ANNUAL REVIEWS

Annual Review of Financial Economics

Generative AI and Finance

Andrea L. Eisfeldt and Gregor Schubert

Anderson School of Management, University of California, Los Angeles, California, USA; email: andrea.eisfeldt@anderson.ucla.edu

https://doi.org/10.1146/annurev-financial-112923-020503

JEL codes: G00, G10, G30, O33, C80

OPEN ACCESS CC BY

Keywords

artificial intelligence, large language models, technology, labor productivity

Abstract

Since ChatGPT's release in 2022, demand for artificial intelligence (AI)–related skills in finance has grown rapidly, as generative AI drives significant technological changes in both the financial research field and the broader economy. We show that financial occupations are highly exposed to the productivity effects of generative AI, 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. Generative AI also holds great potential as a tool for finance researchers and practitioners: 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 in academia and industry interested in using these tools.

1. INTRODUCTION

Generative artificial intelligence (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, demand for AI skills broadly tripled by mid-2024. Moreover, demand for generative AI and LLM skills rose from zero prior to ChatGPT to around 1% of all job postings.

In this review, we focus on generative AI as both a topic of study for researchers in financial economics and a methodological tool for conducting finance research. We focus specifically on recent innovations in LLMs and related deep learning techniques, rather than on the broader set of tools oftentimes grouped under the umbrella term "artificial intelligence," including machine learning, neural nets, machine vision, and robotics, which were adopted in research and practical settings throughout the 2010s. We discuss the impact of these tools on firm value, firm decisions, and 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 corporate finance decisions. For a recent study that summarizes some of these emerging use cases within financial firms, see Aldasoro et al. (2024).[1]

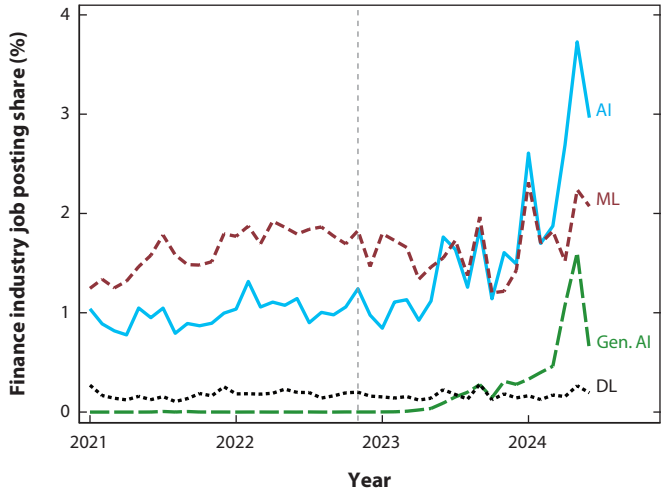

Figure 1

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 code 52) that mention particular skills. The vertical dotted line indicates the ChatGPT release date on November 30, 2022. Abbreviations: AI, artificial intelligence skills; DL, deep learning skills; Gen. AI, generative AI or large language model skills; ML, machine learning skills; NAICS, North American Industry Classification System. Data are from the Lightcast job postings database (see **https://lightcast.io/** for further information).

[1] Other articles have already provided excellent overviews of other aspects of the interaction between technology and finance: For recent perspectives on the broader effects of data on firms and the economy, see Veldkamp & Chung (2024). For use of machine learning in asset pricing, see Nagel (2021), Giglio, Kelly & Xiu (2022), and Kelly & Xiu (2023). Duffie et al. (2022) explore the broader implications of technological changes for payments, data processing, and trading systems, and Jiang & Li (2024) provide an overview of the impact of technology on corporate governance.

In this article, we address the following two broad topics: (*a*) generative AI as a technology shock that affects firm values and corporate policies, and (*b*) generative AI as a technology shock to methods for financial research. For the first topic, in addition to reviewing the existing literature, we highlight fertile areas for future research. For the second topic, we provide examples of recent applications and practical guidance for researchers in academia and industry looking to add generative AI to their research toolbox and discuss best practices for using generative AI research methods.

2. 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. Firm value is expected to change through two channels: varying exposure to the technology shock and firms' strategic responses to the rapidly advancing AI frontier. So far, existing research has only scratched the surface of the ways in which shifts in production processes, changes in productivity, and uncertainty about further innovation are 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 by 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.

2.1. Measuring Generative AI Exposure

How do we know which firms are affected by a technology shock? One approach is to evaluate firms' current technology use through surveys, patent data, and product information, then infer productivity potential from these revealed preferences. 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.

An example of such an alternative measurement approach is provided by 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 generative pre-trained transformer (GPT) 4-level LLM to complete the task. The resulting scores reflect three different levels of exposure: (*a*) direct exposure, access to a ChatGPT-like LLM directly reduces task completion time by ≥50%; (*b*) indirect exposure, when combined with additional software and tools, the LLM could reduce task completion

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

Table 1 Examples of generative AI exposure scores for loan officer tasks[a]

Task	Exposure[b]	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.

Abbreviations: AI, artificial intelligence; GPT, generative pre-trained transformer; LLM, large language model.

[a]This table presents examples of the scoring of loan officer tasks' exposure to generative AI by comparing the tasks' statement to a rubric of generative AI capabilities. The classification was done using the GPT-3.5 Turbo Model accessed through the OpenAI API. The task statements for each occupation were obtained from the O*NET V27.2 database (available at **https://www.onetonline.org/**), and the table shows the explanation provided by the model for the assigned exposure type.

[b]"Direct exposure" means that an off-the-shelf generative AI chatbot could allow a worker to do the task in half the time (as of March 2023), while "indirect exposure" means that this productivity improvement could only be accomplished after adding an additional layer of tools on top of the LLM capabilities, such as Internet search access, or a link to proprietary databases. "None" indicates that generative AI cannot increase the productivity in this task by 50%.

time by ≥50%; and (*c*) 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. These exposure data are publicly available from the authors' websites for use by other researchers.[3]

To illustrate how this approach categorizes different tasks, **Table 1** shows some of the tasks that are done by loan officers [Standard Occupational Classification (SOC) code 13-2072] and the exposure assigned to them: LLMs can perform simple calculations based on existing patterns or write code for 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

[3]The data can be found at **https://sites.google.com/view/gregorschubert**.

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 Section 3.3. 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)[4] 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 (noncore) activities are automated, employees are less likely to be made redundant and might even increase their value to the firm if they can reallocate some of their freed-up time from less essential activities to more productive pursuits.

While Eisfeldt et al. (2023) focus 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.

2.2. 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. **Figure 2*a*** shows that the average generative AI exposure across all occupations (weighting them by their 2022 employment) is 27%. The figure also provides details on selected white-collar occupation groups: Health care occupations have below-average exposure at 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 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.

