AI and Finance*

Andrea L. Eisfeldt & Gregor Schubert

October 15, 2024

Abstract

We provide evidence that the development and adoption of Generative AI is driving a significant technological shift for firms and for financial research. We review the literature on the impact of ChatGPT on firm value and provide directions for future research investigating the impact of this major technology shock. Finally, we review and describe innovations in research methods linked to improvements in AI tools, along with their applications. We offer a practical introduction to available tools and advice for researchers interested in using these tools.

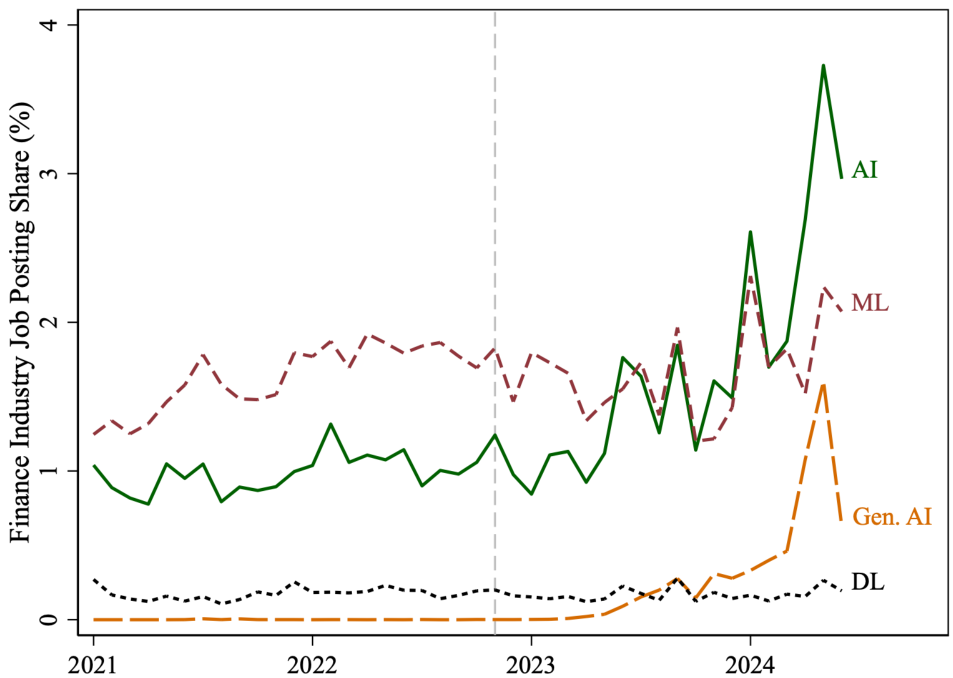
I. Introduction

Generative AI represents a major technology shock to firms and to finance research. In the financial sector, recent advancements, particularly in Generative AI and large language models (LLMs), have sparked a rapid increase in demand for related skills. Figure 1 shows the share of monthly job postings in the finance and insurance sector that mention particular technical skills. Following the release of ChatGPT in November 2022 (denoted by the vertical grey line), demand for "artificial intelligence" skills broadly tripled by mid-2024. Moreover, demand for skills specifically linked to Generative AI or LLMs rose from zero prior to ChatGPT to around 1% of all job postings.

In this review, we focus on Generative AI both as a topic of study for researchers in financial economics, and as a methodological tool for conducting finance research. We focus specifically on recent innovations in large language models and related deep learning techniques, rather than on the broader set of tools that is oftentimes grouped under the umbrella of "artificial intelligence", including machine learning, neural nets, machine vision, robotics etc., which were adopted in research and practical settings throughout the 2010s. We discuss the impact of these tools on firm value and firm decisions, and on *research* in finance. Our goals include both describing and advancing the finance research frontier, and offering practical tips for how generative AI can be used to improve asset management and

*The authors thank Miao Ben Zhang, Keith Chen, Daniel Rock, and Barney Hartman-Glaser, as well as ChatGPT, NotebookLM, Claude, and Perplexity.AI for helpful discussions and research assistance.

Figure 1. Skill demand in the finance and insurance sector. The graph shows the share of online job postings by firms in the finance and insurance sector (NAICS 52) that mention particular skills. "AI" are artificial intelligence skills, "ML" are machine learning skills. "Gen. AI" are Generative AI or large language model skills. "DL" are deep learning skills. The grey drop-down line indicates the ChatGPT release date. Data source: Lightcast.



corporate finance decisions. For a recent study that summarizes some of the emerging use cases within financial firms, see Aldasoro et al. (2024).[1]

Specifically we address the following two broad topics: (1) Generative AI as a technology shock that affects firm values and corporate policies, and (2) Generative AI as a technology shock to methods for financial research. For (1), in addition to reviewing the existing literature, we highlight fertile areas for future research. For (2), we provide practical guidance for researchers looking to add Generative AI to their research toolbox and discuss best practices for using Generative AI research methods.

II. Generative AI: Technology shock to firms

The release of ChatGPT in November 2022 represents a large technology shock that affected firms across all industries. The impact on firm values was generally large and positive, but changes in valuation exhibited substantial cross-sectional variation. In addition to the impact on value from investors' expectations of changes in firms' prospects for growing future free cash flow, as a major technology shock we expect Generative AI to drive future changes in corporate policies. Indeed, firm value is expected to change both due to variation in exposure to the technology shock, and due to changes in firms' decisions conditional on the rapidly advancing AI technology frontier. Existing research has only scratched the surface of the ways in which shifts in production processes, changes in productivity, and uncertainty about further innovation is changing firm behavior.

Despite the relative recency of the Generative AI technology shock, one area in which substantial research progress has been made is on the impact of Generative AI on labor, firm hiring decisions, wages, and ultimately on firm value. In this section we first review the findings in Eisfeldt et al. (2023) on the substantial measured impact of Generative AI. Next, we discuss fruitful directions for future work, which include understanding how innovations in Generative AI are likely to affect other corporate policies, such as capital structure and capital investment.

A. *Measuring Generative AI exposure*

How do we know which firms are affected by a technology shock? One approach to determine exposure to a technology is to use surveys, patent data, or product information by firms to evaluate the *current* use of a technology and infer the productivity potential from the revealed preferences of which firms have deployed the technology. However, for a technology that is rapidly improving, as has been the case for Generative AI, there is bound to be a large gap between the foreseeable productivity potential – which can be priced by financial markets now even if it will only be realized in the future – and current adoption. In those cases, researchers have to rely on current firm characteristics as proxies that are likely to capture the technological potential.

[1]For the the broader effects of data on firms and the economy, see Veldkamp and Chung (2024). For use of machine learning in asset pricing Nagel (2021), Giglio et al. (2022), and Kelly et al. (2023) provide recent perspectives.

An example of such an alternative measurement approach is provided in Eisfeldt et al. (2023). In that work, we build on a study by Eloundou et al. (2023) that assesses task-level exposures of different occupations to the capabilities of LLMs, and measure a firm's exposure by combining an occupation-level score of a firm's Generative AI exposure with information on each firm's employment structure based on LinkedIn profiles. This approach uses data on the task content of different occupations provided by the public O*NET database.[2] We deploy an LLM-powered classification algorithm that assigns each task to a rubric based on its description that distinguishes the likely productivity impacts from using a GPT 4-level LLM to complete the task. The resulting scores reflect three different levels of exposure: (1) Direct Exposure: access to a ChatGPT-like LLM directly reduces task completion time by $\geq$50%; (2) Indirect Exposure: when combined with additional software and tools, the LLM could reduce task completion time by $\geq$50%; (3) No Exposure: access to an LLM does not significantly reduce completion time without significantly impacting execution quality. The approach allows for rapid classification of 19,265 task statements and provides insights into AI's impact on various occupations. Through this method, we found that 14% of occupations tasks are directly exposed, an additional 22% are likely to be exposed when LLMs are combined with appropriate tools, and 64% of tasks are not exposed.

To illustrate how this approach categorizes different tasks, Table I shows some of the tasks that are done by "Loan Officers" (SOC code 13-2072), and the exposure assigned to them: LLMs can perform simple calculations based on existing patterns, or write code tfor more complex calculations, which enables them to "compute payment schedules" for a loan officer directly in the chat window without access to further tools. Similarly, they can fill in generic forms, write business emails, and respond to simple questions by loan applicants through their ability to quickly generate texts in many desired formats based on input data. For tasks that require access to transaction-specific information about the applicant or the lender's product offerings, the LLM would have to be given access to internal databases – for instance through a Retrieval Augmented Generation (RAG) system like the ones described in a later section. Once provided the right access and some additional structure, e.g. through splitting a task into different components with corresponding prompts, state-of-the-art LLMs can be expected to also increase the productivity of tasks like determining loan eligibility or explaining loan options to customers. However, tasks that require decision-making authority, such as giving final loan approval, or a physical body, such as meeting with loan applicants, are unlikely to benefit from LLMs at current levels of capabilities.

This example shows that Generative AI exposure can vary across tasks within occupations, with some tasks becoming easier to do and others being unaffected. As we argue in Eisfeldt et al. (2023), because different tasks are more or less important to the worker's duties at a firm, workers with the same overall share of tasks exposed to the technology can experience very different consequences, depending on which tasks are impacted: If a worker's core tasks (defined by O*NET)[3] can be partially or fully automated, a firm can likely restructure or eliminate some of the associated positions, leading to negative employment outcomes for affected occupations. If mostly "supplemental" (non-core) activities are automated, employees are less likely to be made redundant, and might even increase their value to the

[2] This data can be accessed via the O*NET website at https://www.onetonline.org.

[3] See details at https://www.onetonline.org/help/online/scales#score.

firm if they can reallocate some of their freed-up time from less essential activities to more productive pursuits.

Table I
Examples of Gen. AI Exposure Scores for Loan Officer Tasks

This table presents examples of the scoring of loan officer tasks' exposure to Generative AI by comparing the tasks' statement to a rubric using the GPT 3.5 Turbo Model. We obtain from the O*NET V27.2 database task statements for each occupation, which are then submitted for scoring, and the table shows the explanation provided by the model for the assigned exposure type.

Task	Exposure	Model Explanation
Compute payment schedules.	Direct	The model can perform mathematical calculations and can learn the rules for computing payment schedules.
Contact applicants or creditors to resolve questions about applications or to assist with completion of paperwork.	Direct	The model can assist with writing and responding to emails, including those that involve refuting information or engaging in a negotiation, which is a part of the task described.
Analyze applicants' financial status, credit, and property evaluations to determine feasibility of granting loans.	Indirect	The model can be used to analyze financial data and credit reports, but it may not be able to retrieve up-to-date information from the internet or search an organization's existing knowledge. Additional software could be developed to help with these tasks.
Explain to customers the different types of loans and credit options that are available, as well as the terms of those services.	Indirect	The model can be used to retrieve up-to-date information about loans and credit options, and can be used to generate explanations of those options.
Approve loans within specified limits, and refer loan applications outside those limits to management for approval.	None	The task involves making a final decision based on collected inputs, which cannot be done by the LLM.
Meet with applicants to obtain information for loan applications and to answer questions about the process.	None	The task requires human interaction and communication skills, which cannot be fully replaced by a language model.

