A Scoping Review of ChatGPT Research in Accounting and Finance

December 2023

Michael Dong
College of Business, Missouri State University
michaeldong@missouristate.edu

Theophanis C. Stratopoulos
School of Accounting and Finance, University of Waterloo
tstratopoulos@uwaterloo.ca

Victor Xiaoqi Wang*
College of Business, California State University Long Beach
victor.wang@csulb.edu

ABSTRACT: This paper provides a review of recent publications and working papers on ChatGPT and related Large Language Models (LLMs) in accounting and finance. The aim is to understand the current state of research in these two areas and identify potential research opportunities for future inquiry. We identify three common themes from these earlier studies. The first theme focuses on applications of ChatGPT and LLMs in various fields of accounting and finance. The second theme utilizes ChatGPT and LLMs as a new research tool by leveraging their capabilities such as classification, summarization, and text generation. The third theme investigates implications of LLM adoption for accounting and finance professionals, as well as for various organizations and sectors. While these earlier studies provide valuable insights, they leave many important questions unanswered or partially addressed. We propose venues for further exploration and provide technical guidance for researchers seeking to employ ChatGPT and related LLMs as a tool for their research.

Keywords: ChatGPT, Generative AI, LLMs, audit, financial reporting, tax, AIS, asset pricing, corporate finance

* Corresponding author

I. INTRODUCTION

In the summer of 2022, The Economist (2022) described Large Language Models (LLMs) - like OpenAI's GPT-3 - as uncanny[1] due to their capability to generate human language, which up to that point was considered the pinnacle of intelligence (Wolfram 2023). In less than a year, OpenAI introduced two major updates: GPT-3.5 on November 30, 2022 and GPT-4 on March 14, 2023.[2] Unlike previous technologies primarily designed to automate routine and repetitive tasks, this new technology can potentially replace workers in highly educated, well-compensated white-collar occupations. The most advanced LLMs have exhibited characteristics of general-purpose technologies, suggesting that they could bring about significant economic, social, and policy ramifications (Eloundou et al. 2023). These developments have sparked heated discussions within various industries, including accounting and finance, and an unprecedented rate of adoption. Gartner predicts over 80% enterprise adoption by 2026, a sharp spike from 5% in 2023 (Cooney, 2023). In comparison, enterprise systems and cloud computing took approximately eight and six years, respectively, to achieve a 15% adoption rate (Stratopoulos and Wang 2022).

The reverberations of ChatGPT and other LLMs are keenly felt in academic circles. In a relatively short period (Spring 2022 to Fall 2023), 195 papers related to ChatGPT and other LLMs were uploaded to SSRN within the Accounting, Finance, or Economics networks. Given the surge in scholarly activity and the vital importance of this emerging technology, it becomes imperative to synthesize and analyze this burgeoning body of work. Our literature review serves multiple purposes. First, it seeks to capture the current state of the art in LLM-related research in accounting and finance. By offering a synthesis of current studies, it provides practitioners and researchers with valuable insights into the latest developments and applications. Second, it aims to identify gaps in the existing literature. By critically evaluating existing studies, this review pinpoints areas where further research opportunities are fruitful and abundant. Lastly, this review critically assesses the methodologies employed by researchers using LLMs as research tools and offers guidance on how to appropriately and effectively leverage these models while avoiding potential pitfalls.

Traditionally, literature reviews synthesize well-established and published bodies of knowledge. However, the unprecedented pace of technological advancement, as highlighted earlier, necessitates a shift in our approach. Rather than solely reflecting on where the research has been, we must adopt a forward-looking perspective – one that aligns with the famous adage, "Skate to where the puck is going to be, not where it has been." In this context, our review covers both published works and working papers available on SSRN. While published works remain valuable, they often lag behind the cutting-edge developments reflected in working papers. This combined approach ensures that we capture the evolving landscape of ChatGPT's role in accounting and finance, staying ahead of the curve in our analysis.

---

[1] The term uncanny valley was introduced by Robotics professor Masahiro Mori in 1970 to capture the feelings of eeriness and revulsion in humans when confronted with humanlike machines (Wikipedia 2023).
[2] For more about ChatGPT and the models behind it, see Section II – Background.

This paper complements existing survey studies focusing more on the technical aspects of applying LLMs to related fields. For example, Li et al. (2023) provide an overview of existing LLMs for various finance tasks, as well as on how to finetune pre-trained LLMs or train domain-specific LLMs from scratch. They also offer guidance on key considerations while applying LLMs in finance, such as technical suitability, cost/benefit trade-offs, risks, and limitations. On the other hand, Hadi et al. (2023) discuss fundamental concepts of generative AI, the architecture of GPT, history/evolution of LLMs, how to train LLMs, and their applications. Min et al. (2023) survey recent advancements in pre-trained LLMs, focusing on their capabilities for Natural Language Processing (NLP). Siddik et al. (2023) provide a review of the applications of ChatGPT in Fintech. Ray (2023) provides a review of the background of ChatGPT and its general applications.

To enhance the coherence of this literature review, we use a framework inspired by recent reviews on the adoption of emerging technologies such as blockchain and AI (Lee et al. 2023; Yang Li et al. 2018). At its core, our framework adopts an input-process-output model (Lee et al. 2023), where the *input* relates to motivation for adoption and focus area of application, the *process* to how the technology is used, and the *output* to implications of widespread adoption. More specifically, for the first component of our framework, we delve into the motivations driving the adoption of LLMs. Researchers are natural innovators and are not surprisingly among the earliest adopters who employ the new tool to exploit research opportunities that offer quick returns. Therefore, a systematic analysis of the foci of these studies would serve as a proxy for the areas (e.g., audit, financial reporting) where researchers identify as having the strongest motivations, reflected as initial, most accessible opportunities for applying LLMs. In the *process* phase, we survey how LLMs are employed in the context of accounting and finance. This involves an examination of the specific capabilities of LLMs that researchers leverage in their work (e.g., text generation, classification, and summarization) or how they should be used in accounting or finance practice as proposed by researchers. In the *output* phase, we organize studies into four groups based on their implied adoption maturity (the stage in the adoption cycle): conceptual papers, case studies, potential applications, and value realization. Additionally, we also examine the impact of LLMs on education and labor markets for this last phase, because such impacts result from widespread adoption of the technology.

Approaching this body of work through three interconnected perspectives enables a well-organized and unified categorization of studies. The first perspective reveals that researchers anticipate efficiency and effectiveness gains in nearly all domains of accounting and finance. These studies are highly concentrated in four primary areas: audit, financial reporting, asset pricing and investment, and corporate finance. Early evidence suggests that professionals aided by LLMs are likely to outperform their counterparts who do not use these advanced tools. This trend points towards a potential shift from conventional labor practices to workflows augmented by LLMs. While these studies underscore substantial benefits, they also advise caution regarding the risks associated with adopting these emerging technologies.

A similar message emerges from the second, process-oriented perspective. LLMs often outperform traditional methods in tasks, such as classification, sentiment analysis, and

summarization. The greater efficiency suggests that researchers and professionals using LLMs would be more productive than their counterparts relying on older methods. This is consistent with evidence from educational studies, as well as studies examining the implication of LLMs for the accounting and finance profession. These studies have shown that ChatGPT-4 can pass various professional exams (e.g., CPA, CMA), perform tasks at a level comparable to a human auditor, augment the abilities of financial analysts, and offer effective financial advice. Finally, the output-oriented perspective of our review indicates a shift in focus from conceptual to potential applications of LLMs. This trend not only demonstrates growing confidence in LLM capabilities and a transition to mainstream adoption, but also hints at a potential transformation in task execution across various domains.

In summary, the expanding research on ChatGPT and related LLMs within accounting and finance mirrors the growing enterprise adoption of these technologies. This burgeoning area of inquiry is abundant with unexplored questions, offering a fertile ground for scholarly investigation. Late in the paper, we propose numerous research avenues, which we believe, could yield significant contributions to the theoretical understanding of technology adoption for LLMs as well as their practical applications and implications in various fields of accounting and finance.

The rest of the paper is organized as follows. In Section II, we provide background information for LLMs and ChatGPT. In Section III, we explain the scope of our review and the methodology used. In Section IV, we provide a descriptive analysis of the papers and present a synthesis of them from the lens of input, process, and output. In Section V, we discuss these papers regarding their wider implications across various streams of literature and propose venues for future research. Section VI concludes with closing remarks. The Appendix offers technical guidance on leveraging ChatGPT and LLMs as research tools.

II. BACKGROUND

2.1 LLMs

An LLM is a type of machine learning model trained to understand, generate, and interact with human language.[3] The label "large" comes from the fact that such models have an enormous number of parameters, often on the order of billions or trillions. For example, GPT-3 from OpenAI has 175 billion parameters, whereas its more advanced successor, GPT-4, is estimated to have 1.76 trillion parameters.[4] A parameter can be understood as a coefficient that is learned and tuned during the training process in order to minimize the error in predicting the next token for a given sequence of tokens.[5] To capture the subtleties, intricacies, and grammatical structures of

[3] For comprehensive review of the technical aspects of LLMs, please see Zhao et al. (2023).
[4] https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/
[5] Technically, the parameters of an LLM consist of weights, biases, and word embeddings. Analogous to the coefficient of a linear function, weights determine the strength of connections between neurons in different layers of

human speech and writing, LLMs are trained on an enormous amount of text. For example, GPT-3 was trained on a corpus of approximately 500 billion tokens (Brown et al. 2020).

Even though the history of LLMs can be traced back to early developments in neural networks, a real breakthrough was achieved through the introduction of the transformer architecture (Vaswani et al. 2017). The main innovation of the transformer architecture lies in its self-attention mechanism, which enables the model to weigh the importance of each token in a sequence by considering its interactions with all other tokens in the sequence. This allows the model to understand the context and relationships between tokens in a sequence. The self-attention mechanism is analogous to the way we, as human readers, deduce the meaning of an unfamiliar word by looking at its surrounding words to provide context.

Common LLMs based on the transformer architecture include BERT and GPT models. Even though both leverage the transformer architecture, they are designed with different purposes in mind and function in different ways. BERT, short for Bidirectional Encoder Representations from Transformers, is designed to learn contextual representations of input sequences by considering both the left and right context. On the other hand, GPT, standing for Generative Pre-training Transformer (GPT), is designed as an autoregressive model for language generation, and functions as a sequential language creation model that predicts one word at a time. Its training process follows a one-way (uni-directional) method, where every new word is only influenced by the words that came before it in the text passage. This process, often called causal language modeling, mimics the way humans naturally write or speak in a forward-moving flow.

Thanks to its bidirectional approach, BERT excels at tasks that require a deep understanding of context such as Named Entity Recognition (NER) and Question Answering (QA). However, BERT often requires fine-tuning to improve its performance on domain-specific text. This is because its pre-training, while providing the model with a broad understanding of general language patterns, might not capture all the vocabulary, nuances, and unique characteristics of specialized domains such as finance, legal, medical, or scientific texts.

Designed to generate coherent and contextually relevant text, GPT excels at tasks such as text completion, creative writing, and even code generation. Certain GPT models (e.g., GPT-3 and its successors) have demonstrated impressive capabilities to perform tasks they are not specifically trained for, with zero-shot and few-shot learning. In simple terms, zero-shot learning enables the model to perform a task it is not trained for without seeing any example of how the task should be done. Few-shot learning, on the other hand, allows the model to learn a new task from a few examples. LLMs continually undergo remarkable enhancements. Major LLMs released in 2023 are summarized in the following table. [6]

---

the neural network. Biases allow activation functions to be shifted to the left or right and are similar to the constant of a linear function. Unlike those learned by traditional techniques, such as word2vec or GloVe, word embeddings learned by an LLM during the training process are context-dependent, wherein the final representation of each token is informed by the entire input sequence. This allows the model to capture complexities and subtleties of natural language.

[6] Compiled from multiple sources.

Table 1 Major LLMs from Other Developers Released in 2023

Model Name	Developer	Additional Information	Release Date
LlaMA 1	Meta	Open source, available in various sizes (7, 13, 33, and 65 billion parameters); API available	February 2023
Claude V1	Anthropic	Conversational AI assistant, large context window, estimated to have 175 billion parameters; API also available	March 2023
PaLM 2	Google	340 billion parameters; API available	May 2023
LLaMA2	Meta	Open source, various sizes (7, 13, 70 billion parameters; API available	July 2023
Falcon	TII, UAE	Open source, with 180 billion parameters	September 2023
Mistral 7B	Mistral	Open source, 7.3 billion parameters; outperforms Llama 2 (13B) and Llama 1 (34B) on many benchmarks[7]	September 2023
Grok	xAI	Conversational AI chatbot, 33 billion parameters; API also available	November 2023
Gemini	Google	Available in three sizes (Nano, Pro, and Ultra)	December 2023

The performance of an LLM can be measured using the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al. 2021). This benchmark evaluates the knowledge and problem-solving skills of an LLM across 57 different tasks in subjects such as math, history, and law. At the time of writing, Google Gemini Ultra (CoT) claims to outperform GPT-4 (5-shot) with a score of 90.04, which is even higher than the 89.8 score achieved by expert humans.[8] Since ChatGPT is powered by GPT models, we elaborate on these models developed and released by OpenAI in the next section.

2.2 GPT Models from OpenAI

OpenAI developed a series of LLMs based on its GPT architecture. Table 2 summarizes these models, including their release dates, number of parameters, and context windows. OpenAI introduced GPT-1, its first transformer-based LLM, in June 2018.[9] With 117 million parameters, this model was trained on a large corpus of publicly available text from the Internet and could

[7] https://mistral.ai/news/announcing-mistral-7b/
[8] https://www.newscientist.com/article/2406746-google-says-its-gemini-ai-outperforms-both-gpt-4-and-expert-humans/
[9] Compiled from various sources.

perform various tasks, such as textual alignment, sentiment analysis, and semantic similarity analysis. As an improved iteration of GPT-1, GPT-2 could generate coherent sequences of text and human-like responses to prompts. However, GPT-2 did not perform well at tasks that required more complex reasoning and/or understanding of the context. These limitations led to the development of GPT-3, which demonstrates the ability to perform a wide array of language tasks with little to no task-specific training (known as few-shot and zero-shot learning). It can generate text that is contextually rich and often indistinguishable from human-generated writing. Despite these achievements, GPT-3 still exhibits limitations such as generating factually incorrect information (a phenomenon known as hallucination) and lacking a true understanding of the text it processes.