How does this exposure vary within finance occupations? The higher average exposure among finance occupations[5] aligns with the general pattern that higher-wage occupations are more likely to be exposed to generative AI: Average annual wages among finance occupations were $108,000 in 2022, while the US average was $62,000. However, there is variation in how much particular jobs within finance are likely to be exposed: **Figure 2*b*** shows the breakdown for the

[4]For details, see **https://www.onetonline.org/help/online/scales#score**.

[5]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).

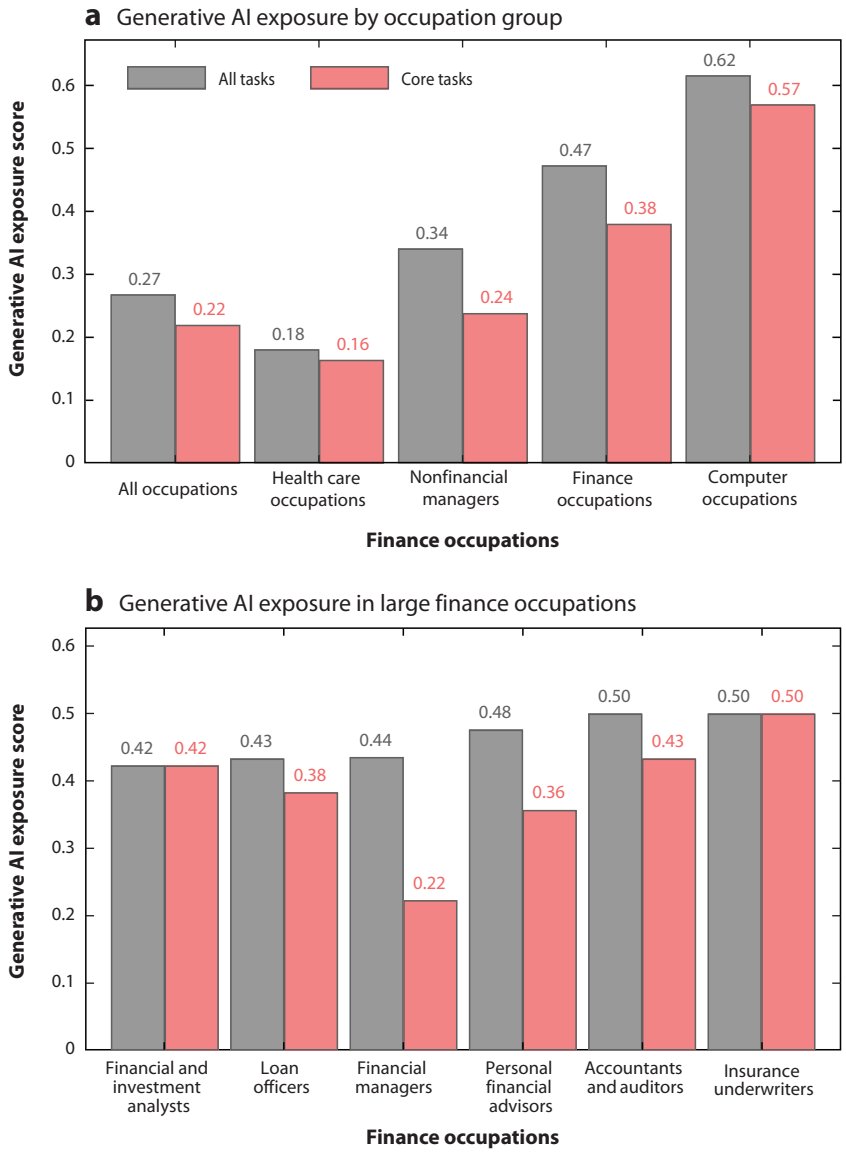

Figure 2

Generative AI exposure by occupation and in large financial occupations. This figure shows the share of tasks in each different white-collar occupation group exposed to generative AI productivity improvements. Panel *a* plots the weighted share of tasks exposed to generative AI in different white-collar occupation groups, where detailed 6-digit occupations are aggregated into weighted averages by occupation groups using data on their relative employment from the BLS. The graph differentiates between overall exposure (*gray*) and exposure only from core tasks (*red*). Core tasks are defined by the Occupational Information Network as tasks that are both relevant and important for an occupation. Panel *b* plots the same data for key large occupations within the financial sector (employment >$100,000), showing that, while exposure in financial occupations is generally high, the share coming from core tasks varies substantially. Abbreviations: AI, artificial intelligence; BLS, Bureau of Labor Statistics. Data from the measure by Eisfeldt et al. (2023).

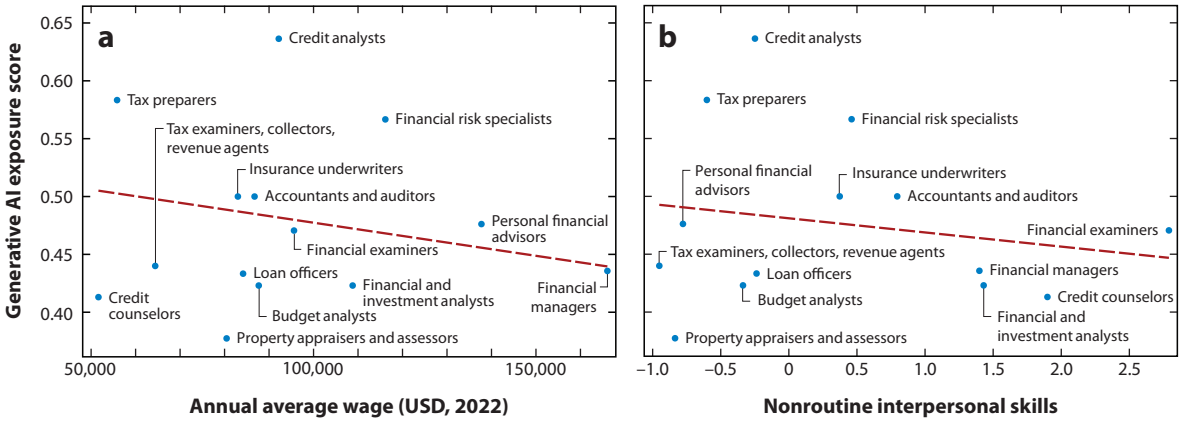

Figure 3

Finance occupations and generative artificial intelligence (AI) exposure, wages, and social skills. The figure shows the relationship between the share of tasks with exposure to generative AI in different finance occupations and (*a*) their 2022 wages from the Bureau of Labor Statistics and (*b*) the interpersonal skills involved in the occupation. Social skills are proxied by the measure of nonroutine cognitive interpersonal skills defined by Acemoglu & Autor (2011) based on O*NET data, which captures to what degree the occupation involves activities like establishing interpersonal relationships, guiding subordinates, or coaching others. Linear fits are weighted by employment.

largest finance-related occupations (employment >$100,000). While overall generative 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 advisors 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*a*** plots generative AI exposure against occupational wages, and **Figure 3*b*** plots exposure against a measure of the occupation's reliance on nonroutine interpersonal skills. **Figure 3*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 of the finance-specific occupations shown here already require a high level of analytical skill. Eisfeldt et al. (2023) show that a need for analytical skill is one of the key predictors of occupational exposure to generative AI. **Figure 3*b*** shows that higher exposure is associated with less occupational need for nonroutine interpersonal skills. Examples of nonroutine 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. **Figure 3*b*** also shows that some of the finance occupations that are shown in **Figure 3*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) also involve high levels of nonroutine interpersonal skills.