While Eisfeldt et al. (2023) focuses on Generative AI exposure, other researchers developed measures for firm exposure to the previous waves of AI-related innovation. For example, Babina et al. (2024) develop a measure of firm-level AI investments based on identifying AI-related skills and their prevalence in different companies during the 2007-2018 period from detailed worker resumes and job postings.

B. *Generative AI exposure across professions*

For finance researchers (in industry or academia) studying the impact of Generative AI it may be important to know how the exposure to this technology varies across different occupations, and how finance-related jobs in particular are likely to be affected.

Figure 2 shows some of these descriptive patterns: Panel A shows that the average Gen. AI exposure across all occupations (weighting them by their 2022 employment) is 0.27. The figure also provides details on selected white collar occupation groups: health care occupations have below average exposure at 0.18, while finance occupations almost double the mean exposure. Moreover, due to LLMs' excellent ability to code, computer-related occupations are among the most exposed with an exposure of 0.62. Managerial occupations outside of finance are above-average exposed – but less than finance and computer occupations. The degree to which this exposure comes from core tasks varies as well, with a higher share of the exposure in computer occupations being driven by core tasks, while managers tend to be more likely to have their modest exposure come from supplemental activities.

Finance occupation exposure. The higher average exposure among finance occupations[4] aligns with the general pattern that higher-wage occupations are more likely to be exposed to Generative AI: average annual wages among the finance occupations were $108K in 2022, while the U.S. average was $62K. However, there is variation in how much particular jobs *within* finance are likely to be exposed: Panel B of Figure 2 shows the breakdown for the largest finance-related occupations (employment > 100K). While overall Gen. AI exposure varies only modestly across these jobs, focusing on core exposure paints a more varied picture: based on this measure, financial managers, loan officers and financial advisers are less likely to be impacted than accountants and auditors, or insurance underwriters. That is, positions where interpersonal skills play a more important role are less likely to have the analytical abilities of LLMs be relevant in their most fundamental duties.

How does higher exposure for occupations within finance relate to other characteristics of the affected jobs? Figure 3 plots Generative AI exposure against occupational wages in panel A, and in panel B plots exposure against a measure of the occupation's reliance on non-routine interpersonal skills. Panel A shows that higher wage finance occupations are *less* likely to be exposed to Generative AI productivity impacts. This finding of a negative relationship between exposure and wages within finance occupations is in stark contrast to the positive relationship between wages and exposure in the economy as a whole. This opposing pattern can be explained by noting that almost all the finance-specific occupations shown here already require a high level of analytical skill. Eisfeldt et al. (2023) showed that a need for analytical skill is one of the key predictors of occupational exposure to Generative AI. Panel B shows that higher exposure is associated with less occupational need for non-routine interpersonal skills. Examples of non-routine interpersonal skills include coaching or directing subordinates, and managing personal relationships. Because those interpersonal skills are less likely to be automated by an LLM or similar technology, they are associated with lower exposure to Generative AI. Panel B also shows that some of the finance occupations that are shown in panel A to have Generative AI exposure that is lower than expected based on their wage levels (e.g. Financial and Investment Analysts and Credit Counselors) involve high levels of non-routine interpersonal skills.

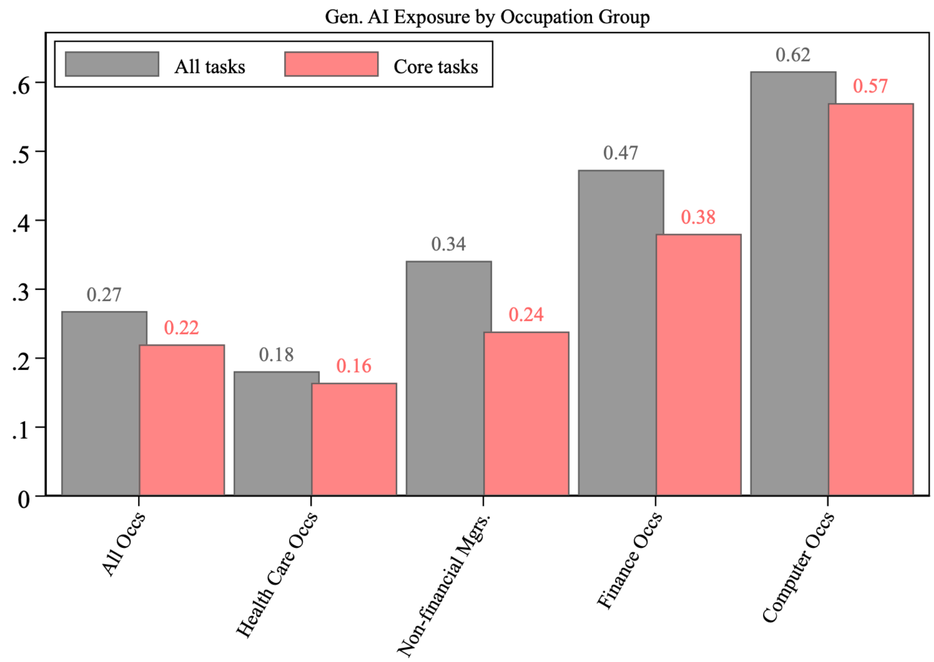
C. Effects of Generative AI on firms

The release and rapid subsequent rise to prominence of Generative AI starting in November 2022 provide an interesting natural experiment for studying how financial markets react to this type of technology shock.

One important research question arising from the event is to what degree financial markets are forecasting that the productivity potential of the new technology will be realized—and what firms are perceived as more likely to benefit. Quantifying this market reaction

[4]We defined finance occupations here as all occupations within the "Financial Specialists" minor occupation group, based on SOC codes, and also include Financial Managers (SOC 11-3031).

Panel A: Share of Tasks Exposed to Generative AI Across Selected Occupation Groups



Panel B: Share of Tasks Exposed to Generative AI in Finance Occupations

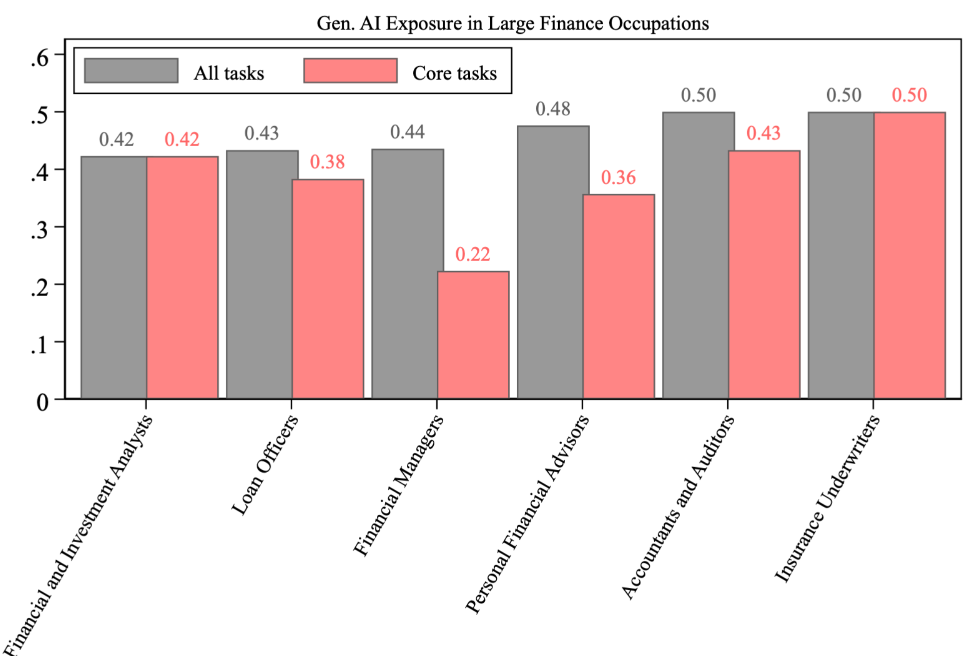
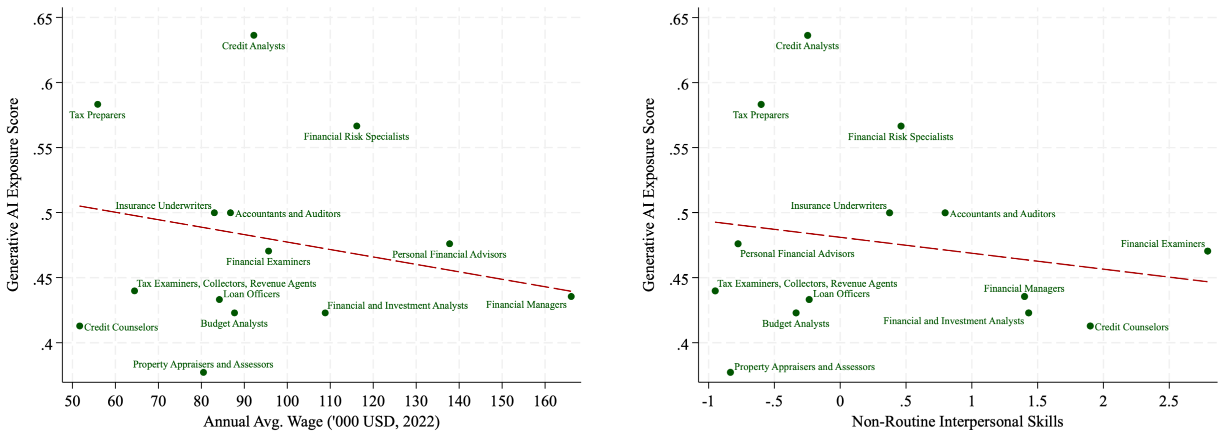


Figure 2. Generative AI Exposure by Occupation. Panel A plots the weighted share of tasks exposed to Generative AI in different white-collar occupation groups. Direct exposure is given a weight of 1, indirect gets a weight of 0.5, and no exposure is scored as 0. Grey bars show the exposure coming from all tasks and red bars show the contribution to the total coming from core tasks. Panel B plots the same data for key large occupations within the financial sector.



(A) **(B)**

Figure 3. **Finance occupation Gen. AI exposure, wages, and social skills.** The figures show the relationship between the exposure to Generative AI of finance occupations and their 2022 wages from the BLS (left panel), as well as the interpersonal skills involved in the occupation (right panel). "Social skills" are proxied by the measure of non-routine cognitive interpersonal skills defined in Acemoglu and Autor (2011) based on O*NET data. Linear fits are weighted by employment.

allows researchers to learn both about the expected magnitude of the value being created, as perceived by sophisticated market participants, and also to learn about barriers or enablers of technology implementation that equity analysts have identified.

In this spirit, in Eisfeldt et al. (2023) we examine the release of ChatGPT on November 30th, 2022, which brought major attention to the potential of the technology, as a catalyzing event that led to a revaluation of the impacted companies based on their productivity potential.

