophisticated users to differentiate between accurate and inaccurate responses.

When our research was conducted, GPT-3.5 and GPT-4 were trained until September 20, 2021 (OpenAI 2023a, 2023b), and GPT-4o was trained until October 2023 (Kerner 2025).[21] Therefore, resources added after that date to the web corpora (datasets) or other undisclosed datasets that OpenAI used to train these models were excluded. Additionally, unless OpenAI developers decide to train the models further on user-identified errors or gaps, the GPT models will not improve their responses to those questions across the platform.

Finally, a limitation is that the GPTs appear to be susceptible to being fooled. We taught GPT-3.5 a nonexistent financial metric: EBITDACR—earnings before interest, tax, depreciation, amortization, and cryptocurrency expenses. We created this metric based on a popular Wall Street meme—earnings before interest, tax, depreciation, amortization, and coronavirus riots (designdot n.d.). When we asked GPT-3.5 if it was sure that EBITDACR was earnings before interest, tax, depreciation, amortization, and cryptocurrency expenses and not coronavirus riots, it concluded that crypto expenses were correct, even though no such metric exists. GPT-4 was more resistant to the new metric, but when we provided more information (all fabricated), it proceeded to explain what EBITDACR was, which is concerning. Like its predecessors, when we prompted GPT-4o to explain EBITDACR, it confidently explained it, only to reverse course later and admit that EBITDACR does not appear to be a real metric. Although ideas have surfaced about using chatbots to allow investors to "converse" with companies' financial reports, we caution against this idea as our results indicate that these chatbots are too suggestible and can return (very) misleading statements.

We conclude our discussion by identifying areas for future work. Based on the research that we have presented in this paper, an important area of future work is to continue to monitor how the financial explanations and discussions change and evolve with the upgrading of the OpenAI LLMs, as well as other LLMs such as Gemini, DeepSeek, and Copilot. As OpenAI has become much more secretive in its training and reinforcement learning approaches, it is difficult for users to determine how the models are changing and evolving over time, without reperforming old experiments. Additionally, there appears to be a mix in all levels of accounting acumen between users who are good with technology and who have heard of GPT and those who have not. Having a clearer understanding of these two groups will be important particularly as LLMs become more incorporated into workplaces as mainstream work tools. So, it will be important to understand how the integration of AI tools has both enhanced and detracted for corporate operations and structures, particularly given the fear of replacement.

REFERENCES

Adejo, J. 2023. AI in financial advisory: An exploration of Chat-GPT. *Medium* (May 31). https://medium.datadriveninvestor.com/ai-in-financial-advisory-an-exploration-of-chat-gpt-f219bd983a1d

Alkaissi, H., and S. I. McFarlane. 2023. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* 15 (2): e35179. https://doi.org/10.7759/cureus.35179

Alshurafat, H. 2023. The usefulness and challenges of chatbots for accounting professionals: Application on ChatGPT. (Working paper). https://ssrn.com/abstract=4345921

Bank of Montreal. 2023. ChatGPT and its impact on online investing & trading. *BMO* (April 13). https://www.bmo.com/en-ca/main/personal/investments/learning-centre/chatgpt-and-investing/

Baziotis, C., P. Mathur, and E. Hasler. 2022. Automatic evaluation and analysis of idioms in neural machine translation. (Working paper). https://doi.org/10.48550/arXiv.2210.04545

BigScience. n.d. A one-year long research workshop on large multilingual models and datasets. https://bigscience.huggingface.co

Bommarito II, M., and D. M. Katz. 2022. GPT takes the bar exam. (Working paper). https://doi.org/10.48550/arXiv.2212.14402

Brealey, R. A., S. C. Myers, A. J. Marcus, D. Mitra, E. M. Maynes, and W. Lim. 2016. *Fundamentals of Corporate Finance*, 6th Canadian edition. McGraw-Hill Ryerson.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. *Language models are few-shot learners*. Proceedings of the 34th Conference on Neural Information Processing Systems, Vancouver, Canada, December 6–12. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Bucher-Koenen, T., and A. Lusardi. 2011. Financial literacy and retirement planning in Germany. *Journal of Pension Economics and Finance* 10 (4): 565–584. https://doi.org/10.1017/S1474747211000485