2.3. Effects of Generative AI on Firms

The release and rapid subsequent rise to prominence of generative AI starting in November 2022 provides 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 allows researchers to learn not only about the expected magnitude of the value being created, as perceived by sophisticated market participants, but also about barriers or enablers of technology implementation that equity analysts have identified.

In this spirit, Eisfeldt et al. (2023) 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. In that study, firms are sorted into five value-weighted portfolios based on their generative AI exposures. The intensity of public attention to the ChatGPT release on platforms like Twitter peaked within 2 weeks after the release. It turns out 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 these two weeks following ChatGPT's release. **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" (AMH) portfolio.

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

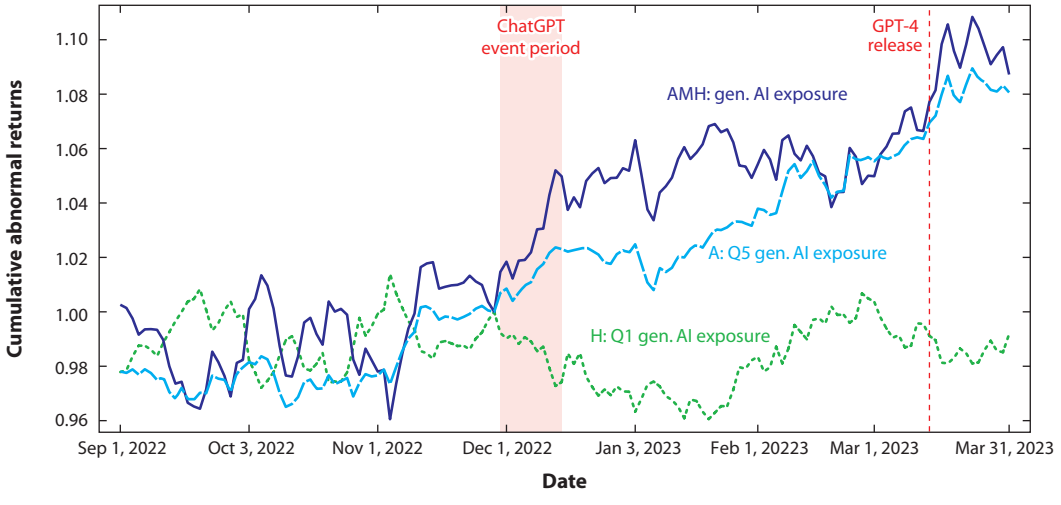

Figure 4

Cumulative abnormal returns by generative (gen.) artificial intelligence (AI) exposure. This figure plots the cumulative abnormal returns of value-weighted quintile portfolios sorted by firms' labor-based generative AI exposure. The graph shows the cumulative abnormal returns of the lowest-exposure quintile, the "Human" (H) portfolio (*green dotted line*); the highest-exposure quintile, the "Artificial" (A) portfolio (*light blue dashed line*); and the zero-investment portfolio that goes long A and shorts H, the "Artificial Minus Human" (AMH) portfolio (*dark blue solid line*). Cumulative returns are relative to November 29, 2022, the day before the release of ChatGPT. Market-adjusted daily abnormal returns are based on factor exposures computed over the 4-month period preceding the period shown in the graph. The underlying daily stock returns are from Yahoo Finance. The red 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. Figure adapted with permission from Eisfeldt et al. (2023).

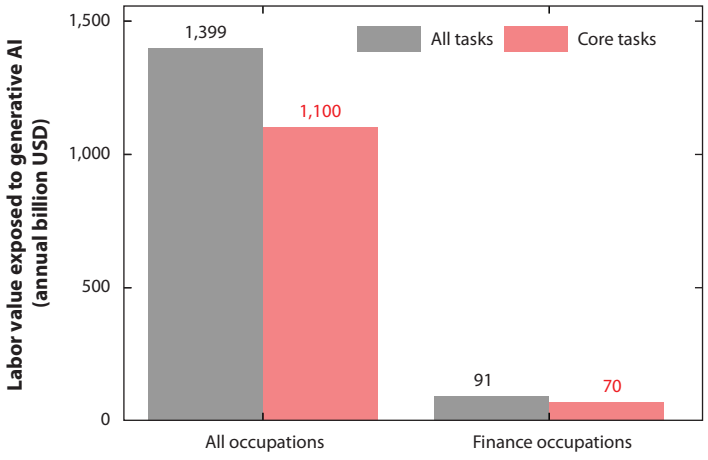

Figure 5

Back-of-the-envelope labor value potential of generative artificial intelligence (AI). The figure shows the results of a simple calculation of the potential labor value that could have been expected to be created around the release of ChatGPT. The calculation uses 2022 Bureau of Labor Statistics 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 large language model availability. The annual labor value potential is then computed as Employment × AnnualWage × 50% × ExposureShare. The resulting value of labor income exposed to productivity changes is in units of annual billion USD. The gray bar shows the exposed value for all tasks, and the red bar for core tasks only. The left side of the graph shows the exposed labor value for all US occupations, and the right side for US finance occupations only.

only apply to the marginal output of each worker, we can compute Employment × AnnualWage × 50% × ExposureShare 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 for expected labor value exposed to generative AI productivity changes is shown in **Figure 5**: Across all occupations, this back-of-the-envelope calculation suggests around $1.4 trillion in annual value created from all tasks and $1.1 trillion just from core tasks. In financial occupations alone, the corresponding numbers are $91 billion and $70 billion, which represents an outsized impact relative to the 2.5% of total US employment that the 3.6 million employees in financial occupations represent.[6] For comparison, the total market capitalization as of November 29, 2022, of the "Artificial" portfolio companies was around $10.2 trillion. 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 $1.4 trillion in annual labor value could generate the 4–5 percentage point increase in relative value of the "Artificial" portfolio that is shown in **Figure 4**.

While this 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 (at that time) was led by the public release of GPT-4 on March 14, 2023. As **Figure 4** shows, the 2 weeks before (the model's release was anticipated) and after the GPT-4 launch coincide with another run-up in the returns to the AMH portfolio.

[6]Total US employment was 148 million in 2022 according to the Bureau of Labor Statistics (2023).