We sorted firms into five value-weighted portfolios based on their Generative AI exposures. Our analysis shows that firms in the highest-exposure quintile, labeled the "Artificial" portfolio, earned 44 basis points higher daily returns than firms in the lowest-exposure quintile, labeled the "Human" portfolio, during the two weeks following ChatGPT's release. This two-week event window was chosen based on the intensity of public attention to the ChatGPT release on Twitter. Figure 4 illustrates the cumulative abnormal returns before and after the event for a zero-investment portfolio that is long on artificial stocks and short on human stocks, which we call the "Artificial Minus Human" portfolio (AMH).

How large are these estimated effects on firm value relative to the expected potential labor value generated by Generative AI? We can do a rough back-of-the-envelope calculation that assumes that the task-level exposure score represents the share of each occupation's labor product that could eventually be made 50% more productive as a result of GPT 4-level LLM availability. Then, assuming that wages represent each worker's marginal product and that the productivity effects only apply to the marginal output of each worker, we can compute Employment $\times$ Annual Wage $\times$ 50% $\times$ Exposure Share as a rough proxy for the value that the new technology might have been expected to create. Of course, there are many margins of adaptation and dynamic effects that are not taken into account here. The result is shown in Figure 5: across all occupations we would expect around \$1.4T in annual value created from

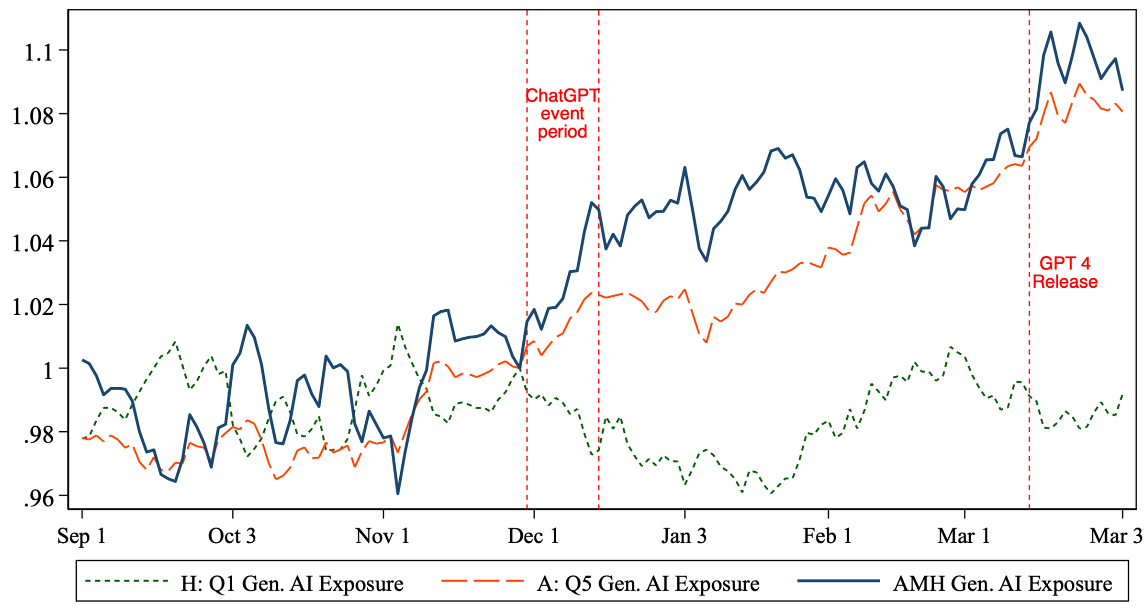


Figure 4. Cumulative Abnormal Returns by Generative AI Exposure This figure, reproduced from Eisfeldt et al. (2023), plots the cumulative abnormal returns (CARs) of value-weighted quintile portfolios sorted by firms' labor-based Generative AI exposure. The graph shows the CARs of the lowest-exposure quintile portfolio, "Human" (H), the highest-exposure quintile portfolio, "Artificial" (A), and the zero investment portfolio that goes long A and shorts H, "Artificial-minus-Human" (AMH). Market-adjusted daily abnormal returns are cumulated from November 29, 2022, the day before the release of ChatGPT, and are based on factor exposures computed over the 4-month period preceding the period shown in the graph. Daily stock returns are from Yahoo Finance. The dashed vertical lines indicate the "ChatGPT event period" from November 30, 2022, to December 14, 2022, and the release of GPT-4 on March 14, 2023.

all tasks, and \$1.1T just from core tasks. In financial occupations alone, the corresponding numbers are \$91B and \$70B, which represents an outsized impact relative to the 2.5% of total U.S. employment that the 3.6M employees in financial occupations represent.[5] For comparison, the total market capitalization as of Nov. 29, 2022, of the "Artificial" portfolio companies was around \$10.2T. While not all the labor value would accrue to this group of companies, it is plausible that the *capitalized* value of even a fraction of the \$1.4T in annual labor value could generate the 4-5 pp increase in relative value of the "Artificial" portfolio that we estimated.

While our original research focused on the release of ChatGPT, this segmentation of stocks into portfolios based on their labor exposure to the new technology is likely to also reflect differential valuation impacts of later updates to the technology. In particular, the most advanced generation of LLMs (as of this writing) was led by the public release of GPT 4 on March 14, 2023. As the figure shows, the two weeks before (its release was anticipated) and after the GPT 4 launch coincide with another run-up in the returns to the AMH portfolio.

[5]Total U.S. employment was 148M in 2022 according to the BLS Occupational Employment Statistics.

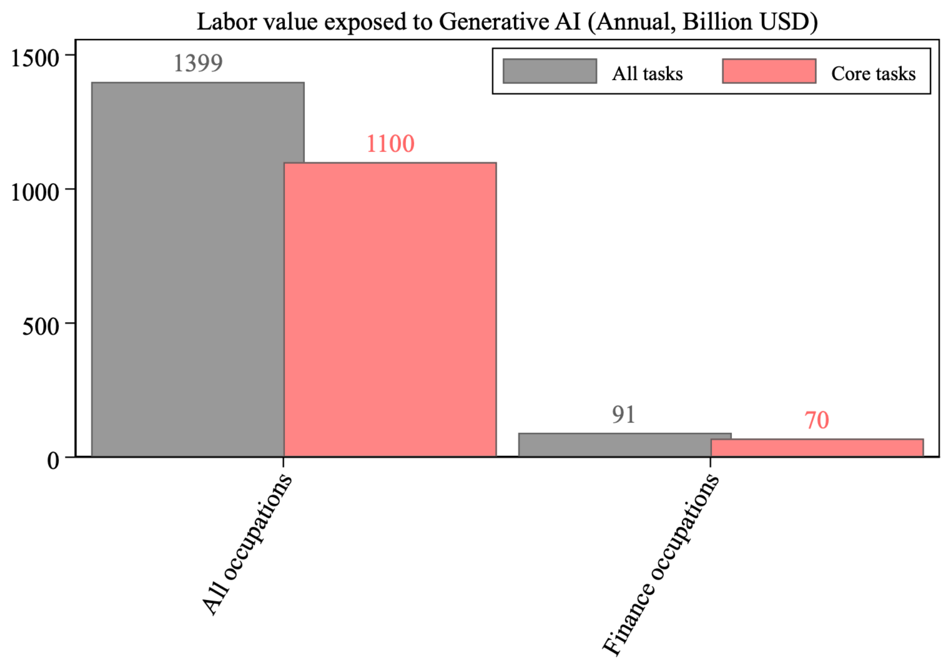


Figure 5. Back-of-the-envelope labor value potential of Generative AI. The figure shows the results of a back-of-the-envelope calculation of the potential labor value that could have been expected to be created around the release of ChatGPT. The calculation uses 2022 BLS data on wages and employment by occupation, and assumes that the task-level exposure score represents the share of each occupation's labor product that could eventually be made 50% more productive as a result of GPT 4-level LLM availability. The annual labor value potential is then computed as Employment $\times$ Annual Wage $\times$ 50% $\times$ Exposure Share

D. Directions for Future Work

In connection to the work on the impact of Generative AI on labor, there are several interesting directions for future research. First, existing research focuses on the relationship between occupations and Generative AI. An interesting question is how Generative AI relates to employees' rank within occupations, since some argue that the first IT revolution of the 1990's had the effect of "hollowing out" middle management. Second, within occupations, access to Generative AI could have two effects on wages. Access to Generative AI could lead to a leveling out effect, reducing wage dispersion by reducing effects of inherent skill differentials. Conversely, these new tools could amplify the effects of skill differentials and exacerbate superstar effects (Rosen (1981)). Finally, existing research takes the set of tasks within occupations as given. New technologies are likely to change occupational tasks and future research can help to understand the resulting changes in what workers do within their jobs.

In addition to its effect on labor markets and firm values, a crucial question for understanding the growth of Generative AI in firms is how such growth will be financed. Indeed, the commercial success of AI relies on intangible assets like software, data, and key organizational talent to drive the Generative AI engine at firms. Prior work on capital structure with intangibles, including Sun and Xiaolan (2019) and Falato et al. (2022) shows that intangible and human-related assets typically have more equity and less debt financing. See also Rampini and Viswanathan (2013) for a general theory of collateral and capital structure, and Graham and Leary (2011) for a broad review of the empirical literature on capital structure.

Generative AI is also expected to change firms' investment decisions. Veldkamp and Chung (2024) emphasize the role of data in forecasting, and, combined with such data, Generative AI has the potential to improve capital budgeting decisions by reducing uncertainty. Crouzet and Eberly (2023) provides a framework which can be used to understand how Generative AI-related intangibles and potential rents from any coincident market power can affect investment incentives. The make-or-buy decision, and how that decision might vary across industries and firm sizes, as well as the ultimate impact of those decisions, is an interesting area for new research. There are large fixed costs to building AI models, but lower costs to adopting externally generated models. One example of how much access to external models matters is provided by Bertomeu et al. (2023), who consider what happens to firms' valuation when *public* access to Generative AI technology is rescinded after it has already been partially adopted. They focus on Italy's decision to ban access to the ChatGPT platform as one popular access point to LLM capabilities, thereby limiting users to open-access models hosted on their own servers. They find that more exposed Italian firms – with exposure measured using a similar methodology to that in Eisfeldt et al. (2023) – saw large drops in market value after the ban, with small and newly established firms being especially negatively impacted.

The investment boom in Generative AI capabilities is the focus of many news articles about the largest US firms (dubbed the 'Magnificent Seven"). One o of them is AI chipmaker Nvidia, which joined the S&P 500 in June of 2024 and is investing heavily to keep up with demand. At the same time, due to its revolutionary nature, Generative AI has created uncertainty and, as a result, some firms may be postponing investment. In contrast to the popular press, academic researchers have not yet systematically documented or modeled the impact of Generative AI on corporate investment.

The Generative AI revolution is also increasing energy needs, and has implications for investment in energy production and for climate change considerations. For example, Microsoft recently signed a deal which would reopen the Three Mile Island nuclear plant. Modeling the effects of, and potential bottlenecks to, Generative AI investment's increase in energy demand will likely be an important issue going forward, to which economists should contribute their expertise in thinking about supply chain connections and interactions between firms.