Table 2 GPT Models from OpenAI

Model Series	**Launch Date**	**Training Data**	**Number of Parameters**	**Context Window**
GPT-1	June 8, 2018	Unknown	117 million	N/A
GPT-2	February 14, 2019	Unknown	1.5 billion	1,024
GPT-3	June 11, 2020	Up to Oct 2019	175 billion	2,048
GPT-3.5	November 30, 2022	Up to Sep 2021	175 billion	4,096
GPT-4	March 14, 2023	Up to Sep 2021	Estimated to be around 1.76 trillion	8,192
GPT-4 Turbo	November 6, 2023	Up to April 2023	Estimated to be around 1.76 trillion	128K

The GPT-3.5 series became widely known when ChatGPT was released in November 2022. This news series builds upon its predecessor by offering enhanced coherence for longer passages, improved contextual understanding for nuanced prompts, refined tone and style adoption, reduced hallucinations and misinformation, and improved instruction following. On November 28, 2022, OpenAI unveiled an enhanced iteration of its GPT model, dubbed "text-davinci-003," building upon the previous "text-davinci-002" model. As of November 30, 2022, both models were categorized by OpenAI under the "GPT-3.5" series. Concurrently, on that day, OpenAI launched ChatGPT, an application driven by a model that was also finetuned for instruction following from "text-davinci-002," making that model another member of the GPT-3.5 series.

On March 14, 2023, OpenAI launched GPT-4, which has a much larger context window of 8,192 tokens and is believed to have more than 1.7 trillion parameters. GPT-4 can take text, speech, and image data as input. In addition, GPT-4 can extract text and other data from web pages when a URL is provided in the prompt. It can also search on the Internet if instructed to do so, and this capability allows it to provide more current information beyond its knowledge cut-off date.

In November 2023, OpenAI introduced GPT-4 Turbo, which has a 128K context window, enabling it to take more than 300 pages of text as a single input. Paying developers can

access this model by passing "gpt-4-1106-preview" through the API. As OpenAI continuously upgrades its development of LLMs, newer and more capable models are widely expected to be released in the not-too-distant future. In fact, OpenAI filed a trademark application for GPT-5 on July 18, 2023. In November 2023, the CEO of OpenAI revealed that OpenAI was working on GPT-5, even though no detail was provided about its new capabilities or the timeline for its release.[10]

Each series represents a family of models. These models have different capabilities, and their capabilities have improved over time.[11] In a nutshell, there are three crucial steps. The first step involves self-supervised pre-training on a large dataset. The second step involves instruction finetuning, a process that enhances the model's ability to accurately interpret and follow human instructions. The third step involves Reinforcement Learning from Human Feedback (RLHF), a process where feedback provided by human evaluators guides the model in understanding the relative quality of various responses.

In addition to GPT models, OpenAI provides several models for other purposes. Researchers have also started to use some of them. Table 3 provides an overview of these additional models.[12] For example, "text-embedding-ada-002" can be used to generate text embeddings, which are necessary for many downstream tasks, such as classification and information retrieval. For another example, the "whisper-1" model can be used to process audio data.

It is worth mentioning that other developers may also use "GPT" to describe their models. Two notable examples are BloombergGPT (Wu et al. 2023) and FinGPT (H. Yang, Liu, and Wang 2023). BloombergGPT is a proprietary LLM from Bloomberg with 50 billion parameters trained on a diverse finance dataset. FinGPT is an open-source LLM framework, developed to democratize access to domain-specific models finetuned on finance data.

Table 3 Other Models Developed by OpenAI

Model	Description	How to Access
DALL·E	A model that can create or modify images in response to text prompts	Labs interface (https://labs.openai.com/) or via API
Whisper	A general-purpose speech recognition model that is capable of multilingual speech recognition and translation	Via API with the "whisper-1" model name or open-source version at https://github.com/openai/whisper

[10] https://arstechnica.com/ai/2023/11/openai-ceo-sam-altman-wants-to-build-ai-superintelligence/
[11] For an excellent summary of how GPT models obtain their capabilities, see https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1
[12] More details about these models can be found at https://platform.openai.com/docs/models.

Embeddings	A set of models that can convert text into numerical representations	Via API with the "text-embedding-ada-002" model name for the most recent version
Moderation	A model that can detect sensitive or unsafe text, involving hate, threatening, self-harm, sexual, or violence content	Via API with the "text-moderation-latest" model name

2.3 ChatGPT

The term "ChatGPT" commonly refers to both the chatbot application created by OpenAI and the underlying models that drive its capabilities. ChatGPT became immensely popular after its release and reached one million users in merely five days. In comparison, it took Instagram 2.5 months to reach the same number of users. Currently, ChatGPT has 180.5 million users.[13] As shown in Figure 1, public interest in ChatGPT is still on the rise, with no sign of slowing down. The Google Trends data also indicates that the general interest in LLMs is not even a fraction of the interest garnered by ChatGPT.

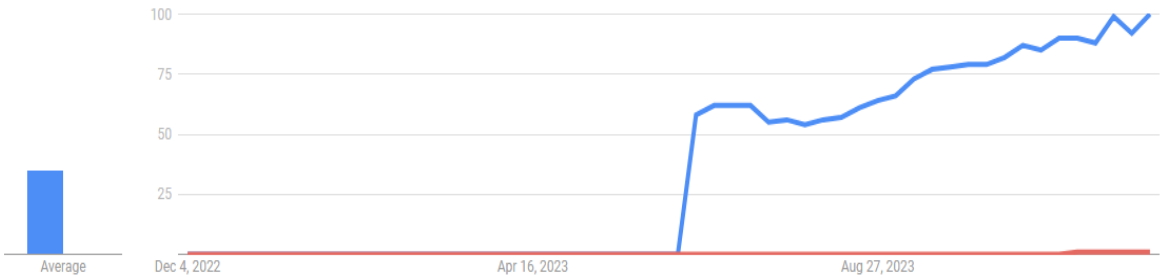

Figure 1 Google Trend of ChatGPT (Blue) and LLMs (Red)

The ChatGPT web interface is easy to use and provides an intuitive way to interact with the model behind it. Other than the free web access to GPT-3.5, OpenAI provides a free playground, where users can try out a large variety of models, including the most advanced "gpt-4-1106-preview".[14] GPT-4 is also available via Microsoft's Bing Chat, which has been rebranded as Copilot as of November 2023. However, for most researchers, the API is the preferred option, especially for processing large data sets. We provide guidance on some of the technical aspects regarding the use of the API in the Appendix.

[13] https://explodingtopics.com/blog/chatgpt-users
[14] Available at: https://platform.openai.com/playground

III. SCOPE AND METHODOLOGY

In assessing the extant literature, researchers opt for one of two primary review methodologies: a systematic review or a scoping/mapping review (Yang Li et al. 2018; Paré et al. 2015).[15] A systematic review is commonly employed to synthesize established literature on a well-researched topic, with the purpose of elucidating "what works." Conversely, a scoping or mapping review aims to explore the breadth and scope of emerging literature related to a new topic. This type of review serves to identify knowledge gaps and inform future research agendas, emphasizing understanding "what has been done" rather than "what has been found."

As informed by the latest Gartner Hype Cycle (Gartner 2023) and supported by anecdotal evidence (Economist 2023a), it is evident that the adoption of AI/LLMs is in its early stages. Given this context, our study employs a scoping review methodology to survey topics that have been the focus of existing research, to identify gaps that remain in the current body of literature, and to propose avenues that hold promise for future exploration.

Guided by established protocols in the literature review (e.g., Paré et al. 2015; Snyder 2019), we design a procedure that consists of four key steps: 1) formulating a framework to organize and guide the review of the literature, 2) developing and implementing a literature search strategy, which includes eligibility criteria, 3) executing a quality assessment, and 4) interpreting, discussing, and synthesizing the results.

3.1 Framework and Research Questions

We develop our framework for organizing and evaluating the studies in our sample by drawing on two streams of research, namely, literature review studies on the adoption of emerging technologies (Lee et al. 2023; Yang Li et al. 2018) and studies on how the state of adoption of an emerging technology influences the type of research that is done or can be done (O'Leary 2008; 2009). Given that ChatGPT is one of the applications of an emerging technology (i.e., AI/LLMs), we propose a framework that builds on approaches used in recent literature reviews related to the adoption of emerging technologies, such as blockchain and AI (Lee et al. 2023; Yang Li et al. 2018). This means that at a high level, studies can be organized using an input-process-output (I-P-O) approach (Lee et al. 2023). Under this approach, *input* relates to motivation for the adoption of a new technology or the focus area of application; *process* relates to how the technology is used, what challenges, difficulties, and problems are involved, and what guidelines and best practices are recommended or available; finally, *output* relates to the implications of widespread adoption of the technology.

We complement the I-P-O approach by incorporating evidence from the current stage of LLM adoption. According to O'Leary (2008), the stage of technology adoption determines the type of academic research that can be (is) done. Using the Gartner Hype Cycle (Fenn and

[15] Snyder (2019) classifies literature reviews as systematic, semi-systematic, and integrative. In this paper, we follow the typology of Paré et al. (2015) because it has a more explicit technology-oriented focus.

Raskino 2008) as a proxy for technology adoption, O'Leary argues that researchers will try to educate themselves about the emerging technology during the early stages when little is known regarding how the technology works and what are their capabilities. Over time, as the technology matures and becomes mainstream,[16] we will start seeing large-scale empirical studies that focus on the financial, market, and competitive payoffs from adoption. Combining the insights from these two streams of research, we develop a systematic approach for organizing and reviewing the studies in our sample by focusing on motivation and application focus, process, and output.

3.1.1 Input: Motivation & Focus Areas

In the technology adoption literature (Rogers 1995), adoption starts with mavericks who can visualize how the new technology could improve the efficiency and effectiveness of their work or their organizations. Extending this logic to accounting and finance research, we argue that the group of researchers who have produced research papers in our sample are those who can visualize the application of LLMs in accounting and finance fields (e.g., financial reporting, audit, and asset pricing). This is consistent with O'Leary (2008), who argues that studies at the early adoption stage will tend to emphasize the positive aspects of the new technology. Therefore, a systematic analysis of the foci of these studies would serve as a proxy for the areas where researchers identify the initial, most accessible opportunities for applying LLMs in accounting and finance.

For this purpose, we first outline the key areas of interest by drawing on the major research areas recognized by the American Accounting Association and American Finance Association, as shown in Table 4. By focusing on motivation and strategic research areas, our framework provides a holistic view of the initial considerations and objectives shaping the utilization of LLMs in accounting and finance. This also allows us to see the state of the art, gaps, and opportunities for future research in each accounting and finance area.

Table 4 Accounting & Finance Areas of Research

Accounting	**Finance**
Accounting Information Systems	Asset Pricing and Investment
Auditing	Corporate Finance
Education	Education
Financial Accounting and Reporting	Risk Management
Management Accounting	
Taxation	

[16] A technology is considered to have become mainstream when it enters the stage of early majority in the adoption cycle, with an adoption rate of approximately 15% (Stratopoulos, Wang, and Ye 2022).

3.1.2 Process: How/Capabilities

From the *process* perspective, we examine the specific capabilities of LLMs that researchers leverage in their work. We start by asking ChatGPT 4.0 what capabilities it possesses. At the prompt of "What are you capable of doing?" ChatGPT generates the following list:[17]

- Answering Questions
- Data Analysis and Assistance
- Language Tasks
- Programming and Coding Help
- Creative Content Generation
- Education and Learning
- General Guidance and Advice

To systematically organize the research papers, we need to dive deeper into these capabilities. While ChatGPT's array of functions provides a broad foundation, our study requires a more focused approach tailored to the unique demands of accounting and finance research. For instance, "Language Tasks" is one of the many capabilities of ChatGPT, and this capability encompasses translation, grammar checks, writing assistance, and summarization, among others. While these are broadly useful in the context of accounting and finance research (Korinek 2023), some of them are notably more relevant than others, e.g., summarization, which aids in extracting key points from lengthy narratives commonly found in corporate disclosures (Kim, Muhn, and Nikolaev 2023a).

To this end, we propose organizing the capabilities of ChatGPT as follows, each ranked by its complexity and practical application in accounting and finance research:

(1) Word embeddings generation.[18]
(2) Information retrieval.
(3) Question-answering (basic) – basic accounting and finance concepts.
(4) Classification and sentiment analysis.[19]
(5) Question-answering (advanced) – application of accounting and finance concepts to complex scenarios.
(6) Text/code generation.
(7) Summarization.

---

[17] Please note that ChatGPT may have customized the answers according to the personal information available in the OpenAI account profile of the author who asked this question.

[18] A word embedding is a numerical representation of a word as a vector of numbers, such that words closer in the vector space have similar meanings. It is also possible to represent a sentence or even an entire document as a numerical vector, and the results are known as sentence embeddings, document embeddings, or simply text embeddings, which are often more useful for many NLP tasks. We use the term "word embeddings" expansively to include also sentence embeddings, document embeddings, and most broadly text embeddings.

[19] Classification of text involves assigning pre-defined labels to words, sentences, or longer blocks of text. Sentiment analysis also involves a classification task, and the labels are often "positive", "negative", and "neutral", or in a more granular format. We single out sentimental analysis, because we find that many papers use ChatGPT for sentiment analysis.

(8) Predictions.
(9) Decision aid by providing recommendations using logical reasoning.

This ranking mirrors the stages and complexity of data analytics, from data management to descriptive, diagnostic, predictive, and prescriptive analytics (Stratopoulos 2018). For example, word embeddings generation and information retrieval represent the initial stage of data collection/management, which is essential for subsequent analysis. Classification and sentiment analysis echo diagnostic analytics, as they make it possible to extract deeper insights from data, facilitating a more nuanced understanding of financial narratives. Lastly, capabilities like predictions and logical reasoning are akin to the advanced stages of predictive and prescriptive analytics, where the focus shifts to forecasting future trends and formulating strategic recommendations based on the analyzed data. Understanding the nuances of the process (i.e., capabilities used and their complexity ranking) is important for gaining insights into the practical applications of LLMs in the field, as well as for identifying gaps in existing literature and opportunities for future research.