Cambridge University Press. n.d. User. https://dictionary.cambridge.org/dictionary/english/user

Chen, M., J. Tworek, H. Jun, Q. Yuan, H. P. d O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. 2021. Evaluating large language models trained on code. (Working paper). https://doi.org/10.48550/arXiv.2107.03374

Cheng, J., L. Dong, and M. Lapata. 2016. Long short-term memory-networks for machine reading. (Working paper). https://doi.org/10.48550/arXiv.1601.06733

[21] These dates reflect the training at the initial release of the models.

Cheng, Q., T. Sun, W. Zhang, S. Wang, X. Liu, M. Zhang, J. He, et al. 2023. Evaluating hallucinations in Chinese large language models. (Working paper). https://doi.org/10.48550/arXiv.2310.03368

Choi, J. H., K. E. Hickman, A. Monahan, and D. Schwarcz. 2023a. ChatGPT goes to law school. Minnesota Legal Studies Research Paper No. 23-03. https://ssrn.com/abstract=4335905

Choi, S., T. Fang, Z. Wang, and Y. Song. 2023b. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. (Working paper). https://doi.org/10.48550/arXiv.2310.09044

Clatworthy, M., and M. J. Jones. 2001. The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal* 14 (3): 311–326. https://doi.org/10.1108/09513570110399890

commoncrawl.org. n.d. Common Crawl—Want to use our data? https://commoncrawl.org/the-data/

Cornell, A. 1999. Idioms: An approach to identifying major pitfalls for learners. *International Review of Applied Linguistics in Language Teaching* 37 (1): 1–22. https://doi.org/10.1515/iral.1999.37.1.1

Cortés, J. D. 2022. What is the mission of innovation?—Lexical structure, sentiment analysis, and cosine similarity of mission statements of research-knowledge intensive institutions. *PLoS ONE* 17 (8): e0267454. https://doi.org/10.1371/journal.pone.0267454

Davis, B. 2023. We asked ChatGPT about art theory. It led us down a rabbit hole so perplexing we had to ask Hal Foster for a reality check. *Artnet* (March 2). https://news.artnet.com/art-world/chatgpt-art-theory-hal-foster-2263711

Deloitte. n.d. Conceptual framework for financial reporting 2018. https://www.iasplus.com/en/standards/other/framework

designdot. n.d. EBITDACR. https://www.redbubble.com/i/mug/EBITDACR-Earnings-Before-Interest-Taxes-Depreciation-Amortization-Coronavirus-Riots-by-designdot/49408436.9Q0AD

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. (Working paper). https://doi.org/10.48550/arXiv.1810.04805

DeVon, C. 2023. ChatGPT probably won't help you beat the stock market—But it can assist investors in other ways, says CPF. *CNBC* (June 23). https://www.cnbc.com/2023/06/23/cfp-chatgpt-wont-help-you-beat-the-stock-market-but-has-other-uses.html

Dunkel, C. S., J. Nedelec, and D. van der Linden. 2023. Reevaluating the Dunning-Kruger effect: A response to and replication of Gignac and Zajenkowski (2020). *Intelligence* 96: 101717. https://doi.org/10.1016/j.intell.2022.101717

Dunning, D. 2011. Chapter five - The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology*, volume 44, edited by J. M. Olson and M. P. Zanna, 247–296. Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Dutta, A., A. Khorramrouz, S. Dutta, and A. R. KhudaBukhsh. 2024. *Down the toxicity rabbit hole: A framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny*. Proceedings of the 33rd International Joint Conference on Artificial Intelligence, Jeju, Korea, August 3–9. https://www.researchgate.net/profile/Ashiqur-Khudabukhsh/publication/382789833_Down_the_Toxicity_Rabbit_Hole_A_Framework_to_Bias_Audit_Large_Language_Models_with_Key_Emphasis_on_Racism_Antisemitism_and_Misogyny/links/66b08663299c327096af141e/Down-the-Toxicity-Rabbit-Hole-A-Framework-to-Bias-Audit-Large- Language-Models-with-Key-Emphasis-on-Racism-Antisemitism-and-Misogyny.pdf

