Hornuf, Lars; Streich, David J.; Töllich, Niklas

**Working Paper**

# Making GenAI Smarter: Evidence from a Portfolio Allocation Experiment

CESifo Working Paper, No. 11862

**Provided in Cooperation with:**
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

*Suggested Citation:* Hornuf, Lars; Streich, David J.; Töllich, Niklas (2025) : Making GenAI Smarter: Evidence from a Portfolio Allocation Experiment, CESifo Working Paper, No. 11862, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/319230

CESifo WORKING PAPERS

11862
2025

April 2025

# Making GenAI Smarter: Evidence from a Portfolio Allocation Experiment

*Lars Hornuf, David J. Streich, Niklas Töllich*

CESifo

## **Impressum**:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.ifo.de/en/cesifo/publications/cesifo-working-papers
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.ifo.de/en/cesifo/publications/cesifo-working-papers

# Making GenAI Smarter: Evidence from a Portfolio Allocation Experiment*

Lars Hornuf† David J. Streich‡ Niklas Töllich§

April 30, 2025

**Abstract**

Retrieval-augmented generation (RAG) has emerged as a promising way to improve task-specific performance in generative artificial intelligence (GenAI) applications such as large language models (LLMs). In this study, we evaluate the performance implications of providing various types of domain-specific information to LLMs in a simple portfolio allocation task. We compare the recommendations of seven state-of-the-art LLMs in various experimental conditions against a benchmark of professional financial advisors. Our main result is that the provision of domain-specific information does not unambiguously improve the quality of recommendations. In particular, we find that LLM recommendations underperform recommendations by human financial advisors in the baseline condition. However, providing firm-specific information improves historical performance in LLM portfolios and closes the gap with human advisors. Performance improvements are achieved through higher exposure to market risk and not through an increase in mean-variance efficiency within the risky portfolio share. Notably, portfolio risk increases primarily for risk-averse investors. We also document that quantitative firm-specific information affects recommendations more than qualitative firm-specific information, and that equipping models with generic finance theory does not affect recommendations.

**Keywords:** Generative artificial intelligence · large language models · domain-specific information · retrieval-augmented generation · portfolio management · portfolio allocation
**JEL-Classification:** G00 · G11 · G40

*Lars Hornuf and David Streich gratefully acknowledge funding from the German Research Foundation (DFG, project number 552870289).

†Dresden University of Technology, Germany. E-Mail: lars.hornuf@tu-dresden.de

‡Catholic University Eichstaett-Ingolstadt, Germany. E-Mail: david.streich@ku.de

§Catholic University Eichstaett-Ingolstadt, Germany. E-Mail: niklas.toellich@stud.ku.de

# 1 Introduction

Large language models (LLMs) have proven remarkably useful in the finance domain. Studies mostly drawing on OpenAI's cutting-edge GPT models demonstrate that LLMs reproduce financial knowledge to correctly answer common financial literacy questions (Niszczota and Abbas, 2023) and pass licensing exams (Fairhurst and Greene, 2024). LLMs have also been shown to correctly predict stock returns and firm-level investment from conference call transcripts, news headlines, and corporate disclosures (Jha et al., 2024; Lopez-Lira and Tang, 2023; Kim et al., 2023). Most recently, a series of studies has documented that LLMs may in fact be well-suited to provide portfolio management recommendations (Fieberg et al., 2023, 2024; Pelster and Val, 2024; Oehler and Horn, 2024; Ko and Lee, 2024).

Studies on the capabilities of LLMs in financial applications have thus far investigated general-purpose models that are not specifically designed for tasks in the finance domain. It stands to reason that performance could be further enhanced by adding domain-specific knowledge to pre-trained LLMs (Wu et al., 2023; Lo and Ross, 2024), which can generally be achieved in one of three ways. First, models can be pre-trained from scratch using both domain-specific and generic training data. Second, domain-specific knowledge can be fed into the models through *fine-tuning*, which refers to the re-training of existing generic models' weights based on domain-specific downstream datasets. Third, knowledge can be injected through *retrieval-augmented generation* (RAG) by implicitly adding domain-specific knowledge to each request without altering the existing weights. RAG has emerged as the most promising way to implement domain-specific knowledge, as it is considerably cheaper than training a finance-specific model from scratch (Wu et al., 2023) and avoids some of the pitfalls (e.g., hallucinations) of fine-tuning existing models using finance-specific downstream datasets.

This study investigates how adding domain-specific investment knowledge affects the performance of several off-the-shelf LLMs in a simple portfolio allocation task. The task involves assigning portfolio weights to a pre-determined and tractable universe of securities in accordance with relevant investor characteristics. We assess how performance in that task differs when we add: (i) basic investment theory (as taught in undergraduate finance courses); (ii) quantitative firm-level indicators (proxies for commonly used asset pricing risk factors); (iii) qualitative information on previous firm performance (as described in the firms' most recent 10-K filings); and (iv) all three additional information types. These treatments reflect the most commonly used information types contained in pre-trained (Wu et al., 2023) and fine-tuned (Yang et al., 2023) LLMs in the finance context, allowing us to offer practical recommendations regarding the effectiveness of adding the various types of information. We assess the change in LLM performance relative to a baseline condition in which no additional information was provided. For a realistic human benchmark, we also obtain portfolio recommendations for the same portfolio allocation task from actual US financial advisors through an incentivized online survey.

Our key results can be summarized as follows. First, LLM recommendations are largely unaffected by the provision of basic investment theory, but do respond to firm-specific information. This result is reassuring, as it is highly likely that basic financial concepts are already included in the training data of most state-of-the-art models (cf. Fieberg et al., 2024).

Second, in the baseline condition, LLMs recommend portfolios with a significantly lower proportion of risky assets than the human financial advisors. Adding firm-specific information increases the proportion of risky assets to the levels observed in human portfolios. This finding permits two interpretations. One possibility is that providing firm-specific information increases LLMs' tendency to recommend riskier securities simply because *more* firm-specific information is available. In this case, the LLMs' recommendations suggest a form of *ambiguity aversion* (Ellsberg, 1961). Alternatively, LLMs may interpret the inclusion of firm-specific information as an implicit request to include more stocks in a portfolio (*desirability bias,* cf. Zhao et al., 2023; Lou and Sun, 2024).

Third, LLM portfolios are less likely than human portfolios to include stocks with high investor attention. Adding firm-specific information further decreases the tendency to recommend high-attention stocks. This result suggests that—intuitively and in line with human decision-makers (Barber and Odean, 2008)—LLMs tend to pick stocks with higher investor attention in the absence of stock-specific information. Providing stock-specific information decreases this tendency.

Fourth, when quantitative information is provided, LLMs emphasize ESG scores and follow a momentum strategy. When qualitative information is provided, LLM portfolios shift towards stocks with more positive sentiment in the provided 10-K filing sections. When both qualitative and quantitative information is provided, LLM recommendations are more affected by quantitative information. This finding reflects the fact that numbers may be better suited for direct comparison of a pair of securities than textual information.

Fifth, LLMs tailor portfolios more to an investor's ESG preferences when firm-specific information is provided. Again, this result could either be driven by more precise knowledge of the individual securities' ESG scores or by LLMs interpreting the provision of ESG scores as a request to use them as decision criteria.

Finally, LLM portfolios underperform human portfolios and a naive 1/N diversification strategy. Adding firm-specific information significantly improves historical performance, closing the gap between LLM and human recommendations. Performance improvement through firm-specific information is achieved through an increase in the proportion of risky assets and not through an increase in efficiency within the risky portion of the portfolios. The performance-enhancing effect of firm-specific information is not fully replicated in a short-window out-of-sample test.

We contribute to the emerging literature on LLM capabilities by investigating how injecting domain-specific information impacts LLM performance in a controlled environment using an applied task with no objectively correct solution. Although it seems intuitive that adding domain expertise improves performance, findings regarding the performance implications of

fine-tuning models and including a RAG pipeline should be taken with a grain of salt. This is because researchers and developers, who may have an interest in reporting favorable results, have considerable freedom in the types of tests and benchmark models they use in their studies.[1] Thus, we design an applied portfolio allocation experiment that tests applied skill rather than reproduction of factual knowledge (Fairhurst and Greene, 2024; Niszczota and Abbas, 2023). By controlling the exact information to be provided to LLMs (rather than leaving the selection of information to a *retriever* model, as is typical in RAG applications), we can directly link changes in portfolio recommendations to the injected information. Finally, by investigating several recently released LLMs, our capability results can be generalized to a wider range of models. Taken together, our results suggest that adding firm-specific information improves the suitability and performance of LLM recommendations, but also substantially increases portfolio risk for risk-averse investor profiles, which may result in a misalignment of portfolio risk and investor preferences.

The remainder of the paper is structured as follows. Section 2 provides background on the various methods of achieving domain expertise in LLMs, as well as an overview of existing empirical studies on the performance impact of domain-specific knowledge. Section 3 details the experimental design. Section 4 describes how we construct key variables for our analyses and reports descriptive statistics. Section 5 elaborates on the main results. Section 6 discusses the implications of our findings and concludes.

## 2 Background

There are three ways of achieving domain-specific knowledge in LLMs. First, models can be pre-trained from scratch using domain-specific training data. This approach is by far the most comprehensive way of injecting domain-specific information. For example, BloombergGPT (Wu et al., 2023) combines general-purpose training data with proprietary domain-specific data from Bloomberg's vast data sources to train a model specifically designed for natural language processing tasks in the financial domain. Similarly, FinBERT (Yang et al., 2020; Liu et al., 2021) is a domain-specific adaptation of the BERT model, one of the earliest LLMs, developed by Google (Devlin, 2018). Second, an existing pre-trained model (e.g., LLaMA) can be fine-tuned for a specific context. This is achieved through a continuation of training using a smaller downstream dataset after pre-training to adjust the model to the particular task or knowledge domain (Gururangan et al., 2020). In the process, some or all of the model's parameters are adjusted to the introduced data depending on the specific fine-tuning method (Jurafsky and Martin, 2024). Third, domain-specific knowledge can be directly injected into the generation process by providing it as context to the prompt. This is operationalized

[1] We discuss the results of existing studies on the performance implications of knowledge injection in section 2).

through retrieval-augmented generation (RAG), which was first introduced by Lewis et al. (2020).

In its most basic form, RAG is a two-phase process (Gao et al., 2023). In the first step, knowledge relevant to the query is fetched from external databases using a retriever model (*retrieval*). In the second step, the retrieved knowledge is used to enhance the user query and the LLM generates output given the information-rich prompt (*generation*, cf. Gao et al., 2023; Ram et al., 2023; Muhlgay et al., 2023). This approach facilitates the introduction of domain-specific knowledge to LLMs without additional training, which keeps the parameter weights of the original model unaltered and allows for swifter information updates. Our experimental design most closely resembles the augmented generation part of RAG, given that we define the context data to add to requests rather than have it retrieved by a retriever model.

There is mixed evidence on the performance of pre-trained domain-specific LLMs compared to benchmark models. Perhaps not surprisingly, the models' developers report substantial performance improvements over benchmark models (e.g., Wu et al., 2023; Liu et al., 2021, for finance-specific LLMs). However, large general-purpose models such as GPT-3.5 and GPT-4 have been shown to outperform domain-specific models such as BloombergGPT and FinBERT in financial sentiment analysis, the very task they were designed to excel in (Li et al., 2023a; Zhang et al., 2024). Because pre-training domain-specific models from scratch entails considerable development costs and foregoes the competitive advantage of general-purpose model developers, recent efforts have focused on fine-tuning existing models and adding context through RAG.

Table 1 lists recent studies investigating the performance implications of injecting domain-specific information via fine-tuning and RAG, along with their domains of application, injection type (fine-tuning, RAG, or both), author type (academic institution, model developer, or both), and whether or not the new model displays a performance improvement over its base model (i.e., the existing model used for fine-tuning or RAG) and unrelated benchmark models (i.e., other general-purpose or domain-specific models).