Finally, from an asset pricing perspective, major technology shocks are expected to change the composition of the economy and to thereby change what constitutes market risk. Cochrane et al. (2008) provides a model of expected returns in a framework with two sectors that change in size over time. Consistent with these ideas, Babina et al. (2023) provides evidence suggesting that firms that have hired more AI talent have experienced increases in systematic risk over time. These ideas are very topical as they are related to trends in equity market concentration and the relative dominance of the so-called "Magnificent Seven".

III. Generative AI: Technology shock to research

At the same time as these new technologies are becoming an increasingly important tool for finance companies and the economy as a whole, they are also starting to be used by researchers in finance and economicsin creative and exciting ways. Generative AI, large language models, and other deep learning methods can both lower the time and monetary cost of existing research designs in finance and enable new types of analyses.

An early impact of the Generative AI technology shock is that some existing activities that are part of academic research are likely to become more productive (Korinek, 2023): programming to clean data or conduct statistical analyses can be done in less time and with faster iteration cycles by using LLMs for code generation and debugging; classification of text data can be done faster and usually at a much lower cost, using deep learning techniques or LLMs rather than human labelers (see Dell (2024) for detailed advice on when and how to use these methods, for example to construct economic measures from historical documents); LLMs can assist in the drafting of text, and provide near-instant proofreading services for papers and presentations alike – and there are many other applications.

Perhaps more revolutionary effects will come from the fact that the technological innovation of Generative AI enables novel approaches to research: LLMs can be used to generate new research ideas and hypotheses (perhaps by listing plausible alternative explanations); generative models can be used to simulate survey participants and test research designs before deployment; and large-scale deployment of Generative AI models enables the analysis of qualitative data at a scale that would previously have been prohibitively expensive for academic researchers. As a result, we expect a wave of novel and impactful new research.

Overall, we are optimistic that, for academic researchers, these new tools are more likely to automate—or increase the productivity of—what we label as "supplemental" tasks in Eisfeldt et al. (2023). As opposed to directly contributing to the core objectives of the role, supplemental tasks are less critical to a job's primary function but can still increase occupational output. As a result, using Generative AI for supplemental tasks can help to free up more time for researchers to focus on their core tasks, such as research design, mentoring, teaching, and communication.

However, as with all new research methods, deploying Generative AI and LLMs is not without pitfalls: many of the applications listed below have a lower upfront cost in terms of engineering and technical skill development as a result of the natural language interface with which many of these new technologies can be deployed. But this greater ease of access often comes at the cost of less control over the output. Ascertaining whether a particular Generative AI-based research method achieves results of acceptable accuracy still requires careful validation of its results against "ground truth" data sets, and an exploration of the sensitivity of results to different design choices (e.g. prompt engineering). Ultimately, these powerful new models cannot be treated as "magical black boxes" and do not obviate the need for researchers to steer and carefully supervise their analyses. This is particularly true because many of the applications (some of which are discussed below) are often using LLMs for a particular purpose for the first time, and thus there is yet little knowledge about how reliable these methods are in those contexts.

We highlight some of the important issues to consider in writing and refereeing papers using these methods in the applications that we discuss below. We would also like to encourage other researchers (and editors) to support the development of these methodological innovations by writing (and publishing) studies of how these tools behave in different settings, and which design choices matter for the results. The goal should be to eventually build up a repertory of "canonical" methods and tests. Such tests would be in the spirit of the standard diagnostic tools that exist, for example, for applied econometric methods like difference-in-difference designs. These diagnostic tools will help to reassure readers and referees that the results of particular Generative AI analyses can be trusted.[6]

We discuss a number of different applications of Generative AI (and related deep learning techniques) in financial research in the following sections, covering: embeddings of high-dimensional data; text classification using LLMs; using LLMs as tools for simulating survey responses; and how LLMs can help in finding new research ideas and hypothesis generation. While not narrowly a "research" application, we will also briefly discuss some potential uses of LLMs in teaching finance, which can indirectly affect the productivity of many academic researchers. See Table II for an overview. We end the section with a subjective list of subjective advice for researchers and reviewers.

A. Embeddings

In finance and economics, we often need to analyze large sets of high-dimensional data. For example, we might study earnings calls from U.S. companies to determine if they provide information for predicting stock returns. But how do we convert these texts into data suitable for return prediction? Since there are endless ways words can form sentences, trying to track all possible phrases would be impossible and create too many variables. Instead, we can extract a lower-dimensional numerical representation that captures the key elements or "essence" of the text.

[6]See the recent review by Dell (2024) of deep learning methods in economics as a good example of this type of contribution.

Table II
Applications of Generative AI in Finance Research

Application	Question Types	Examples
Embeddings	• How to represent complex data concisely? • What are semantic relationships in data? • How to cluster similar entities?	• Gabaix et al. (2023): Asset embeddings • Chen and Sarkar (2020): 10-K filings embeddings • Kim et al. (2024): Labor market clusters
Text Classification	• What is the sentiment of financial text? • How to categorize documents? • What topics are discussed?	• Chang et al. (2024): Earnings call sentiment • Krockenberger et al. (2024): Covenant violations • Caragea et al. (2020): FinTech patent classification
Retrieval-Augmented Generation (RAG)	• How to find relevant information in large datasets? • How to use LLMs for classification based on many potential sources? • How to retrieve similar documents from a corpus?	• Bartik et al. (2023): Housing regulation classification • Chen and Wang (2024): AI patent assingment to functionalities
Simulating Agent Behavior	• Can LLMs replicate human preference heterogeneity? • What are expected survey responses? • What would human expectations be in counterfactual scenarios?	• Fedyk et al. (2024): Asset class preferences • Bybee (2023): Macroeconomic expectations • Hewitt et al. (2024): Experimental outcome predictions
Hypothesis Generation	• How to generate new research ideas or business ideas? • How to conduct qualitative research at scale?	• Si et al. (2024): Novel research idea generation • Girotra et al. (2023): New product ideas • Ludwig and Mullainathan (2024): Feature importance in neural network predictions
Teaching Finance	• How to create engaging course materials? • How to use LLMs for course management? • How to design interactive simulations?	• Mollick et al. (2024): Interactive simulations • Exam preparation and grading schemes • Chatbots for course material review

To achieve this, an important technique used in natural language processing and Generative AI is called "embedding", which is done using so-called encoder models. Embeddings are a way to represent different types of data, like words, texts, portfolio holdings, or market events, as numbers in a vector form. They capture the meaning or important features of the data by turning complex information into a set of numbers, which makes it easier to analyze non-numerical and understand relationships between such novel sources of data. More technically, embeddings are dense vector representations of discrete data (like text) in a continuous, lower-dimensional space. They capture semantic relationships (if the underlying data is text) or other key features of the data, mapping complex and high-dimensional information into a numerical space of desired dimensions.

Consider a study of general price inflation using news headlines: for example, "inflation is high and rising fast", "high inflation is the cause of rising prices", and "fast growth in high-rises". We would want a numerical representation that shows the first two headlines as more similar to each other than to the third, since they both discuss general price inflation, without getting confused by the use of superficially similar words in the third headline. Embeddings would solve this issue by assigning numerical vectors to each headline, where these vectors capture the meaning or context of the text. The similarity between the vectors depends on how the embedding model was trained, what the training data set was, and the target notion of similarity (eg. identifying patterns for predicting next sentences, or key concepts). Intuitively, embeddings can be thought of as a type of principal component analysis, where the principal components are recovered through neural networks that have been trained to extract the components most relevant for a particular prediction task.

Recently, researchers have begun using embeddings generated by models based on "transformers", which are particularly effective at capturing subtle differences in meaning between texts. Transformers are a specific type of neural network module, consisting of layers of nodes that apply mathematical transformations to input data and pass the output to subsequent nodes. Introduced by Vaswani et al. (2017), transformers process language by capturing complex dependencies and contextual relationships within the text. Modern encoder models rely on transformers to map input data into embeddings that account for a word's context by analyzing its surrounding words in both directions. Transformer models can process entire sequences of words and capture their meaning in context, as well as the semantic relationships between ideas within a text. This allows them to detect subtle differences in language use. For example, where a simple word-list approach might fail to distinguish between a firm lamenting that its competitors "increasingly profit" due to its lack of investment, and a firm celebrating its "increasing profit," a transformer-based model can leverage context to differentiate these meanings.

These models are typically trained using sequences of tokens, which represent basic units of text—such as words, parts of words, or even individual characters, depending on how the text is processed. For instance, in many models, each word in a sentence is treated as a token. A common training approach is the masked token model, where the model is asked to predict a missing (masked) token in a sentence, as well as the sequence of sentences within a text. During this process, the model's neural network adjusts its parameters, gradually transforming input tokens until it can predict the missing ones with high accuracy. This training creates embedding vectors that capture the semantic relationships necessary for these predictions.

This mechanism forms the basis for many innovations in model architecture, leading to the development of Generative Pre-Trained Transformers (GPT), such as those used in models like ChatGPT. See Wolfram (2023) for an accessible in-depth exploration of transformers and embeddings and Radford et al. (2018) for the introduction of general pre-trained transformer (GPT) models.

In financial research, embeddings can be particularly useful for summarizing and translating high-dimensional data, e.g. in the context of earnings call transcripts. The recovered embedding vectors can then be used as inputs into other analyses, for instance more traditional machine learning prediction models such as default prediction or sentiment analysis. We discuss such classification applications further in the following section.

Embeddings can also be utilized to create "Semantic Axes," as proposed by An et al. (2018). This involves embedding opposing concepts, such as "risky" versus "safe," and using the difference between their embeddings to define an axis in the embedding space. This axis effectively captures the extent to which embeddings differ regarding the concept of interest. For instance, this approach can be applied to text data concerning firms to characterize their alignment with specific industry clusters. Kim et al. (2024) create a novel representation of labor markets using combined data describing industries, occupations, skills and firms.

The initial application of the semantic axis technique in the finance literature appears to be Fedyk et al. (2024), who employ it to analyze survey results where respondents justify their preferences for various asset classes. In this context, explanations are categorized along "risk" and "return" axes derived from embeddings of opposing statements about these concepts. Overall, this methodology allows researchers to derive intuitive, continuous, and quantitative measures of specific content within text data, moving beyond mere binary indicators. By leveraging semantic axes, researchers can gain deeper insights into complex financial phenomena and better understand the nuances of language in financial contexts.

Embeddings can also be created from data other than text. An innovative application of embeddings using financial holdings data is presented by Gabaix et al. (2023). The authors develop a transformer-based method for generating "asset embeddings." This approach is similar to the masked language modeling employed by other transformer models, which predict missing words in a sentence based on context. In this case, the authors predict assets within an ordered list of an investor's portfolio based on the context of other assets the investor holds.