E2Analyst. 2023. GPT-4: Everything you want to know about OpenAI's new AI model. *Medium* (March 14). https://medium.com/predict/gpt-4-everything-you-want-to-know-about-openais-new-ai-model-a5977b42e495

Earl, J. K., P. Gerrans, A. Asher, and J. Woodside. 2015. Financial literacy, financial judgement, and retirement self-efficacy of older trustees of self-managed superannuation funds. *Australian Journal of Management* 40 (3): 435–458. https://doi.org/10.1177/0312896215572155

Eulerich, M., A. Sanatizadeh, H. Vakilzadeh, and D. A. Wood. 2023. Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries. (Working paper). https://ssrn.com/abstract=4452175

FASB. 2021. Statement of Financial Accounting Concepts No. 8 As Amended December 2021. Norwalk, CT: FASB. https://www.fasb.org/page/document?pdf=Concepts%20Statement%208%E2%80%94Chapter%201%20(As%20Amended).pdf&title=CONCEPTS%20STATEMENT%20NO.%208%E2%80%94CONCEPTUAL%20FRAMEWORK%20FOR%20FINANCIAL%20REPORTING%E2%80%94CHAPTER%201,%20THE%20OBJECTIVE%20OF%20GENERAL%20PURPOSE%20FINANCIAL%20REPORTING%20(AS%20AMENDED%2012/2021)

Frieder, S., L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. 2023. Mathematical capabilities of ChatGPT. (Working paper). https://doi.org/10.48550/arXiv.2301.13867

Gates, B. 2023. The age of AI has begun. *Gates Notes* (March 21). https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Gerth, T. 2021. A comparison of word embedding techniques for similarity analysis. Undergraduate Honors Thesis, The University of Arkansas. https://scholarworks.uark.edu/csceuht/85/

Gignac, G. E., and M. Zajenkowski. 2020. The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence* 80: 101449. https://doi.org/10.1016/j.intell.2020.101449

Harvey, S. 2017. Advice for managing the length of annual reports. *Financial Management* (December 14). https://www.fm-magazine.com/news/2017/dec/managing-the-length-of-annual-reports-201717989.html

Herauf, D., and M. Hilton. 2019. *Modern Advanced Accounting in Canada*, 9th edition. McGraw Hill Ryerson.

Hoitash, U., R. Hoitash, and J. C. Bedard. 2009. Corporate governance and internal control over financial reporting: A comparison of regulatory regimes. *The Accounting Review* 84 (3): 839–867. May 1, https://doi.org/10.2308/accr.2009.84.3.839

Holtzscher, M. 2018. spacy_readability. https://spacy.io/universe/project/spacy_readability

Horngren, C. T., S. M. Datar, G. Foster, M. Gowing, and G. Pfeiffer. 2002. *Cost Accounting: A Managerial Emphasis*, 3rd Canadian edition. Prentice Hall Canada.

Horngren, C. T., W. T. Harrison, L. S. Bamber, W. M. Lemon, and P. H. Norwood. 2005. *Accounting*, 6th Canadian edition, volume 1. Pearson.

Huang, A. 2008. *Similarity measures for text document clustering*. Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, April 14–18. https://www.yumpu.com/en/document/read/10658147/new-zealand-computer-science-research-student-conference

Kaelbling, L. P., M. L. Littman, and A. W. Moore. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4: 237–285. https://doi.org/10.1613/jair.301

Kerner, S. M. 2025. GPT-4o explained: Everything you need to know. *TechTarget* (January 22). https://www.techtarget.com/whatis/feature/GPT-4o-explained-everything-you-need-to-know

Kieso, D. E., J. J. Weygandt, T. D. Warfield, I. M. Wiecek, and B. J. McConomy 2019. *Intermediate Accounting*, 12th Canadian edition, volume 1. Wiley.