The results suggest that fine-tuning models on domain-specific information improves performance over their base models across domains. For example, InvestLM, an instruction-tuned LLM based on LLaMA, demonstrates improved performance over LLaMA on standard financial evaluation metrics (Yang et al., 2023). Likewise, in the legal domain, fine-tuning LLaMA on Chinese legal data produces a model that excels in tasks such as charge prediction and answering legal exam questions, outperforming its base model in both tasks (Huang et al., 2023b). The same does not hold true for other benchmark models, where three of the six studies reported that GPT-4 outperformed the fine-tuned model in question (GPT-4 was used as a benchmark model in four studies). For example, Roziere et al. (2023) demonstrate that GPT-4 outperforms Code LLaMA in coding problems on the HumanEval benchmark. Similarly, Yang et al. (2023) show that GPT-4 achieves the best performance in six out of nine finance-specific NLP tasks.

In addition, fine-tuning models entails the risk of *catastrophic forgetting*, where models lose previously acquired knowledge upon learning new information, leading to performance degradation outside the domains they were fine-tuned for (Luo et al., 2023; Huang et al., 2024). Fine-tuning can also increase the susceptibility of models to hallucinations. This phenomenon occurs because the introduction of new factual information during fine-tuning destabilizes the model's reliance on its pre-existing knowledge (Gekhman et al., 2024) or because models are taught to provide responses that satisfy human evaluators (cf. Zhao et al., 2023; Huang et al., 2023a).

Looking at the performance implications of RAG, all studies report performance improvements over their base models. Incorporating context-specific knowledge in LLMs through RAG improves open-domain question-answering ability (Lewis et al., 2020; Ram et al., 2023; Alawwad et al., 2024) and increases the factual accuracy of responses (Muhlgay et al., 2023). In medical applications, injecting medicine-specific information enhances performance on widely used medical evaluation benchmarks (Xiong et al., 2024) as well as on open-ended clinical questions crafted by medical practitioners (Zakka et al., 2024).[2] In the financial domain, applying RAG strengthens LLMs' capabilities in sentiment analysis tasks (Zhang et al., 2023) and supports more accurate financial question-answering (Li et al., 2024a).[3] Further, RAG limits the problem of hallucinations in LLM-generated content (Shuster et al., 2021) while mitigating the risk of *catastrophic forgetting*, as pre-trained weights are not altered.[4] However, RAG can decrease performance when a model uses external information that is irrelevant to the task at hand (Petroni et al., 2020; Shi et al., 2023; Yoran et al., 2023; Wu et al., 2024). For example, Shi et al. (2023) demonstrate that performance in complex problem-solving tasks decreases when irrelevant information is included in the task description.

Most studies listed in Table 1 conclude that domain-specific information in LLMs improves performance over some benchmark models. However, the results of the studies should be interpreted with caution as both academic scholars and developers have an incentive to overstate their models' performance. They also enjoy considerable discretion in how performance is measured, including the choice of tasks and benchmark models. Notably, three studies fail to report their models' performance relative to their base models. Since these are the very models they adjust to improve performance in a particular task, it seems natural to report the performance of the adjusted model relative to the base model. In addition, the number and

[2] Injecting domain-specific knowledge via RAG allowed the LLMs to outperform their respective non-RAG counterpart in the Medical Information Retrieval-Augmented Generation Evaluation (MIRAGE) benchmark, achieving up to an 18 percent improvement in accuracy by integrating relevant medical information into the model's reasoning process, as shown by Xiong et al. (2024).

[3] Including RAG in the sentiment analysis on Twitter postings improves performance for GPT-4 and a fine-tuned LLaMA in accuracy and F1-score (Zhang et al., 2023).

[4] In a human evaluation of Wizards of Wikipedia test questions, RAG improved the rate at which correct and knowledgeable information was included in the generated outputs and substantially reduced hallucinations (Shuster et al., 2021).

type of benchmark models used varies widely across studies. Thus, in this study, we design a controlled experimental setting to assess the impact of providing domain-specific information through augmented prompts in the financial context. We are particularly interested in the effectiveness of different types of information on LLM performance.

[Table 1 around here]

# 3 Experimental design

This section details our experimental design. In particular, it elaborates on our considerations in defining investor profiles to assess whether portfolio recommendations reflect investor characteristics (section 3.1), in determining an investment universe for which weights are to be assigned (section 3.2), and in eliciting LLM-generated (3.3) and human (3.4) portfolio recommendations. We have pre-registered our experimental design for both the LLM prompts and the survey involving human financial advisors (Hornuf et al., 2024).

## 3.1 Investor profiles

Regulatory bodies provide guidelines for investment recommendations made to retail investors, emphasizing that these recommendations must align with the individual's unique personal and financial characteristics. In the United States, the Financial Industry Regulatory Authority (FINRA) specifies in Rule 2111 (suitability rule) that recommendations must align with the customer's investment profile, which contains, among other things, investment objectives, experience, time horizon, liquidity needs, and risk tolerance (FINRA, 2020).[5] Similarly, the European Securities and Markets Authority (ESMA) requires investment advisors to consider their clients' investment horizon (or holding period) and information related to risk tolerance (ESMA, 2018, p. 6). Thus, we define investor profiles as differing on three key dimensions: risk tolerance, investment horizon, and ESG preference (cf. Fieberg et al., 2024; Streich, 2023).[6]

We vary *risk tolerance*, given that, according to *modern portfolio theory*, it should be the only investor characteristic governing portfolio allocation (Tobin, 1958). Empirical studies consistently document a significant correlation between individuals' risk preferences and their propensity to hold stocks (Dohmen et al., 2011) and the share of risky assets in their portfolios (Corter and Chen, 2006; Cardak and Wilkins, 2009; Nguyen et al., 2019).

In addition to the investor's subjective level of risk tolerance, we also vary investment horizon as an objective measure of *risk capacity* (cf. Frey et al., 2017; Davies, 2017). As the end of the investment horizon (and thus, disinvestment) approaches, an investor's capacity

[5] The SEC defines a similar retail customer investment profile in *Regulation Best Interest* (SEC, 2019).

[6] Table A1 in the Appendix provides a full list of all investor profiles.

to bear losses decreases. Academic studies have found investment horizons to be related to allocation to risky investments (e.g. Hansson and Persson, 2000; Cocco et al., 2005). In line with this view, practitioners typically suggest heuristics such as the "100 minus age" rule, which suggests that as individuals' investment horizon shortens, they should reduce the risky portion of their portfolio in favor of safer investments (Cocco et al., 2005).

We also account for investors' *sustainability preferences*, as recent regulation compels investment advisors to account for their clients' sustainability preferences when making portfolio recommendations (ESMA, 2023, p. 23). While ESG investing does not improve risk-return characteristics (Pedersen et al., 2021), investors seem to be willing to forego expected returns in their pursuit of sustainable investments (e.g., Hong and Kacperczyk, 2009; Borgers et al., 2015; Siemroth and Hornuf, 2023).[7]

### 3.2 Investment universe

To investigate the effect of providing domain-specific information on LLM performance, we design a simple portfolio allocation task. In addition to general domain-specific knowledge, we are particularly interested in the effect of firm-specific information on performance. Thus, we define an investment universe consisting of 12 actual US stocks and a broad US bond fund. We include a bond fund to allow recommendations to differ with respect to their risky share (i.e., proportion of the portfolio allocated to stocks) and to avoid allocations within the risky share to reflect differences in risk tolerance. We choose 12 stocks to strike a balance between the complexity of the task and the tractability of information injections and analyses. Specifically, because some treatment conditions contain substantial pieces of text for each stock in addition to the base prompts, this imposes a technical restriction in the form of token limits. We use US stocks to ensure that firm-level information is consistently structured across stocks (e.g., 10-K filings) and because the United States features most prominently among model developers in both the training data and in global capital markets (cf. Fieberg et al., 2024).

In identifying the 12 stocks to be used in our experiment, we employ a stratified sampling approach. The goal is to avoid selection effects while ensuring sufficient variation in cross-sectional stock characteristics to facilitate distinct investment strategies. In particular, we use both small and large stocks (as measured by market capitalization), as well as growth and value stocks (as measured by the book-to-market value of equity, cf. Fama and French, 1992). Specifically, we start out with all stocks in the S&P 500 and S&P 600 indices for which all information for our analysis is available in the Thomson Reuters Eikon database (929 stocks). We then sort these stocks into four buckets by size and book-to-market value (median splits).

[7] Pedersen et al. (2021) introduce the concept of the ESG-efficient frontier, demonstrating that integrating ESG factors can enhance the Sharpe ratio for investors prioritizing these criteria, optimizing the trade-off between risk, return, and sustainability. However, their findings also indicate that strong preferences for high ESG score investments may result in lower expected returns, driven by increased demand for ESG-friendly assets.

Finally, we randomly draw three stocks out of each of the four buckets, resulting in 12 stocks that differ in terms of size and book-to-market value. Table 2 lists the 12 resulting stocks, which include household names (e.g., Berkshire Hathaway and PepsiCo) as well as less well-known securities (e.g., Air Lease and St. Joe). The table reveals substantial variation in the stocks' market betas (ranging from 0.16 to 1.61), book-to-market ratios (0.02 to 1.49), and size (ranging from $1 billion to $974 billion). Finally, we use Vanguard's Total Bond Market ETF (BND), one of the most commonly used US bond funds with net assets exceeding $100 billion, as the fixed-income option.

[Table 2 around here]

### 3.3 LLM-generated recommendations

This section details our methodological choices in identifying models for inclusion in our analysis (section 3.3.1), selecting and constructing the additional information provided to the LLMs in the four treatment and two placebo conditions (sections 3.3.2 through 3.3.6), and developing standardized prompts for data collection based on related literature investigating prompt design (3.3.7). Table 3 provides an overview of the various experimental conditions, as well as the information that is provided in each condition. In all specifications, we communicate the investor's circumstances and the eligible investment universe. In treatment condition 1, we add general investment theory as context. In treatment condition 2, we add firm-specific quantitative metrics. In treatment condition 3, we add firm-specific qualitative data distilled from each firm's latest annual report. In treatment condition 4, we add all three pieces of additional information to assess their joint impact on portfolio recommendations.

[Table 3 around here]

#### 3.3.1 Models

We use seven of the most recent LLMs developed by the leading makers of text-based artificial intelligence (OpenAI, Meta, Mistral AI, Microsoft, and Alibaba; see Table 4). For models by developers other than OpenAI, we use smaller versions of the most recent models (with up to 10 billion parameters), as the injection of domain-specific information introduces constraints regarding memory capacity and processing power. Thus, while most of our analyses will compare the performance of recommendations across different configurations of a given LLM, not all of our results may be generalizable to the largest versions of these models.

One of the key concerns in empirical research on the capabilities of LLMs, especially in the finance context, is look-ahead bias resulting from information leakage from models' training data (Sarkar and Vafa, 2024; Alonso, 2024). In our study, this affects in-sample performance analyses, as well as whether the firm-specific information we inject is indeed novel to the models. To account for this potential bias, we document for each model the date at which it is cut off from information. While most developers report the information cut-off date for their

models, others do not. In addition, reported information cut-off dates have proven unreliable at times (Cheng et al., 2024). We therefore verify the models' information cut-offs through an additional analysis based on the approach outlined by Cheng et al. (2024). Specifically, we test the models' knowledge of significant global events that occurred between January 2023 and June 2024. For each month, we identify three notable events and formulate a falsifiable question for each event. We then pose these questions to each model and assess the accuracy of each response. We define as the revealed cut-off date the most recent month for which a model has correctly answered at least one of the three questions.[8]

Table 4 displays the models alongside their reported and revealed information cut-off dates. Notably, the earliest revealed information cut-off is observed for Llama-3.1-8B-Instruct, which incorporates no new information beyond May 2023, while Meta reports December 2023 as its information cut-off date. In contrast, GPT-4o-mini evidently has the most recent revealed cut-off date of January 2024, which postdates its reported cut-off date of October 2023. Importantly, all models' information cut-off dates predate the publication dates of the 10-K filings we use in treatment 3 (see Section 3.3.4) and the data used to construct the quantitative metrics used in treatment 2 (see Section 3.3.3).