3.1.3 Output: Adoption Maturity & Implications

For the *output* perspective, we analyze the outcomes as reported in studies leveraging ChatGPT and similar LLMs in accounting and finance. To discern between expected outcomes, where LLMs are utilized in controlled environments, and actual outcomes, where their intended user base adopts LLMs, we propose the following four groups based on implied adoption maturity (the stage in the adoption cycle):

(1) Conceptual papers, which represent the researchers' conceptualization and visualization of how the technologies can be applied and how to best apply them, based on general theories and practices.
(2) Case studies from early adopters of the technology.
(3) Potential applications, which delve into forward-looking analyses of how the technology could be utilized by demonstrating its capability through large-scale experiments or designing a framework that guides the application of the technology in a specific field.
(4) Value realization studies, which assess the impact of technology adoption on organizations and professions. They examine whether integrating the technology has led to tangible value creation for users and organizations.

Mapping these studies into groups of different adoption maturity helps capture the unique aspects of LLMs. Initially, as observed by O'Leary (2008), the adoption of emerging technologies tends to be low, leading to early studies that are conceptual and exploratory, often centered on pilot projects within individual firms. However, with the Economist (2023b) predicting that "Generative AI will go mainstream in 2024," we expect a quicker transition to large-scale empirical studies. These studies will provide insights into the financial and market implications of LLMs, reflecting the evolving nature of this technology.

Another facet of the Output perspective concerns the interplay between education and the labor market, influenced by LLMs. Several studies have highlighted LLMs' impact on education, suggesting potential effects on students' future earnings (Huseynov 2023). Coupled with a

multitude of studies on the impact of LLMs on white-collar jobs and their broader implications for the labor market, it becomes imperative to review studies that investigate the implications and impacts of LLMs specifically for and on the accounting and finance sectors. This comprehensive approach is necessary to understand the broader implications of LLMs in our domain. By synthesizing these diverse studies, we aim to provide both a comprehensive and a nuanced view of how LLMs are not only transforming educational and professional landscapes but also redefining the future trajectory of accounting and finance as a field.

3.2 Literature Search and Selection

The aim of a literature review on an emerging technology is to create initial conceptualization and theoretical models rather than review old models. Thus, "this type of review often requires a more creative collection of data" (Snyder 2019, 336). We endeavor to capture this preliminary conceptualization by reviewing both work-in-progress and published research.

We rely on SSRN to identify existing research since most studies are preprints or work-in-progress. SSRN allows researchers to quickly disseminate their research in social sciences, humanities, and other disciplines. It is a common practice for researchers in accounting, finance, and economics to upload their working papers or recently published papers to SSRN, so that the research community can benefit from their most recent or ongoing research findings. This is especially true for studies on emerging and timely topics like LLMs.

To obtain our initial list of papers, we searched on SSRN for papers that [1] have "ChatGPT" or "GPT" in the title, abstract, or keyword list; [2] are in "Accounting", "Finance", or "Economics" networks; and [3] are uploaded during the period from the beginning of 2022 to the end of October 2023. We further supplement this list of papers by including published papers cited by them. We exclude working papers with five pages or fewer since such brevity often indicates that the associated studies are likely to be too preliminary or lack rigor. 195 papers were uploaded to SSRN during our sample period. Among these, 72 have an accounting/finance focus, 40 an economic focus, and the remaining another focus (e.g., law). Within the subset of accounting/finance papers, 29 have an accounting focus and 43 are more related to finance.

To identify published papers, we conduct a search on the World of Science (WoS) for papers in business economics that contain "ChatGPT" or "GPT" in their titles or abstracts. The classification of business economics on WoS encompasses accounting, finance, economics, and other disciplines in business and economics. This search yields an initial list of 71 published papers up to October 31, 2023. Subsequently, we manually review each paper based on its title and abstract, retaining 15 papers specifically in the fields of accounting or finance.

3.3 Descriptive Statistics

Figure 2 depicts the timeline when the working papers were initially uploaded to SSRN. The activity started in late 2022 and has since experienced significant growth, peaking with a notable surge in Spring 2023, followed by some fluctuations thereafter.

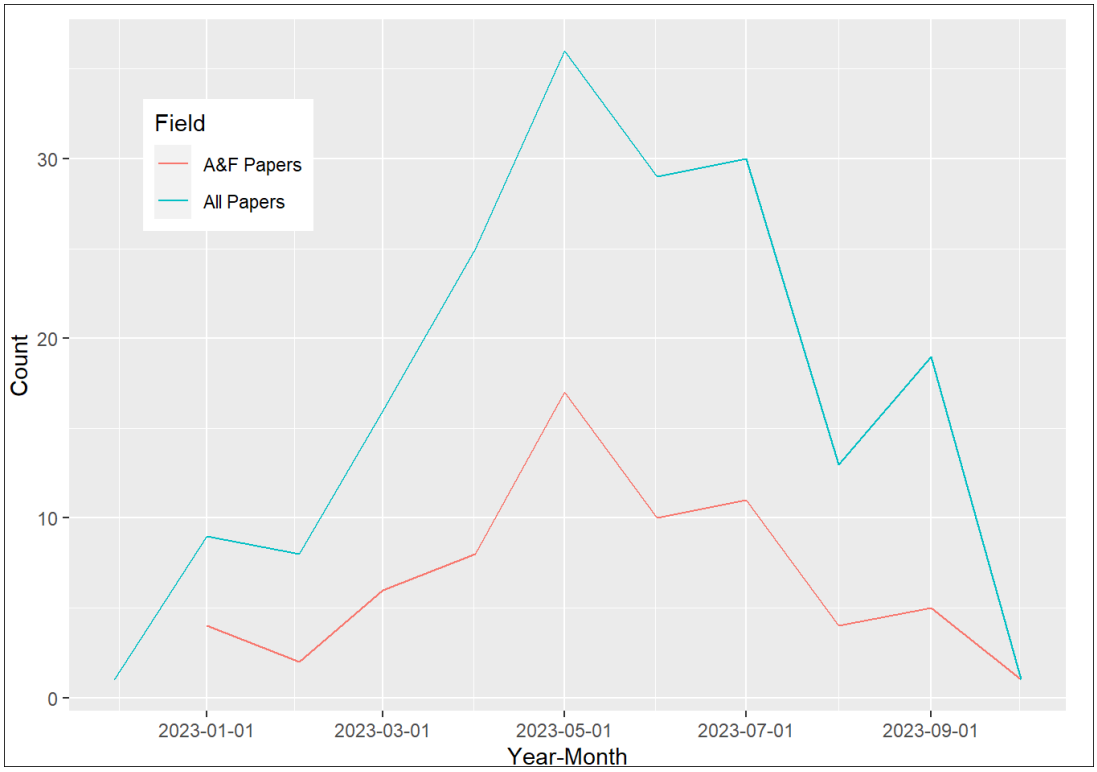

Figure 2 Uploaded Papers per Month

The observed pattern from the monthly distribution of uploads is consistent with the unfolding of key developments in LLMs, particularly the release of ChatGPT 3.5 on November 30, 2022, and ChatGPT 4.0 on March 14, 2023. The pivotal role of ChatGPT in this body of research is reflected in the frequency of papers using ChatGPT, as shown in Table 5. When a paper discloses the specific model version of ChatGPT used, we make a further distinction between GPT-3.5 and GPT-4. In cases where the model version is not disclosed, we use the blanket term "ChatGPT", which, given the timing of these papers, predominantly refers to GPT-3.5. When a paper uses the OpenAI API to access a model, the specific version of the model is often disclosed. Some papers use multiple models, often comparing their performance on certain tasks.

Table 5 Frequency of Papers per Model (Top Ten Models)

Models	Count	Percent
ChatGPT	21	17.6
GPT-4	15	12.6
GPT-3.5	12	10.1
GPT-3	7	5.9
FinBERT	5	4.2
GPT-3.5-turbo	4	3.4
Google Bard	4	3.4
BERT	3	2.5
Bing Chat	2	1.7
GPT-1	2	1.7

Regarding activity within each accounting and finance field (Figure 3), we observe that in the former most papers focus on financial accounting and reporting (33%) and auditing (19%), while in the latter most papers focus on either Asset Pricing and Investment (35%) or Corporate Finance (25%). From the perspective of our framework, this may indicate that ChatGPT can be more easily applied to these four areas or the benefit of ChatGPT adoption is more evident in these areas. It could also be the case that more researchers work in financial accounting and reporting than in other areas of accounting research.

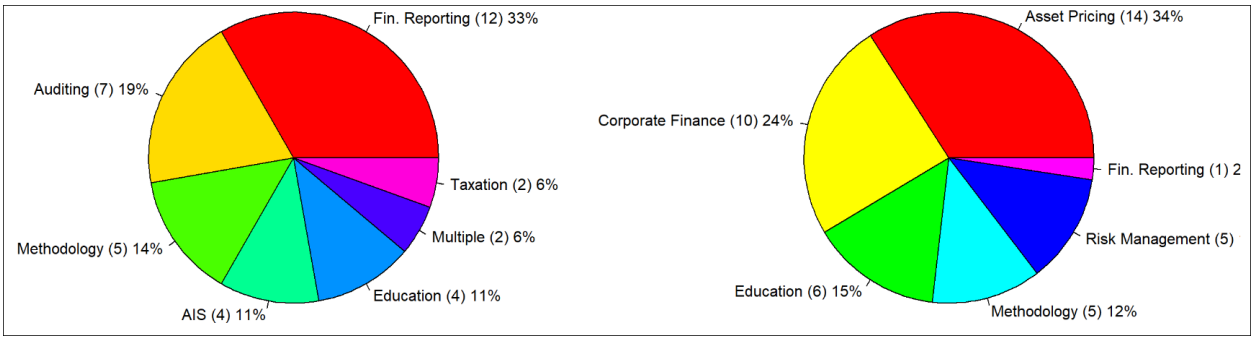

Figure 3 Most Active Accounting and Finance Fields

For studies wherein ChatGPT is used as a practical tool, either by researchers or within a corporate setting, we examine the specific capabilities of ChatGPT being leveraged. Table 6 lists the top five within accounting and finance. It is important to note that these statistics exclude papers in which the authors merely suggest or hypothesize potential uses of ChatGPT or related LLMs without actively or meaningfully employing a model in their research.

Table 6 Most Frequently Used Tasks in Accounting & Finance

Task	Count	Percent
Text generation	6	23
Question-answering	5	19
Classification	4	15
Logical reasoning	3	12
Sentiment analysis	3	12

Finally, Tables 7 and 8 capture the frequency of output categories in accounting and finance respectively. Consistent with the current state of LLM adoption, we see a good representation of conceptual papers. Interestingly, we observe that studies exploring potential applications of ChatGPT and other LLMs in accounting and finance constitute the majority of studies in accounting (67%) and represent over 40% in finance.

Table 7 Most Frequent Output Category in Accounting

Output_Category	Count	Percent
Potential applications	16	67
Conceptual	7	29
Case study	1	4

Table 8 Most Frequent Output Category in Finance

Output_Category	Count	Percent
Potential applications	14	41
Conceptual	12	35
Value realization	8	24

IV. SYNTHESIS OF EXTANT LITERATURE

4.1 Input Phase: Applications of ChatGPT in Accounting

Practical applications of a new technology start with an awareness and understanding of its functions and capabilities. Researchers can play a significant role in facilitating the widespread adoption of new technologies by demystifying them and envisioning their practical applications. Consistent with the observation of O'Leary (2008), during the early stages, numerous studies aim to educate readers about ChatGPT and LLMs in general. These studies describe and demonstrate their capabilities and explore their potential applications in various accounting fields, such as financial reporting, audit, and tax. For example, Zhao and Wang (2023) discuss the prospective applications of ChatGPT across diverse accounting tasks.[20] Similarly, Street, Wilck, and Chism (2023) encourage CPAs to improve their productivity by experimenting with LLMs for language generation tasks, and additionally provide some general principles for safely and effectively doing so. Two studies, Fotoh and Mugwira (2023) and Street and Wilck (2023a), provide domain-specific guidelines. The former discusses the potential benefits, pitfalls, and ethical considerations of integrating ChatGPT into external audits, while the latter introduces ChatGPT to forensic accounting professionals.

According to these studies, ChatGPT holds potential to reshape various accounting processes by automating routine mundane tasks that require modest professional judgment. Such tasks may include drafting various types of documents, ranging from internal memos to correspondence with audit or tax clients. LLMs can offer valuable assistance to CPAs working at firms more constrained in resources, helping to address staff shortages exacerbated by the profession's diminishing appeal to new talent (Boritz and Stratopoulos 2023).

However, these potential benefits do not come without risks. The risks or challenges include output accuracy, data security, client privacy, integration complexities, accountability,

[20] For existing and proposed applications of ChatGPT in business in general, please see Singh and Singh (2023).

and intellectual property rights. As such, the authors of these studies urge accounting professionals to maintain a high level of vigilance over these risks. Additionally, the authors caution accounting professionals against an over-reliance on these new technological tools, warning that such over-dependence could potentially undermine their critical thinking skills. Consistent with the findings of Dell'Acqua et. al (2023), they recommend that CPAs exercise professional skepticism and use LLMs to enhance rather than replace their own expertise.

4.1.1 Auditing

Several studies focus on the use of LLMs in audit settings. For example, Wei, Wu, and Chu (2023) investigate the cognitive capabilities of ChatGPT in comparison to human auditors by asking an identical set of questions to ChatGPT and human auditors. They find that the answers from ChatGPT demonstrate a high similarity in sentiment, diction, and linguistic complexity to those from human auditors. The researchers interpret these parallels as indicative of analogous cognitive processes between human auditors and ChatGPT and posit that ChatGPT could potentially undertake some of the tasks traditionally performed by human auditors.

Using a design science approach, Gu et al. (2023) proposes the integration of LLMs into auditing tasks through a process referred to as "Artificial Intelligence Co-Piloted Auditing." Under this new audit paradigm, LLMs take on a more substantial role by supporting human auditors in performing various auditing procedures. The authors demonstrate the use of this new approach by fine-tuning GPT-4 to conduct three audit tasks, which include journal entry testing, financial analysis, and text mining.