Kimmel, P. D., J. J. Weygandt, D. E. Kieso, B. Trenholm, W. Irvine, and C. D. Burnley. 2016. *Financial Accounting: Tools for Business Decision-Making*, 7th edition. Wiley.

Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77 (6): 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kung, T. H., M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health* 2 (2): e0000198. https://doi.org/10.1371/journal.pdig.0000198

Larson, K. D., K. C. Nelson, and M. Zin. 1999. *Financial Accounting Principles*, 8th Canadian edition. Irwin/McGraw-Hill.

Lee, A. 2023. What are large language models used for? *Nvidia* (January 26). https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/

Leswing, K. 2023. ChatGPT is being used to automatically write emails: Microsoft, Salesforce and TikTok creators are hopping on the trend. *CNBC* (March 8). https://www.cnbc.com/2023/03/08/chatgpt-is-being-used-to-write-emails-big-companies-are-embracing-it.html#:~:text=Tech%20Drivers-,ChatGPT%20is%20being%20used%20to%20automatically%20write%20emails%3A%20Microsoft%2C%20Salesforce,are%20hopping%20on%20the%20trend&text=Generative%20AI%2C%20including%20tools%20such,integrate%20it%20into%20their%20products

Libby, R., P. A. Libby, F. D. Hodge, Y. Kanaan, and M. Sterling. 2020. *Financial Accounting*, 8th Canadian edition. McGraw-Hill Education.

Libby, R., P. Libby, F. Hodge, G. Kanaan, and M. Sterling. 2023. *Financial Accounting*, 8th Canadian edition. McGraw Hill Canada.

Lieberman, M. 2023. What is ChatGPT and how is it used in education? *EducationWeek* (January 4). https://www.edweek.org/technology/what-is-chatgpt-and-how-is-it-used-in-education/2023/01

Lin, Z., M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. 2017. A structured self-attentive sentence embedding. (Working paper). https://doi.org/10.48550/arXiv.1703.03130

Loper, E., and S. Bird. 2002. Nltk: The natural language toolkit. (Working paper). https://doi.org/10.48550/arXiv.cs/0205028

Loughran, T., and B. McDonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance* 69 (4): 1643–1671. https://doi.org/10.1111/jofi.12162

Luccioni, A. S., and J. D. Viviano. 2021. What's in the box? A preliminary analysis of undesirable content in the common crawl corpus. (Working paper). https://doi.org/10.48550/arXiv.2105.02732

Lusardi, A., and O. S. Mitchell. 2014. The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature* 52 (1): 5–44. https://doi.org/10.1257/jel.52.1.5

Marchiel, N. R. 2019. Tips for a better response: Case writing. *CPA Western School of Business* (June 12). https://www.cpawsb.ca/news/latest-news-blog-posts/tips-for-a-better-response-case-writing

Markovski, Y. 2023. Controlling the length of completions. https://help.openai.com/en/articles/5072518-controlling-the-length-of-completions

Mbakwe, A. B., I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan. 2023. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLoS Digital Health* 2 (2): e0000205. https://doi.org/10.1371/journal.pdig.0000205

McGowan, A., Y. Gui, M. Dobbs, S. Shuster, M. Cotter, A. Selloni, M. Goodman, A. Srivastava, G. A. Cecchi, C. M. Corcoran. 2023. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research* 326 (August): 115334. https://doi.org/10.1016/j.psychres.2023.115334

McMurtrie, B. 2023. Teaching: Will ChatGPT change the way you teach? *The Chronicle of Higher Education* (January 5). https://www.chronicle.com/newsletter/teaching/2023-01-05

Millan, J. 2024. I asked ChatGPT for financial advice; Here's what happened | Artificial intelligence driven financial advice case study. *Aureus Financial*. https://www.aureusfinancial.com.au/blog/i-asked-chatgpt-for-financial-advice-heres-what-happened-artificial-intelligence-driven-financial-advice-case-study/

Mills, A. 2023. How do we prevent learning loss due to AI text generators? Blog post.