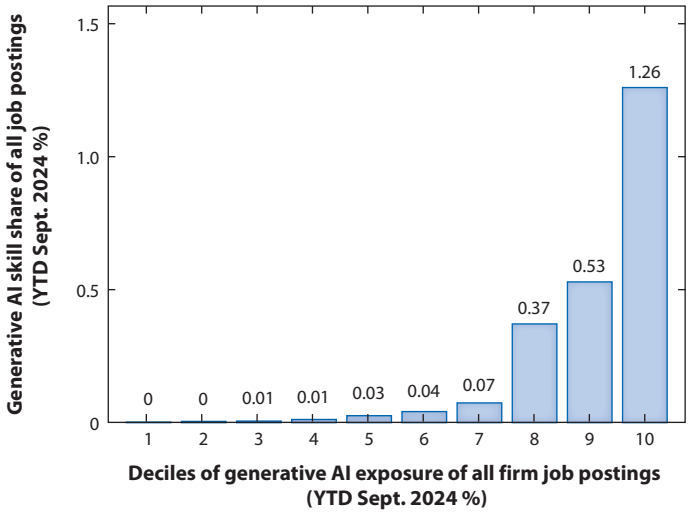

Figure 6

Generative artificial intelligence (AI) exposure and generative AI adoption by firms. This graph shows the relation between the share of job postings at the firm level that mention generative AI skills in the YTD September 2024 period, and the Eisfeldt et al. (2023) measure of generative AI exposure of the firm's job postings in the same period. Firms are sorted into deciles by exposure, weighted by total job postings, and each bar shows the total job posting–weighted mean share of all posted jobs at firms in that decile of exposure that mention generative AI. Figure adapted with permission from Schubert (2025).

Schubert (2025) provides evidence that the expected exposure to generative AI technologies based on a firm's employment composition is likely to have implications for firms even beyond the market reaction during the initial ChatGPT release. **Figure 6** shows that actual mentions of generative AI tools in a firm's job postings as of October 2023–September 2024 increase monotonically with the generative AI exposure of hiring in the same period [measured using the approach from Eisfeldt et al. (2023)]. This suggests that labor-based firm exposure to the technology continues to be a good proxy for firm behaviors in response to the generative AI boom.

2.4. 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 1990s 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. Some evidence with regard to such skill changes is provided by Schubert (2025), who finds that firms that adopted generative AI were more likely to upskill their hiring and to hire for roles that require greater decision-making skills.

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 that by Sun & Xiaolan (2019) and Falato et al. (2022), shows that intangible and human-related assets typically have more equity and less debt financing. For a general theory of collateral and capital structure, see also Rampini & Viswanathan (2013). For a broad review of the empirical literature on capital structure, see Graham & Leary (2011).

Generative AI is also expected to change firms' investment decisions. Veldkamp & Chung (2024) emphasize the role of data in forecasting; combined with such data, generative AI has the potential to improve capital budgeting decisions by reducing uncertainty. Crouzet & Eberly (2023) provide a framework that 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, how that decision might vary across industries and firm sizes, and the ultimate impact of those decisions represent 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 used by 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 of them is AI chip-maker 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 attention in 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 that 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, Longstaff & Santa-Clara (2008) provide 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) provide 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.

The wide availability of LLMs to financial investors may also change the speed and accuracy with which new information is incorporated into asset prices. For example, Hansen et al. (2024) suggest that synthetic LLM forecasting agents can effectively replicate some professional judgment in financial markets. Their study shows that LLMs can provide forecasts of macroeconomic variables based on real-time data that are at least as good as those that are provided with a lag by human professional forecasters. As evidence that LLM-generated information is already

being incorporated into financial markets, Sheng et al. (2024) show that hedge funds increasingly started to change their portfolio composition in response to the release of information of the kind that could be extracted from earnings calls using LLMs. However, this shift in technology does not necessarily equalize the playing field: Sheng et al. (2024) find that larger hedge funds, and those with AI-skilled employees, generate higher returns from incorporating this information into their portfolio adjustments than other investors. This contrasts with other findings—for example, Brynjolfsson, Li & Raymond (2023) found that generative AI use compresses the productivity distribution for customer support agents.

One potential resolution of these conflicting findings may be that generative AI effects on productivity depend on whether workers in an occupation require complementary skills to benefit from generative AI tools. Evidence in this direction is provided by Schubert (2025), who shows that generative AI adoption tends to be accompanied by an upskilling of a firm's workers, as jobs that use generative AI are more likely to require new hires to have experience and decision-making skills.

Additional evidence is needed of how generative AI affects knowledge disparities across diverse professional groups within the financial sector—and also how AI tools use by some professionals affect the productivity of other professionals with whom they interact. Moreover, more research is needed to understand to what degree widespread adoption of generative AI to gather and analyze information diminishes the importance of the Grossman & Stiglitz (1980) paradox and affects rents in financial markets. Li (2024) argues that there will be continued rewards for human creativity as an input for LLM training, even as LLM use decreases the cost of information processing, but empirical research is needed to evaluate how returns and spreads in financial markets are affected by the use of generative AI.

3. 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 economics in creative and exciting ways. Generative AI, LLMs, 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;[7] 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.

[7]For detailed advice on when and how to use these methods, e.g., to construct economic measures from historical documents, see Dell (2024).

Overall, we are optimistic that, for finance researchers, these new tools are more likely to automate—or increase the productivity of—supplemental tasks. 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 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).

Despite their power, these technologies should not be treated as magical black boxes—researchers must still carefully supervise and direct their analyses. This is particularly true because applications that are using LLMs for a particular purpose for the first time (some of which are discussed below) often cannot build on existing knowledge about how reliable these methods are in a new context. Moreover, just as for the ongoing trend of using big data in finance research (Goldstein, Spatt & Ye 2021), successfully deploying some of these new methodological approaches may require interdisciplinary collaborations with other academic disciplines, such as computer science.

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.[8]

We discuss a number of different applications of generative AI (and related deep learning techniques) in financial research in the following sections, covering the following topics: embeddings of high-dimensional data; text classification using LLMs; 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 also briefly discuss some potential uses of LLMs in teaching finance, which can indirectly affect the productivity of many academic researchers. For an overview of the different use cases that we discuss, see **Table 2**. We end the section with a list of subjective advice for researchers and reviewers.

3.1. Embeddings

In finance and economics, researchers often need to analyze large sets of high-dimensional data. For example, we might study earnings calls from US 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

[8]For a recent review of deep learning methods in economics as a good example of this type of contribution, see Dell (2024).