Two studies focus on the application of LLMs to internal audit settings. Eulerich and Wood (2023) demonstrate with specific examples how internal auditors can use ChatGPT in all aspects of the internal audit process. Through a case study analysis, Emett et al. (2023) provide an in-depth examination of how Uniper, a large multinational company, has started to use ChatGPT to assist various internal audit processes, harvesting significant efficiency gains estimated to range from 50% to 80%. Additionally, the authors discuss the risks and challenges that Uniper has identified as well as the rules and practices that have been put into place to address them. These insights, grounded in a real use case, should prove valuable to other companies planning to adopt ChatGPT to enhance their internal auditing processes.

Within certain jurisdictions, e.g., the European Union (EU), companies are required to obtain assurance for their sustainability reporting. Föhr et al. (2023) propose incorporating LLMs into the auditing of sustainability reporting. Using a design science approach, they propose a framework for integrating LLMs into the auditing workflows and demonstrate the application of the framework through an experimental case study. Specifically, they use GPT-4 to analyze ESG reports to evaluate whether they comply with the EU Taxonomy. Their experiments suggest that ChatGPT can be a useful tool that auditors can rely on for improved efficiency.

4.1.2 Financial Accounting and Reporting

LLMs are well suited for analyzing unstructured textual information, e.g., forward-looking disclosures in annual reports and narrative disclosures in sustainability reports. The study of de

Villiers, Dimes, and Molinari (2023) proposes the use of LLMs for sustainability reporting. Based on a conceptual framework covering the four stages of non-financial reporting (i.e., management information gathering, report generation, assurance/auditing, and consumption by users), they discuss how generative AI can improve efficiency in each of these stages. The authors, however, raise a concern regarding the potential misuse of generative AI to facilitate greenwashing, a deceptive practice involving the production and dissemination of misleading information to portray a company as more environmentally friendly or sustainable than it actually is.

Extracting pertinent information from annual reports can be laborious and time-consuming, given the often excessive length and poor readability of these documents. Gupta (2023) shows that the process can be automated with the aid of LLMs. The author designs a tool for using LLMs to extract innovation-related information for thousands of public companies. The proposed method can benefit users of annual reports such as financial analysts and other investors, as well as researchers.

Ni et al. (2023) build a system to automatically analyze sustainability reports and evaluate whether the reports comply with Task Force for Climate-Related Financial Disclosures (TCFD) recommendations. Currently, this system relies on "text-embedding-ada-002" for generating text embeddings and uses ChatGPT for summarizing and assessing compliance with the TCFD framework. In prompt engineering, they incorporate human expert feedback to improve accuracy and reduce hallucination. This system also supports customized Q&A, through which users can ask questions about the contents in the sustainability report.

4.1.3 Taxation

Based on current evidence, the ability of ChatGPT in tax is still quite limited, perhaps due to the high technicality of tax law. Zhang (2023) evaluates the ability of ChatGPT (GPT-4) and other LLMs (Bing Chat and Google Bard) to answer tax-related questions. He finds that at the current stage, these LLMs are only able to answer open-ended questions without definitive right or wrong answers. The author argues that it may take several years before LLMs can perform tax research in a way as capable as junior tax associates. Alarie et al. (2023) evaluate the ability of ChatGPT (GPT-3.5 and GPT-4) and Blue J (a proprietary LLM trained on tax data) to answer tax-related questions and demonstrate some advantages of Blue J over ChatGPT, e.g., higher accuracy. One caveat is that the authors are employees of the company that develops and markets Blue J.

4.1.4 Accounting Information Systems (AIS)

Given the prediction of exponential growth in the adoption of LLMs at the enterprise level (Cooney 2023; Economist 2023b), it is not surprising that several studies have discussed their role in the firm's IT infrastructure/AIS systems. Seeing LLMs as part of the infrastructure of an enterprise system, O'Leary (2022) compares common LLMs and discusses emerging issues. Using an approach like penetration testing (i.e., a simulated cyberattack on a computer system performed to evaluate the security of the system), O'Leary (2023a) compares the performance of ChatGPT, BlenderBot, and LaMDA and discusses the characteristics, risks, and limitations

regarding the adoption of enterprise LLMs (O'Leary 2023b). Analyzing the role of LLMs from a systems development standpoint, Beerbaum (2023) argues that Generative AI-enabled RPA has the potential to be applied to accounting practice, freeing accountants from repetitive routine tasks and allowing them to focus on tasks that require more judgment.

4.2 Input: Applications of ChatGPT in Finance

Numerous studies argue that LLMs have the potential to transform the way financial professionals do their daily work, given their capabilities of understanding intricate patterns and automating routine and even certain complex processes. Treating ChatGPT as a domain-specific expert is a common method used to visualize the scope and applications of this form of new technology (O'Leary 2008). For example, Zaremba and Demir (2023) explore the applications of ChatGPT in finance and their potential for enhancing NLP-based financial analysis. In multiple related studies, Krause (2023d; 2023b; 2023c) delves into the benefits of applying generative AI to finance, ranging from improved operational efficiency to enhanced analytical accuracy. Krause emphasizes the role of generative AI in identifying key trends and themes in financial data and explores the capabilities of LLMs. The benefits that ChatGPT can bring to the finance field come with potential risks and other challenges. These may include information inaccuracy, privacy and security concerns, opacity in decision-making processes, labor displacement, and legal considerations (Khan and Umer 2023).

4.2.1 Asset Pricing & Investment

Multiple studies create portfolios using ChatGPT to see if they can beat those created using traditional methods. For example, Romanko et al. (2023) use ChatGPT to select stocks from S&P 500 companies and find that its stock selection is overall effective. However, ChatGPT lacks the ability to assign appropriate weights to stocks in the portfolio. For improved performance, the authors recommend enhancing ChatGPT's stock selection with established portfolio optimization methods. Lu, Huang, and Li (2023) show that ChatGPT can generate portfolios of high alpha based on its analysis of news announcements regarding the Chinese government's economic policy.

Several studies leverage ChatGPT to create customized investment portfolios that align with clients' risk profiles and preferences (e.g., for sustainability investing). Remarkably, the suggested portfolios exhibit high risk-adjusted returns that are on par with those managed by professional portfolio managers. For example, Cheng and Tang (2023) use ChatGPT to create factors and find the factor portfolios generate significantly high Sharpe ratios and alphas beyond those explainable by Fama-French 3-factor and 5-factor models. Goyenko and Zhang (2022) demonstrate the ability of LLMs to actively time premium realizations of factors, dynamically re-balance, and diversify between factors. Jain et al. (2023) create an ESG classifier using GPT-3.5 and find that the classifier can identify ESG factors, allowing investors to make investment choices that align with their values. Additionally, the model is useful for evaluating the ESG performance of companies within various industries, helping investors select sustainable and socially responsible investments.

The evidence from these studies, to some extent, complements Niszczota and Abbas (2023) and Fieberg et al. (2023). These two studies evaluate the potential application of ChatGPT in financial advising. Both studies find that ChatGPT can effectively serve as a financial advisor by offering appropriate financial advice. However, ChatGPT exhibits weaknesses such as home bias and its tendency to overlook investment horizon (Fieberg, Hornuf, and Streich 2023).

Wang (2023) shows that ChatGPT can also be used to assist proxy voting for small passive investment funds. Traditionally relying on proxy advisors, these funds may now make more informed voting decisions with the help of ChatGPT. This improvement can better serve shareholders and help these funds navigate the competitive market environment.

Leippold (2023) presents an interview with GPT-3 on climate finance, revealing the model's impressive knowledge in this domain. The persuasive responses showcase the potential application of GPT-3 in climate finance. It is noteworthy that these findings are based on an older GPT model. Newer models (e.g., GPT-4) may have even greater capabilities for climate finance applications.

4.2.2 Corporate Finance

The use of ChatGPT in firm valuation is another area that has attracted the attention of finance researchers. Jha et al. (2023) leverage ChatGPT to analyze corporate disclosures to assess a company's investment policies. They create an investment score according to managers' future capital expenditure plans revealed at conference calls. They find that these scores can predict future capital expenditure for up to nine quarters and that firms with high investment scores experience substantial future abnormal returns.

Stock holdings and trading data of investors contain useful information. Borrowing the idea of word embeddings as with LLMs, Gabaix, Koijen, and Yogo (2023) generate "asset embeddings" from investors' stock holdings and trading data. They demonstrate that the so-called asset embeddings are useful for improving the accuracy of firm valuation, explaining return co-movements, and identifying asset substitution patterns. They also show that "investor embeddings", generated as a by-product of asset embeddings, can serve as a metric for measuring investor similarity.

Krause (2023a) discusses the limitations of AI models when evaluating private companies. For instance, AI models might struggle to grasp the intricacies of a private company's business model or its competitive landscape. The author argues that it is crucial to complement AI models with additional due diligence methods like traditional financial analysis and industry research.

Chen, Wu, and Zhao (2023) provide an overview of the latest advancements in generative AI within business and finance, covering its practical applications and the challenges and limitations associated. Additionally, they test the ability of ChatGPT to capture the sentiment of environmental policies as disclosed in corporate reports. They find that the sentiment as assessed

by ChatGPT provides valuable insights into companies' risk management capabilities and can predict future stock returns.

4.2.3 Risk Management

A few papers examine how ChatGPT can be used for risk management. For example, Hofert (2023a) examines ChatGPT's understanding of key concepts associated with quantitative risk management. In another study by the same author, Hofert (2023b) tests the ability of ChatGPT to understand correlation pitfalls in risk management. Both studies note that while ChatGPT displays adeptness in grasping non-technical concepts, it encounters challenges in comprehending complex mathematical models underlying highly technical concepts. In another related study, Wang (2023) explores the transformative potential of generative AI, such as ChatGPT, for operations risk management (ORM). The study highlights the technology's capability to analyze large data sets, simulate scenarios, and automate tasks.

In summary, numerous researchers have suggested scenarios/use cases for leveraging LLMs in practically all accounting and finance areas. The consensus emerging from these studies indicates that LLMs can improve the efficiency and effectiveness of various accounting and finance tasks.

4.3 Process: How ChatGPT is Used in Accounting and Finance Research

A large number of studies use ChatGPT and related LLMs as research tools. This is not surprising since a significant amount of information provided by or available to accounting/finance professionals comes in the form of unstructured textual data. LLMs, due to their powerful capability in NLP, provide new tools that enable researchers to analyze and draw insights from such textual data. Despite a growing number of studies using ChatGPT as a research tool, ChatGPT and other LLMs are new to most researchers.

To popularize this new type of tool, De Kok (2023) provides comprehensive guidance on how to use LLMs for textual analysis, demonstrated using ChatGPT. Even though the guidance focuses on analyzing text (e.g., earnings conference call transcripts), which traditionally fall into the accounting and finance domain, many guidelines also apply to research in other disciplines. For example, the author proposes a framework for effectively using LLMs, covering tasks such as model selection, prompt engineering, and construct validity tests. However, researchers should recognize that each project has its special considerations, and no guideline can be universally applied to all contexts. Furthermore, given that LLMs represent a new and rapidly advancing technology, effectively harnessing their capabilities to address diverse research questions remains an ongoing challenge.

Several studies explore and propose ways that LLMs can be used to enhance research productivity for researchers in economics, finance, and other related disciplines (Dowling and Lucey 2023; Feng, Hu, and Li 2023; Korinek 2023). Illustrated using ChatGPT, these studies provide use cases ranging from ideation and literature review to data analysis, coding, and mathematical derivation. The consensus seems to be that LLMs excel in idea generation and data

identification, but they exhibit limitations in literature synthesis and the development of suitable testing frameworks.

Next, we discuss the studies that rely on ChatGPT as a research tool by using one or more of its capabilities. While most studies use ChatGPT without any fine-tuning, one study in our sample fine-tunes a base model to enhance its capability for domain-specific tasks (B. Zhang, Yang, and Liu 2023). These authors adapted LLaMA-7B using instruction tuning for sentiment analysis of finance text. They find that the fine-tuned model outperforms common LLMs such as FinBERT, ChatGPT, and original LlaMAs in this specific task on finance text, especially when numerical information and contextual understanding are crucial for determining the sentiment. There are other fine-tuned models available for finance text. For a comprehensive discussion of these models and the technical aspects regarding their fine-tuning, see Li et al. (2023).

In what follows, we review the capabilities leveraged in accounting and finance studies.

4.3.1 Word Embeddings

Word embeddings are mathematical representations of words or tokens as real-valued vectors in a high-dimensional space, capturing the intricate semantic relationships between them. LLMs use context-dependent word embeddings, wherein the representation of a word depends on its context, thus better capturing the nuanced meanings of and intricate relations between words. As described in the background, word embeddings are part of the parameters within an LLM and are learned during the training of the LLM. Pre-trained LLMs can be used to generate word embeddings. OpenAI provides a dedicated model, "text-embedding-ada-002", for this purpose.

Breitung and Müller (2023) demonstrate the power of this technique by developing a new measure of global business networks, which assesses the economic interconnectedness between firms. Drawing inspiration from Hoberg and Phillips (2010), their approach differs in that they utilize word embeddings generated by LLMs such as "ada-002", Luminous and T5-XXL to calculate the cosine similarity between company descriptions. The authors show that the new measure outperforms traditional industry classifications in identifying relationships such as customer, supplier, and competitor networks, highlighting the new possibilities of enhancing downstream tasks using word embedding generated by state-of-the-art LLMs.

Bandara, Flannery, and Chandak (2023) evaluate the performance of word embeddings generated by several LLMs and traditional models for two downstream tasks, namely company identification and earnings surprise forecast. For the first task, they generate word embeddings for 10-Ks and earnings conference calls (the presentation section only) of the same company and classify 10-Ks and earnings calls based on cosine similarity scores. They find that "ada-002" from OpenAI outperforms all other models (including BERT, FinBERT, LSI, Word2Vec, and Doc2Vec) for this classification task. For the second task, they use word embeddings of earnings conference calls to predict earnings surprises and they find that BERT performs best among all models.

Yang (2023) uses "ada-002" to generate word embeddings for patent applications and finds that such context-dependent word embeddings have a much higher power in predicting the

economic value of patents. The author also shows that it is possible to develop an improved trading strategy based on these predictions.