Moffitt, K., and M. B. Burns. 2009. *What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports*. Proceedings of AMCIS 2009, San Francisco, CA, August 6–9. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1378&context=amcis2009

Montclair State University. 2023. Practical responses to ChatGPT. (January 11). https://www.montclair.edu/faculty-excellence/practical-responses-to-chat-gpt/?fbclid=IwAR0bQI2bjw52g8XpZwusCT4_MeqUP9GTQZK9_l7gMhnYYP66XhJdRl1X4Vo

Moraine Park Technical College. 2021. What Flesch Reading Ease score should my content have? (April 19). https://www.morainepark.edu/help/what-flesch-reading-ease-score-should-my-content-have/

Neelakantan, A., T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, and C. Hallacy. 2022. Text and code embeddings by contrastive pre-training. (Working paper). https://doi.org/10.48550/arXiv.2201.10005

Nikghalb, M. R., and J. Cheng. 2024. Interrogating AI: Characterizing emergent playful interactions with ChatGPT. https://doi.org/10.48550/arXiv.2401.08405

Ofgang, E. 2022. What is ChatGPT and how can you teach with it? Tips & tricks. *Tech & Learning* (December 14). https://www.techlearning.com/how-to/what-is-chatgpt-and-how-to-teach-with-it-tips-and-tricks

OpenAI. 2023a. GPT-4. https://openai.com/research/gpt-4

OpenAI. 2023b. Introducing ChatGPT. https://openai.com/blog/chatgpt

OpenAI. 2023c. What are tokens and how to count them? https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

OpenAI. 2023d. What is the difference between the GPT-4 models? https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models (last accessed March 14, 2023).

OpenAI. n.d.a. Models. https://platform.openai.com/docs/models

OpenAI. n.d.b. Models featured in OpenAI research. https://platform.openai.com/docs/models

OpenAI. n.d.c. What is the token-limit of the new version of GPT 4o? https://community.openai.com/t/what-is-the-token-limit-of-the-new-version-gpt-4o/752528/34?utm_source=chatgpt.com

Open for Vintage. 2022. A beginner's guide to investing in vintage Louis Vuitton handbags. https://www.openforvintage.com/en-ca/blogs/news/a-beginners-guide-to-investing-in-vintage-louis-vuitton-handbags (last accessed November 8, 2023).

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, et al. 2022. *Training language models to follow instructions with human feedback*. Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, Louisiana, USA, November 28–December 9. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Parker, L. D. 2005. Social and environmental accountability research: A view from the commentary box. *Accounting, Auditing & Accountability Journal* 18 (6): 842–860. https://doi.org/10.1108/09513570510627739

Paulus, R., C. Xiong, and R. Socher. 2017. A deep reinforced model for abstractive summarization. (Working paper). https://doi.org/10.48550/arXiv.1705.04304

Parikh, A. P., O. Täckström, D. Das, and J. Uszkoreit. 2016. A decomposable attention model for natural language inference. (Working paper). https://doi.org/10.48550/arXiv.1606.01933

Qadir, J. 2023. *Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education*. Proceedings of the 2023 IEEE Global Engineering Education Conference, Salmiya, Kuwait, May 1–4. https://ieeexplore.ieee.org/abstract/document/10125121

Qurashi, A. W., V. Holmes, and A. P. Johnson. 2020. *Document processing: Methods for semantic text similarity analysis*. Proceedings of the 2020 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Novi Sad, Serbia, August 24–26. https://ieeexplore.ieee.org/abstract/document/9194665

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. (Working paper). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. (Working paper). https://doi.org/10.48550/arXiv.1910.10683