The resulting embeddings cluster assets in semantic space according to their likelihood of being held by similar types of investors and in comparable contexts. Consequently, these asset embeddings enable finance researchers to characterize investors based on their portfolio choices in a manner that transcends traditional observable characteristics. Furthermore, asset embeddings can predict co-movement among clusters of semantically similar assets. This capability could be valuable for understanding how macroeconomic events that specifically impact certain stocks propagate through equity markets, providing deeper insights into market dynamics and investor behavior.

B. *Text classification*

As noted above, one important category of financial research, concerns the analysis of text data generated by firms or about firms (e.g. earnings call, annual reports, SEC filings, press releases, news headlines). There is a large literature in which researchers use this data as inputs to extract information about which category companies fall into. That is, the goal is to *classify* firms by labeling text. For example, a researcher might want to classify if a firm is likely to be impacted by a particular new regulation, and therefore wants to label which firms mention this regulation in their earnings calls, and what the sentiment (positive / negative) of any such mentions is.

Before the rise of large-scale data analysis with Generative AI, researchers conducting text analysis often focused on specific sets of words or phrases within their texts. This approach typically resulted in a vector of binary indicators showing whether particular words were present, which could then be assigned ”positive” or ”negative” valences based on pre-defined lists created by the researchers. These indicators were then often combined into overall sentiment scores (see Loughran and McDonald (2020) for an overview of these techniques and Jurafsky and Martin (2024) for a broad reference on Natural Language Processing). However, this method can fall short when dealing with more complex texts, as the meaning of words can vary significantly depending on context. For example, distinguishing between a company "leaving an earnings slump behind" and "entering an earnings slump" would be problematic, as individual word valences alone would not reliably indicate whether these statements represent positive or negative updates. This limitation highlights the need for more sophisticated approaches that account for contextual nuances in text analysis.

As previously noted, transformer-based models excel at capturing subtle nuances in language by considering the context of the document in question. A particularly influential transformer model in finance research is BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin et al. (2018). Unlike the more recent Generative AI models, which generate output text from input text, BERT functions as an *encoder*. It processes chunks of text as input, and outputs either a vector of lower-dimensional numbers or a numerical classification label. BERT can be utilized directly as a pre-trained encoding model, effectively serving as an "off-the-shelf" solution for converting text chunks into semantic embeddings. These semantic embeddings are low-dimensional vector representations that capture the semantic content of the text based on the language patterns learned during BERT's original training. This ability to produce meaningful embeddings makes BERT a valuable tool for financial researchers seeking to analyze text data in a sophisticated manner.

Additionally, for analyses involving domain-specific text classification, models like BERT can be fine-tuned by adjusting their parameters to optimize an objective function relevant to the specific task at hand. This process involves training the model on a specialized dataset to generate embeddings that are tailored for predicting a particular outcome variable. The goal of fine-tuning is to refine the model's understanding of how words relate to one another, aligning it more closely with the specific domain of the text being classified, such as 10-K filings. This approach is especially beneficial when the vocabulary or syntax used in the text of interest significantly differs from the language found in the original training data for the encoder. For example, BERT was initially trained on datasets that included Wikipedia and self-published e-books, which may not capture the specialized language or structure present in financial documents. Fine-tuning helps ensure that the model accurately reflects the nuances and specific terminology of the domain, leading to improved classification performance. See Araci (2019) for the development of FinBERT using financial news articles as well as a sample of sentences from such news articles annotated by financial experts.

Importantly, by incorporating an additional classification layer, BERT and similar models can be fine-tuned to produce classification labels instead of merely generating generic embeddings. This enhancement further reduces the dimensionality of the text representation, making these models effective data-labeling tools. BERT's ability to process large chunks of text is particularly advantageous for handling extensive volumes of unstructured text data.

This model has been effectively applied in finance research for analyzing patent data. For instance, Caragea et al. (2020) trained a BERT-based model to classify the abstracts of millions of patent filings according to a taxonomy of FinTech-related inventions. Similarly, Chen and Wang (2024) utilized a transformer model to embed patent abstracts, enabling them to compare the proximity of these filings in semantic vector space to groups of reference patents. These reference patents were selected to represent AI-based systems with specific functionalities. By comparing individual patents to this reference group, the researchers can classify other patents at scale based on the similarity of the described technologies to the functionalities of interest. The motivation for this analysis stems from the idea that the granting of certain types of patents may differentially impact firm valuations and hiring behaviors, highlighting the importance of understanding how innovations are categorized and perceived in financial contexts.

Researchers have also developed models similar to BERT to analyze financial news, regulatory filings, analyst reports, and call transcripts. One example includes Krockenberger et al. (2024), which develops CovenantAI. Specifically, that research builds a database of covenant violations using 10-K and 10-Q reports filed by firms by applying the MPNET Sentence Transformer (a model designed to improve upon BERT that returns a single vector embedding for a text chunk rather than separate embeddings for each word) and training a classifier to identify text semantics that suggest covenant violations. For the development of MPNet, see Song et al. (2020).

Chen and Sarkar (2020) use an off-the-shelf BERT model on firm 10-K filings with the Security and Exchange Commission (SEC), use the mean of each firm's texts' embeddings to characterize each firm, and then cluster firms based on whether they have similar average embeddings. They show that firms clustered based on this text information demonstrate greater cross-group dispersion in firm fundamentals than that generated by existing industry definitions. This suggests that the 10-K text embeddings indeed identify similar firms. Clustering firms based on this type of qualitative data may enable researchers to generate better industry definitions that more closely resemble firms' "soft" characteristics. Improved industry classifications might also increase the fit of asset pricing models or help researchers to find appr See also Bonne et al. (2022) for use of an early embedding model, Doc2Vec, to combine textual 10-K data with numerical data on firm characteristics and stock returns to create industry groupings that align firms along risk and return dimensions better than industry classification systems commonly used in asset management and asset pricing research. Finally, see Hoberg and Phillips (2016) for a successful early approach to industry classification based text using the cosine similarity of the words firms use in the business descriptions of their annual reports.

More recently, the latest generation of large language models, starting with the release of ChatGPT in November 2022, made another set of advanced capabilities accessible for finance researchers. This type of "generative" model can respond to user queries about the content of texts and is able to comprehend and interpret complex texts. This allows researchers to go beyond mapping a text into semantic vectors—which would then have to be interpreted or used as an input into a classification model—and instead allows them to extract direct classifications or labels from raw text using an LLM. This way, researchers do not have to train a separate machine learning model that maps from embeddings to labels, as the LLM directly returns a label. This likely lowers the difficulty for researchers looking to conduct such classification analyses. However, researchers will likely still want to validate the output generated by an LLM in this way – see related comments in the "practical considerations" section below.

There are many potential applications of this methodology of using LLMs to directly classify text content at scale: For instance, in corporate finance and asset pricing it allows for the generation of new data sets and trading signals from corporate communications, where the scale of the text corpus, or the lack of machine learning skills or suitable training data, previously made analysis across large sets of firms impractical. The ability of LLMs to classify text "zero-shot"—without requiring additional training data or examples—simply based on their understanding derived from massive training data and a large number of parameters, enables them to perform such analyses without any need to invest costly effort upfront in creating training data.[7]

One recent example of such an application is Chang et al. (2024), who use gpt-3.5-turbo-16k to evaluate hundreds of thousands of earnings calls. The large context window of this LLM allows the researchers to increase the amount of textual data, measured in tokens (i.e. words or parts of words), that the model can consider at once. As a result, they are able to consider the earnings calls in their entirety and to map them into a sentiment score from -10 to 10. This score strongly predicts subsequent returns. The ability of LLMs to consider the context and semantic links between words is likely helpful to reliably interpret the tone of an earnings call. Additional evidence for this increase in capabilities is provided by Lopez-Lira and Tang (2023), who show that LLMs can predict next-day returns out-of-sample when given news headlines. However, they show that this requires sufficiently advanced LLMs: the return predictability only becomes significant when later generations of LLMs (like GPT-3.5 or GPT-4) are used to interpret the news headlines, and this is particulary true for more complex news sources. Similarly, Chen et al. (2023) show that a GPT-3.5-level model can predict aggregate U.S. stock returns in the following month from news headlines, while a fine-tuned BERT model or word count sentiment scores do not predict returns in the same setting. We expect that these methods will find further use in finance research by being applied to the large variety of texts that are generated in and around financial activities: press releases, news articles, patent texts, speeches by policymakers, product descriptions, job postings, websites, earnings calls, customer reviews, resumes, etc. can all potentially be transformed into quantitative information about firms. Addressing classical problems such as p-hacking and replicability are key to the ultimate success of this type of research (see, for example, Harvey (2017) and Harvey (2019)).

[7] However, as we argue below, researchers may often still want to obtain or create a sample with "ground truth" labels in order to validate the accuracy of the output from an LLM.

B.1. Retrieval-augmented generation

In many applications, it is too computationally expensive, or simply technically infeasible, to provide *all* the potentially relevant documents to an LLM. For example, when asking an LLM to evaluate how a firm's products are affected by different regulatory documents, which can run into the thousands of pages in length, it will be necessary to first identify the most relevant sections and include only a subset of the full text in the prompt to the LLM. One popular method for doing this is a technique called retrieval-augmented generation (RAG).

RAG is an analytical approach that combines the strengths of large language models in responding to text prompts with the retrieval of relevant knowledge from a database. This retrieved information can be integrated into the prompt to provide context for the request. For instance, a researcher interested in extracting infrequent mentions of environmental issues in firms' manufacturing processes could utilize RAG to search through a vast database of historical earnings calls, enabling a more targeted and contextually informed analysis. By leveraging both the language model's capabilities and the specific information retrieved, RAG enhances the efficiency and accuracy of information extraction in complex datasets.

A typical RAG workflow can be summarized by the following steps:

1. Database Creation: Start by creating a database of texts that might provide relevant context for queries directed at a large language model (LLM).

2. Text Chunking: Split the input texts into smaller, manageable segments (chunks).

3. Embedding: Convert these text chunks into semantic vectors using an embedding model. This allows for efficient searching and retrieval based on how similar the chunks are to a given query.

4. Query Processing: When a user asks a question, it's also converted into an embedding.

5. Retrieval: Use the query embedding to find the most relevant text chunks from the database.

6. Filtering/Re-ranking: The retrieved chunks may be filtered or re-ranked to determine which ones should be included in the final analysis.

The final step is helpful because after converting the user query into an embedding and retrieving a set of text chunks that are semantically similar to that query, you often end up with multiple chunks that may not all be equally relevant. Thus, filtering is applied to eliminate any chunks that do not meet specific criteria, such as those with low relevance scores based on their embeddings. Next, the remaining chunks undergo re-ranking to determine their order of relevance, often involving a calculation of more nuanced similarity scores or considering additional context about the query.

Once these steps are completed, the selected (and potentially ranked) text chunks are combined with a prompt that presents them as input alongside a request to a large language model (LLM) for analysis. This prompt structure enables the LLM to incorporate the retrieved information when formulating its response. Consequently, the model can effectively answer the question while relying on the relevant details contained in the selected text chunks, ensuring a more informed and contextually relevant output.