4.3.2 Classification

Classification involves assigning pre-defined labels to the input data based on certain characteristics or criteria. Several studies use ChatGPT to classify textual data or identify text of certain topics. For example, Bernard et al. (2023) develop a new measure of business complexity by using a GPT-3 to classify XBRL tags. Specifically, they use OpenAI's GPT-3 Babbage model to predict XBRL tag names for numbers in footnote disclosures and use the "confidence" of the model's prediction as part of the input for constructing the business complexity measure. The idea is that if the model is more confident in predicting a tag, then the number behind the tag captures a more common business transaction, hence a smaller business complexity.[21]

A few studies use ChatGPT to identify whether corporate communications contain certain topics. For example, Li, Peng, and Yu (2023) examine whether companies increase ESG disclosures at M&A conference calls after such disclosures are mandated in many countries. To identify ESG-related disclosures, they ask ChatGPT whether text from conference call transcripts covers ESG topics. Kuroki, Manabe, and Nakagawa (2023) use GPT-3.5-turbo to classify management presentations at earnings conference calls of Japanese companies into "facts" or "opinions". They find that presentations from companies with lower profit margins or higher market-to-book ratios contain more "opinion" statements.

Notably, Föhr, Marten, and Schreyer (2023) use ChatGPT to classify interview transcripts. More specifically, they propose a framework for using LLMs and task-specific AI models in risk-based auditing procedures. To develop this framework, the authors rely on focus group discussions and interviews with subject matter experts to identify Deep Learning (DL) requirements (i.e., how DL can be used) at each stage of the risk-based auditing process. The authors analyze the qualitative transcribed data using ChatGPT to extract feature attributes (i.e., common themes) through the "Chain-of Thought Prompting" (J. Wei et al. 2023). This is the first paper we encountered that uses ChatGPT for processing qualitative data from field studies. LLMs hold promise as a tool for assisting survey-based research that involves questionnaires (Jansen, Jung, and Salminen 2023).

4.3.3 Sentiment Analysis

Sentiment analysis is a computational technique used to determine the emotional tone behind a body of text.[22] For this type of analysis, a piece of text is categorized as having a positive, negative, or neutral tone. Further refinement is possible by assigning numeric values to

[21] It is worth mentioning that a GPT model generates its output (i.e., a sequence of text) by predicting what tokens most likely follow the input (prompt). When the API is used, it is possible for researchers to extract the probability of each token in the output. However, it is not clear whether such probabilities can be used to construct a novel measure that captures the construct of business complexity, since probabilities in the context of an LLM are ultimately based on frequencies. We provide guidance on how to extract such probabilities in the Appendix.
[22] Fundamentally, sentiment analysis is a classification task. Because many studies use ChatGPT for sentiment analysis, we group these papers together and discuss them under this dedicated section.

differentiate the intensity of the tone. Automatic sentiment analysis can be performed using dictionaries or machine learning (ML) approaches. The word list developed by Loughran and McDonald (2011) is a commonly used dictionary for accounting and finance text. For ML-based approaches, BERT or its fine-tuned version for finance text, i.e., FinBERT (Huang, Wang, and Yang 2023) becomes popular in recent years.

Multiple studies examine whether ChatGPT's capability in sentiment analysis is superior to other approaches. Using MD&A disclosures in the Chinese language, Hu, Liang, and Yang (2023) compare the performance of GPT-3, FinBERT, and the word list of Loughran and McDonald (2011). They find that both GPT-3 and FinBERT outperform the wordlist approach. However, they find GPT-3 underperforms FinBERT, despite its larger parameter size, suggesting that larger-scale LLMs do not automatically guarantee better performance. Also focusing on a non-English language, Nakano and Yamaoka (2023) compares ChatGPT and traditional word-list approaches using news headings in the Japanese language, and they find that ChatGPT is superior.

Using titles and subtitles of news articles in Financial Times, Zhang (2023) finds that sentiment scores generated by a GPT-3.5 model outperform those generated by traditional BERT for predicting broad stock market movement. Using news headlines for individual companies (mostly press releases), Lopez-Lira and Tang (2023) run a horserace in sentiment analysis among multiple generations of GPT models and various variants of BERT models. They find that GPT-4 outperforms both earlier generations of GPT models and various BERT models in that its scores can more accurately predict individual stock returns.

Two studies use ChatGPT for tasks broadly related to sentiment analysis. Andreou, Lambertides, and Magidou (2023) use ChatGPT to validate whether forward-looking R&D disclosures appear to be overly optimistic, even though they do not construct a sentiment score and include it in subsequent analyses. Leippold (2023) uses GPT-3 to generate adversarial attacks for comparing the robustness in sentiment analysis between dictionary-based approaches and those based on context-aware ML models. They find that dictionary-based approaches are more vulnerable to adversarial attacks, whereas context-aware models like BERT are much more robust.[23]

4.3.4 Text Generation

Text generation involves automated creation of coherent, contextually appropriate text using advanced algorithms. Generative AI models excel in this task because they are specifically designed for this task. In a practical application, Bai et al. (2023) quantify the extent of new information from executives during Q&A sessions of earnings conference calls using various LLMs. To construct their measure of new information, they ask LLMs such as ChatGPT and Google Bard to answer questions from equity analysts and compare executives' answers with

[23] An adversarial attack involves deliberately and carefully modifying the input in order to deceive a model or algorithm. Leippold (2023) instructs GPT-3 to generate synonyms for token keywords in a sentence that is on the L&M word list and further asks it to replace these token keywords with one of the synonyms. It is not surprising that the L&M word list is more susceptible to such attacks because it is applied in a static manner without any adaption.

AI-generated answers. If executives' answers are highly similar to those generated by LLMs, they consider that executives provided little new information. An *implicit assumption* (which may be debatable) is that AI is not able to provide new information beyond that used in training.

4.3.5 Summarization

Summarization, traditionally a challenging task in NLP, involves condensing extensive textual content into a concise, coherent form while retaining its core information and intent. This process has historically lacked satisfactory solutions due to the complexity of accurately capturing and representing the nuances of large text bodies. Most advanced LLMs can create reliable summaries that mirror the depth and tone of the original text, when instructed to do so. This capability may enable investors and financial analysts to quickly digest information from corporate disclosures and promote market efficiency.

Kim, Muhn, and Nikolaev (2023a) use GPT-3.5-turbo to summarize MD&As in 10-Ks and conference call transcripts. They find that GPT-generated summaries are richer in information content, as they are better able to explain stock market reactions to the disclosed text. Based on the idea that a shorter summary indicates a smaller information density in the original disclosure text, they create a measure of "disclosure bloat" based on the length of the summary relative to the length of the original length. They find that bloated disclosures can slow down price discovery and increase information asymmetry. They also find that the tone of summaries tends to be amplified (i.e., the summary has an even more positive (negative) tone if the original document has a positive (negative) tone). However, this result may depend on the parameters the authors have chosen. It is possible to ask an LLM to maintain the same tone.

Using a similar approach, Kim, Muhn, and Nikolaev (2023b) ask GPT-3.5-turbo to summarize the exposure of a company to political, climate, and AI-related risks from corporate disclosures at earning conference calls. They additionally ask the model to make its own assessment of these risks based on the content from conference call disclosures. They find that such risk summaries and assessments are informative in that they outperform existing risk measures in predicting stock return volatilities and firms' investment and innovation policies. They also find that GPT's own risk assessments are even more informative than risk summaries.

4.3.6 Prediction

Several studies have focused on the capability of LLMs to make predictions. Li, Tu, and Zhou (2023) find that predictions of future earnings generated by GPT-4 exhibit greater forecast errors than analyst consensus, and GPT-4 is more optimistic than financial analysts. In sub-sample analyses, they find that GPT-4 generates more accurate forecasts for firms that have a better information environment, or which provide disclosures of higher quality in earnings press releases (e.g., more words or sentences, and high specificity). This is consistent with the notion that GPT-4 can utilize more background or contextual information for its forecast. Another interesting finding from this study is that the performance of GPT-4 is not impacted by the readability of the text, as measured by the Fog index. This could indicate that GPT-4 has a superior capability of comprehending complex text. It could also be the case that there is significant measurement error in this commonly used readability measure in extant literature.

Comlekci et al. (2023) use ChatGPT to forecast future financial performance (i.e., sales and net income) and dividends of public companies included in the Borsa Istanbul 100 Index of Turkey. They find that the performance of ChatGPT significantly improves when recent industry news is provided in addition to historical financial data. This finding highlights the importance of supplying context information to ChatGPT for improved performance.

In summary, evidence from the *process* perspective indicates that newer LLMs outperform traditional methods in many tasks. We note that there are more studies utilizing ChatGPT for tasks of lower complexity (e.g., word embeddings, classification) than for tasks of higher complexity (e.g., summarization and prediction).

4.4 Output: Adoption Maturity

Given the current early state of LLM adoption, it is not surprising that numerous conceptual studies aim to educate readers on how to apply LLMs in accounting (e.g., Street and Wilck 2023b; Street, Wilck, and Chism 2023; J. (Jingwen) Zhao and Wang 2023) and finance settings (e.g., Zaremba and Demir 2023; Romanko, Narayan, and Kwon 2023; Jha et al. 2023). While these papers offer initial and preliminary insights grounded in limited concrete evidence, they are valuable at the early stage of a new technology when guidance or best practice is scant. Drawing on their own expert opinions, they visualize the technology's potential scope and applications, critically assessing both its strengths and limitations. Such endeavors are essential for promoting a foundational understanding of the technology, covering both theoretical underpinnings and practical applications.

With the notable exception of Emett et al. (2023), who offer a comprehensive discussion of the integration of ChatGPT into various internal audit processes within a large multinational company, we have not seen any other case studies. Case studies are important because they provide real-world insights into how organizations have implemented the technology, including the challenges faced and the lessons learned. These papers are invaluable for understanding the practical aspects of technology adoption and offer a glimpse into the initial stages of its integration into business processes.

Interestingly, there has been a surge in forward-looking studies on how the technology can be utilized, wherein the authors demonstrate its capability through large-scale experiments or develop frameworks that guide the application of the technology to specific fields. We have seen such studies in practically every major field of accounting and finance, e.g., audit (e.g., Gu et al. 2023), financial reporting (e.g., de Villiers, Dimes, and Molinari 2023), asset pricing (e.g., Y. Cheng and Tang 2023), and corporate finance (e.g., Krause 2023a). This group of studies explores potential use cases and envisions how the new technology may create value for users and organizations. In doing so, they serve as a bridge between theoretical understanding and large-scale practical application, highlighting innovative ways technology can be leveraged to improve productivity.

We have seen several finance studies that examine the impact of LLMs on firm valuation. At the current stage, most of these studies use market reaction tests to infer the potential impacts on firms or industries from differential stock price returns observed on major event dates

throughout the development of LLMs. For example, Eisfeldt, Schubert, and Zhang (2023) investigate the impact of recent developments in Generative AI on the market value of U.S. public companies. Using their firm-level measure of workforce exposure to Generative AI, they find that higher-exposure firms earned higher returns than lower-exposure firms following the release of ChatGPT. Using an industry measure of workforce substitutability by automation, Blomkvist, Qiu, and Zhao (2023) document that companies within industries of high substitutability exhibit notably negative stock returns. Bertomeu et al. (2023) examine the effect of Italy's temporary ban on ChatGPT and they find that the stock prices of Italian firms with greater exposure to Generative AI underperform those with smaller exposure during the period of the ban. Pietrzak (2023) evaluates the short-term market response to public companies' mentioning of ChatGPT in 8-K and 6-K reports and finds that the stock market barely reacts to the release of such information. Based on survey data, Bughin (2023) investigates how large global corporations generate returns from AI investments and finds that only companies investing heavily in AI can generate significant returns from such investments. Overall, these studies do not present a consistent picture of the impact of LLMs on firm values. More studies are necessary to understand the long-term benefits of the new technology. With increasing adoption, future studies may employ actual adoption data to assess the ROI from LLMs.

In summary, our review has revealed a small number of conceptual papers and a substantial body of work focused on exploring potential applications. The substantial volume of the latter body of work may signify an accelerated adoption of LLMs. Prior literature on technology adoption (e.g., Alexopoulos 2011; Stratopoulos and Wang 2022; Stratopoulos, Wang, and Ye 2022) has introduced several proxies (e.g., book titles, news articles, Google Trend, and firm disclosures) to evaluate the stage of technology adoption. Based on our analysis and evidence related to LLM adoption (Cooney (2023); Economist (2023b)), the relatively large volume of working papers related to potential applications may serve as another proxy for predicting the stage of adoption.

4.5 Output: Impact on Education & Profession

Concerns and debates regarding the impact of technology on the labor market date back to the early 19th century marked by the Luddite movement and have resurfaced with each new wave of innovation. What sets LLMs apart from prior technologies is their potential to replace highly educated, well-compensated white-collar jobs. This has prompted a debate on job augmentation versus job replacement, and the evidence – at least for now and from the economics literature - is still equivocal (Allen et al. 2023; Hui, Reshef, and Zhou 2023; Kausik 2023; J. Liu et al. 2023). Given that education prepares students for such high-paying jobs, it is not surprising that some college students, especially those from non-STEM majors, feel pessimistic about their future job prospects (Huseynov 2023). Their concerns are consistent with the findings of Haugom et al. (2023), who show that companies in the edtech sector have been negatively affected after the public release of ChatGPT.

Within the realm of accounting and finance, we have observed studies focusing on how LLMs can be used to enhance education or on demonstrating the ability of LLMs to master accounting and finance concepts. Liu et al. (2023) integrate ChatGPT into a Python

programming course for data analytics in finance. Yang and Stivers (2023) assess the ability of ChatGPT and Google Bard to solve undergraduate finance problems, and they find that GPT-4 significantly outperforms Bard-1.0. In an early crowdsourced study, Wood et al. (2023) evaluate the ability of ChatGPT-3.5 to answer accounting questions on various topics and find that ChatGPT overall underperforms accounting students. Bommarito et al. (2023) evaluate the ability of ChatGPT-3.5 and earlier versions to answer CPA exam questions (multiple-choice questions) on zero-shot prompts. They find that ChatGPT-3.5 significantly underperforms human takers on a sample CPA REG exam, which is heavy in computation, with a correct rate of less than 15%. Cheng et al. (2023) find the ability of ChatGPT-3.5 and ChatGPT-4 to solve accounting business cases depends on the type of questions asked. These two models perform better in explaining concepts, applying rules, and evaluating ethical issues involved than in making journal entries, preparing financial statements, and using software.