[Table 4 around here]

### 3.3.2 Treatment 1: General investment theory

While LLMs are increasingly capable of reproducing advanced financial knowledge (Niszczota and Abbas, 2023), they are less effective at applying that knowledge to real-world conditions (Smith, 2024).[9] To ensure that the LLMs' portfolio recommendations are grounded in sound finance theory, we provide the models with seminal investment theory. Specifically, we provide the models with a summary of key chapters of a commonly used textbook for undergraduate finance courses (Berk and De Marzo, 2020, Chapters 9–11), which provide a comprehensive overview of important knowledge to consider in portfolio allocation tasks. The book chapters discuss methods for stock valuation such as the dividend discount model, explore the influence of corporate decisions on stock prices, and introduce measures of risk and return in capital markets, as well as the concepts of idiosyncratic and systematic risk. Building on this, the chapters introduce mean-variance optimization, the efficient frontier, and the capital asset pricing model (CAPM) as foundations of portfolio optimization.

Because we cannot use the entire text from the three chapters due to context length restrictions, we use GPT-4 (which is not part of our LLM sample) to create summaries of

[8] Table A2 in the Appendix contains the events and questions used. Table A3 in the Appendix reports the results per model.

[9] Niszczota and Abbas (2023) find that GPT-4 correctly answers 99% of a standard battery of financial literacy questions, while the corresponding figure is less than half for the general US population. However, Smith (2024) finds that ChatGPT fails to take into account the time value of money (one of the concepts elicited in the financial literacy questions) in hypothetical household finance scenarios.

the chapters. The ability of LLMs to extract the most important information from financial text has been documented in several studies.[10] We follow related studies (Zhang et al., 2022; Subbiah et al., 2024) and separately upload the three chapters and have GPT-4 summarize them in two steps. We use the following prompt to elicit summaries (cf. Subbiah et al., 2024):

> Please summarize the following [excerpts from a finance textbook/chapter summaries] into a concise, coherent, and continuous narrative. Base your summary solely on the provided text, focusing on key concepts. Ensure the summary is clear, accurate, and unbiased, providing a foundational understanding that can inform portfolio recommendations. Avoid bullet points, opinions, interpretations, or any external information beyond what is explicitly stated in the text.

The resulting summary is displayed in Table A4 in Appendix A.

#### 3.3.3 Treatment 2: Quantitative firm-specific financial metrics

A considerable body of work has identified more than 400 factors that are alleged to predict cross-sectional stock returns (Harvey et al., 2016; Hou et al., 2020; Aghassi et al., 2023; Jensen et al., 2023). To assess whether providing firm-level information on commonly used risk factors affects LLM-generated portfolio recommendations, we use a total of six quantitative indicators. We start with the three-factor model proposed by Fama and French (1992), which includes market risk, size, and value, and add the commonly used momentum factor proposed by Carhart (1997). Both the Fama-French three-factor model and the Carhart four-factor model have been the subject of extensive empirical research and are among the most cited in academic finance (Aghassi et al., 2023). In addition to these four factors, we include the earnings-to-price (E/P) ratio as a stock fundamental that reflects a simple version of the dividend discount model. It is one of the most commonly reported figures in practice.[11] Finally, to assess the portfolio strategy's sensitivity to sustainability ratings, we add each company's Thomson Reuters ESG score as a sixth metric. Table 2 reports the respective metrics for each firm in our sample.

We downloaded all of the metrics described above from Thomson Reuters Eikon on September 20, 2024. Thomson Reuters computes market betas from CAPM regressions using

[10] Pu et al. (2023) show that LLMs are capable of creating text summaries that are preferred by human evaluators over summaries created by humans and fine-tuned models. LLM summaries also surpass those of humans in terms of factual accuracy. Similar findings have been reported by Goyal et al. (2022) and Zhang et al. (2024). Finally, Kim et al. (2023) find that sentiment scores of LLM-generated summaries of corporate disclosures are better for predicting market reactions than sentiment scores of the original disclosures.

[11] Both Yahoo Finance and Bloomberg report the E/P ratio (or its inverse) on the summary page for each stock. Brokerages typically also include E/P ratios in their stock summary pages.

monthly return data for the previous five years.[12] Firm size is computed as the respective stock's market capitalization as of September 20, 2024. As an indicator of the value anomaly (Rosenberg et al., 1985; Fama and French, 1992), the book-to-market ratio is computed as the ratio of each stock's book value of equity per share (as reported in its most recent SEC filing) to its current share price as of September 20, 2024. As an indicator of the momentum anomaly (Jegadeesh and Titman, 1993; Carhart, 1997), we use the cumulative return over the previous twelve months. Earnings-to-price ratios are computed as the earnings per share over the previous twelve months by a stock's current share price as of September 20, 2024. Finally, we use the Thomson Reuters ESG score, which quantifies a company's self-reported performance in the environmental, social and corporate governance pillars. The score is reported on a scale from 0 to 100, where higher scores indicate stronger adherence to ESG principles.

#### 3.3.4 Treatment 3: Qualitative firm-specific information (10-K summaries)

To allow for fundamental analysis of a firm's future profit potential, we inject qualitative (i.e., textual) information from the stock's latest 10-K filings. The reports are filed during the first quarter of 2024 and typically refer to the fiscal year ending December 31, 2023.[13] The reports provide a comprehensive view of each firm's performance in the previous year and strategy for the future. The annual reports were filed after the information cut-off date of all models, suggesting that they represent novel information not contained in the models' training data.[14]

Because we cannot use the entire 10-K document due to context length restrictions, we focus on the management discussion and analysis (MD&A) section of the filing, which offers investors a critical perspective on the company's financial conditions and future outlook (Kim et al., 2023; Cao et al., 2023). Again, we have GPT-4 (which is not part of our LLM sample) summarize the MD&A content into concise, 500-token summaries. In addition to ensuring that we remain within the context length limits of our models, providing summaries instead of the entire MD&A section likely improves its information content (Kim et al., 2023).[15] The resulting summaries of the MD&A sections are presented in Table A7 in Appendix A.

---

[12] Thomson Reuters uses either the S&P 500 Index or the NASDAQ Composite Index as the market portfolio.

[13] Pepsi's fiscal year ended December 30, 2023. See Table A6 in Appendix A for an overview of report and publication dates of each filing.

[14] The only exception is Lockheed Martin's 10-K filing in January 2024, which may coincide with the information access of GPT-4o-mini, which likely includes data from January 2024 (see Table A2 in the Appendix).

[15] The prompt we used to summarize the MD&A content is the following: "Please summarize the following excerpts from a company's 10-K annual report into a concise summary of no more than 500 tokens. Base your summary solely on the provided text, focusing on key financial metrics, business strategies, market conditions, risks, and other essential details relevant for making investment decisions. Ensure the summary is clear, accurate, and unbiased, providing a solid foundation for generating portfolio recommendations. Do not add any opinions, interpretations, or external information beyond what is explicitly stated in the text."

#### 3.3.5 Treatment 4: Combined information

Treatment 4 provides the LLMs with all information contained in treatments 1 through 3. It is intended to assess how the inclusion of the combined information impacts LLM performance.

#### 3.3.6 Placebo treatments

To ensure that any adjustments in the LLMs' recommendations are indeed driven by the information content of the treatments (rather than the mere addition of information), we define two placebo treatments that provide the LLMs with information that is entirely irrelevant to the portfolio allocation task. To test for potential differences arising from adding qualitative information versus quantitative information, we provide, in two separate specifications: a summary of a seminal information systems model (*technology acceptance model*, cf. Davis, 1989, placebo 1); and 2023 weather data for each US state capital (placebo 2).[16] The technology acceptance model is the theoretical foundation of thousands of empirical technology adoption studies. It holds no implication for the portfolio allocation experiment and should thus not systematically affect recommendations. The same is true of the weather data, which contain the average daily precipitation levels for the year 2023.[17]

#### 3.3.7 Prompt design

To elicit portfolio recommendations from the LLMs, we carefully design prompts for each treatment condition and investor profile. We use a standardized prompt structure as outlined in Table 6. The structure comprises seven important components.

First, we apply role prompting by assigning the LLM a specific role. The use of detailed expert identities to guide LLM responses has been shown to improve the accuracy of a model's responses by aligning outputs with a specific persona (White et al., 2023; Kong et al., 2023). Second, we include general instructions, which include the details of the investor profile, as well as the general task of recommending a portfolio for that particular investor. Third, we introduce the eligible security universe. The securities are presented in randomized order to mitigate order effects that could potentially bias the model response. Fourth, we introduce additional information based on the respective treatment condition. To ensure that recommendations are not influenced by the specific wording, we use similar wording for each treatment.[18] Whenever firm-specific information is displayed, the order of firms is randomized

[16] We use cross-sectional instead of time series weather data to avoid any implicit reference to climate change, which could bias recommendations toward high-ESG-score securities.

[17] Data for the placebo specifications (as well as the baseline specification) were collected on November 21, 2024, using the same prompts as used in the initial data collection on November 4, 2024. Online Appendix D reports details and results from the placebo analyses. Because the placebo analyses were recommended to us ex-post, they are not contained in the pre-registration.

[18] "Additionally, you may consider [information]."

to avoid order effects. Fifth, we specify the output format for the models' responses. Sixth, to mitigate errors in generating portfolio recommendations—errors commonly encountered in related studies (Fieberg et al., 2023, 2024)—we explicitly instruct the models to ensure that portfolio weights sum to 100%. Finally, we use chain-of-thought (CoT) prompting, which causes models to break down complex tasks into smaller sub-tasks and has been shown to improve reasoning and arithmetic performance (Wei et al., 2022). Specifically, we use zero-shot CoT prompting (Kojima et al., 2022), which induces step-by-step reasoning without pre-crafted examples, enabling systematic task decomposition.

[Table 6 around here]

### 3.4 Human recommendations

To obtain a human benchmark for the portfolio allocation task, we elicit portfolio recommendations for each of the 12 investor profiles from human US-based financial advisors through *Prolific*, an online survey platform tailored for use in academic research (Palan and Schitter, 2018). The platform is well-established for providing access to a large and high-quality pool of survey participants and has been used to recruit study participants in several recent, high-quality finance publications (e.g., Döttling and Kim, 2024; Chapkovski et al., 2024; Cordes et al., 2023; Huber and Huber, 2020).

To ensure high-quality answers from the relevant demographic, we (i) restrict participation to US-based financial advisors, (ii) include two attention checks in the survey,[19] (iii) document the time taken to complete the survey as a measure of effort, and (iv) provide a financial incentive to ensure high-quality portfolio recommendations. Specifically, in addition to the fixed remuneration of GBP 3 ($3.90), respondents receive a variable remuneration of up to GBP 3 based on the performance of their portfolio recommendations. In particular, one of the 12 investor profiles is randomly selected after all survey responses are collected, and all survey respondents are ranked according to the risk-adjusted performance of their portfolios for this scenario. Those respondents with higher risk-adjusted performance receive higher variable payments. Thus, because the survey takes approximately 20 minutes to complete, participants could expect to be paid GBP 13.50 ($17.55) per hour.[20]

The structure of the survey is as follows. Following a brief introduction explaining the structure of the survey and the details of the fixed and variable remuneration components, participants are briefly introduced to each of the 12 scenarios and asked to provide portfolio weights for the 13 pre-determined securities. Both the scenarios and the securities are displayed in a randomized order to avoid learning or order effects. Following the recommendation part of the survey, we include two manipulation checks: First, we ask for the risk tolerance

[19] Participants are asked (i) to state the risk tolerance level of the final investor profile they are shown and (ii) to briefly describe the rationale underlying their recommendations.