To illustrate this workflow more concretely, let's revisit the example involving environmental issues in corporate earnings calls. The process begins by dividing the earnings call transcripts into smaller text segments, potentially allowing for some word overlap to ensure that statements are not split mid-sentence. Each of these segments is then embedded into a unique semantic vector using a model like Sentence BERT and stored in a vector database.

For the LLM-based analysis focused on how firms discuss environmental concerns in their earnings calls, the researcher might first impose a hard filter to limit the transcripts to a specific firm and year. Next, they would search for chunks of that firm's earnings calls that are semantically similar to relevant texts about environmental issues that the researcher has selected, such as environmental impact press releases or environmental regulations. In the initial refinement step, the researcher may apply a filter to retain only those text chunks that explicitly mention the word "environment."

Subsequently, the researcher could employ an LLM to assess whether the selected excerpts discuss the environmental issues of other firms or those of the firm holding the call, keeping only the latter. Finally, the filtered set of excerpts would be combined with a prompt that includes a rubric for scoring the level or type of environmental concern expressed by the firm. This combined input would then be submitted to a sufficiently capable LLM with a request to return an appropriate score.

After parsing the LLM's response to extract the assigned label, the researcher would ultimately compile a dataset that indicates which firms express environmental concerns and in which years.

A recent application of Retrieval-Augmented Generation (RAG) is found in the real estate literature, where Bartik et al. (2023) utilize this approach to classify municipal housing regulations within a comprehensive nationwide database of zoning codes. They extract relevant text segments from this database and evaluate them using a large language model (LLM), in order to determine which zoning rules, if any, apply in different locations. To validate their methodology, the authors compare the results of their RAG-based classifications to a small existing sample of manual zoning classifications for 187 municipalities in Massachusetts. Their findings confirm that the RAG approach demonstrates high accuracy, particularly in binary classifications such as whether accessory dwellings are allowed.

This methodology can be expected to work well in applications where finding specific instances of relevant information in a large corpus of text is required to generate a classification or summary of interest—that is, to *find* the metaphorical needle in the haystack. It will not provide useful results in cases where the entirety of a long text needs to be considered—that is, to count *how many* needles there are in the haystack. Questions like "what is the most common issue discussed in company A's earnings calls" cannot be answered this way, as a retrieved subset of chunks of earnings calls would not provide sufficient information to the LLM.

C. *Simulation of agent behavior and expectations*

Another potential use of LLMs in research is as a quick, cheap, and always available survey respondent that can "simulate" human responses and function as a stand-in for them, for instance in preliminary survey testing, in product or user experience testing in companies, or in settings where the human subjects might not be available.

For instance, Hewitt et al. (2024) show that GPT-4 can surpass human experts in predicting experimental outcomes. That paper uses the model to simulate individual responses to treatments and uses the simulated responses to estimate treatment effects. Comparing GPT-4 to human forecasters, the model outperformed humans in predicting relative effect sizes of experiments. However, GPT-4 systematically overestimated absolute effect size, such that accurate predictions would require downscaling of the magnitudes it provided. The model's consistency across subgroups suggests broad applicability. The authors note that this approach of consulting an LLM for the expected outcome of an experiment could be used to enhance research designs, run simulated pilot studies, or generate priors for Bayesian analyses.

Similarly, LLMs could be used to generate tailored financial advice based on the demographics of a human counterpart, or take over interviewer tasks in data collection efforts. The utility of asking an LLM a question rather than a corresponding human subject depends on how well it can match the human responses (with precision and without bias), the replicability and robustness of the LLM-based method, and the costs of deployment of an LLM-based system relative to a survey of human respondents. As the latter is likely to be almost always favorable to the LLM, as costs per response tend to be small and declining, the first two issues are more likely to be important for researchers weighing whether or not to use an LLM-based system.

As an example of this type of application, Bybee (2023) shows that forward-looking expectations of macroeconomic time series and stock returns generated by LLMs based on contemporaneous news headlines closely track the corresponding expectations observed in real-world surveys, such as the Survey of Professional Forecasters (SPF) and that LLM-generated beliefs replicate behavioral biases exhibited by the human respondents.

Similarly, Fedyk et al. (2024) show that LLMs can behave similarly to human survey respondents in their preferences over asset classes (stocks, bonds, and cash) when prompted to respond from the perspective of someone with particular demographic characteristics (here: age, income, gender). Some differences exist, however: the researchers find that LLM responses are more likely to show transitive preferences than those of human subjects.

We expect that these methods can also be used in empirical applications to generate control variables in the spirit of propensity weights. If treatment is not random, propensity score assigned by an LLM could be used to match treated units to comparable units in the control group, or to re-weight observations. Alternatively, an LLM prediction might be a plausible proxy for how treated households would most likely have behaved without an intervention, e.g. in studying the effect of an information intervention on personal finance decisions. Such LLM-simulated behavior might be used as a synthetic control group in some settings.

In theoretical papers, LLMs might serve to generate realistic narrative examples of agents with heterogeneous characteristics either for illustrative purposes or to serve as inputs into simulations of outcomes. For example, a narrative might describe the following three agents: a high-income individual who is risk-averse and prefers stable investments, a small business owner who is more risk-seeking and willing to invest in innovative projects, or a low-income family facing employment challenges and making decisions based on limited resources. Then, an LLM could both extrapolate likely characteristics of these agents, and generate "plausible" model parameters to use in simulations of these agents' behavior. Similarly, LLMs can generate actions for heterogeneous agents in simulated scenarios and thereby allow researchers to generate hypotheses for how observed outcomes may result from complex strategic interactions—which may then be tested in real-world settings (Horton, 2023; Tranchero et al., 2024).

In time series settings, LLMs could also serve as predictability benchmarks: if the goal is to distinguish the expected from the unexpected variation in an outcome, an LLM prediction based on time t information could be used as the benchmark for what part of time $t+1$ outcomes could be rationally expected, and which parts were surprises. For example, studies that measure monetary policy "surprises" could use LLM-based predictions of changes in central bank policy based on commentary and news coverage during the run-up to a decision both to calibrate which policy changes were not expected, and to provide qualitative expectations of what *reasoning* observers would ex ante have expected for different policy choices and whether actual central bank justifications aligned with them. This ability of LLMs to interpret motivations and sentiment can allow researchers to measure not just changes in policy rates, but also in the broader policy regime and softer aspects of forward guidance by policymakers.

D. Hypothesis generation

An important part of academic research is the generation of new research ideas. As many graduate students and other researchers can attest, new research ideas are not generated "out of thin air", but rather arise from interaction with existing ideas and human feedback. As we can attest ourselves, the around-the-clock availability and broad knowledge of LLMs can make them fruitful partners for discussing early-stage ideas. In addition to kicking the tires on human-generated ideas, LLMs can also be useful in generating new ideas in the first place: Girotra et al. (2023) show that GPT 4 can generate new product ideas with minimal prompting that, on average, elicit a higher purchase intent than those that students at a top-ranked MBA program generated. Si et al. (2024) show that a system of LLMs can generate more novel research ideas than expert natural language processing researchers in a blind evaluation of their generated idea write-ups. However, they also find that LLMs are not reliable *judges* of the quality of research ideas. This suggests that LLMs can play an important role in complementing human researchers in the ideation process. Moreover, even the LLM-based system the paper uses to generate ideas leaves a role for human researchers as designers as it requires careful chaining of LLMs to generate and pre-select ideas, as well as structuring a write-up for human evaluation.

Conversely, sometimes humans and algorithms can work in "hybrid" teams, where humans can generate hypotheses that assist in understanding the "black box" analyses generated by machine learning or artificial intelligence. Ludwig and Mullainathan (2024) provide an example: they found that mugshots of defendants predicted pre-trial jail decisions. They used a Convolutional Neural Network (CNN), which is a specialized type of deep learning model designed primarily for processing grid-like data (such as images) to make the predictions. Then, to understand which facial features mattered for this prediction, they morphed images to become more "jail-able" and used *human* respondents to provide free-form comments to assess what features the model was changing to achieve that. These comments could then be used, in turn, to generate interpretable features of mugshots that could be tested for their predictive power. Another fruitful example of a "hybrid" team is Batista and Ross (2024), which generates testable hypotheses by combining the results of marketing experiments with an LLM that can generate suggestions for why humans engage with particular headlines. This method can be used to generate alternative explanations for effects and help design tests to rule these alternatives out.

Another aspect of idea generation in economics and finance is often the collection and analysis of qualitative data through structured interviews, which can then be used to validate or create new theories to explain the behavior of economic actors. Here, Generative AI can be used to scale researchers' ability to conduct "in-person" interviews with subjects, by using LLMs as affordable *interviewers* with human interviewees. Geiecke and Jaravel (2024) develop and validate an LLM-based tool for conducting qualitative interviews. They show that it can deliver high-quality responses in applications which, for instance, survey humans to elicit political preferences or discuss a complex topic such as having a meaningful life. Thus, these new survey methods can help to scale both hypothesis development and survey data collection.

It is also possible that, rather than being used for independent idea generation, the greater effect of LLMs on ideas in finance will be as a continuous sounding board, conversation partner, or constructive critic in the iterative (and admittedly often messy) process of academic innovation. In conjunction with the LLM, the researcher becomes a "cyborg" (to use the term coined by Mollick (2024)), with blurred lines of intellectual authorship and creative control. For a balanced view, it is also useful to consider Felin and Holweg (2024), who argue that AI cannot generate genuinely new knowledge and study how human reasoning is distinct from computer prediction.

E. Teaching finance with Generative AI

One channel through which Generative AI makes academic researchers more productive is by making time devoted to teaching more productive.

We mention just a few potential LLM applications that we have found useful in teaching MBAs. Generative AI can help to convert loose notes into scripts in Tex or Markdown formats that can be compiled into slides can save time in class preparation. The more advanced versions of LLMs available at the time of this writing (e.g. OpenAI's o1-preview model) can check derivations on slides for typos, and can quickly generate and test problem sets that are aimed at reinforcing particular concepts. A conversation with a chatbot can also provide helpful practice for classroom discussions and feedback or inspiration for how to engage students.Models can additionally execute code to quickly produce graphs illustrating equations derived in class. Chatbots can also be used directly as instructional tools: we have used transcripts of class sessions together with an LLM to create a chatbot that can answer questions about the material covered in an MBA course, for the benefit of both prospective students and current students reviewing course materials before an exam.

Another application in teaching finance (and other subjects) is to enable faculty to "trial-run" exams: by prompting LLMs with relevant personas (e.g. "assume you are an MBA student"), or even explicitly asking for suggestions to improve the exam, we have had success in generating realistic-sounding responses to in-class exams. This sometimes served to highlight potential misunderstandings of questions, and allows refinement of exam questions before releasing the exam to students. A related advantage is that model-generated responses can be used as starting points for developing grading schemes for exams.

As Mollick et al. (2024) note, LLMs can also be used to design interactive simulations for business school students that might lead to more engaging interactions with case studies covered in class. One can also students to design their own LLM-based chatbots: this activity requires them to understand the potential use cases for Generative AI in the workplace and the need to carefully design prompts and constrain functionality to reliably elicit a desired behavior from—at times mercurial—model personas.