Table 2 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 & Sarkar (2020): 10-K filings embeddings ■ Kim, Ahn & Park (2024): labor market clusters
Text classification	■ What is the sentiment of financial text? ■ How to categorize documents? ■ What topics are discussed?	■ Chang et al. (2023): earnings call sentiment ■ Krockenberger et al. (2024): covenant violations ■ Caragea et al. (2020): FinTech patent classification
RAG	■ How to find relevant information in large data sets? ■ How to use LLMs for classification based on many potential sources? ■ How to retrieve similar documents from a corpus?	■ Bartik, Gupta & Milo (2023): housing regulation classification ■ Li et al. (2024): corporate culture based on analyst reports
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, Yang & Hashimoto (2024): novel research idea generation ■ Meincke et al. (2023): new product ideas ■ Ludwig & 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

Abbreviations: AI, artificial intelligence; FinTech, financial technology; LLMs, large language models; RAG, retrieval-augmented generation.

a lower-dimensional numerical representation that captures the key elements or essence of the text.

To achieve this, an important technique used in natural language processing (NLP) 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 nonnumerical 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 the following news headline examples: "Inflation Is High and Rising Fast," "High Inflation Is the Cause of Rising Bread 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 (the notion of "high-rises" in a real estate setting). Embeddings would solve this issue by assigning numerical vectors to each headline, where these vectors capture the meaning or context of the text. One could then cluster these vectors based on their similarity to determine which headlines belong together.

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. 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, eventually leading to the development of GPT, such as those used in models like ChatGPT. For an accessible in-depth exploration of transformers and embeddings, see Wolfram (2023), and for the introduction of GPT models, see Radford et al. (2018).

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, e.g., more traditional machine learning prediction models such as default prediction or sentiment analysis. For example, Chen, Kelly & Xiu (2022) generate embeddings of news articles using different LLMs and then compare their ability to predict the sign of stock returns, finding that LLMs outperform simpler NLP methods. We discuss text classification applications as one prominent class of use cases further in the following section.

Embeddings can also be used to create semantic axes, as proposed by An, Kwak & Ahn (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, Ahn & Park (2024) create a novel representation of labor markets using combined data describing industries, occupations, skills, and firms.

One early application of the semantic axis technique in the finance literature is by 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 predicts missing words in a sentence based on context. In this case, the authors predict assets within an ordered list of an investor's portfolio holdings 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 comovement 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.

3.2. 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 calls, annual reports, US Securities and Exchange Commission (SEC) filings, press releases, news headlines]. There is a large literature in which researchers use these text data as inputs to extract information about different categories that companies fall into. For example, a researcher might want to classify if a firm is likely to be impacted by a particular new regulation; therefore, they want to label which firms mention this regulation in their earnings calls and what the sentiment (positive/negative) of any such mention 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 predefined lists created by the researchers. These indicators were then often combined into overall sentiment scores.[9] However, word-based methods fall short when analyzing complex texts, since word meanings can vary significantly based on context. For example, distinguishing between a company "leaving an earnings slump behind" and one "entering an earnings slump" would be difficult, 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 then outputs either a vector of lower-dimensional numbers or a numerical classification label. BERT can be utilized directly as a pretrained 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

[9]For an overview of these techniques, see Loughran & McDonald (2020); for a broad reference on NLP, see Jurafsky & Martin (2025).

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 data set 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 data sets 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. For example, Araci (2019) develops a model called FinBERT that is trained specifically on 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 makes these models effective data-labeling tools. BERT's ability to process large chunks of text can be easily scaled to large volumes of unstructured text data.

One area where this model has been effectively applied in finance research is patent data analysis. 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 & Wang (2024) utilize 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. Acikalin et al. (2022) fine-tune a model on patent applications that were rejected after a court decision made it harder to patent business processes. They then use this fine-tuned transformer to assign a risk of being affected by the court decision to firms' portfolio of patents that were granted before the court decision.

Researchers have also developed models similar to BERT to analyze financial news, regulatory filings, analyst reports, and call transcripts. One example includes CovenantAI, developed by Krockenberger et al. (2024). Specifically, the study builds a database of covenant violations using 10-K and 10-Q reports filed by firms: First, it applies 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 then trains a classifier to identify text semantics that suggest covenant violations.[10]

Chen & Sarkar (2020) apply an off-the-shelf BERT model to firm 10-K filings with the SEC, and extract the mean of each firm's texts' embeddings. They show that clusters of firms based on this text information are more differentiated with regard to firm fundamentals from other clusters than groups defined using existing industry definitions. 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

[10]For more information on MPNet, see Song et al. (2020).

asset pricing models or help researchers to find appropriate control groups in empirical studies. Bonne et al. (2022) use 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. Similarly, Hoberg & Phillips (2016) develop a successful early approach to industry classification based on text by using the cosine similarity between vector representations of the words that firms use in the business descriptions of their annual reports.

More recently, the latest generation of LLMs, 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 Section 3.7.

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), which is derived from massive data encountered during their initial training and their large number of parameters, enables them to be used off-the-shelf in many settings. That is, they can perform classification analyses without any need for the researcher to invest costly effort upfront in creating large training data themselves.[11]

One recent example of such an application is Chang et al. (2023), 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, which is then used to predict 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 & 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 particularly true for more complex news sources. Similarly, Chen et al. (2023) show that a GPT-3.5–level model can predict aggregate US stock returns in the month following news headlines, while 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

[11] However, as we argue in Section 3.7.3, researchers may often still want to obtain or create at least a small sample with ground truth labels in order to validate the accuracy of the output from an LLM.

articles, patent texts, speeches by policy makers, 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 is key to the ultimate success of this type of research (see, for example, Harvey 2017, 2019).

3.3. Retrieval-Augmented Generation (RAG)

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 the RAG technique.

RAG is an analytical approach that combines the strengths of LLMs 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 ability to focus on specific subsets of retrieved data, RAG improves the ability to extract information from complex data sets.

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 an 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 search and retrieval based on how similar the chunks are to a given query.
4. Query processing: When a user asks a question, it is 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, additional filtering is applied to eliminate any chunks that do not meet specific criteria, such as those missing particular keywords. 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 an 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 would begin 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 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 an additional 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 be able to compile a data set that indicates which firms express environmental concerns and in which years.

A recent application of RAG is found in the real estate literature, where Bartik, Gupta & Milo (2023) utilize this approach to classify municipal housing regulations based on a comprehensive nationwide database of zoning codes. They extract relevant text segments from this database and evaluate them using an 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 achieves 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, where it is necessary to find the metaphorical needle in the haystack. Li et al. (2024) provide an application of RAG where this ability to retrieve fragmented bits of information matters, as they analyze corporate culture through culture-related excerpts from analyst reports. When encountering segments of text that seem to lack sufficient context, they allow the model to retrieve additional related segments from the same report. This targeted retrieval strategy helps ensure semantic coherence of the analyzed text while staying within token limits. This approach may also reduce inference costs relative to an approach that indiscriminately includes the full report text in the analysis.

In contrast, RAG 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. For instance, 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.

3.4. 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. They use the model to simulate individual responses to treatments and use these 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.