[20] (GBP 3 + GBP 1.5) * 3

level of the last displayed scenario.[21] Second, we ask participants to provide a brief explanation of the rationale behind their recommendations. Finally, we ask for the respondents' level of experience in the financial advice industry[22] and three of the "Big Five" questions commonly used to measure financial literacy (Lusardi and Mitchell, 2011).

The survey was published on November 2, 2024. The targeted number of respondents (100) was reached on November 3, 2024. Fixed payments were immediately distributed to participants upon completion. Variable payments were based on risk-adjusted performance (Sharpe ratios) from November 4, 2024, to December 19, 2024, and were paid out on December 22, 2024.[23] Online Appendix B describes in detail the pre-processing of the survey data, which yielded the full sample of 89 respondents who passed our screening criteria (working as financial advisors) and 68 respondents who passed stricter attention checks. The participants in our sample are mostly male financial advisors with substantial self-reported experience in the industry and substantially higher levels of financial literacy than the general population (see Table B1 in Online Appendix B).

# 4 Construction of key variables

## 4.1 Dependent variables

To assess the degree to which recommended portfolios are diversified, we compute three measures. First, we use the number of securities included in a portfolio as a rough diversification measure. Second, to account for the distribution of portfolio weights among the included securities, we compute the Herfindahl–Hirschman Index (HHI):

$$HHI_p = \sum_i (w_{ip})^2 \tag{1}$$

where $w_{ip}$ is the weight of security $i$ in portfolio $p$. Third, to distinguish shifts between risky assets and the bond fund from diversification within the equity portion of the portfolio, we also compute the HHI only within the equity portion of the portfolio.

To capture portfolio risk, we use four measures. First, we use the risky share (i.e., the proportion of the portfolio allocated to stocks). Second, we compute portfolio volatility from a monthly time series of portfolio values from April 2013 to September 2023 and assume

[21] The choices are: ["high," "medium," "low," and "I do not recall."]. Because risk tolerance can be either "high" or "low," but never "medium," we exclude respondents answering "medium" in robustness tests.

[22] Question 1: "How many years of experience do you have in the financial advice industry? [Less than 1 year · 1–5 years · 5–10 years · More than 10 years]"; question 2: "Approximately how many clients have you served in your role as a financial advisor throughout your career? [Manual entry]"

[23] Only participants who passed strict attention checks (N = 68) received variable compensation. The top quartile of participants (N = 17) received GBP 3, the second quartile (N = 17) received GBP 2, the third quartile (N = 17) received GBP 1, and the fourth quartile (N = 17) received GBP 0.

monthly rebalancing. We use data from April 2013—when price information first becomes available for all stocks—to September 2023 in order to avoid overlap between the historical data and the momentum measure we provide to the LLMs.[24] Third, as a measure of market risk, we use the market beta ($\beta_p^M$) obtained from the following six-factor regression model:

$$\begin{aligned} r_{p,t} - r_{f,t} = \alpha_p + \beta_p^{\mathrm{M}}(R_{mkt,t} - r_{f,t}) + \beta_p^{\mathrm{SMB}} \times \mathrm{SMB}_t + \beta_p^{\mathrm{HML}} \times \mathrm{HML}_t \\ + \beta_p^{\mathrm{RMW}} \times \mathrm{RMW}_t + \beta_p^{\mathrm{CMA}} \times \mathrm{CMA}_t + \beta_p^{\mathrm{WML}} \times \mathrm{WML}_t + \epsilon_{p,t} \end{aligned} \tag{2}$$

where $r_{p,t} - r_{f,t}$ is portfolio $p$'s excess return in month $t$.[25] The excess return is regressed on six commonly used asset pricing portfolios: The market excess return ($R_{mkt,t} - r_{f,t}$), the small-minus-big size portfolio (SMB), the high-minus-low value portfolio (HML), the robust-minus-weak operating profitability portfolio (RMW), the conservative-minus-aggressive investment portfolio (CMA), and the winners-minus-losers momentum factor (WML). Data on monthly US factor portfolios and the risk-free rate are obtained from the Kenneth French Data Library. Fourth, we use idiosyncratic volatility as a measure of portfolio-specific risk, which is computed as the standard deviation of the error term $\sigma(\epsilon_p)$.

To measure a portfolio's risk-adjusted performance, we compute three measures. First, we use the average monthly excess returns. Second, we use the annualized Sharpe ratio (excess return divided by monthly volatility). Third, we use alphas ($\alpha_p$) to the six-factor model. To account for the fact that sustainability-oriented investors may be willing to forego risk-adjusted returns to pursue sustainability goals (Pedersen et al., 2021), we also investigate mean-variance optimization separately for an ESG-efficient frontier.

To approximate the transaction costs required to maintain the suggested portfolio allocation, we compute two measures. First, because most brokers charge a fixed transaction cost component, we measure the number of trades required per year as the number of stocks that have to be bought or sold:

$$\text{No. trades p.a.} = \left( \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{1}\left(w_{i,t} - w_{i,0} \neq 0\right) \right) \times 12 \tag{3}$$

where $w_{i,t}$ represents the weight of security $i$ in month $t$ and $w_{i,0}$ denotes the recommended portfolio weight for security $i$. Second, because some brokers charge variable transaction costs, we measure the average turnover required to maintain the suggested portfolio:

$$\text{Annual turnover} = \left( \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} |w_{i,t} - w_{i,0}| \right) \times 12 \tag{4}$$

[24] The momentum measure is computed as the cumulative stock return from October 2023 to September 2024.

[25] The risk-free rate is approximated by the one-month US Treasury bill rate.

Finally, to assess whether portfolio recommendations accurately reflect an investor's sustainability preferences, we compute a portfolio-level ESG score according to equation (5):

$$\text{ESG score} = \sum_i \frac{w_i}{w_p^R} \text{ESG}_i \tag{5}$$

where $ESG_i$ is the Thomson Reuters ESG score for stock $i$, $w_i$ is stock $i$'s portfolio weight in portfolio $p$, and $w_p^R$ is portfolio $p$'s risky share.

[Table 6 around here]

## 4.2 Stock-level variables

### 4.2.1 Investor attention

Given that an important part of LLMs' training data is made up of news archives, it is conceivable that their portfolio recommendations are more likely to contain securities with high investor attention (cf. Fieberg et al., 2024). We use four measures to capture cross-sectional differences in investor attention. First, the Securities and Exchange Commission (SEC) provides daily log files of download requests on a document basis as part of its Electronic Data Gathering and Retrieval (EDGAR) database. Following related studies, we use the number of downloads of a respective stock's SEC documents as a direct measure of investor attention (cf. Chen et al., 2020; Andrei et al., 2023). Second, we use the Google search volume index (SVI) for a specific company (or its ticker) as an alternative measure of investor attention. We use three approaches to ensure the Google SVI measures capture as little noise as possible.[26] We use the companies' names and the exchange-linked stock tickers to ensure we pick up attention by investors rather than, e.g., customers. Third, following the recommendations by deHaan et al. (2024), we include "stock" in the keyword to reduce noise.[27] We compute an aggregate Google SVI by using the median of the average SVI over the three specifications in 2023. Fourth, we use company-related news coverage as another measure of investor attention (cf. Barber and Odean, 2008). Specifically, we use the number of news articles reporting on the respective company and which have been published in major financial media outlets (*Bloomberg*, *Economist*, *Kiplinger*, *Wall Street Journal*) in 2023 as a measure of investor attention.[28] Finally, we use the number of analyst reports published in 2023 as an alternative measure of investor attention (cf. Mola et al., 2013; Claussen et al., 2020). Online Appendix C details the construction of our investor attention measures and provides some descriptive statistics.

[26] Studies investigating the within-stock variation in Google search volume frequently normalize Google SVI to account for stock-specific noise (Da et al., 2011). Because we are interested in between-stock differences, we do not use abnormal Google SVI.

[27] We omit St. Joe Co. in this specification to avoid noise from the ambiguous ticker ("JOE").

[28] We obtain articles using the literature database *Business Source Complete* to identify news articles.

#### 4.2.2 10-K sentiment measures

We further measure the sentiment in each company's 10-K summary to assess whether including the summaries sways recommendations toward stocks with positive sentiment. We follow the method proposed by Loughran and Mcdonald (2011), who develop a dictionary specific to the text corpus used in 10-K filings. Specifically, we count for each 10-K MD&A summary the number of positively and negatively connoted words according to their dictionaries and compute the tone as the difference between positive and negative words divided by the sum of positive and negative words. As Figure A1 in Appendix A shows, this simple dictionary-based sentiment measure strongly correlates with a classification provided by OpenAI's GPT-4, which has been shown to be able to predict stock returns and corporate investments from corporate disclosures and news reports (Jha et al., 2024; Lopez-Lira and Tang, 2023).

## 5 Results

### 5.1 Security selection

This section investigates what drives human advisors' and LLMs' decision to include specific securities in their portfolios. Table 7 displays average portfolio weights for the 13 securities by experimental condition. Several patterns emerge. First, human advisors recommend substantially greater risky shares than LLMs (84% vs. 70% in the baseline condition). While adding generic finance information (treatment $T_1$) does not affect the risky share, adding firm-specific information (treatments $T_2$ through $T_4$) is associated with an increase in the average risky share in LLM recommendations to between 78% and 84%. Second, human advisors distribute portfolio weights much more evenly across stocks than LLMs. While average portfolio weights for stocks range from 6% to 10% for human advisors, they range from 1% to 15% for LLMs. Third, LLM portfolios are concentrated in large-cap growth stocks and have substantially less small-cap exposure than human portfolios. Adding general finance theory (treatment $T_1$) is associated with a shift toward small-cap stocks, while adding firm-specific information is associated with an increased allocation to large-cap growth stocks.

[Table 7 around here]

**Result 1.** *Providing firm-specific information increases the risky shares in LLM recommendations, aligning them with human recommendations.*

To identify the stock-level determinants of security selection, Table 8 reports average fundamentals, investor attention proxies, and 10-K sentiment scores for stocks that are recommended versus stocks that are not recommended by the respective advice source. For both human- and LLM-generated advice, the table reports averages for the baseline conditions. The results suggest that human advisors and LLMs generally pick similar stocks. In particular, stocks included in the recommendations have lower market betas, lower book-to-market

ratios, higher market capitalization, higher ESG scores, and more investor attention than stocks not included in the recommendations. However, the differences in means imply that LLMs differentiate based on these drivers to a greater extent than do human advisors. For example, the size difference between recommended and omitted stocks is $25 billion ($131 billion minus $106 billion) in human portfolios and $76 billion ($161 billion minus $85 billion) in LLM portfolios. This pattern suggests that LLM portfolios are more concentrated in large stocks, which is consistent with the lower number of securities observed in LLM portfolios (Table 6).