However, a more recent study by Eulerich et al. (2023) tests whether ChatGPT is capable of passing major accounting certification exams, including CPA, CMA (Certified Management Accountant), CIA (Certified Internal Auditor), and EA (Enrolled Agent). They find that while ChatGPT-3.5 is not able to pass any of these exams, ChatGPT-4 can pass all of them. They also find that the performance of ChatGPT improves when it is shown some examples or when it is allowed to use a calculator or other resources. These findings are consistent with those from numerous studies that have demonstrated the ability of ChatGPT to perform tasks at a level comparable to a human auditor (e.g., T. Wei, Wu, and Chu 2023) or augment the abilities of financial analysts (e.g., Gupta 2023).

Evidence from finance studies has shown that ChatGPT can be used for financial advising. For example, Niszczota and Abbas (2023) evaluate the potential of GPT to function as a widely accessible financial robo-advisor. The assessment involves a combination of a financial literacy test and an advice-utilization task known as the Judge-Advisor System. GPT models achieved a satisfactory score. Fieberg et al. (2023) demonstrate that ChatGPT-4 is capable of offering effective financial advice. It can recommend customized investment portfolios tailored to an investor's specific circumstance, including risk tolerance, risk capacity, and sustainability preferences.

Based on the findings from these studies spanning the supply chain of talents (from education to accounting and investment firms), it is evident that ChatGPT and LLMs in general have the potential to disrupt the educational system and the accounting/finance industry.

V. DISCUSSION AND RESEARCH OPPORTUNITIES

The main purpose of this paper is to analyze and synthesize the expanding body of working papers and recent publications that focus on the applications and implications of ChatGPT and other LLMs in accounting and finance. Our approach is guided by a forward-looking perspective, with the aim to thoroughly understand the current state of the art and to identify promising avenues for future research. In our review, we employ a well-structured framework, informed by the literature on emerging technology adoption, to organize and synthesize the vast body of work covering diverse topics. Specifically, we approach the papers from three distinct

yet interconnected angles: *input*, which delves into the area of focus and underlying motivations; *process*, exploring the methodologies and capabilities employed; and *output*, assessing the level of adoption maturity and its broader implications.

Approaching this body of work from the *input* lens, we have seen studies in practically all accounting and finance areas. While there is a higher concentration in audit, financial reporting, asset valuation, and corporate finance, we note an absence of studies focusing on the potential applications in management accounting/behavioral research. We feel that this may be a promising area of future research for a couple of reasons. First, future studies may examine how ChatGPT can be integrated with Business Intelligence (BI) tools to enhance management accounting by combining complex data analysis with NLP. For example, ChatGPT can interpret outputs from BI systems—like visualizations or statistical data—to generate comprehensive narratives, which explain trends and business implications. These narratives may allow decision-makers to understand the context of the numbers. Second, there is already some interest in using LLMs in behavioral economics (e.g., Bauer et al. 2023; Tsuchihashi 2023).

While we have seen studies that propose applications of ChatGPT to the reporting of textual data, we have not seen any studies explore the application of this new technology to the reporting of financial numbers. This lack of study may be partly attributable to the inadequate capability of ChatGPT in financial reporting, e.g., making journal entries (X. (Joyce) Cheng et al. 2023). As the capability of ChatGPT undergoes continuous improvement, there is an opportunity to examine how ChatGPT can be integrated into the accounting information system of a company for automating certain “recording” tasks.

Focusing on the *process* aspect of ChatGPT adoption, we recognize multiple opportunities for future research. In accounting and finance, LLMs can serve as invaluable tools for textual generation. For example, an LLM can effectively define or explain financial concepts like "income statement" or "dividend" in an understandable manner. Such capabilities offer educational advantages and can significantly reduce the time and effort many users spend on manual content creation, providing a cost-effective solution for educational platforms or knowledge bases. On the other hand, the utility of LLMs extends far beyond mere textual generation. They can also provide support for decision making. For such applications, LLMs can extract useful information from a large text corpus to assist with decision making. For instance, an LLM can be trained to analyze earnings announcements, capturing the management’s sentiment, and thereby providing critical input for investment decisions. Unlike text generation applications, decision analytics requires a more nuanced understanding of the context and relies heavily on the model's capabilities in pattern recognition, inference, and prediction.

Another fruitful area of future research lies in the utilization of ChatGPT and other related LLMs as a tool for textual analysis. The evidence from existing studies seems to suggest that ChatGPT has a superior ability as a classifier, which can be used to generate measures for empirical tests. However, most existing studies have applied ChatGPT or other related models to a small volume of textual data. Taking sentiment analysis as an example, the existing studies focus on short pieces of text such as news headlines or press releases. Notably absent from our reviewed literature are studies where ChatGPT has been employed for sentiment analysis on

more extensive documents, such as annual reports or quarterly reports. We advocate for studies on larger scales to provide stronger and more conclusive evidence regarding the capability of LLMs for sentiment analysis on accounting and finance text. In addition, as the global economy becomes increasingly more integrated, there is a growing need to extend sentiment analysis across diverse languages to gain insights into how sentiments in regional markets affect the global financial market. We encourage further research focusing on less commonly studied languages to provide a more comprehensive understanding of the multilingual capability of ChatGPT and other LLMs.

Currently, most researchers have used ChatGPT for classification tasks in their studies. ChatGPT and other LLMs are inherently designed for generating text. Text generation thus arguably represents their most significant advantage over more traditional NLP tools. Creating metrics with ChatGPT for empirical testing typically necessitates the classification of text by assigning categorical labels. However, the creation of truly novel metrics necessary for addressing intriguing and hitherto unexplored research questions requires a creative application of ChatGPT's capabilities. One way to approach this is to convert a classification task into a text generation task. We have seen two papers that have applied this strategy (Kim, Muhn, and Nikolaev 2023a; Bai et al. 2023), which holds promise as a path forward for novel research applications using ChatGPT.

There are other ChatGPT capabilities not explored in existing studies within our review's scope. Two notable examples are Named Entity Recognition (NER) and translation. For example, the capability of ChatGPT in NER may be used to refine the disclosure specificity measure initially proposed by Hope, Hu, and Lu (2016). Using the translation capability, researchers may perform a comparative analysis of the disclosure practices of multinational companies across different geographical areas and jurisdictions.

No studies in our review explore the ability of ChatGPT to process images. Certain studies assessing the ability of ChatGPT to answer domains-specific questions purposely exclude questions with pictures (e.g., Eulerich et al. 2023). Future studies can evaluate the ability of ChatGPT by including questions with images. Additionally, multimodal LLMs can be used to process corporate disclosures, e.g., conference call slide decks, which are fraught with images and other infographics, and examine how such visuals affect investors' processing of information. An understanding of how investors integrate infographics, data visualizations, and other non-textual elements into their assessments provides valuable guidance for firms in shaping their disclosure practices. Such insights may also inform market regulators of potential new disclosure regulations that benefit investors and promote overall market efficiency.

Regarding the *output* aspect of LLM adoption, two key research streams have emerged. The first stream pertains to studies that explore expected/anticipated outcomes arising from adoption. This stream includes descriptive or conceptual research, postulating the potential benefits of LLM adoption. It also includes studies where the utility of LLMs is hypothesized, and the conjectured advantages are validated through small-scale experiments using researcher-generated data instead of real-life data. The second stream is rooted in archival research, concentrating on real-world scenarios where practitioners, such as auditors or portfolio

managers, have integrated LLMs into their operations. In these instances, the research aims to discern any performance disparities between LLMs users and non-users, offering a pragmatic perspective on the tangible impact of LLM adoption. The second stream also includes studies where the researchers infer the benefits of LLMs adoption from the stock market reactions to major advancements in LLM development.

Several studies propose or design a framework for applying ChatGPT to various fields of accounting and finance. As the technology continues to develop and becomes more integrated into accounting practices, researchers can empirically test whether the anticipated benefits of these tools are realized. For example, researchers may investigate the impact of ChatGPT and other LLMs on audit quality or financial reporting quality, or professional skepticism of practitioners. Before actual data becomes available, researchers may explore the perceived impacts of LLM adoption through surveys among accounting and finance professionals.

There is only one case study among the papers within our review, namely, Emett et al. (2023), which discusses how a multinational company has adopted ChatGPT in its internal auditing process. Even though case studies tend to have low external validity, they provide nuanced understanding and contextual insights into real-world applications. We encourage more case studies on how generative AI models are used in accounting and finance practices. Case studies allow researchers to document and study unique applications of LLMs in different organizations, capturing varied patterns of implementation, challenges faced, and innovative strategies employed. In-depth case analyses reveal how organizational contexts, such as firm size, industry, regulatory environment, and corporate culture, may influence the adoption and impact of LLMs. Case studies may also uncover unintended consequences of technology adoption. In addition, detailed case studies play a pivotal role in theory development and refinement pertaining to technology adoption, and insights gained from case studies can also inform future empirical research directions.

Another fruitful area of research is to investigate the actual impact of LLMs on firm performance. This can be done by searching corporate disclosures for announcements of LLMs adoption and linking the adoption to firm performance such as stock returns, ROA, or other operational outcomes. To facilitate this type of research, LLMs themselves can serve as a powerful tool. Researchers can utilize LLMs to sift through vast amounts of textual data contained in corporate disclosures, press releases, annual reports, and other public communications. Through their advanced text analytics, LLMs can assist in pinpointing which firms have integrated these technologies into their business processes, as well as the context and extent of their adoption. Through such research, academics and practitioners could gain a clearer understanding of the strategic value that LLMs bring to firms. Findings from such studies will not only inform corporate decision-making regarding AI investments but also offer important insights for policymakers, investors, and regulators concerned with the broader economic impacts of LLMs integration into business processes.

In summary, research into the application of ChatGPT and related LLMs within accounting and finance is in its infancy. This burgeoning area of inquiry is abundant with unexplored questions, offering fertile ground for scholarly investigation. To date, the literature

has only begun to scratch the surface of the potential applications and impacts of these advanced LLMs on the practice of accounting and finance. Given the transformative capabilities of LLMs, there is a clear opportunity for a deeper inquiry into a myriad of pertinent topics. We have proposed a great variety of exciting and promising research avenues, which, if pursued, could yield significant contributions to the theoretical understanding of technology adoption for LLMs as well as their practical applications and implications in various fields of accounting and finance.

VI. CONCLUDING REMARKS

At a high level, our review of these early studies suggests that integrating LLMs in practically every accounting and finance field can significantly enhance efficiency and effectiveness and that users assisted by LLMs may work more productively than those without such assistance. This points to a plausible trend towards substituting traditional labor with LLM-enhanced workflows. Additionally, the *process* aspect of our review shows that LLMs often outperform traditional methods in tasks like classification, sentiment analysis, and summarization. These superior capabilities of LLMs suggest that researchers and professionals using LLMs may potentially outperform their counterparts relying on older methods. Notably, the *output* aspect of our review indicates a shift in focus from conceptual to practical applications of LLMs. This shift not only demonstrates the growing confidence of researchers and professionals in the capabilities of LLMs but also suggests a potential acceleration in the adoption of LLMs across various domains.

Notwithstanding the pioneering role and contribution of the studies in our review, their primary focus has been proposing ways to perform existing tasks more effectively and efficiently using LLMs like ChatGPT. This pattern is consistent with the early stage of the new technology. However, as highlighted in the seminal work of David (1990), the true transformative potential of a new technology often emerges not merely through the improvement of existing tasks but through a paradigm shift that sees new applications or processes being created. Much like the incremental efficiency gains achieved through the initial use of the dynamo in factories for merely replacing steam engines, over-emphasis on enhancing existing functions with LLMs may yield smaller improvements than what these advanced models are capable of. David's analysis suggests that much greater advancements and productivity gains can be achieved when technology is used to reimagine and reengineer processes, rather than used to merely enhance existing ones. Beyond the purview of the studies in our review, the next stride in harnessing LLMs might come from a paradigm shift towards innovative use cases, which entail the creative deployment of LLMs in ways that not only redefine current practices but also unleash new possibilities.

Appendix Technical Guide

In this appendix, we provide some guidance on how to use ChatGPT as a research tool. These guidelines are based on our experience using ChatGPT, our reading of official OpenAI documentation, posts on OpenAI Developer Forum, and blog posts from ML practitioners. We also incorporate some good practices we observed from the papers covered in this survey.

Choice of Models

In Section 2.2, we provide an overview of the GPT models offered by OpenAI. These models have different capabilities and cost points. When you use the OpenAI API, you can choose a model that best suits your needs and budget.[24] When you choose a model, you should be aware of the context window to make sure that the length of the input and the intended output fits the context window. Another consideration is whether the performance of the model meets your needs. You can try out different models on the free playground.[25] Once you have chosen a model, it is a good practice to fully disclose the model used, including its series number and other details. For example, if "gpt-3.5-turbo" is specified as the model at an API call, the request will currently be routed to the "gpt-3.5-turbo-0613" model variant behind the scenes. Disclosing the full name of the exact model variant used enables better comparison and reproducibility of research findings over time.

Context Window

Context window is the amount of information that an LLM can actively consider when it generates a response. It is measured in the number of tokens. For example, the most advanced GPT-4 model has a context window of 128K tokens. However, this does not mean that users can provide a prompt of 128K tokens to the OpenAI API. This is because this window is shared between the prompt (input) and the response (output). In other words, the context window limit applies to the total length of the prompt and the response. If the length of the prompt reaches the limit of the context window, then there is no room left for the model to generate a response.