F. Practical considerations for using Generative AI in research

As the methodologies discussed above have only recently been developed and are in the process of being used in finance research for the first time, there may be be limited consensus about how to validate that these models have been applied correctly, what aspects of such analyses should be reported in papers, and what pitfalls to look out for. To advance this process of communal learning by our profession, below we provide some subjective views of what better—albeit likely not yet best—practice might look like. This draws on our own experiences deploying some of the methods mentioned above, as well as our experiences as referees of papers in this field.

The methodological advice provided below should *not* be interpreted as criticisms of the mentioned papers. To the contrary, the dearth of existing applications of these methods means that researchers developing new applications and experimenting with Generative AI tools are providing a valuable service to the profession, as they are not just pushing the boundary of the topics that they are studying but are also providing new tools for other researchers to build on.

F.1. Obtaining numerical scores from LLMs

In many applications, one may not just want to assign a binary label to a text (e.g. "CEO talks about ESG = yes/no"), but rather wants to assign more fine-grained numerical distinctions (e.g. "CEO attitude towards ESG on a scale from 1 to 10"), or absolute quantifications ("CEO mentions ESG frequently or *very* frequently"). Obtaining such numerical distinctions from an LLM is not always straightforward. For example, as mentioned above, Chang et al. (2024) use an LLM to map earnings calls into a sentiment score from -10 to 10. Similarly, Jha et al. (2024) ask an LLM to assess whether an earnings call suggests that a firm plans to increase or decrease its capital spending and whether it plans to do so "substantially". What should we interpret the magnitudes provided by LLMs in response to these kinds of prompts to mean?

Note that the LLM is *not* necessarily evaluating each text, e.g. an earnings call transcript, relative to other earnings calls, either in the sample, or in general. The reason is that it is usually only provided one input text (e.g. a single earnings call) and does not have a memory that would allow it to look "across" the sample of calls, other than what general patterns have been retained from its training data. As a result, the comparison set for what a "4 out of 10" or "substantial increase" mean can depend heavily on what the distribution of data in the training set looks like. Even if the right kind of comparison data is in the training set, the model is not "considering" that training data - rather it will respond in a way that is the most likely continuation, which is likely to be driven by the words used in other (potentially erroneous) descriptions of similar data sets (for instance other papers about earnings calls), rather than anything resembling an evaluation of the data at hand relative to tranining data. To make that concrete: if most studies in the training data report CEOs that use the word "climate change" to be more than average concerned about the climate, while in recent years the same words have been used mostly by those who oppose actions to mitigate climate change, then a sample of recent earnings calls will likely be classified based on the old mapping between words and intentions and not the more applicable recent usage.

While the desired comparison set can be narrowed somewhat by explicitly asking the LLM to restrict its assessment to an evaluation relative to a benchmark (e.g. "U.S. earnings calls from the last ten years"), whether this constraint is effective will, among other things, depend on the extent to which this category was represented in the training data for the LLM, and how the fine-tuning of the LLM determined its inclination to obey such instructions. As a result, the exact reference group for such large vs. small comparisons is not just unclear in any particular analysis, but can also vary substantially across different model vintages. As an example of how this lack of calibration for a use case can go wrong, consider an example provided by Dhinakaran (2023): when a text is deliberately modified to include continuously varying levels of spelling errors or sadness indicators, and different LLMs are asked to provide continuous scores from 0 to 100% for the degree of spelling errors and sadness in the text, the resulting score *levels* are not comparable across models, e.g. what one considers a 100% score might be 50% in a different model. Moreover, within a model, the assigned score gradient is often flat or sometimes even decreasing for large ranges of the true level of the text characteristic.

A related issue to that of unclear numerical scores arises in promising application of LLMs in financial research that involve categorization or clustering of entities, for example, the clustering of firms into groups based on their corporate communication. One good example of this approach is Beckmann et al. (2024) who are trying to uncover the effect of *unusual* communication in firms' earnings calls on their subsequent financial market outcomes.

They first ask an LLM to categorize companies' earnings calls as "unusual" and to justify these labels. Then, they use an LLM to infer high-level category clusters of what can make an earnings call unusual by reviewing the different justifications provided by the initial LLM calls. Finally, they again ask an LLM to revisit the initial task of labeling unusual communications and to assign each earnings call binary usual / unusual labels in each of the 25 derived categories.

This is a clever use of multi-level LLM prompting to derive insights from financial text data. It raises some interesting questions that researchers should consider when doing this type of analysis: For example, labels assigned by LLMs may be ambiguous and depend on the context provided to the model. Without further definition, "unusual" will refer to an earnings call being different in *some* dimension from the model's implicit and unstated benchmark. It will not be defined relative to the sample of earnings calls at hand - as the model is only provided one of the them at a time and has no way of considering the corpus as a whole.

As noted in the case of numerical categories, there is no reason why a model's implicit categories would correspond to the categories that a financial researcher cares about. For example, the study finds that the LLM sometimes applies "unusual" labels when an earnings call discusses a narrow set of topics or contains lengthy responses, focusing on deviations from expected patterns in style rather than content, which may or may not align with the issues that the researchers care about.

Similarly, the second step of clustering justifications implicitly requires the LLM to apply a distance metric to the justification statements to decide which of them are "similar" to one another. Different distance metrics will lead to different outcomes and are usually a deliberate choice by the researcher in classical machine learning (see, e.g. Ghazal (2021)). So is the number of desired clusters. When asking an LLM to cluster, the implicit distance metric and desired cluster number are neither transparent nor guided by the researchers. To relate the outcome of such an LLM-guided clustering analysis to hypotheses that are relevant to finance research, it can be helpful to structure the number and type of categories ex ante to align with characteristics that are of interest as a result of theoretical predictions or the researcher's intuition about mechanisms that matter for the topic at hand.

To address these types of issues, it can be helpful to provide examples of how particular inputs should be scored, rubrics for how different quantitative categories are defined, definitions of classification categories, and constraints on the desired outcomes of clustering tasks. The correct intuition is usually that if the correct execution of a task would not be obvious to a human research assistant and would predictably vary across different research assistants, then an LLM will also be confused. In most cases, a researcher defining what subjective terms like "substantial" mean or how numerical scales are calibrated (e.g. "a 7 out of 10 in sadness involves the use of mostly negative adjectives, few mentions of happy moments, etc...") will ensure greater replicability and likely lead to less measurement error.

For example, the approach of defining a rubric narrows the problem from one where the interpretation of the provided categories is left to the LLM to one where LLMs have to simply correctly map from the provided rubric to the sample data, which should reduce the variability in outcomes. Moreover, to detect issues with the scoring, researchers should report details on the overall distribution of scores that the LLM provided in their write-up. Unusual bunching of scores may indicate issues with the LLM's mapping. In general, treating the LLM like a diligent but inexperienced research assistant is more likely to lead to a correct research design than assuming it is an oracle with mystical powers.

However, we want to stress that, while the ability of LLMs to execute this kind of quantitative mapping and categorization consistently without a given rubric or benchmark is still to be shown, this type of procedure will nevertheless often generate more reliable results than more traditional methods based on counts of words (see Loughran and McDonald (2020)), because LLMs are more likely reliably understand different usages and context. Thus, in spite of their flaws, using LLMs in these settings holds great promise: the relevant benchmark for comparison are labels and scores applied by human evaluators and older natural language processing methods that serious shortcomings of their own.

F.2. Explanation and evaluation of LLM reasoning

One common technique for evaluating whether the output of an LLM reliably aligns with human intuition and the request made in the prompt is to ask the LLM to provide a step-by-step explanation for its reasoning – which is sometimes referred to as "zero-shot chain-of-thought prompting". Then, the researcher can evaluate whether the reasoning is as intended or shows a misunderstanding of the task (e.g. due to ambiguous wording or missing context) and can adjust the prompt accordingly. For instance, Eisfeldt et al. (2023) asks the LLM to provide an explanation for why a task was classified into a particular rubric for whether Generative AI can be applied. This explanation helps to understand whether the rubric was consistently applied to the task descriptions in the expected way. However, caution is required in applying this technique: Wang et al. (2023) note that LLMs have positional bias - the order in which information and requests are provided can substantively change the responses. When asking an LLM to explain a score that it provided, Wang et al. (2023) note, that "due to the nature of the auto-regressive model, the conclusions generated by the model are not supported by the explanation generated afterward." That is, the model will rationalize the most likely explanation for the previously generated score, but there is no direct sense in which this ex post reasoning causally preceded the score. An LLM will enthusiastically justify its alleged reasoning for any score that the user claims it generated. Wang et al. (2023) therefore suggest that, for researchers interested in the justification for a response, simply asking a model to *first* provide an explanation and only then a quantitative score is more likely to yield responses where these two aspects are consistently linked.[8]

F.3. Implementing retrieval-augmented generation

RAG approaches normally involve cleaning and chunking the texts that can be retrieved in advance. Removing meta-data or irrelevant formatting information is often helpful as it might distort the embedding and lead to noise when identifying the most relevant texts. *How* to best divide a larger text into smaller chunks is an important issue, as arbitrary splits by word counts might render longer arguments incomprehensible to the model that is trying to parse the meaning contained in each chunk. There are a number of different solutions to this issue (e.g. splitting the text based on semantic break points detected by an embedding model that parses the semantics of a moving text window). Each method fares differently along dimensions of relevant trade-offs, e.g. between the sensitivity to changes in topics and computational cost.

While hard evidence is scarce, it does appear that LLMs are sensitive to the inclusion of irrelevant information in long prompts. As a result, RAG approaches do not just run the risk of missing relevant texts that should have been included in a prompt, but also that irrelevant text added to the prompt confuses the LLM. The latter problem can be alleviated by further filtering or re-ranking the text chunks selected after an initial search over the database. Elaborating on the discussion of RAG workflows in the previous section, adding simple positive or negative keyword filters before or after retrieving the best matches

[8] See page 4 "Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment. Then, output two lines indicating the scores".

can ensure that superficially similar but not relevant topics are not included, or that all retrieved chunks mention a particular entity by name. Keyword filters have a relatively low computational cost relative to large language models, so this is an easy way to ensure that the input tokens in an LLM query are not wasted on irrelevant text. A more computationally expensive method for filtering the retrieved chunks may be to submit them to an LLM (which might be a less capable model than the one used for the final analysis) and ask the LLM to remove texts that are not germane to the user query in terms of their content.

One important input into creating valid RAG pipelines is evaluating each step of the process for whether it performs as expected. In most cases, this will involve either formally or informally sampling outcomes separately from the retrieval and the generation steps and comparing them to a validation sample of researcher-determined expected outcomes. Returning to the earlier example of determining a CEO's attitude to ESG from earnings calls, researchers should manually sample and review retrieved earning call chunks for whether they actually capture ESG-related topics, and compare LLM-generated scores to how a human researcher would have labeled the chunks. This is likely an iterative process - if a query does not result in the retrieval of relevant chunks, researchers could then consider adding additional filters.