To assess how domain-specific information affects the LLMs' recommendations, Table A8 in the Appendix reports the corresponding univariate differences for the various treatment conditions. The findings can be summarized as follows. First, the univariate relationships between recommendation and stock fundamentals observed in the baseline condition remain qualitatively unchanged when domain-specific information is added. The only exception to this is momentum, which is positively related to inclusion in portfolios when fundamentals (including momentum) are explicitly provided to the LLMs (treatments $T_2$ and $T_4$). Consistent with the baseline condition, LLMs still generally include higher-attention stocks in their portfolio recommendations.[29] Second, adding general finance theory (treatment $T_1$) is associated with a smaller size and attention difference between recommended and omitted stocks, which is consistent with a more even distribution of portfolio weights within the risky portfolio portion (see Table 7). Third, adding quantitative firm-specific information (treatment $T_2$) is associated with an increased tendency to include stocks with low market betas, book-to-market ratios, and ESG scores, and a decreased tendency to include large and high-attention stocks. Notably, recommended and omitted stocks differ significantly in terms of their momentum (cumulative annual returns of 19% vs. 13% , $p < 0.01$). Thus, when information on past returns is provided, LLMs seem to follow a momentum strategy. Fourth, LLMs are more likely to include stocks with positive sentiment when qualitative information in the form of 10-K MD&A summaries is provided (treatment $T_3$). As a consequence, the tendency to recommend large, high-attention stocks observed in the baseline condition is reduced. This finding suggests that LLMs can not only correctly infer stock price movements from sentiment in corporate disclosures (e.g., Pelster and Val, 2024), but they also use the sentiment to form portfolio recommendations. Fifth, when all domain-specific information (treatment $T_4$) is provided, size and attention matter less than in the baseline, while momentum and ESG scores matter more than in the baseline. The qualitative information does not seem to affect security selection in treatment $T_4$. Thus, quantitative information has a greater impact on the LLMs' recommendations than qualitative information.

To ensure that the results we obtain are indeed causally driven by the additional information contained in the various treatments, Table A9 in the Appendix assesses univariate

[29] The only exception is the analyst reports measure, for which recommended stocks in treatment condition $T_2$ display higher values than stocks not recommended ($p < 0.1$).

differences for the two placebo conditions. The results suggest that adding irrelevant information does not affect the stock-level determinants of portfolio recommendation.[30]

[Table 8 around here]

Finally, to assess the degree to which the various drivers contribute to explaining the variation in portfolio weights, we conduct a relative weight analysis (cf. Blaseg and Hornuf, 2024). Table 9 displays standardized dominance statistics for regressions of a stock's portfolio weight on its fundamentals, an investor attention proxy, and the sentiment in its latest 10-K MD&A section, estimated separately for each experimental condition. The dominance statistics measure the extent to which an independent variable contributes to the regression model's ability to explain variation in the dependent variable. Thus, it captures the relative importance of a variable in determining portfolio weights. The table highlights four key findings. First, unexplained variation is substantially higher for human recommendations ($R^2$ = 0.02) than for LLM recommendations ($R^2$ = 0.21). This suggests that human portfolio recommendations are driven to a larger extent by factors other than the fundamentals we use. For LLMs, the highest explanatory power for the model is obtained when quantitative information is provided to the LLMs (treatment $T_2$, $R^2 = 0.30$). This is consistent with higher portfolio concentration in this condition (e.g., 40% allocation to large-cap growth stocks, see Table 7). Second, investor attention accounts for a smaller share of the explained variation in LLM-recommended portfolios than human-recommended portfolios (11% vs. 34% in the baseline conditions). When domain-specific information is provided to the LLMs, the relative contribution of the investor attention measure decreases further. Third, stock fundamentals contribute most to explained variation in treatment condition $T_2$ and sentiment in 10-Ks contributes most in treatment condition $T_3$. This pattern confirms that LLMs respond to the firm-specific information provided to them when determining portfolio weights. Fourth, among the fundamental metrics, a company's ESG score is the largest contributor to explained variation for LLM portfolios, but not human portfolios, for which size is most important. ESG scores matter most when they are explicitly provided to the LLMs (i.e., in $T_2$ and $T_4$). Intuitively, when investor profiles without sustainability preferences are considered (see Table A10 in the Appendix), the contribution of a stock's ESG score to explaining variation in portfolio weights decreases.

Table A11 in the Appendix reports the corresponding results for the placebo analyses. The standardized weights suggest that when irrelevant information is provided to the models, investor attention accounts for a larger share of the explained variation, while overall goodness-

[30] The results further suggest that the univariate differences observed in the initial baseline are generally replicated in the repeated baseline collection (see Tables A8 and A9 in the Appendix.)

of-fit remains mostly unchanged.[31] This finding suggests that the differences observed between treatments is in fact driven by the contents of the treatments.

[Table 9 around here]

**Result 2.** *Providing domain-specific information decreases allocations to high-attention stocks in LLM recommendations.*

## 5.2 Suitability of portfolio recommendations

This section investigates whether the recommended portfolios accurately reflect an investor's risk tolerance and sustainability preference. Table 10 investigates the correlation of a portfolio's risky share with the investor's risk tolerance and investment horizon, separately by experimental condition. In line with standard economic theory, there is a significant positive correlation between risk tolerance and risky shares across all treatment conditions. However, there are differences between the various conditions. First, LLMs differentiate to a greater extent between profiles with high and low risk tolerance than human advisors (46 vs. 16 percentage points), which is due to the lower baseline risky share in LLM portfolios (43 vs. 75% for investors with low risk tolerance, 3-month horizon, and without sustainability preference). Second, when quantitative firm-specific information is provided (treatments $T_2$ and $T_4$), the difference in equity shares between investors with high and low risk tolerance decreases as the baseline equity share increases. No such change is observed for the two placebo conditions (see Table A13 in the Appendix).

[Table 10 around here]

**Result 3.** *LLMs take into account investors' risk tolerance when recommending portfolios. When firm-specific information is provided, sensitivity to risk tolerance decreases as baseline portfolio risk increases.*

Another important aspect for financial advisors to consider is their clients' sustainability preferences. To investigate whether the recommended portfolios reflect an investor's sustainability preference, Table 11 reports the coefficients resulting from a regression of the portfolio-level ESG score on a dummy variable indicating that an investor has a sustainability preference, controlling for all other investor characteristics. The coefficients suggest that LLMs recommend portfolios with higher average ESG scores to investors with a preference for sustainability, while human advisors do not. The latter result may be a consequence of the financial incentive structure we employ when eliciting human advice. Specifically, variable compensation is based on risk-adjusted returns. Assuming that sustainable investment

[31] The table further suggests that relative weights are highly robust across portfolios in the two LLM baseline conditions. Recall that we elicit portfolio recommendations for the LLM baseline condition twice to assess robustness (see Online Appendix D for details).

strategies forego risk-adjusted returns in favor of sustainability goals (cf. Pedersen et al., 2021), human advisors might rationally choose not to reflect sustainability preferences in their recommendations given the prevailing incentive structure. For LLM recommendations, sensitivity to sustainability preferences is highest when ESG scores are explicitly provided (treatments $T_2$ and $T_4$). This can be interpreted either as LLMs having more precise knowledge of the ESG scores of the individual stocks, or as LLMs interpreting the provision of an ESG score as a request to incorporate it in their portfolio recommendation. Finally, LLMs distinguish between investors with and without sustainability preferences more when they are exposed to general finance theory (treatment $T_1$).[32]

[Table 11 around here]

**Result 4.** *Providing domain-specific information increases the sensitivity of LLM recommendations to sustainability preferences. Sensitivity is highest when quantitative metrics are provided.*

### 5.3 Diversification and portfolio risk

This section investigates differences in portfolio diversification and risk between human advisors and LLMs, as well as across treatments. Tables 12 and 13 report the coefficients for linear regressions of diversification and risk measures on a dummy variable indicating that a recommendation has been provided by an LLM (vs. a human advisor, $\mathbb{1}(LLM)$) or dummy variables indicating that a recommendation has been provided by an LLM under treatment condition $i$ ($\mathbb{1}(\mathbb{T}_\mathbb{i})$). For the comparison between human advisors and LLMs, we only use portfolios for the baseline conditions (i.e., with no additional information provided).

The findings from these two analyses can be summarized as follows. First, LLMs recommend 1.3 fewer securities on average. Adding information does not affect the number of securities included in a portfolio. Second, portfolios are less concentrated when firm-specific information is provided (treatments $T_2$ and $T_4$, see Table 12). The increase in diversification is achieved through an increase in the risky share, as the Herfindahl index within the equity portion is not affected. Third, human portfolios display significantly higher equity allocation (14 percentage points, $p < 0.01$, see Table 13). Adding firm-specific information increases the risky share in LLM recommendations by between 8 and 14 percentage points ($p < 0.01$ for all firm-specific treatments), closing the gap between LLM and human portfolios. Adding quantitative information is associated with a stronger impact on recommendations than qualitative information. Finally, higher equity shares translate to higher volatility, IVOL, and market risk.

[Tables 12 and 13 around here]

[32] While we observe a slight placebo effect, the point estimates in the treatment specifications are higher than those obtained from the placebo specifications (see Table A14 in the Appendix).

**Result 5.** *Providing firm-specific information increases portfolio risk in LLM recommendations. The impact is more pronounced when quantitative metrics are provided.*

## 5.4 Performance

We conduct two types of performance analyses. First, to make sure that our performance measures are based on a representative set of price data, we conduct in-sample tests based on monthly time series data from April 2013 to September 2023.[33] To address recent literature suggesting that ESG-conscious investors trade off financial performance and investment-related sustainability goals (Pedersen et al., 2021), which implies that the quality of portfolios recommended to investors with ESG preferences cannot simply be assessed based on financial performance, we also assess mean-variance optimization separately for the entire security universe and an ESG-screened sub-universe. Second, to address concerns over look-ahead bias (i.e., LLMs recommending securities that have historically done well), we conduct out-of-sample performance analyses based on a small number of time series observations. In addition to the human baseline constructed from the recommendations of US financial advisors, we compare the performance of LLM recommendations to that of a simple 1/N diversification strategy.

### 5.4.1 In-sample tests

Table 14 reports regression coefficients of performance measures on the various treatment conditions, while controlling for investor characteristics. The performance measures are based on historical price data for the period April 2013 to September 2023. The results suggest that LLM portfolios significantly underperform human portfolios. Excess returns are 15 basis points lower, annual Sharpe ratios are 0.07 lower, and FF6 alphas are 10 basis points lower in LLM portfolios than in human portfolios. The difference is economically sizable as it represents more than half of a standard deviation in excess returns, one-third of a standard deviation in Sharpe ratios, and more than three-quarters of a standard deviation in six-factor alphas (see Table 6). At the same time, transaction costs as proxied by annual turnover are significantly lower in LLM portfolios, which is driven by the lower number of securities in LLM portfolios.

**Result 6.** *LLM recommendations underperform human recommendations and a naive 1/N diversification strategy in in-sample tests.*

Within LLM portfolios, adding firm-specific information significantly improves excess returns and Sharpe ratios, but not six-factor alphas. The increase in excess returns and Sharpe ratios is economically significant and suggests that adding firm-specific information closes the

[33] April 2013 is the first month for which we obtain data for all 13 securities. September 2023 is the first month that is not included in the momentum measure provided to the LLMs.

performance gap between LLM and human portfolios. Notably, the impact of firm-specific information on performance is stronger for quantitative than for qualitative information. The performance improvement is driven by the increase in the risky share associated with firm-specific information (see Table 12). This explains why there is no improvement in six-factor alphas, which account for market risk. As Figure A5 in the Appendix shows, adding firm-specific information brings LLM performance closer to that of human portfolios and a naive 1/N diversification strategy, which yields the best performance outcomes.

[Table 14 around here]

**Result 7.** *Providing firm-specific information significantly improves the historical performance of LLM recommendations. Performance improvements are achieved through higher exposure to market risk and are greatest for quantitative information.*

Next, we address potential performance differences between portfolios recommended to investors with and without sustainability preferences. Specifically, ESG-oriented investors may prioritize ESG-related investment outcomes over conventional mean-variance considerations (Hong and Kacperczyk, 2009; Borgers et al., 2015). As a consequence, the performance of portfolios recommended to ESG-oriented investors cannot be measured using standard risk-adjusted performance measures. Thus, following the reasoning in Pedersen et al. (2021), we assess mean-variance optimization separately for non-ESG investors and ESG investors. The conventional mean-variance frontier is based on the entire universe of stocks in our experiment. The ESG mean-variance frontier is based on a restricted sample of stocks whose ESG score exceeds 40.