This approach can also be useful in financial service settings: For example, LLMs could be used to generate tailored financial advice based on the demographics of a human counterpart. However, this type of use is likely to be fraught with ethical issues, and it might require updating current regulatory frameworks to grapple with the notion of fiduciary duty as applied to LLMs (Lo & Ross 2024).

When deciding whether to use an LLM instead of human responses, researchers should consider three factors: how well LLMs match human responses in precision and bias, the replicability and robustness of LLM-based methods, and relative deployment costs. 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—bias and replicability—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, and that LLM-generated beliefs replicate behavioral biases exhibited by the human respondents. Zarifhonarvar (2024) finds that LLMs' inflation expectations in simulated survey experiments closely mirror patterns observed in human surveys, including responses to information treatments and even partisan biases when prompted after assuming different political personas.

These applications highlight a key LLM strength—their ability to assume different personas that mirror responses from specific demographic groups. For example, 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 a particular age, income, or gender. Some differences exist, however: The same study also finds that preferences expressed in LLM responses are more likely to obey transitivity compared to responses by human subjects. In a study focusing on financial institution stability, Kazinnik (2023) employs GPT-4 to simulate depositor behavior during bank runs and argues that the results align with aggregate patterns for how different groups respond during financial crises. Studies like this may provide an important link between empirical analyses on aggregate data in finance and field experiments evaluating individual preferences—either by generating new hypotheses for heterogeneity tests or as a prognosis for what is reasonable to expect when well-understood preferences interact with complex real-world scenarios that cannot be studied in an experimental setting.

An important methodological takeaway from the range of studies discussed above is that it is important to determine in advance whether the generative AI model in question has the same behavioral biases as humans. Depending on the application, replicating human biases may be desirable (e.g., to get realistic predictions of responses) or undesirable (e.g., when using LLMs as advisors or to capture the responses of rational agents). The degree to which an LLM exhibits biases cannot be stated unconditionally: Ross, Kim & Lo (2024) show that whether

state-of-the-art models exhibit human behavioral biases or rational choices may actually depend on the exact application in which they are deployed.

We expect that LLM-based methods can also be used in empirical applications to generate control variables in the spirit of propensity weights. If treatment is not random, propensity scores assigned by an LLM could be used to match treated units to comparable units in the control group or to reweight 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 also be used as a synthetic control group in some settings.

For theoretical papers, LLMs can generate realistic examples of heterogeneous agents, either for illustration or as inputs for outcome simulations. 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 changes not only in policy rates but also in the broader policy regime and softer aspects of forward guidance by policy makers.

3.5. 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. 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: Meincke et al. (2023) show that GPT-4 generated new product ideas with minimal prompting that, on average, elicited a higher purchase intent than those that students at a top-ranked master of business administration (MBA) program generated. Si, Yang & Hashimoto (2024) show that a system of LLMs can generate more novel research ideas than expert NLP 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. Moreover, even the LLM-based system in this study leaves a role for humans as research designers, as it requires careful chaining of LLMs to generate and preselect ideas as well as structuring a write-up for human evaluation. This suggests that LLMs can play an important role in complementing human researchers in the

ideation process: While LLMs can generate many candidate ideas, human researcher judgment will still be needed to select good ideas and design the idea-creation process for the LLM to follow.

LLMs can also be used in developing ideas for how to test a given research hypothesis: Han (2024) shows how LLMs can be used to systematically search for instrumental variables through helping the researcher with counterfactual reasoning and simulating potential determinants of outcomes in different contexts. Where researcher ingenuity and knowledge of institutional details used to be necessary to uncover natural experiments in the past, LLMs' encyclopedic knowledge can perhaps accelerate the creative process.

Sometimes, humans and algorithms can work in hybrid teams, where humans generate hypotheses that assist in understanding the black box analyses generated by machine learning or AI. Ludwig & Mullainathan (2024) provide an example: Using a convolutional neural network (CNN), which is a specialized type of deep learning model designed primarily for processing grid-like data (such as images), they found that mug shots of defendants predicted pretrial jail decisions. Then, to understand which facial features mattered for this prediction, they used the model to morph images to become more jail-able and then 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 mug shots that could be tested for their predictive power. Another fruitful example of a hybrid team is by Batista & Ross (2024), who generate 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 & Jaravel (2024) develop and validate an LLM-based tool for conducting qualitative interviews. They show that it can deliver high-quality responses in applications that, for instance, survey humans to elicit political preferences or discuss a complex topic such as having a meaningful life. These automated 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 & Holweg (2024), who argue that AI cannot generate genuinely new knowledge and study how human reasoning is distinct from computer prediction.

The possibility of LLMs being able to automate large parts of the research process may be exciting for quantitative researchers at hedge funds, as it might allow them to better scale the process of testing different signals. However, it portends substantial challenges for the academic publication process. For example, Novy-Marx & Velikov (2024) generate 288 research papers on stock return predictability by first data mining accounting data for predictors that pass statistical tests and then using LLMs to write entire papers with plausible hypotheses and economic stories that rationalize each predictor ex post. The potential for large-scale generation of papers raises serious questions for editors with regard to the need for a proof of work by human researchers and the ability of referees to effectively identify true results based on conventional standards for review.

3.6. Teaching Finance with Generative AI

One channel through which generative AI makes academic researchers more productive is by making more effective use of their time devoted to teaching. In this section, we mention just a few potential LLM applications that we have found useful in teaching MBA students:

1. Generative AI can help to convert loose notes into scripts in LaTeX or Markdown formats that can then be compiled into slides to save time in class preparation.
2. The more advanced versions of LLMs available at the time of this writing (e.g., OpenAI's o1 model) can check derivations on slides for typos and can quickly generate and test problem sets that are aimed at reinforcing particular concepts.
3. A conversation with a chatbot can also provide helpful practice for classroom discussions and feedback or inspiration for how to engage students.
4. Models can execute code to quickly produce graphs illustrating equations derived in class.
5. Chatbots can 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.
6. Faculty can 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 allowed refinement of exam questions before releasing the exam to students. A related advantage is that model-generated responses can be edited for use as example answers or as starting points for developing grading schemes.
7. As Mollick et al. (2024) note, LLMs can be used to design interactive simulations for business school students that might lead to more engaging interactions with case studies covered in class.
8. Instructors can also ask students to design their own LLM-based chatbots: For example, we have used an activity that required students to consider the potential use cases for generative AI in the workplace and to then carefully design prompts and constraints on the chatbot functionality to reliably elicit a desired behavior from—at times mercurial—model personas.

3.7. Practical Considerations for Using Generative AI in Research

As the methodologies discussed above have only recently been developed and are often still in the process of being used in finance research for the first time, there may 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.