Given that the performance of particular prompts varies with the exact formulation of a request as well as the foundation model used and can depend on small variations, e.g. the exact order in which examples are presented, it is also important that researchers closely monitor and optimize the performance of their prompts and retrieval queries. For instance, in the example given, a researcher might be asking an LLM to evaluate whether an earnings call implies a yes or no response to the question "are environmental concerns discussed?" However, they should likely not simply write that exact question as stated and hope for the best. Instead, providing context, defining relevant terms, and adding examples of what might differentiate "concerns" from mere "mentions" might all improve the performance of the model. Conversely, offering this same context might distort the embedding of the query used to search for similar text chunks in the vector database. It can therefore be optimal for the retrieval query to not be phrased in the same way as the ultimate prompt that will be submitted to the LLM in the final step of the RAG.

In general, it is important for researchers to inspect the performance of the RAG system and adjust the tools used as needed, as "standard" tools may not work for all types of text, and prompts usually need to be adjusted, e.g. to include additional examples of correct labels, to ensure that the LLM responds as intended. Just as in classical machine learning, the researchers should score a small random subset of "ground truth" examples themselves and ensure that different iterations of the RAG workflow and prompt perform adequately for this subset before scaling the analysis to the full data set. Researchers should document these design decisions in their papers in the same way that an econometric method, such as a difference-in-differences design, or an estimating equation, would be included.

Reviewers or editors should look for whether papers document and explain their design choices so they can assess whether the choices appear reasonable (e.g. choosing a prompt that is unambiguous) and show evidence that the researchers took care to prevent the most likely issues in execution. Documenting these approaches will also allow for easier replication, and for the profession to be able to learn what choices might affect the performance of these systems as we accumulate more examples of their use.

IV. Conclusion

Generative AI represents a major technology shock in the research and practice of finance. Eisfeldt et al. (2023) shows the immediate and large impact on firm values from the release of ChatGPT, suggesting that future research on the impact of Generative AI on corporate policies will prove fruitful. A growing body of innovative research utilizes Generative AI tools to study classic problems in corporate finance and asset pricing. As with previous technological innovations, such as the wide availability of financial market data that was made possible in the late 20^{th} century, and the advances in computing power in the early 21^{st} century, we expect to see a far-reaching impact of Generative AI on research in finance. In addition to reviewing innovative existing studies, our hope is that this review provides useful tools for successful future work.

REFERENCES

Acemoglu, Daron, and David Autor, 2011, Skills, tasks and technologies: Implications for employment and earnings, in *Handbook of labor economics*, volume 4, 1043–1171 (Elsevier).

Aldasoro, Iñaki, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti, and Merlin Stein, 2024, Intelligent financial system: how ai is transforming finance, Technical report, Bank for International Settlements.

An, Jisun, Haewoon Kwak, and Yong-Yeol Ahn, 2018, Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment, *arXiv preprint arXiv:1806.05521* .

Araci, D, 2019, Finbert: Financial sentiment analysis with pre-trained language models, *arXiv preprint arXiv:1908.10063* .

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, 2024, Artificial intelligence, firm growth, and product innovation, *Journal of Financial Economics* 151, 103745.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson, 2023, Artificial intelligence and firms' systematic risk, *Available at SSRN 4868770* .

Bartik, Alexander, Arpit Gupta, and Daniel Milo, 2023, The costs of housing regulation: Evidence from generative regulatory measurement, *Available at SSRN* .

Batista, Rafael M, and James Ross, 2024, Words that work: Using language to generate hypotheses, *Available at SSRN 4926398* .

Beckmann, Lars, Heiner Beckmeyer, Ilias Filippou, Stefan Menze, and Guofu Zhou, 2024, Unusual financial communication-evidence from chatgpt, earnings calls, and the stock market .

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, *Available at SSRN* 4452670.

Bonne, George, Andrew W Lo, Abilash Prabhakaran, Kien Wei Siah, Manish Singh, Xinxin Wang, Peter Zangari, and Howard Zhang, 2022, An artificial intelligence-based industry peer grouping system .

Bybee, J Leland, 2023, The ghost in the machine: Generating beliefs with large language models, *arXiv preprint arXiv:2305.02823* .

Caragea, Doina, Mark Chen, Theodor Cojoianu, Mihai Dobri, Kyle Glandt, and George Mihaila, 2020, Identifying fintech innovations using bert, in *2020 IEEE International Conference on Big Data (Big Data)*, 1117–1126, IEEE.

Chang, Anne, Xi Dong, and Changyun Zhou, 2024, Does democratized information access demand democratized ai? a tale of the rich, the poor, and the ai (chatgpt) in earnings calls .

Chen, Jiafeng, and Suproteem Sarkar, 2020, A semantic approach to financial fundamentals, in *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, 22–26.

Chen, Jian, Guohao Tang, Guofu Zhou, and Wu Zhu, 2023, Chatgpt, stock market predictability and links to the macroeconomy, *Available at SSRN 4660148* .

Chen, Mark A, and Joanna Xiaoyu Wang, 2024, Displacement or augmentation? the effects of ai on workforce dynamics and firm value, *The Effects of AI on Workforce Dynamics and Firm Value (April 8, 2024)* .

Cochrane, John H, Francis A Longstaff, and Pedro Santa-Clara, 2008, Two trees, *The Review of Financial Studies* 21, 347–385.

Crouzet, Nicolas, and Janice Eberly, 2023, Rents and intangible capital: A q+ framework, *The Journal of Finance* 78, 1873–1916.

Dell, Melissa, 2024, Deep learning for economists, Technical report, National Bureau of Economic Research.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Dhinakaran, Aparna, 2023, Llm as a judge: Numeric score evals are broken!!!, https://twitter.com/aparnadhinak/status/1616065794623197185, Accessed: 2024-07-23.

Eisfeldt, Andrea L, Gregor Schubert, Miao Ben Zhang, and Bledi Taska, 2023, The labor impact of generative ai on firm values, *Available at 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 preprint arXiv:2303.10130* .

Falato, Antonio, Dalida Kadyrzhanova, Jae Sim, and Roberto Steri, 2022, Rising intangible capital, shrinking debt capacity, and the us corporate savings glut, *The Journal of Finance* 77, 2799–2852.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier, 2024, Chatgpt and perception biases in investments: An experimental study, *Available at SSRN 4787249* .

Felin, Teppo, and Matthias Holweg, 2024, Theory is all you need: Ai, human cognition, and decision making, *Human Cognition, and Decision Making (February 23, 2024)* .

Gabaix, Xavier, Ralph SJ Koijen, Robert Richmond, and Motohiro Yogo, 2023, Asset embeddings, *Available at SSRN 4507511* .

Geiecke, Friedrich, and Xavier Jaravel, 2024, Conversations at scale: Robust ai-led interviews with a simple open-source platform, Technical report, London School of Economics (LSE) and CEPR, Working Paper.

Ghazal, Taher M, 2021, Performances of k-means clustering algorithm with different distance metrics, *Intelligent Automation & Soft Computing* 30, 735–742.

Giglio, Stefano, Bryan Kelly, and Dacheng Xiu, 2022, Factor models, machine learning, and asset pricing, *Annual Review of Financial Economics* 14, 337–368.

Girotra, Karan, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich, 2023, Ideas are dimes a dozen: Large language models for idea generation in innovation, *Available at SSRN 4526071* .

Graham, John R, and Mark T Leary, 2011, A review of empirical capital structure research and directions for the future, *Annu. Rev. Financ. Econ.* 3, 309–345.

Harvey, Campbell R, 2017, Presidential address: The scientific outlook in financial economics, *The Journal of Finance* 72, 1399–1440.

Harvey, Campbell R, 2019, Replication in financial economics, *Available at SSRN 3409466* .

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer, 2024, Predicting results of social science experiments using large language models, Unpublished working paper.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of political economy* 124, 1423–1465.

Horton, John J, 2023, Large language models as simulated economic agents: What can we learn from homo silicus?, Technical report, National Bureau of Economic Research.

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang, 2024, Chatgpt and corporate policies, Technical report, National Bureau of Economic Research.

Jurafsky, Daniel, and James H Martin, 2024, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Kelly, Bryan, Dacheng Xiu, et al., 2023, Financial machine learning, *Foundations and Trends® in Finance* 13, 205–363.

Kim, Seongwoon, Yong-Yeol Ahn, and Jaehyuk Park, 2024, Labor space: A unifying representation of the labor market via large language models, in *Proceedings of the ACM on Web Conference 2024*, 2441–2451.

Korinek, Anton, 2023, Generative ai for economic research: Use cases and implications for economists, *Journal of Economic Literature* 61, 1281–1317.

Krockenberger, Vanessa S, Anthony Saunders, Sascha Steffen, and Paulina M Verhoff, 2024, Covenantai-new insights into covenant violations, Technical report, SSRN.

Lopez-Lira, Alejandro, and Yuehua Tang, 2023, Can chatgpt forecast stock price movements? return predictability and large language models, *SSRN* 69, Posted: 10 Apr 2023, Last revised: 14 Apr 2024.

Loughran, Tim, and Bill McDonald, 2020, Textual analysis in finance, *Annual Review of Financial Economics* 12, 357–375.

Ludwig, Jens, and Sendhil Mullainathan, 2024, Machine learning as a tool for hypothesis generation, *The Quarterly Journal of Economics* 139, 751–827.

Mollick, Ethan, 2024, *Co-Intelligence* (Random House UK).

Mollick, Ethan, Lilach Mollick, Natalie Bach, LJ Ciccarelli, Ben Przystanski, and Daniel Ravipinto, 2024, Ai agents and education: Simulated practice at scale, *arXiv preprint arXiv:2407.12796* .

Nagel, Stefan, 2021, *Machine learning in asset pricing*, volume 1 (Princeton University Press).

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., 2018, Improving language understanding by generative pre-training .

Rampini, Adriano A, and Sridhar Viswanathan, 2013, Collateral and capital structure, *Journal of Financial Economics* 109, 466–492.

Rosen, Sherwin, 1981, The economics of superstars, *The American economic review* 71, 845–858.

Si, Chenglei, Diyi Yang, and Tatsunori Hashimoto, 2024, Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, *arXiv preprint arXiv:2409.04109* .

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu, 2020, Mpnet: Masked and permuted pre-training for language understanding, *Advances in neural information processing systems* 33, 16857–16867.

Sun, Qi, and Mindy Z Xiaolan, 2019, Financing intangible capital, *Journal of Financial Economics* 133, 564–588.

Tranchero, Matteo, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj, 2024, Theorizing with large language models, Technical report, National Bureau of Economic Research.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017, Attention is all you need, *Advances in neural information processing systems* 30.

Veldkamp, Laura, and Cindy Chung, 2024, Data and the aggregate economy, *Journal of Economic Literature* 62, 458–484.

Wang, Peiyi, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui, 2023, Large language models are not fair evaluators, *arXiv preprint arXiv:2305.17926* .

Wolfram, Stephen, 2023, *What Is ChatGPT Doing:... and Why Does It Work?* (Wolfram Media).