Figure 1 depicts the risk-return profiles of LLM-generated portfolio recommendations in the baseline condition, as well as the two mean-variance frontiers. Intuitively, restricting the investment universe through ESG screening results in an inferior mean-variance frontier as portfolios do not make full use of the diversification potential. To measure a portfolio's mean-variance efficiency subject to ESG constraints, we use the shortest Euclidean distance from the portfolio's risk-return profile to the relevant mean-variance frontier. The Euclidean distance measure serves as an inverse measure of risk-adjusted performance (i.e., the closer to the respective mean-variance frontier, the more efficient).

[Figure 1 around here]

Table 15 compares excess returns, volatility, and the Euclidean distance measures by treatment condition and investor type. All three measures are computed only for the risky portion of the portfolios. A few things are worth noting. First, the numbers confirm our previous finding that human advisors do not account for ESG preferences in their portfolios, as there is no significant difference in returns (panel A) or risk (panel B). As a consequence, Euclidean distance to the frontier is significantly larger for non-ESG profiles (panel C). Second, the portfolios LLMs recommend to ESG-oriented investors display significantly lower returns

(panel A) than portfolios recommended to non-ESG investors, but mostly similar risk levels (panel B). This suggests that LLMs take into account ESG-oriented investors' preference for sustainable investments at the cost of financial returns. Given these preferences, there is no significant difference in the efficiency of the risky portfolio share as measured by the Euclidean distance from the relevant frontier (panel C). Third, even though human portfolio recommendations include more securities, the risky portion of the portfolios is less efficient than that of LLM recommendations (as measured by Euclidean distance from the relevant frontier, panel C). This is true for both ESG investors and non-ESG investors and implies that the performance difference between human advisors and LLMs stems from a greater, but no more diversified, risky share. Fourth, when domain-specific information is provided, Euclidean distances increase for both ESG investors and non-ESG investors (panel C). This suggests that the performance improvements observed in Table 14 is attributable only to an increase in the portfolios' risky shares and not to an increase in efficiency within the risky portfolio share (see Table 16).

[Tables 15 and 16 around here]

**Result 8.** *Providing domain-specific information does not increase mean-variance efficiency in LLM recommendations.*

### 5.4.2 Out-of-sample tests

In addition to the in-sample tests, we perform out-of-sample tests to avoid look-ahead bias in our performance analyses. Because data collection was completed by November 3, 2024, we compute out-of-sample performance as the average daily performance from November 4, 2024, to March 21, 2025 (the most recent data at the time of analysis).

Table 17 reports regression coefficients for the out-of-sample tests. Consistent with our in-sample results, LLM-generated portfolios underperform human portfolios by 6 basis points in terms of daily (unadjusted) returns. This pattern holds true independent of the investment horizon we specify for the investor profile. However, in line with the substantially lower risky share in LLM portfolios, daily volatility is significantly lower in LLM portfolios.

The performance improvements of adding domain-specific information observed in the in-sample tests are not fully replicated in the out-of-sample tests. While returns are significantly higher when qualitative firm-specific information (treatment $T_3$) is provided, returns are lower when quantitative information (treatment $T_2$) or all types of information (treatment $T_4$) are provided. In line with a greater risky share, adding domain-specific information significantly increases portfolio volatility.

[Table 17 around here]

**Result 9.** *Providing firm-specific information does not generally improve performance in out-of-sample tests. While returns increase when qualitative information is provided, risk increases for all firm-specific information.*

# 6 Conclusion

In this study, we investigate the effect of additional domain-specific information on LLM performance in an applied portfolio allocation task. We use a controlled environment to introduce various types of financial information to seven LLMs from the most highly regarded, state-of-the art developers. We find that while LLM recommendations generally underperform recommendations by human financial advisors, providing firm-specific information improves historical performance in LLM portfolios and closes the gap with human advisors. Performance improvements are achieved through higher exposure to market risk and not through an increase in mean-variance efficiency within the risky portfolio share. We further document that quantitative firm-specific information affects recommendations more than qualitative firm-specific information and that providing generic finance theory does not affect recommendations. The performance results are not fully reproduced using a short out-of-sample test window.

Our results have implications for both research and practice. Improving LLM performance through the addition of domain-specific information may seem like a foregone conclusion. However, our results suggest that more domain expertise does not lead to an unambiguous quality improvement in LLM responses in the financial advice context. The availability of firm-specific information does reduce attention-based buying and aligns LLM recommendations with those of professional financial advisors in terms of both exposure and historical performance. Firm-specific information nevertheless substantially increases risk, severely limiting the suitability of these portfolios for risk-averse investors. To avoid such a mismatch, financial advisors could restrict the LLM's autonomy in assigning portfolio weights based on the investor's risk profile, trading off the benefits of more individualized recommendations against the need to limit misalignment between portfolio exposure and investor preferences. A similar argument can be made with regard to allocation within the risky portfolio share. Our results suggest that providing firm-specific information does not increase mean-variance efficiency. A potential remedy could be to pre-determine the asset mix of the risky portfolio share, reducing the role of the applied LLM to a slightly more sophisticated version of the risk profiling and portfolio allocation algorithms already used by robo-advisors (Jung et al., 2018; Rühr et al., 2019; Litterscheidt and Streich, 2020). This, in turn, may jeopardize the higher user acceptance typically associated with intelligent systems (Dietvorst et al., 2015; Berger et al., 2021).

From a development perspective, we add to a number of studies from various domains documenting that LLMs inherit prevalent human biases, most likely from their training data (e.g., Fieberg et al., 2024; Lou and Sun, 2024). Specifically, we find that the addition of firm-specific information increases risk-seeking in the LLMs' responses. Although we cannot pin down the exact reason for this relationship due to the black-box nature of LLMs, we offer two interpretations. First, this pattern might resemble the ambiguity aversion identified in human decision-makers. Ambiguity-averse decision-makers are less willing to accept a risky gamble if

the probability distribution of its outcome is unknown. Thus, by providing more information (in this case on the distribution of future returns), ambiguity is reduced, and LLMs are more willing to accept the risky gamble. Alternatively, the provision of firm-specific information may be interpreted by the LLMs as a hint to include more stocks in a portfolio. This behavior is in line with desirability bias and can arise either because the training data of an LLM are biased in a similar way or because reinforcement learning from human feedback induces this pattern as a desirable property of the LLM (cf. Lou and Sun, 2024). Note that, while the ambiguity aversion interpretation assumes that the provided information has merit because it allows LLMs to form return expectations, the desirability interpretation does not make this assumption. In any case, our finding challenges the notion that AI will automatically provide an improvement over human decision-making without further refinement.

Several avenues for further research emerge. First, future studies might investigate in a systematic manner whether LLMs can distinguish between relevant and irrelevant domain-specific information, which could occur at the retriever stage or the augmented response stage (or both) of the RAG pipeline. If LLMs are unable to identify relevant pieces of information for their specific task, relevance has to be ensured through the curation of the domain-specific datasets used for RAG. Second, based on our finding that portfolio risk in LLM recommendations increases when firm-specific information is provided, a thorough analysis of the drivers of risk tolerance in LLM recommendations is warranted.

## References

Aghassi, M., Asness, C., Fattouche, C., and Moskowitz, T. (2023). Fact, fiction, and factor investing. *The Journal of Portfolio Management*, 49(2):57–94.

Alawwad, H. A., Alhothali, A., Naseem, U., Alkhathlan, A., and Jamal, A. (2024). Enhancing textbook question answering task with large language models and retrieval augmented generation. *arXiv preprint arXiv:2402.05128*.

Alonso, M. N. i. (2024). Look-ahead bias in large language models (llms): Implications and applications in finance. *Available at SSRN*.

Andrei, D., Friedman, H., and Ozel, N. B. (2023). Economic uncertainty and investor attention. *Journal of Financial Economics*, 149(2):179–217.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2):785–818.

Berger, B., Adam, M., Rühr, A., and Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, 63(1):55–68.

Berk, J. and De Marzo, P. (2020). Corporate Finance, Global Edition (5th edition).

Blaseg, D. and Hornuf, L. (2024). Playing the business angel: The impact of well-known business angels on venture performance. *Entrepreneurship Theory and Practice*, 48(1):171–204.

Borgers, A., Derwall, J., Koedijk, K., and Ter Horst, J. (2015). Do social factors influence investment behavior and performance? evidence from mutual fund holdings. *Journal of Banking & Finance*, 60:112–126.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of ai. *The Review of Financial Studies*, 36(9):3603–3642.

Cardak, B. A. and Wilkins, R. (2009). The determinants of household risky asset holdings: Australian evidence on background risk and other factors. *Journal of Banking and Finance*, 33(5):850–860.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82.

Chapkovski, P., Khapko, M., and Zoican, M. (2024). Trading gamification and investor behavior. *Management Science*.

Chen, H., Cohen, L., Gurun, U., Lou, D., and Malloy, C. (2020). Iq from ip: Simplifying search in portfolio choice. *Journal of Financial Economics*, 138(1):118–137.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., and Van Durme, B. (2024). Dated data: Tracing knowledge cutoffs in large language models. *arXiv preprint arXiv:2403.12958*.

Claussen, J., Litterscheidt, R., and Streich, D. J. (2020). Seeking Analysts: User-Generated Analyst Coverage and Stock Market Quality.

Cocco, J. F., Gomes, F. J., and Maenhout, P. J. (2005). Consumption and portfolio choice over the life cycle. *The Review of Financial Studies*, 18(2):491–533.

Cordes, H., Nolte, S., and Schneider, J. C. (2023). Dynamics of stock market developments, financial behavior, and emotions. *Journal of Banking & Finance*, 154:106711.

Corter, J. E. and Chen, Y.-J. (2006). Do investment risk tolerance attitudes predict portfolio risk? *Journal of business and psychology*, 20:369–381.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.

Davies, G. B. (2017). New Vistas in Risk Profiling.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS quarterly*.

deHaan, E., Lawrence, A., and Litjens, R. (2024). Measuring investor attention using google search. *Management Science*.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Döttling, R. and Kim, S. (2024). Sustainability preferences under stress: Evidence from covid-19. *Journal of Financial and Quantitative Analysis*, 59(2):435–473.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669.

ESMA (2018). Guidelines on certain aspects of the MiFID II suitability requirements 06/11/2018 — ESMA35-43-1163. https://www.esma.europa.eu/sites/default/files/library/esma35-43-1163_guidelines_on_certain_aspects_of_mifid_ii_suitability_requirements_0.pdf. [online; accessed 19 August 2024].

ESMA (2023). Guidelines on certain aspects of the MiFID II suitability requirements 06/11/2018 — ESMA35-43-1163. https://www.esma.europa.eu/sites/default/files/2023-04/ESMA35-43-3172_Guidelines_on_certain_aspects_of_the_MiFID_II_suitability_requirements.pdf. [online; accessed 21 November 2024].

Fairhurst, D. J. and Greene, D. (2024). How much does chatgpt know about finance? *Financial Analyst Journal*. forthcoming.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Fieberg, C., Hornuf, L., and Streich, D. (2024). Using large language models for financial advice. *Available at SSRN 4850039*.

Fieberg, C., Hornuf, L., and Streich, D. J. (2023). Using GPT-4 for financial advice. *Available at SSRN 4499485*.

FINRA (2020). FINRA Rule 2111: Suitability. https://www.finra.org/rules-guidance/rulebooks/finra-rules/2111. Accessed: 2024-11-12.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., and Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10):1–13.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., and Herzig, J. (2024). Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Goyal, T., Li, J. J., and Durrett, G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Hansson, B. and Persson, M. (2000). Time diversification and estimation risk. *Financial Analysts Journal*, 56(5):55–62.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hong, H. and Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics*, 93(1):15–36.