3.7.1. Obtaining numerical scores from LLMs. In many applications, one may want not only to assign a binary label to a text [e.g., "CEO talks about Environmental, Social, and Governance (ESG) topics = yes/no"] but also 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. (2023) 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 when evaluating a text (like an earnings call transcript), the LLM is not fully able to compare it to other texts, either within the sample or more broadly. 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 of the requested response, which could be driven by the words used in other (potentially erroneous) descriptions of similar data sets (for instance, other papers about earnings calls) rather than by a process resembling an evaluation of the data at hand relative to training data. To make that concrete: If most studies in the training data report CEOs that use the word "climate change" to have greater than average levels of concern 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., "US earnings calls from the last 10 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 versus 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 & Jolley (2024): 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 applications of LLMs in financial research that involve the categorization or clustering of entities, e.g., the clustering of firms into groups based on their corporate communication. One good example of this approach is by 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 multilevel 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 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 also no reason why a model's implicit categories would correspond to the categories that a finance researcher cares about. For example, the study finds that the LLM sometimes applies "unusual" labels when an earnings call discusses only 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 (e.g., see 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 & McDonald 2020), because LLMs are more likely to reliably understand different usages and context. Thus, in spite of their flaws, using LLMs in these settings holds great promise: The relevant benchmarks for comparison are labels and

scores applied by human evaluators and older NLP methods that often have serious shortcomings of their own.

3.7.2. 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) ask the LLM to provide an explanation for why a task was classified into a particular rubric for whether generative AI can be applied to it. This explanation helps the researcher 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" (p. 4). 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 would 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.[12]

3.7.3. Implementing RAG. RAG approaches typically begin by preprocessing texts—cleaning them and dividing them into retrievable chunks. Removing metadata 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 run the risk not only 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 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 LLMs, 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

[12] "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" (Wang et al. 2023, p. 4).

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 earnings 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.

Prompt performance can vary significantly based on subtle factors: exact wording, choice of foundation model, and even the order of examples. Therefore, 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 enables replication and helps the profession understand how different choices affect system performance as we gather more examples.

4. CONCLUSION

Generative AI represents a major technology shock in the research and practice of finance. Eisfeldt et al. (2023) show 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 twentieth century and the advances in computing power in the early twenty-first 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.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

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

LITERATURE CITED

Acemoglu D, Autor D. 2011. Skills, tasks and technologies: implications for employment and earnings. In *Handbook of Labor Economics*, Vol. 4, ed. D Card, O Ashenfelter, pp. 1043–171. Elsevier

Acikalin U, Caskurlu T, Hoberg G, Phillips GM. 2022. Intellectual property protection lost and competition: an examination using large language models. Work. Pap. 4023622, Tuck School of Business

Aldasoro I, Gambacorta L, Korinek A, Shreeti V, Stein M. 2024. *Intelligent financial system: How AI is transforming finance.* Tech. Rep., Bank for International Settlements

An J, Kwak H, Ahn YY. 2018. SemAxis: a lightweight framework to characterize domain-specific word semantics beyond sentiment. Preprint, arXiv:1806.05521[cs.CL]

Araci D. 2019. FinBERT: financial sentiment analysis with pre-trained language models. Preprint, arXiv:1908.10063[cs.CL]

Babina T, Fedyk A, He AX, Hodson J. 2023. Artificial intelligence and firms' systematic risk. SSRN Work. Pap. 4868770

Babina T, Fedyk A, He A, Hodson J. 2024. Artificial intelligence, firm growth, and product innovation. *J. Financ. Econ.* 151:103745

Bartik A, Gupta A, Milo D. 2023. The costs of housing regulation: evidence from generative regulatory measurement. SSRN Work. Pap. 4627587

Batista RM, Ross J. 2024. Words that work: using language to generate hypotheses. SSRN Work. Pap. 4926398

Beckmann L, Beckmeyer H, Filippou I, Menze S, Zhou G. 2024. Unusual financial communication: ChatGPT, earnings calls, and financial markets. Res. Pap. 2024/02, Olin Business School Center for Finance and Accounting

Bertomeu J, Lin Y, Liu Y, Ni Z. 2023. Capital market consequences of generative AI: early evidence from the ban of ChatGPT in Italy. SSRN Work. Pap. 4452670

Bonne G, Lo AW, Prabhakaran A, Siah KW, Singh M, et al. 2022. An artificial intelligence-based industry peer grouping system. *J. Financ. Data Sci.* 4(2):9–36

Brynjolfsson E, Li D, Raymond LR. 2023. Generative AI at work. NBER Work. Pap. 31161

Bureau of Labor Statistics. 2023. *Occupational employment and wages—May 2022.* News Release USDL-23-0794, Apr. 25. Department of Labor. **https://www.bls.gov/news.release/archives/ocwage_04252023.pdf**

Bybee JL. 2023. The ghost in the machine: generating beliefs with large language models. Preprint, arXiv:2305.02823[econ.GN]

Caragea D, Chen M, Cojoianu T, Dobri M, Glandt K, Mihaila G. 2020. Identifying FinTech innovations using BERT. In *2020 IEEE International Conference on Big Data*, pp. 1117–26. IEEE

Chang A, Dong X, Martin X, Zhou C. 2023. AI democratization, return predictability, and trading inequality. SSRN Work. Pap. 4543999

Chen J, Sarkar S. 2020. A semantic approach to financial fundamentals. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pp. 22–26