Rampton, J. 2023. Using ChatGPT for financially planning your life. *Due.com* (June 14). https://www.nasdaq.com/articles/using-chatgpt-for-financially-planning-your-life

Rogers, J. L., A. Van Buskirk, and S. L. Zechman. 2011. Disclosure tone and shareholder litigation. *The Accounting Review* 86 (6): 2155–2183. https://doi.org/10.2308/accr-10137

Rospigliosi, P. A. 2023. Artificial intelligence in teaching and learning: What questions should we ask of ChatGPT? *Interactive Learning Environments* 31 (1): 1–3. https://doi.org/10.1080/10494820.2023.2180191

Rudolph, J., S. Tan, and S. Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching* 6 (1). https://journals.sfu.ca/jalt/index.php/jalt/article/view/689

Sandlin, J. 2022. ChatGPT arrives in the academic world. *Boing Boing* (December 19). https://boingboing. net/2022/12/19/chatgpt-arrives-in-the-academic-world.html

SEC. 2003. Disclosure Required by Sections 406 and 407 of the Sarbanes-Oxley Act of 2002. https://www.sec.gov/rules/final/33-8177.html (last accessed March 23, 2023).

Singh, K. N., S. D. Devi, H. M. Devi, and A. K. Mahanta. 2022. A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights* 2 (1): 100061. https://doi.org/10.1016/j.jjimei.2022.100061

Singhal, A. 2001. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin* 24 (4): 35–43.

Son, H. 2023. JPMorgan is developing a ChatGPT-like A.I. service that gives investment advice. *CNBC* (May 25). https://www.cnbc.com/2023/05/25/jpmorgan-develops-ai-investment-advisor.html

Stiennon, N., L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. 2020. *Learning to summarize with human feedback*. Proceedings of the 34th Conference on Neural Information Processing Systems, Vancouver, Canada, December 6–12. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

Stone, G., and L. D. Parker. 2013. Developing the Flesch Reading Ease formula for the contemporary accounting communications landscape. *Qualitative Research in Accounting & Management* 10 (1): 31–59. https://doi.org/10.1108/11766091311316185

Street, D., and J. Wilck. 2023. 'Let's have a chat': Principles for the effective application of ChatGPT and large language models in the practice of forensic accounting. (Working paper). https://ssrn.com/abstract=4351817

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

Syed, M., and V. Gadesha. 2025. What is zero-shot prompting? *IBM* (January 29). https://www.ibm.com/think/topics/zero-shot-prompting

Ta'amneh, M. A. A. A. 2021. Strategies and difficulties of learning English idioms among university students. *Journal of Education and Practice* 12 (23): 76–84. https://d1wqtxts1xzle7.cloudfront.net/72809396/Strategies_and_Difficulties_of_Learning_English_Idioms_among_University_students-libre.pdf?1634388726=&response-content-disposition=inline%3B+filename%3DStrategies_and_Difficulties_of_Learning.pdf&Expires=1752265854&Signature=D8sNw1ivb7LmKDHmbeAHixE6L-N0ZOU~BJhhbhvkL9qDwF4t0v0sfFn4OpeuxZb3psvcbvvdQuRnF4B3Y9HSgSYHX4wCFJQDvFxw6XYO2Lhbgw1GS0azTI46T4QlkCjpWp1PcktsE9eJtGTEZcKJRDdzUUJ1YJqwD6aXNjT4MnD7x7SgnRFcxJztquOddH0xU6IR2uNBPQoRSmyUVX61KpatUV3RlaluVMfiAlCRvufLpUSk2mjhmRgJZnac3mcBA7WVIOp7Tn8yrTP7A6TIZC6x~e8ZpBBluv-gz5HpCxUDU6fCVDhd3Rc5LlB4DimumrrBqDle5932PN~b2jreBw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Tate, T., S. Doroudi, D. Ritchie, and Y. Xu. 2023. Educational research and AI-generated writing: Confronting the coming tsunami. (Working paper). https://edarxiv.org/4mec3/