Hornuf, L., Streich, D., and Töllich, N. (2024). Domain-specific knowledge and LLM performance. https://www.socialscienceregistry.org/trials/14755.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *The Review of financial studies*, 33(5):2019–2133.

Huang, J., Cui, L., Wang, A., Yang, C., Liao, X., Song, L., Yao, J., and Su, J. (2024). Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023a). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., and Feng, Y. (2023b). Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Huber, C. and Huber, J. (2020). Bad bankers no more? truth-telling and (dis) honesty in the finance industry. *Journal of Economic Behavior & Organization*, 180:472–493.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Jha, M., Qian, J., Weber, M., and Yang, B. (2024). ChatGPT and corporate policies. *Available at SSRN 4521096*.

Jung, D., Dorner, V., Glaser, F., and Morana, S. (2018). Robo-advisory: Digitalization and automation of financial advisory. *Business and Information Systems Engineering*, 60(1):81–86.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 20, 2024.

Kim, A., Muhn, M., and Nikolaev, V. (2023). Bloated disclosures: Can ChatGPT help investors process financial information? *arXiv preprint arXiv:2306.10224*.

Ko, H. and Lee, J. (2024). Can chatgpt improve investment decisions? from a portfolio management perspective. *Finance Research Letters*, 64:105433.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. (2023). Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., and Shah, S. (2023a). Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422.

Li, X., Li, Z., Shi, C., Xu, Y., Du, Q., Tan, M., Huang, J., and Lin, W. (2024a). Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. (2023b). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H.-T., Xie, P., Huang, F., and Jiang, Y. (2024b). Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590.

Litterscheidt, R. and Streich, D. J. (2020). Financial education and digital asset management: What's in the black box? *Journal of Behavioral and Experimental Economics*, 87.

Liu, Z., Huang, D., Huang, K., Li, Z., and Zhao, J. (2021). Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

Lo, A. W. and Ross, J. (2024). Generative AI from theory to practice: A case study of financial advice. https://mit-genai.pubpub.org/pub/l89uu140/release/2#:~:text=A%20finance%2Dspecific%20LLM%20will,agents%20within%20the%20financial%20system. [online, accessed 19 August 2024].

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Lou, J. and Sun, Y. (2024). Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. (2023). An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Lusardi, A. and Mitchell, O. S. (2011). Financial Literacy around The World: An Overview. *Journal of Pension Economics and Finance*, 10(4):497–508.

Lusardi, A. and Mitchell, O. S. (2023). The importance of financial literacy: Opening a new field. *Journal of Economic Perspectives*, 37(4):137–154.

Markey, N., El-Mansouri, I., Rensonnet, G., van Langen, C., and Meier, C. (2024). From rags to riches: Using large language models to write documents for clinical trials. *arXiv preprint arXiv:2402.16406*.

Mola, S., Rau, P. R., and Khorana, A. (2013). Is there life after the complete loss of analyst coverage? *Accounting Review*, 88(2):667–705.

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., and Shoham, Y. (2023). Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Nguyen, L., Gallery, G., and Newton, C. (2019). The joint influence of financial risk perception and risk tolerance on individual investment decision-making. *Accounting & Finance*, 59:747–771.

Niszczota, P. and Abbas, S. (2023). GPT has become financially literate: Insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice. *Finance Research Letters*, 58:104333.

Oehler, A. and Horn, M. (2024). Does chatgpt provide better advice than robo-advisors? *Finance Research Letters*, 60:104898.

Palan, S. and Schitter, C. (2018). Prolific.ac — a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Pedersen, L. H., Fitzgibbons, S., and Pomorski, L. (2021). Responsible investing: The esg-efficient frontier. *Journal of financial economics*, 142(2):572–597.

Pelster, M. and Val, J. (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters*, 59:104786.

Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., and Riedel, S. (2020). How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Pu, X., Gao, M., and Wan, X. (2023). Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Rosenberg, B., Reid, K., and Lanstein, R. (1985). Persuasive evidence of market inefficiency. *Journal of portfolio management*, 11(3):9–16.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al. (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Rühr, A., Streich, D., Berger, B., and Hess, T. (2019). A classification of decision automation and delegation in digital investment management systems. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pages 1435–1444.

Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. *Available at SSRN*.

SEC (2019). SEC Final Rule Release No. 34-86031: Regulation Best Interest: The Broker-Dealer Standard of Conduct. https://www.sec.gov/files/rules/final/2019/34-86031.pdf. Accessed: 2024-11-12.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., and Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Siemroth, C. and Hornuf, L. (2023). Why do retail investors pick green investments? a lab-in-the-field experiment with crowdfunders. *Journal of Economic Behavior & Organization*, 209:74–90.

Smith, G. (2024). LLMs can't be trusted for financial advice. *Journal of Financial Planning*, 37(5).

Streich, D. J. (2023). Risk preference elicitation and financial advice taking. *Journal of Behavioral Finance*, 24(3):259–275.

Subbiah, M., Zhang, S., Chilton, L. B., and McKeown, K. (2024). Reading subtext: Evaluating large language models on short story summarization with writers.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *Review of Economic Studies*, 25(2):65–86.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Wu, S., Xie, J., Chen, J., Zhu, T., Zhang, K., and Xiao, Y. (2024). How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., and Huang, J. (2023). Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., and Mao, Z. (2023). Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Yang, Y., Tang, Y., and Tam, K. Y. (2023). Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

Yang, Y., Uy, M. C. S., and Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Yoran, O., Wolfson, T., Ram, O., and Berant, J. (2023). Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., et al. (2024). Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Zhang, B., Yang, H., Zhou, T., Ali Babar, M., and Liu, X.-Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 349–356.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhang, Y., Zhang, X., Wang, X., qing Chen, S., and Wei, F. (2022). Latent prompt tuning for text summarization.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Table 1: Domain-specific knowledge injection and LLM performance (literature overview)

| Author & Year | Domain | Injection type | Author type | Improvement over base model? | Base models | Improvement over benchmark model? | Benchmark models |
|---|---|---|---|---|---|---|---|
| Roziere et al. (2023) | Code | Fine-tuning | Academics + developers | ✓ | LLaMA 2 | × | GPT-4 |
| Roziere et al. (2023) | Code | Fine-tuning | Academics + developers | ✓ | LLaMA 2 | ✓ | GPT-3.5, StarCoder, PaLM, Codex, GPT-Neo, AlphaCode |
| Chen et al. (2021) | Code | Fine-tuning | Developer | ✓ | GPT-3 | ✓ | GPT-Neo, TabNine, GPT-J |
| Li et al. (2024b) | E-commerce | Fine-tuning | Academics + developers | ✓ | BLOOMZ | ✓ | ChatGPT, BLOOM |
| Xie et al. (2023) | Finance | Fine-tuning | Academics | | | × | GPT-4 |
| Xie et al. (2023) | Finance | Fine-tuning | Academics | | | ✓ | ChatGPT, GPT NeoX, OPT, BLOOM, BloombergGPT |
| Yang et al. (2023) | Finance | Fine-tuning | Academics | ✓ | LLaMA | × | GPT-4 |
| Yang et al. (2023) | Finance | Fine-tuning | Academics | ✓ | LLaMA | ✓ | BloombergGPT |
| Zhang et al. (2023) | Finance | Fine-tuning | Academics | ✓ | LLaMA | ✓ | FinBERT, BloombergGPT, ChatGLM2, ChatGPT 4.0 |
| Zhang et al. (2023) | Finance | RAG | Academics | ✓ | ChatGPT 4.0, instruction-tuned LLaMA | | |
| Xiong et al. (2024) | Medicine | RAG | Academics | ✓ | GPT-4, GPT-3.5, Mistral, LLaMA 2, MEDITRON, PMC-LLaMA | | |
| Zakka et al. (2024) | Medicine | RAG | Academics | ✓ | GPT-4, Bard, Bing | | |
| Markey et al. (2024) | Medicine | RAG | Developers | ✓ | GPT-4 | | |
| Lewis et al. (2020) | Generic | RAG | Academics + developers | ✓ | BART | ✓ | T5 |
| Muhlgay et al. (2023) | Generic | RAG | Developers | ✓ | GPT-Neo, OPT | | |
| Ram et al. (2023) | Generic | RAG | Developers | ✓ | GPT-2, GPT-Neo, GPT-J, OPT, LLaMA | | |
| Li et al. (2024a) | Finance | Fine-tuning + RAG | Academics + developers | | | ✓ | FinMA, ChatGLM, FinGPT, ChatGPT (3.5 Turbo) |
| Huang et al. (2023b) | Law | Fine-tuning + RAG | Academics | ✓ | Chinese LLaMA | | |
| Li et al. (2023b) | Medicine | Fine-tuning + RAG | Academics | | | ✓ | ChatGPT |
| Alawwad et al. (2024) | Generic | Fine-tuning + RAG | Academics | ✓ | LLaMA 2 | | |

Table 2: Stock universe and firm-level quantitative information

| Company | Stock ticker | Market beta | Book-to-market ratio | Market cap (B USD) | Momentum | Earnings-to-price ratio | ESG score |
|---|---|---|---|---|---|---|---|
| *Panel A: Large-cap value stocks* | | | | | | | |
| Berkshire Hathaway Inc | BRK.B | 0.88 | 0.62 | 974 | 0.30 | 0.07 | 25 |
| Cincinnati Financial Corp | CINF | 0.66 | 0.60 | 21 | 0.27 | 0.10 | 57 |
| Eastman Chemical Co | EMN | 1.45 | 0.46 | 12 | 0.34 | 0.07 | 83 |
| *Panel B: Large-cap growth stocks* | | | | | | | |
| PepsiCo Inc | PEP | 0.53 | 0.08 | 243 | 0.01 | 0.04 | 86 |
| Lockheed Martin Corp | LMT | 0.46 | 0.05 | 137 | 0.40 | 0.05 | 70 |
| Kimberly-Clark Corp | KMB | 0.40 | 0.02 | 49 | 0.17 | 0.05 | 75 |
| *Panel C: Small-cap value stocks* | | | | | | | |
| Air Lease Corp | AL | 1.61 | 1.49 | 5 | 0.11 | 0.11 | 54 |
| S&T Bancorp Inc | STBA | 0.78 | 0.80 | 2 | 0.56 | 0.08 | 45 |
| Sturm Ruger & Company Inc | RGR | 0.16 | 0.45 | 1 | -0.18 | 0.05 | 42 |
| *Panel D: Small-cap growth stocks* | | | | | | | |
| Alkermes Plc | ALKS | 0.44 | 0.28 | 5 | -0.06 | 0.10 | 58 |
| St Joe Co | JOE | 1.26 | 0.21 | 3 | 0.04 | 0.02 | 41 |
| Evertec Inc | EVTC | 1.15 | 0.22 | 2 | -0.10 | 0.03 | 24 |

**Note:** Data were retrieved from Thomson Reuters in September 2024. Market beta is the regression coefficient of a CAPM regression of the stock's excess returns on the excess returns of a market portfolio (S&P 500 index/ NASDAQ Composite Index). Book-to-market ratio is the ratio of the book value of equity per share to its share price. Market cap is the market value of all outstanding shares. Momentum is defined as the cumulative stock return over the previous twelve months. Earnings-to-price ratio is the earnings per share over the last twelve months divided by the current share price. The ESG score is a self-reported measure of the firm's adherence to ESG principles.