Chen J, Tang G, Zhou G, Zhu W. 2023. ChatGPT, stock market predictability and links to the macroeconomy. SSRN Work. Pap. 4660148
Chen MA, Wang JX. 2024. Displacement or augmentation? The effects of AI on workforce dynamics and firm value. SSRN Work. Pap. 4787286
Chen Y, Kelly BT, Xiu D. 2022. Expected returns and large language models. SSRN Work. Pap. 4416687
Cochrane JH, Longstaff FA, Santa-Clara P. 2008. Two trees. *Rev. Financ. Stud.* 21(1):347–85
Crouzet N, Eberly J. 2023. Rents and intangible capital: A Q+ framework. *J. Finance* 78(4):1873–916
Dell M. 2024. Deep learning for economists. NBER Work. Pap. 32768
Devlin J, Chang MW, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint, arXiv:1810.04805[cs.CL]
Dhinakaran A, Jolley E. 2024. Why you should not use numeric evals for LLM as a judge. *Arize AI Blog*, Mar. 3. **https://arize.com/blog-course/numeric-evals-for-llm-as-a-judge/**
Duffie D, Foucault T, Veldkamp L, Vives X. 2022. *Technology and finance: the future of banking.* Rep., Centre for Economic Policy Research, London
Eisfeldt AL, Schubert G, Zhang MB, Taska B. 2023. Generative AI and firm values. SSRN Work. Pap. 4436627
Eloundou T, Manning S, Mishkin P, Rock D. 2023. GPTs are GPTs: an early look at the labor market impact potential of large language models. Preprint, arXiv:2303.10130[econ.GT]
Falato A, Kadyrzhanova D, Sim J, Steri R. 2022. Rising intangible capital, shrinking debt capacity, and the US corporate savings glut. *J. Finance* 77(5):2799–852
Fedyk A, Kakhbod A, Li P, Malmendier U. 2024. ChatGPT and perception biases in investments: an experimental study. SSRN Work. Pap. 4787249
Felin T, Holweg M. 2024. Theory is all you need: AI, human cognition, and decision making. SSRN Work. Pap. 473765
Gabaix X, Koijen RS, Richmond R, Yogo M. 2023. Asset embeddings. SSRN Work. Pap. 4507511
Geiecke F, Jaravel X. 2024. Conversations at scale: robust AI-led interviews with a simple open-source platform. Discuss. Pap. 19705, Centre for Economic Policy Research
Ghazal TM. 2021. Performances of K-means clustering algorithm with different distance metrics. *Intell. Autom. Soft Comput.* 30(2):735–42
Giglio S, Kelly B, Xiu D. 2022. Factor models, machine learning, and asset pricing. *Annu. Rev. Financ. Econ.* 14:337–68
Goldstein I, Spatt CS, Ye M. 2021. Big data in finance. *Rev. Financ. Stud.* 34(7):3213–25
Graham JR, Leary MT. 2011. A review of empirical capital structure research and directions for the future. *Annu. Rev. Financ. Econ.* 3:309–45
Grossman SJ, Stiglitz JE. 1980. On the impossibility of informationally efficient markets. *Am. Econ. Rev.* 70(3):393–408
Han S. 2024. Mining causality: AI-assisted search for instrumental variables. Preprint, arXiv:2409.14202[econ.EM]
Hansen AL, Horton JJ, Kazinnik S, Puzzello D, Zarifhonarvar A. 2024. Simulating the survey of professional forecasters. SSRN Work. Pap. 5066286
Harvey CR. 2017. Presidential address: the scientific outlook in financial economics. *J. Finance* 72(4):1399–440
Harvey CR. 2019. Replication in financial economics. SSRN Work. Pap. 3409466
Hewitt L, Ashokkumar A, Ghezae I, Willer R. 2024. Predicting results of social science experiments using large language models. Work. Pap., Stanford University, New York University
Hoberg G, Phillips G. 2016. Text-based network industries and endogenous product differentiation. *J. Political Econ.* 124(5):1423–65
Horton JJ. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? NBER Work. Pap. 31122
Jha M, Qian J, Weber M, Yang B. 2024. ChatGPT and corporate policies. NBER Work. Pap. 32161
Jiang W, Li T. 2024. Corporate governance meets data and technology. SSRN Work. Pap. 4746141
Jurafsky D, Martin JH. 2025. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models. Online manuscript, 3rd draft, Jan. 12. **https://web.stanford.edu/jurafsky/slp3/**

Kazinnik S. 2023. Bank run, interrupted: modeling deposit withdrawals with generative AI. SSRN Work. Pap. 4656722
Kelly B, Xiu D. 2023. Financial machine learning. *Found. Trends Finance* 13(3–4):205–363
Kim S, Ahn YY, Park J. 2024. Labor space: a unifying representation of the labor market via large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 2441–51. Association for Computing Machinery
Korinek A. 2023. Generative AI for economic research: use cases and implications for economists. *J. Econ. Lit.* 61(4):1281–317
Krockenberger VS, Saunders A, Steffen S, Verhoff PM. 2024. CovenantAI—new insights into covenant violations. SSRN Work. Pap. 4640653
Li J. 2024. Is generative AI an existential threat to human creatives? Insights from financial economics. Preprint, arXiv:2407.19586[cs.HC]
Li K, Mai F, Shen R, Yang C, Zhang T. 2024. Dissecting corporate culture using generative AI–insights from analyst reports. SSRN Work. Pap. 4558295
Lo AW, Ross J. 2024. Can ChatGPT plan your retirement? Generative AI and financial advice. SSRN Work. Pap. 4722780
Lopez-Lira A, Tang Y. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. SSRN Work. Pap. 4412788
Loughran T, McDonald B. 2020. Textual analysis in finance. *Annu. Rev. Financ. Econ.* 12:357–75
Ludwig J, Mullainathan S. 2024. Machine learning as a tool for hypothesis generation. *Q. J. Econ.* 139(2):751–827
Meincke L, Girotra K, Nave G, Terwiesch C, Ulrich KT. 2023. Using large language models for idea generation in innovation. SSRN Work. Pap. 4526071
Mollick E. 2024. *Co-Intelligence: Living and Working with AI*. Penguin Random House
Mollick E, Mollick L, Bach N, Ciccarelli L, Przystanski B, Ravipinto D. 2024. AI agents and education: simulated practice at scale. Preprint, arXiv:2407.12796[cs.CY]
Nagel S. 2021. *Machine Learning in Asset Pricing*. Princeton Lectures in Finance. Princeton University Press
Novy-Marx R, Velikov M. 2024. AI-powered (finance) scholarship. SSRN Work. Pap. 5103553
Radford A, Narasimhan K, Salimans T, Sutskever I, et al. 2018. Improving language understanding by generative pre-training. Preprint
Rampini AA, Viswanathan S. 2013. Collateral and capital structure. *J. Financ. Econ.* 109(2):466–92
Rosen S. 1981. The economics of superstars. *Am. Econ. Rev.* 71(5):845–58
Ross J, Kim Y, Lo AW. 2024. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. Preprint, arXiv:2408.02784[cs.CL]
Schubert G. 2025. Organizational technology ladders: remote work and generative AI adoption. SSRN Work. Pap. 5094265
Sheng J, Sun Z, Yang B, Zhang AL. 2024. Generative AI and asset management. SSRN Work. Pap. 4786575
Si C, Yang D, Hashimoto T. 2024. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. Preprint, arXiv:2409.04109[cs.CL]
Song K, Tan X, Qin T, Lu J, Liu TY. 2020. MPNet: masked and permuted pre-training for language understanding. *Adv. Neural Inform. Proc. Syst.* 33:16857–67
Sun Q, Xiaolan MZ. 2019. Financing intangible capital. *J. Financ. Econ.* 133(3):564–88
Tranchero M, Brenninkmeijer CF, Murugan A, Nagaraj A. 2024. Theorizing with large language models. NBER Work. Pap. 33033
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. *Adv. Neural Inform. Proc. Syst.* 30
Veldkamp L, Chung C. 2024. Data and the aggregate economy. *J. Econ. Lit.* 62(2):458–84
Wang P, Li L, Chen L, Cai Z, Zhu D, et al. 2023. Large language models are not fair evaluators. Preprint, arXiv:2305.17926[cs.CL]
Wolfram S. 2023. What is ChatGPT doing … and why does it work? *Stephen Wolfram Writings*, Febr. 14. **https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/**
Zarifhonarvar A. 2024. Experimental evidence on large language models. SSRN Work. Pap. 4825076