Users should also be aware that tokens are not words. Tokens can include words, punctuation, special characters, line breaks, and even word pieces. As a rule of thumb, 1000 tokens are equivalent to 750 words. OpenAI provides a tool that allows users to find out the exact number of tokens, and that number varies with models due to their different definitions of tokens. Users can intuitively see how words differ from tokens on this web interface (https://platform.openai.com/tokenizer). More conveniently, users can use the "tiktoken" tokenizer available as a Python library (https://github.com/openai/tiktoken) to count the number of tokens. Knowing the exact number of tokens is important for properly sizing the prompt to avoid exceeding the context window. This information is also useful for estimating the cost

[24] The current pricing of various models is available at https://openai.com/pricing. The pricing is based on number of tokens rather than words, and both input tokens and output tokens count as billable tokens.
[25] https://platform.openai.com/playground

based on the pricing scheme. For how to count tokens using the "tiktoken" tokenizer, see the OpenAI notebook.[26]

Newer LLMs have increasingly larger context windows. However, even the model with the largest context window currently available may not be sufficient for some tasks. For example, the most capable GPT-4 model has a context window of 128K, which is equivalent to approximately 100K words. This massive capacity may not be enough to take lengthy 10-K filings of some companies in a single pass, e.g., for a summarization task. A common workaround is to break a large document into chunks and feed one chunk to the model at a time. The outputs from these individual passes can then be aggregated into a combined output, as in Gupta (2023) and Kim, Muhn, and Nikolaev (2023a).

The chunkization approach should work well for classification tasks. It is not clear whether this approach works equally well for summarization tasks. Kim, Muhn, and Nikolaev (2023a) take this approach for summarizing MD&A and earnings conference call transcripts. They use the length of the summary relative to that of the original document to capture a construct that they refer to as "disclosure bloat." It is not clear whether chunkization may introduce bias for long documents. It is plausible that the length of the combined summaries from multiple passes is longer than the length of a single-pass summary from a model with a context window long enough.

Major Parameters

The OpenAI API offers some parameters that allow users to exercise control over the output. We encourage researchers to fully disclose their parameter settings. Such transparency can help reconcile differences in findings from studies on similar topics and improve reproducibility.

Completion length (max_tokens): The "max_tokens" parameter allows users to control the length of the output. The model will stop generating output when the length of the output reaches the maximum number of tokens set by "max_tokens". A small value of "max_tokens" may result in a truncated output, which is undesired for certain tasks, e.g., summarization and text generation. For such tasks, a preferred way to control the length of the output is to expressly tell the model the length limit in the prompt. The model will also stop when the context window limit is reached. It is important to assess how this may affect the quality of the output.

Temperature: This parameter controls the creativity (i.e., randomness) of the output generated by a GPT model. According to the official OpenAI API documentation, it ranges from 0 to 2, and is defaulted to 1. A higher temperature (e.g., 0.8) allows the model to generate more diverse output, whereas a lower temperature results in more deterministic or focused output. A temperature of zero leads to completely deterministic results. For the papers in this review, authors most often choose a temperature of zero to ensure that the output is as deterministic as possible. This choice is consistent with the nature of the research questions. For classification, summarization, and information extraction tasks, a zero temperature is recommended. There is

[26] https://github.com/openai/openai-cookbook/blob/main/examples/How_to_count_tokens_with_tiktoken.ipynb

empirical evidence suggesting that a lower temperature is preferred for annotation or classification task, because a lower temperature helps increase consistency without decreasing accuracy (Gilardi, Alizadeh, and Kubli 2023).

TOP_P: This parameter adjusts the behavior of nucleus sampling, where the model only considers next tokens within the "top_p" probability mass. This parameter ranges from 0 to 2 and is defaulted to 1. A higher value increases the pool of possible next tokens and leads to more creative and unpredictable output. Conversely, a smaller value reduces the pool of possible next tokens and yields outputs that are more predictable and less diverse. OpenAI generally recommends altering "temperature" or "top_p" from its default, but not both.

Frequency and Presence Penalty: "frequency_penalty" and "presence_penalty" impose a penalty on the next token depending on how many times or whether the token has already appeared in the output. Both parameters have a range of -2 to 2 and are defaulted to zero. A negative (positive) value encourages (discourages) repetition. A high value of "frequency_penalty" can help to avoid repetition in longer texts, and a high value of "presence_penalty" can encourage the introduction of new concepts. For summarization tasks, a slight frequency penalty could be beneficial for reducing redundancy. OpenAI generally recommends altering one, but not both of these two parameters from their defaults. These two parameters are less relevant to classification or information extraction tasks.

Logprobs: If this parameter is set to True, the model will output the log probability of each output token. The default is False. The probability is useful for determining the confidence of the model in its prediction of a classification label. For example, Bernard et al. (2023) extract the log probabilities from the initial classification tasks and use them as inputs for constructing their measure of business complexity. For classification tasks, the probabilities can also be used to create precision-recall curves for determining appropriate thresholds. For information extraction tasks, the probabilities are useful to gauge how likely the text contains the information extracted or how likely the model has made up the content. For more about how to use log probabilities, see this OpenAI notebook.[27]

Seed: This is a new parameter recently introduced by OpenAI to make the output more deterministic. It takes an integer and works like the seed of a random number generator, even though complete determinism is not guaranteed. Using the same seed and exactly the same parameters would likely produce the same output across different requests. Users can verify or monitor whether this is true by examining the "system_fingerprint" parameter in the response. We encourage researchers to use this new parameter for greater reproducibility of their work. For more about how to use this parameter for more consistent outputs, see this official guide.[28]

Look-Ahead Bias

When ChatGPT or related LLMs are used for prediction tasks, it is important to avoid look-ahead bias by being cognizant of the knowledge cut-off date of the model used. For example,

[27] https://cookbook.openai.com/examples/using_logprobs
[28] https://cookbook.openai.com/examples/deterministic_outputs_with_the_seed_parameter

GPT-4 has a knowledge cut-off date of September 2021, which means that GPT-4 was trained on data available up to that date and it thus has no knowledge of what happened thereafter. For example, Li, Tu, and Zhou (2023) assess the ability of GPT-4 to forecast future earnings of companies and limit their sample of earnings press releases to those announced on or after September 2021. This choice helps to reduce look-ahead bias. However, it may not completely avoid look-ahead bias because the management and/or financial analysts may provide long-term earnings forecasts. OpenAI continuously updates its models and periodically deprecates older iterations, which will become inaccessible typically several months later. Studies using a GPT model for a task that is sensitive to the knowledge cut-off date should be aware that the model used may not still be available for additional analyses in the later review process.

Prompt Engineering

LLMs take instructions from users in natural human language. Such instructions are known as prompts. How an instruction is framed may affect the output from the model. To obtain the desired output, users often need to try out different framings of an instruction. This gives rise to a technique known as prompt engineering, which involves crafting prompts to guide the model's generation of desired outputs. The purpose is to enhance output quality for specific tasks by influencing the model's behavior. Prompt engineering is a crucial technique for optimizing LLM performance. Yet, prompt engineering is both an art and a science, and there is no universal rule that fits all contexts. However, there are some good practices that can help users get started. Below are some useful sources for learning prompt engineering:

Best practices for prompt engineering with OpenAI API: https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

OpenAI guide on prompt engineering: https://platform.openai.com/docs/guides/prompt-engineering

Prompt examples: https://platform.openai.com/examples

Prompt engineering guide: https://github.com/dair-ai/Prompt-Engineering-Guide

Academic papers on prompt-based tuning for pre-trained LLMs: https://github.com/thunlp/PromptPapers?tab=readme-ov-file#promptpapers

Other Useful Resources

OpenAI official documentation: https://platform.openai.com/docs/overview

OpenAI official API reference: https://platform.openai.com/docs/api-reference

OpenAI codebook: https://github.com/openai/openai-cookbook/tree/main/examples

OpenAI Blog Posts: https://openai.com/blog

OpenAI Developer Forum: https://community.openai.com/

REFERENCES

Alarie, Benjamin, Kim Condon, Susan Massey, and Christopher Yan. 2023. "The Rise of Generative AI for Tax Research." SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4476510.

Alexopoulos, Michelle. 2011. "Read All about It!! What Happens Following a Technology Shock?" *American Economic Review* 101 (4): 1144–79. https://doi.org/10.1257/aer.101.4.1144.

Allen, Darcy W. E., Chris Berg, Nataliya Ilyushina, and Jason Potts. 2023. "Large Language Models Reduce Agency Costs." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4437679.

Andreou, Panayiotis C., Neophytos Lambertides, and Marina Magidou. 2023. "Stock Price Crash Risk and the Managerial Rhetoric Mechanism: Evidence from R&D Disclosure in 10-K Filings." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3891736.

Bai, John (Jianqiu), Nicole M. Boyson, Yi Cao, Miao Liu, and Chi Wan. 2023. "Executives vs. Chatbots: Unmasking Insights through Human-AI Differences in Earnings Conference Q&A." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4480056.

Bandara, Wachi, Brandon Flannery, and Anshuma Chandak. 2023. "Can AI Explain Company Performance: A Horserace." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4480665.

Bauer, Kevin, Lena Liebich, Oliver Hinz, and Michael Kosfeld. 2023. "Decoding GPT's Hidden 'Rationality' of Cooperation." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4576036.

Beerbaum, Dirk Otto. 2023. "Generative Artificial Intelligence (GAI) Ethics Taxonomy-Applying Chat GPT for Robotic Process Automation (GAI-RPA) as Business Case." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4385025.

Bernard, Darren, Elizabeth Blankespoor, Ties de Kok, and Sara Toynbee. 2023. "Confused Readers: A Modular Measure of Business Complexity." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4480309.

Bertomeu, Jeremy, Yupeng Lin, Yibin Liu, and Zhenghui Ni. 2023. "Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4452670.

Blomkvist, Magnus, Yetaotao Qiu, and Yunfei Zhao. 2023. "Automation and Stock Prices: The Case of ChatGPT." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4395339.

Bommarito, Jillian, Michael James Bommarito, Jessica Ann Mefford Katz, and Daniel Martin Katz. 2023. "Gpt as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4322372.

Boritz, J. Efrim, and Theophanis C. Stratopoulos. 2023. "AI and the Accounting Profession: Views from Industry and Academia." *Journal of Information Systems* 37 (3): 1–9. https://doi.org/10.2308/ISYS-2023-054.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Cheng, Xu (Joyce), Ryan Dunn, Travis Holt, Kerry Inger, J. Gregory Jenkins, Jefferson Jones, James H. Long, et al. 2023. "Artificial Intelligence's Capabilities, Limitations, and Impact on Accounting Education: Investigating ChatGPT's Performance on Educational Accounting Cases." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4431202.

Cheng, Yuhan, and Ke Tang. 2023. "GPT's Idea of Stock Factors." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4560216.

Comlekci, İstemi, Serkan Unal, Ali Ozer, and Mehmet Akif Oncu. 2023. "Can AI Technologies Estimate Financials Accurately? A Research on Borsa Istanbul with ChatGPT." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4545954.

Cooney, Michael. 2023. "Gartner: Top Strategic Technology Trends for 2024." Network World. October 16, 2023. https://www.networkworld.com/article/3708635/gartner-top-strategic-technology-trends-for-2024.html.

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *The American Economic Review* 80 (2): 355–61.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Rochester, NY. https://doi.org/10.2139/ssrn.4573321.

Dowling, Michael, and Brian Lucey. 2023. "ChatGPT for (Finance) Research: The Bananarama Conjecture." *Finance Research Letters* 53 (May): 103662. https://doi.org/10.1016/j.frl.2023.103662.

Economist. 2022. "Huge 'Foundation Models' Are Turbo-Charging AI Progress." *The Economist*, June 11, 2022. https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress.

———. 2023a. "Your Employer Is (Probably) Unprepared for Artificial Intelligence." *The Economist*, 2023. https://www.economist.com/finance-and-economics/2023/07/16/your-employer-is-probably-unprepared-for-artificial-intelligence.

———. 2023b. "Generative AI Will Go Mainstream in 2024." *The Economist*, November 13, 2023. https://www.economist.com/the-world-ahead/2023/11/13/generative-ai-will-go-mainstream-in-2024.

Eisfeldt, Andrea L., Gregor Schubert, and Miao Ben Zhang. 2023. "Generative AI and Firm Values." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4436627.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." arXiv.Org. March 17, 2023. https://arxiv.org/abs/2303.10130v4.

Emett, Scott A., Marc Eulerich, Egemen Lipinski, Nicolo Prien, and David A. Wood. 2023. "Leveraging ChatGPT for Enhancing the Internal Audit Process – A Real-World Example from a Large Multinational Company." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4514238.

Eulerich, Marc, Aida Sanatizadeh, Hamid Vakilzadeh, and David A. Wood. 2023. "Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA?" SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4452175.

Eulerich, Marc, and David A. Wood. 2023. “A Demonstration of How ChatGPT Can Be Used in the Internal Auditing Process.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4519583.

Feng, Zifeng, Gangqing Hu, and Bingxin Li. 2023. “Unleashing the Power of ChatGPT in Finance Research: Opportunities and Challenges.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4424979.

Fenn, Jackie, and Mark Raskino. 2008. *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*. Harvard Business Press.

Fieberg, Christian, Lars Hornuf, and David Streich. 2023. “Using GPT-4 for Financial Advice.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4499485.

Föhr, Tassilo Lars, Kai-Uwe Marten, and Marco Schreyer. 2023. “Deep Learning Meets Risk-Based Auditing: A Holistic Framework for Leveraging Foundation and Task-Specific Models in Audit Procedures.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4488271.

Föhr, Tassilo Lars, Marco Schreyer, Tatjana Alexandra Juppe, and Kai-Uwe Marten. 2023. “Assuring Sustainable Futures: Auditing Sustainability Reports Using AI Foundation Models.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4502549.

Fotoh, Lazarus, and Tatenda Mugwira. 2023. “The Use of ChatGPT in External Audits: Implications and Future Research.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4453835.

Gabaix, Xavier, Ralph S. J. Koijen, and Motohiro Yogo. 2023. “Asset Embeddings.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4507511.

Gartner. 2023. “What’s New in Artificial Intelligence From the 2023 Gartner Hype Cycle.” Gartner. August 17, 2023. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. “ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks.” *Proceedings of the National Academy of Sciences* 120 (30): e2305016120. https://doi.org/10.1073/pnas.2305016120.

Goyenko, Ruslan, and Chengyu Zhang. 2022. “Multi-(Horizon) Factor Investing with AI.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4187056.

Gu, Hanchi, Marco Schreyer, Kevin Moffitt, and Miklos A. Vasarhelyi. 2023. “Artificial Intelligence Co-Piloted Auditing.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4444763.