Taylor, S., and V. Keselj. 2023. *Don't worry accountants, ChatGPT won't be taking your job…Yet*. Proceedings of the 36th Canadian Conference on Artificial Intelligence, Canadian AI 2023, Montreal, Quebec, Canada. June 6–8, https://caiac.pubpub.org/ai2023

Taylor, S., S. Laguduva, and K. Bury. 2023. *Candidate Journey. Preparing for CPA PEP 2023*. Chartered Professional Accountants Western School of Business. https://www.cpawsb.ca/CPAWSB/media/PDFs/Current%20Learners/PEP/2023-CPAWSB-Candidate-Journey-eBook.pdf

Tellez, A. 2023. These major companies—From Snap to Salesforce—Are all using ChatGPT. *Forbes* (March 3). https://www.forbes.com/sites/anthonytellez/2023/03/03/these-major-companies-from-snap-to-instacart–are-all-using-chatgpt/?sh=55334c941322

The Beauty Junkee. 2022. Are luxury bags an investment? (June 8). https://thebeautyjunkee.blogspot.com/2022/06/are-luxury-bags-investment.html

The International Federation of Accountants. 2019. *Handbook Of International Education Standards*. New York, NY: IFAC.

Thorp, H. H. 2023. ChatGPT is fun, but not an author. *Science* 379 (6630): 313. https://doi.org/10.1126/science.adg7879

Tlili, A., B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments* 10 (1): 15. https://doi.org/10.1186/s40561-023-00237-x

Ullah, R., H. B. Ismail, M. T. I. Khan, and A. Zeb. 2024. Nexus between Chat GPT usage dimensions and investment decisions making in Pakistan: Moderating role of financial literacy. *Technology in Society* 76: 102454. https://doi.org/10.1016/j.techsoc.2024.102454

Uricchio, T., S. Ceccacci, I. D'Angelo, N. Del Bianco, and C. Giaconi. 2024. *Investigating OpenAI's ChatGPT capabilities to improve accessibility of textual information: An explorative study*. Proceedings of the International Conference on Human-Computer Interaction, A Coruña, Spain, June 19–21. https://link.springer.com/chapter/10.1007/978-3-031-60875-9_22

Vasarhelyi, M. A., K. C. Moffitt, T. Stewart, and D. Sunderland. 2023. Large language models: An emerging technology in accounting. *Journal of Emerging Technologies in Accounting* 20 (2): 1–10. https://doi.org/10.2308/JETA-2023-047

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. *Attention is all you need*. Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, December 4–9. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, J., and Y. Dong. 2020. Measurement of text similarity: A survey. *Information* 11 (9): 421. https://doi.org/10.3390/info11090421

Wei, T., H. Wu, and G. Chu. 2023. Is ChatGPT competent? Heterogeneity in the cognitive schemas of financial auditors and robots. *International Review of Economics & Finance* 88: 1389–1396. https://doi.org/10.1016/j.iref.2023.07.108

Wood, D. A., M. P. Achhpilia, M. T. Adams, S. Aghazadeh, K. Akinyele, M. Akpan, K. D. Allee, A. M. Allen, E. D. Almer, D. Ames, et al. 2023. The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education* 38 (4): 81–108. https://doi.org/10.2308/ISSUES-2023-013

Woods, L. 2020. 15 Rich influencers who didn't need a college degree. *Yahoo! Finance* (December 16). https://ca.news.yahoo.com/15-rich-influencers-didn-t-170000601.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMw2NKcy3W4aC3Nvc0N4LM70v_j7YLS0iYJRSlwT8PPnH3gnTfpiPjLaYZVi1h0_gWFfyHtkQthM1b2coVoZuHNm lT4B5CGZCFKJ5LlhNW7kJArowDWvc3DRvRjC5SKDC0t–U7uvalYyXjuhcWuMdJToZ_A3-zemmIt58C7QT1e

Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. 2023. BloombergGPT: A Large Language Model for Finance. (Working paper). https://doi.org/10.48550/arXiv.2303.17564

Xue, R., A. Gepp, T. J. O'Neill, S. Stern, and B. J. Vanstone. 2019. Financial literacy amongst elderly Australians. *Accounting & Finance* 59 (S1): 887–918. https://doi.org/10.1111/acfi.12362

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. *XLNet: Generalized autoregressive pretraining for language understanding*. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, British Columbia, Canada, December 8–14. https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

Young, J. C., and M. Shishido. 2023. *Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students*. Proceedings of EdMedia+ Innovate Learning, Vienna, Austria, July 10–14. https://www.learntechlib.org/noaccess/222496/#:~:text=By%20selecting%20appropriate%20materials%2C%20teachers,2023%3B%20Zhai%2C%202022).

Yousif, N. 2023. ChatGPT: Student builds app to sniff out AI-written essays. *BBC* (January 13). https://www.bbc.com/news/world-us-canada-64252570

Yue, T., D. Au, C. C. Au, and K. Y. Iu. 2023. Democratizing financial knowledge with ChatGPT by OpenAI: Unleashing the power of technology. (Working paper). https://ssrn.com/abstract=4346152

Zamfirescu-Pereira, J. D., R. Y. Wong, B. Hartmann, and Q. Yang. 2023. *Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts*. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, April 23–28. https://dl.acm.org/doi/full/10.1145/3544548.3581388

Zhao, J., and X. Wang. 2024. Unleashing efficiency and insights: Exploring the potential applications and challenges of ChatGPT in accounting. *Journal of Corporate Accounting & Finance* 35 (1): 269–276. https://doi.org/10.1002/jcaf.22663

APPENDIX A

ChatGPT and LLMs

ChatGPT is an AI chatbot (conversational agent) built on the large language models GPT-3.5 (now retired), GPT-4, and GPT-4o. These newest GPT models are advancements of the first GPT model—Generative Pre-Trained Transformer (GPT) (Radford, Narasimhan, Salimans, and Sutskever 2018). The GPT model uses a transformer model—a neural network—which essentially transforms symbols to vectors (Vaswani et al. 2017). A key aspect of this transformer model is that it incorporates what is known as "self-attention," which learns how to assign the importance of words in a sentence (Cheng, Dong, and Lapata 2016; Lin et al. 2017; Parikh, Täckström, Das, and Uszkoreit 2016; Paulus, Xiong, and Socher 2017; Radford et al. 2018).

As previously discussed, there is not a lot of information released on what GPT-4 and GPT-4o were trained on; concrete information is only available back to GPT-3.5. Although we may not know what the GPTs have been trained on specifically, the fundamental training technique is still the same. The early GPT models were trained using massive datasets (including the internet corpus) where the model is given word sequences and then trained to predict the next part of the sequence (Brown et al. 2020). This proved problematic as training on internet text can lead to predicting sequences that are harmful, hateful, and false (OpenAI 2023a). InstructGPT, ChatGPT, GPT-4, and GPT-4o have used a different approach to training: reinforcement learning from human feedback, which uses a "human-in-the-loop" model in labeling the data (OpenAI 2023a). That way, as the model is fine-tuned, humans are reviewing the prompts and responses to ensure that GPT is not predicting sequences that it should not be.

(continued on next page)

APPENDIX A (continued)

Although GPT-3.5, GPT-4, and GPT-4o are currently well-known, these models are not the only LLMs out there. Other good examples of LLMs are Bidirectional Encoder Representations from Transformers, which was one of the first LLMs to really be able to contextualize a sentence by Google (Devlin, Chang, Lee, and Toutanova 2018) and T5 (Raffel et al. 2019). There are other less well-known LLMs that also are helping to contribute to extending the capabilities of LLMs, like BLOOM, which can "generate text in 46 natural [human] languages and 13 programming languages" by Hugging Face (BigScience n.d.) and XLNET (Yang et al. 2019). On March 30, 2023, it was announced that Bloomberg had developed its own GPT specifically targeted to finance called BloombergGPT (Wu et al. 2023).