Gupta, Udit. 2023. “GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4568964.

Hadi, Muhammad Usman, Qasem Al-Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, et al. 2023. *Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects*. https://doi.org/10.36227/techrxiv.23589741.v1.

Haugom, Erik, Stefan Lyocsa, and Martina Halousková. 2023. “The Financial Impact of Chatgpt for the Higher Education Industry in the U.S.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4573714.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. “Measuring Massive Multitask Language Understanding.” arXiv. https://doi.org/10.48550/arXiv.2009.03300.

Hoberg, Gerard, and Gordon Phillips. 2010. “Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis.” *The Review of Financial Studies* 23 (10): 3773–3811. https://doi.org/10.1093/rfs/hhq053.

Hofert, Marius. 2023a. “Assessing ChatGPT’s Proficiency in Quantitative Risk Management.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4444104.

———. 2023b. “Correlation Pitfalls with ChatGPT: Would You Fall for Them?” *Risks* 11 (7): 115. https://doi.org/10.3390/risks11070115.

Hope, Ole-Kristian, Danqi Hu, and Hai Lu. 2016. “The Benefits of Specific Risk-Factor Disclosures.” *Review of Accounting Studies* 21 (4): 1005–45. https://doi.org/10.1007/s11142-016-9371-1.

Hu, Nan, Peng Liang, and Xu Yang. 2023. “Whetting All Your Appetites for Financial Tasks with One Meal from GPT? A Comparison of GPT, FinBERT, and Dictionaries in Evaluating Sentiment Analysis.” SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4426455.

Huang, Allen H., Hui Wang, and Yi Yang. 2023. “FinBERT: A Large Language Model for Extracting Information from Financial Text.” *Contemporary Accounting Research* 40 (2): 806–41. https://doi.org/10.1111/1911-3846.12832.

Hui, Xiang, Oren Reshef, and Luofeng Zhou. 2023. “The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4544582.

Huseynov, Samir. 2023. “ChatGPT and the Labor Market: Unraveling the Effect of AI Discussions on Students’ Earnings Expectations.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4444728.

Jain, Yash, Shubham Gupta, Serhan Yalciner, Yashodhan Nilesh Joglekar, Parth Khetan, and Tony Zhang. 2023. “Overcoming Complexity in ESG Investing: The Role of Generative AI Integration in Identifying Contextual ESG Factors.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4495647.

Jansen, Bernard J., Soon-gyo Jung, and Joni Salminen. 2023. “Employing Large Language Models in Survey Research.” *Natural Language Processing Journal* 4 (September): 100020. https://doi.org/10.1016/j.nlp.2023.100020.

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang. 2023. “ChatGPT and Corporate Policies.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4521096.

Kausik, B. N. 2023. “Long Tails & the Impact of GPT on Labor.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4525008.

Khan, Muhammad Salar, and Hamza Umer. 2023. “Chatgpt in Finance: Addressing Ethical Challenges.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4439967.

Kim, Alex G., Maximilian Muhn, and Valeri V. Nikolaev. 2023a. “Bloated Disclosures: Can ChatGPT Help Investors Process Information?” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4425527.

———. 2023b. “From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4593660.

Kok, Ties de. 2023. “Generative LLMs and Textual Analysis in Accounting: (Chat)GPT as Research Assistant?” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4429658.

Korinek, Anton. 2023. “Generative AI for Economic Research: Use Cases and Implications for Economists.” *Journal of Economic Literature* 61 (4): 1281–1317. https://doi.org/10.1257/jel.20231736.

Krause, David. 2023a. “ChatGPT and Other AI Models as a Due Diligence Tool: Benefits and Limitations for Private Firm Investment Analysis.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4416159.

———. 2023b. “ChatGPT and Generative AI: The New Barbarians at the Gate.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4447526.

———. 2023c. “Proper Generative AI Prompting for Financial Analysis.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4453664.

———. 2023d. “A Rumsfeldian Framework for Understanding How to Employ Generative AI Models for Financial Analysis.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4455916.

Kuroki, Yutaka, Tomonori Manabe, and Kei Nakagawa. 2023. “Fact or Opinion? – Essential Value for Financial Results Briefing.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4430511.

Lee, Maggie C. M., Helana Scheepers, Ariel K. H. Lui, and Eric W. T. Ngai. 2023. “The Implementation of Artificial Intelligence in Organizations: A Systematic Literature Review.” *Information & Management* 60 (5): 103816. https://doi.org/10.1016/j.im.2023.103816.

Leippold, Markus. 2023. “Sentiment Spin: Attacking Financial Sentiment with GPT-3.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4337182.

Li, Edward Xuejun, Zhiyuan Tu, and Dexin Zhou. 2023. “The Promise and Peril of Generative AI: Evidence from ChatGPT as Sell-Side Analysts.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4480947.

Li, Tong, Qilin Peng, and Luping Yu. 2023. “ESG Considerations in Acquisitions and Divestitures: Corporate Responses to Mandatory ESG Disclosure.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4376676.

Li, Yang, Thierry Marier-Bienvenue, Alexis Perron-Brault, Xinyi Wang, and Guy Paré. 2018. “Blockchain Technology in Business Organizations: A Scoping Review.” In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 4474–83. Hawaii. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/c7c562e3-0a26-4c0a-bc14-e767cf458d7f/content.

Li, Yinheng, Shaofei Wang, Han Ding, and Hang Chen. 2023. “Large Language Models in Finance: A Survey.” arXiv. https://doi.org/10.48550/arXiv.2311.10723.

Liu, Jin, Xingchen Xu, Yongjun Li, and Yong Tan. 2023. “‘Generate’ the Future of Work through AI: Empirical Evidence from Online Labor Markets.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4529739.

Liu, Yang, Laura K. Miller, and Xu Niu. 2023. “Incorporating ChatGPT into a Financial Data Science Course with Python Programming.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4412371.

Loughran, Tim, and Bill McDonald. 2011. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” *The Journal of Finance* 66 (1): 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Lu, Fangzhou, Lei Huang, and Sixuan Li. 2023. “ChatGPT, Generative AI, and Investment Advisory.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4519182.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. “Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey.” *ACM Computing Surveys* 56 (2): 30:1-30:40. https://doi.org/10.1145/3605943.

Nakano, Masafumi, and Takuya Yamaoka. 2023. “Enhancing Sentiment Analysis Based Investment by Large Language Models in Japanese Stock Market.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4511658.

Ni, Jingwei, Julia Bingler, Chiara Colesanti Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, et al. 2023. “chatReport: Democratizing Sustainability Disclosure Analysis through LLM-Based Tools.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4476733.

Niszczota, Paweł, and Sami Abbas. 2023. “GPT Has Become Financially Literate: Insights from Financial Literacy Tests of GPT and a Preliminary Test of How People Use It as a Source of Advice.” *Finance Research Letters* 58 (December): 104333. https://doi.org/10.1016/j.frl.2023.104333.

O’Leary, Daniel E. 2008. “Gartner’s Hype Cycle and Information System Research Issues.” *International Journal of Accounting Information Systems* 9 (4): 240–52. https://doi.org/10.1016/j.accinf.2008.09.001.

———. 2009. “The Impact of Gartner’s Maturity Curve, Adoption Curve, Strategic Technologies on Information Systems Research, with Applications to Artificial Intelligence, ERP, BPM, and RFID.” *Journal of Emerging Technologies in Accounting* 6 (January): 45–66. https://doi.org/10.2308/jeta.2009.6.1.45.

O’Leary, Daniel E. 2022. “Massive Data Language Models and Conversational Artificial Intelligence: Emerging Issues.” *Intelligent Systems in Accounting, Finance and Management* 29 (3): 182–98. https://doi.org/10.1002/isaf.1522.

O’Leary, Daniel E. 2023a. “An Analysis of Three Chatbots: BlenderBot, ChatGPT and LaMDA.” *Intelligent Systems in Accounting, Finance and Management* 30 (1): 41–54. https://doi.org/10.1002/isaf.1531.

———. 2023b. “Enterprise Large Language Models: Knowledge Characteristics, Risks, and Organizational Activities.” *Intelligent Systems in Accounting, Finance and Management* 30 (3): 113–19. https://doi.org/10.1002/isaf.1541.

Paré, Guy, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. “Synthesizing Information Systems Knowledge: A Typology of Literature Reviews.” *Information & Management* 52 (2): 183–99. https://doi.org/10.1016/j.im.2014.08.008.

Pietrzak, Marcin. 2023. “A Trillion Dollars Race – How Chatgpt Affects Stock Prices.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4586428.

Ray, Partha Pratim. 2023. “ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope.” *Internet of Things and Cyber-Physical Systems* 3 (January): 121–54. https://doi.org/10.1016/j.iotcps.2023.04.003.

Romanko, Oleksandr, Akhilesh Narayan, and Roy Kwon. 2023. “ChatGPT-Based Investment Portfolio Selection.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4538502.

Siddik, Abu Bakkar, Yong Li, Arshian Sharif, and Javier Cifuentes-Faura. 2023. “The Role of Artificial Intelligence and Chatgpt in Fintech: Prospects, Challenges, and Research Agendas.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4439965.

Singh, Harjit, and Avneet Singh. 2023. “ChatGPT: Systematic Review, Applications, and Agenda for Multidisciplinary Research.” *Journal of Chinese Economic and Business Studies* 21 (2): 193–212. https://doi.org/10.1080/14765284.2023.2210482.
Snyder, Hannah. 2019. “Literature Review as a Research Methodology: An Overview and Guidelines.” *Journal of Business Research* 104: 333–39. https://doi.org/10.1016/j.jbusres.2019.07.039.
Stratopoulos, Theophanis C. 2018. “Business Value of Information Technology - A Data Analytics Approach.” SSRN Scholarly Paper ID 3186759. Rochester, NY: Social Science Research Network. https://dx.doi.org/10.2139/ssrn.3186759.
Stratopoulos, Theophanis C., and Victor Xiaoqi Wang. 2022. “Estimating the Duration of Competitive Advantage from Emerging Technology Adoption.” *International Journal of Accounting Information Systems* 47 (December): 100577. https://doi.org/10.1016/j.accinf.2022.100577.
Stratopoulos, Theophanis C., Victor Xiaoqi Wang, and Hua (Jonathan) Ye. 2022. “Use of Corporate Disclosures to Identify the Stage of Blockchain Adoption.” *Accounting Horizons* 36 (1): 197–220. https://doi.org/10.2308/HORIZONS-19-101.
Street, Daniel, and Joseph Wilck. 2023a. “Let’s Have a Chat: Principles for the Effective Application of ChatGPT and Large Language Models in the Practice of Forensic Accounting.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4351817.
———. 2023b. “‘Let’s Have a Chat’: Principles for the Effective Application of ChatGPT and Large Language Models in the Practice of Forensic Accounting.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4351817.
Street, Daniel, Joseph Wilck, and Zachariah Chism. 2023. “Six Principles for the Effective Use of ChatGPT and Other Large Language Models in Accounting.” SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4551289.
Tsuchihashi, Toshihiro. 2023. “Do Ais Dream of Homo Economicus? Answers from Chatgpt.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4498882.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” arXiv. https://doi.org/10.48550/arXiv.1706.03762.
Villiers, Charl de, Ruth Dimes, and Matteo Molinari. 2023. “How Will AI Text Generation and Processing Impact Sustainability Reporting? Critical Analysis, a Conceptual Framework and Avenues for Future Research.” *Sustainability Accounting, Management and Policy Journal* ahead-of-print (ahead-of-print). https://doi.org/10.1108/SAMPJ-02-2023-0097.
Wang, Chen. 2023. “Can ChatGPT Personalize Index Funds’ Voting Decisions?” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4413315.
Wang, Yanqing. 2023. “Generative AI in Operational Risk Management: Harnessing the Future of Finance.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4452504.
Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” arXiv. https://doi.org/10.48550/arXiv.2201.11903.
Wei, Tian, Han Wu, and Gang Chu. 2023. “Is ChatGPT Competent? Heterogeneity in the Cognitive Schemas of Financial Auditors and Robots.” *International Review of*

Economics & Finance 88 (November): 1389–96. https://doi.org/10.1016/j.iref.2023.07.108.
Wikipedia. 2023. "Uncanny Valley." In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Uncanny_valley&oldid=1180349546.
Wolfram, Stephen. 2023. "What Is ChatGPT Doing … and Why Does It Work?" Stephen Wolfram Writings. February 14, 2023. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.
Wood, David A., Muskan P. Achhpilia, Mollie T. Adams, Sanaz Aghazadeh, Kazeem Akinyele, Mfon Akpan, Kristian D. Allee, et al. 2023. "The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions?" *Issues in Accounting Education*, April, 1–28. https://doi.org/10.2308/ISSUES-2023-013.
Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. "BloombergGPT: A Large Language Model for Finance." arXiv. https://doi.org/10.48550/arXiv.2303.17564.
Yang, Changyu, and Adam Stivers. 2023. "Investigating AI Languages' Ability to Solve Undergraduate Finance Problems." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4460814.
Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang. 2023. "FinGPT: Open-Source Financial Large Language Models." arXiv. https://doi.org/10.48550/arXiv.2306.06031.
Yang, Stephen. 2023. "Predictive Patentomics: Forecasting Innovation Success and Valuation with ChatGPT." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4482536.
Zaremba, Adam, and Ender Demir. 2023. "ChatGPT: Unlocking the Future of NLP in Finance." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4323643.
Zhang, Boyu, Hongyang Yang, and Xiao-Yang Liu. 2023. "Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4489831.
Zhang, Christopher L. 2023. "Feel the Market: An Attempt to Identify Additional Factor in the Capital Asset Pricing Model (CAPM) Using Generative Pre-Trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT)." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4521946.
Zhang, Libin. 2023. "Four Tax Questions for ChatGPT and Other Language Models." SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4458628.
Zhao, Joanna (Jingwen), and Xinruo Wang. 2023. "Unleashing Efficiency and Insights: Exploring the Potential Applications and Challenges of ChatGPT in Accounting." *Journal of Corporate Accounting & Finance*. https://doi.org/10.1002/jcaf.22663.
Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. "A Survey of Large Language Models." arXiv. https://doi.org/10.48550/arXiv.2303.18223.