Table 3: Experimental conditions

| | Provided information | | | | |
|---|---|---|---|---|---|
| | Investor profile | Investment universe | General investment theory | Quantitative stock metrics | Qualitative company information (10-K) |
| Baseline | ✓ | ✓ | | | |
| Treatment 1 | ✓ | ✓ | ✓ | | |
| Treatment 2 | ✓ | ✓ | | ✓ | |
| Treatment 3 | ✓ | ✓ | | | ✓ |
| Treatment 4 | ✓ | ✓ | ✓ | ✓ | ✓ |

**Note:** Basic investment theory comprises a summary of the relevant chapters of a standard finance textbook used for undergraduate-level finance courses (Berk and De Marzo, 2020, Chapters 9–11, see Section 3.3.2 for details). Quantitative financial metrics comprise each stock's CAPM market beta, book-to-market ratio of equity, market capitalization, momentum (previous 12-month cumulative return), earnings-to-price ratio, and Thomson Reuters ESG score (see Table 2). All data are retrieved from Thomson Reuters Eikon. In addition to the stock-level metrics, a brief definition of each metric is provided (see Table A5 in the Appendix). Qualitative company information (10-K) comprises for each stock a summary of the respective company's latest 10-K management discussion & analysis (MD&A) section (see Section 3.3.4 for details).

Table 4: Model overview

| | | | Information cut-off date | |
|---|---|---|---|---|
| Model | Developer | Size (B) | Reported | Revealed |
| GPT-4o | OpenAI | Not reported | October 2023 | December 2023 |
| GPT-4o-mini | OpenAI | Not reported | October 2023 | January 2024 |
| GPT-4-turbo | OpenAI | Not reported | December 2023 | November 2023 |
| Llama-3.1-8B-Instruct | Meta | 8.0 | December 2023 | May 2023 |
| Mistral-7B-Instruct-v0.3 | MistralAI | 7.3 | Not reported | July 2023 |
| Phi-3-small-128k-instruct | Microsoft | 7.4 | October 2023 | May 2023 |
| Qwen2.5-7B-Instruct | Qwen (Alibaba) | 7.6 | Not reported | October 2023 |

**Note:** The table contains the 7 models used in our experiment. Size refers to the number of parameters in billions. Reported cut-off dates are taken from the respective model's documentation. Revealed cut-off dates are obtained by assessing each model's knowledge of specific events in the period from January 2023 to June 2024 (see Tables A3 and A2 in Online Appendix A for details).

Table 5: Prompt design

| # | Prompt component | Text | Comment |
|---|---|---|---|
| 1 | Role prompting | You are a financial expert specializing in personalized portfolio management. With a deep understanding of your client's individual circumstances, you can craft tailored investment recommendations. | Assigning the role of a financial expert may increase the accuracy of the model's responses (White et al., 2023; Kong et al., 2023; Xu et al., 2023). |
| 2 | General instruction | Please create a portfolio recommendation for an investor with a [high/low] risk tolerance, [preference/no preference] for sustainable investment, and a [1 month/ 6 month/ 12 month] investment horizon. | The instruction is standardized across profiles, with the placeholders replaced by the actual investor characteristic (12 profiles). |
| 3 | Security universe | The available security universe includes the following securities: [+security names + tickers]. | The security universe (12 stocks, 1 bond fund) is included as a JSON file. Securities are sorted in a randomized order to avoid order effects. |
| 4 | Additional information (by treatment condition) | | Additional information is provided according to the experimental conditions. |
| | *Baseline* | - | |
| | *Treatment 1 (investment theory)* | Additionally, you may consider the following summary of basic investment theory: [+textbook summary]. | |
| | *Treatment 2 (quantitative indicators)* | Additionally, you may consider the following financial metrics for the 12 stocks: [+stock names / tickers + metrics]. The metrics are defined as follows: [+definitions]. | Information is provided as a JSON file. Stocks are sorted in randomized order to avoid order effects. |
| | *Treatment 3 (10-K summaries)* | Additionally, you may consider the following summaries of the 12 stocks' annual reports: [+stock names / tickers + 10-K MD&A summaries]. | Information is provided as a JSON file. Stocks are sorted in randomized order to avoid order effects. |
| | *Treatment 4 (investment theory + quantitative indicators + 10-K summaries)* | Additionally, you may consider the following information. First, you may consider the following summary of basic investment theory: [+ textbook summary]. Second, you may consider the following financial metrics for the 12 stocks: [+stock names / tickers + metrics]. The metrics are defined as follows: [+definitions]. Third, you may consider the following summaries of the 12 stocks' annual reports: [+ stock names / tickers + 10-K MD&A summaries]. | Information is provided as a JSON file. Stocks are sorted in randomized order to avoid order effects. |
| 5 | Specific instruction | Please provide the recommendation in a table format, including the ticker symbol and the recommended percentage for each security. | Specific instructions ensure responses are recorded in the desired way. Table format allows for simple data processing. |
| 6 | Error correction | Ensure that the total allocation sums up to 100 percent. | Correction prompt to avoid common portfolio allocation error (cf. Fieberg et al., 2023). |
| 7 | Chain of thought | Let's think step by step. | Chain-of-thought prompt may improve accuracy of responses, especially for quantitative tasks (Wei et al., 2022; Kojima et al., 2022). |

Table 6: Summary statistics: Portfolio recommendations

| | N | Mean | SD | p5 | p50 | p95 |
|---|---|---|---|---|---|---|
| *Panel A: LLM-generated recommendations* | | | | | | |
| No. securities | 419 | 6.37 | 2.3 | 3.0 | 6.0 | 11.0 |
| Herfindahl index (HHI) | 419 | 0.25 | 0.1 | 0.1 | 0.2 | 0.5 |
| Fixed income share | 419 | 0.23 | 0.2 | 0.0 | 0.1 | 0.7 |
| Risky share | 419 | 0.77 | 0.2 | 0.3 | 0.9 | 1.0 |
| Monthly volatility (%) | 419 | 3.80 | 1.4 | 1.9 | 3.9 | 6.0 |
| Idiosyncratic volatiliy (%) | 419 | 2.11 | 0.7 | 1.2 | 2.0 | 3.3 |
| FF6 market beta | 419 | 0.67 | 0.2 | 0.3 | 0.7 | 1.0 |
| Monthly excess return (%) | 419 | 0.47 | 0.3 | 0.0 | 0.5 | 0.8 |
| Annual Sharpe Ratio | 419 | 0.38 | 0.2 | 0.0 | 0.4 | 0.6 |
| Monthly FF6 alpha (%) | 419 | -0.22 | 0.1 | -0.4 | -0.2 | 0.0 |
| No. trades p.a. | 419 | 76.32 | 27.9 | 36.0 | 72.0 | 132.0 |
| Annual turnover | 419 | 0.41 | 0.1 | 0.2 | 0.4 | 0.6 |
| Avg. ESG score | 416 | 62.58 | 8.9 | 49.4 | 63.3 | 76.9 |
| *Panel B: Human recommendations* | | | | | | |
| No. securities | 1,068 | 7.76 | 4.0 | 2.0 | 7.0 | 13.0 |
| Herfindahl index (HHI) | 1,068 | 0.25 | 0.2 | 0.1 | 0.2 | 0.7 |
| Fixed income share | 1,068 | 0.16 | 0.2 | 0.0 | 0.1 | 0.7 |
| Risky share | 1,068 | 0.84 | 0.2 | 0.3 | 0.9 | 1.0 |
| Monthly volatility (%) | 1,068 | 4.33 | 1.3 | 1.8 | 4.3 | 6.6 |
| Idiosyncratic volatiliy (%) | 1,068 | 2.48 | 1.1 | 1.2 | 2.2 | 4.5 |
| FF6 market beta | 1,068 | 0.72 | 0.2 | 0.3 | 0.8 | 1.0 |
| Monthly excess return (%) | 1,068 | 0.56 | 0.2 | 0.0 | 0.6 | 0.8 |
| Annual Sharpe Ratio | 1,068 | 0.42 | 0.2 | 0.0 | 0.5 | 0.6 |
| Monthly FF6 alpha (%) | 1,068 | -0.12 | 0.1 | -0.4 | -0.1 | 0.1 |
| No. trades p.a. | 1,068 | 92.56 | 49.1 | 24.0 | 84.0 | 156.0 |
| Annual turnover | 1,068 | 0.47 | 0.2 | 0.2 | 0.5 | 0.6 |
| Avg. ESG score | 1,038 | 55.36 | 9.5 | 39.5 | 55.4 | 70.9 |

**Note:** The restricted sample only draws on the recommendations of human financial advisors that have passed stricter attention checks than the full sample (N = 89, see Appendix C for details).

Table 7: Portfolio weights by experimental condition

| | Human | LLM | | | | |
|---|---|---|---|---|---|---|
| | Baseline | Baseline | Treatment 1 | Treatment 2 | Treatment 3 | Treatment 4 |
| | | | *Stocks* | | | |
| *Panel A: Large-cap value stocks* | | | | | | |
| Berkshire Hathaway Inc | 0.09 | 0.08 | 0.06 | 0.05 | 0.07 | 0.06 |
| Cincinnati Financial Corp | 0.07 | 0.06 | 0.05 | 0.08 | 0.08 | 0.10 |
| Eastman Chemical Co | 0.06 | 0.05 | 0.06 | 0.10 | 0.05 | 0.09 |
| | 0.22 | 0.19 | 0.17 | 0.22 | 0.20 | 0.25 |
| *Penal B: Large-cap growth stocks* | | | | | | |
| PepsiCo Inc | 0.10 | 0.12 | 0.10 | 0.15 | 0.11 | 0.15 |
| Lockheed Martin Corp | 0.06 | 0.06 | 0.05 | 0.12 | 0.06 | 0.09 |
| Kimberly-Clark Corp | 0.06 | 0.11 | 0.11 | 0.13 | 0.09 | 0.12 |
| | 0.22 | 0.29 | 0.26 | 0.40 | 0.26 | 0.36 |
| *Panel C: Small-cap value stocks* | | | | | | |
| Air Lease Corp | 0.06 | 0.04 | 0.05 | 0.06 | 0.07 | 0.07 |
| S&T Bancorp Inc | 0.06 | 0.01 | 0.03 | 0.03 | 0.03 | 0.04 |
| Sturm Ruger & Company Inc | 0.06 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 |
| | 0.18 | 0.09 | 0.11 | 0.11 | 0.12 | 0.12 |
| *Panel D: Small-cap growth stocks* | | | | | | |
| Alkermers Plc | 0.08 | 0.05 | 0.06 | 0.05 | 0.06 | 0.04 |
| St Joe Co | 0.07 | 0.04 | 0.05 | 0.03 | 0.04 | 0.05 |
| Evertec Inc | 0.07 | 0.04 | 0.04 | 0.01 | 0.09 | 0.03 |
| | 0.22 | 0.13 | 0.15 | 0.09 | 0.19 | 0.11 |
| ***Total stocks*** | ***0.84*** | ***0.70*** | ***0.70*** | ***0.82*** | ***0.78*** | ***0.84*** |
| | | | *Bonds* | | | |
| Vanguard Bond Market ETF | 0.16 | 0.30 | 0.30 | 0.18 | 0.22 | 0.16 |

**Note:** The table displays the average portfolio weights of the 13 securities making up the investable universe, separately by treatment condition.

Table 8: Univariate determinants of stock recommendations (baseline condition)

| | (1) Human | | | | | (2) LLM | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Stocks not recommended | Stocks recommended | Δ | t stat. | | Stocks not recommended | Stocks recommended | Δ | t stat. | |
| *Panel A: Fundamentals* | | | | | | | | | | |
| Market beta | 0.83 | 0.81 | -0.02 | 2.67 | *** | 0.84 | 0.79 | -0.05 | 1.67 | * |
| Book-to-market ratio | 0.46 | 0.43 | -0.03 | 4.63 | *** | 0.48 | 0.39 | -0.09 | 3.78 | *** |
| Market cap (B USD) | 105.82 | 130.58 | 24.77 | -5.12 | *** | 84.70 | 160.68 | 75.98 | -4.56 | *** |
| Momentum | 0.16 | 0.15 | 0.00 | 0.52 | | 0.15 | 0.16 | 0.01 | -0.42 | |
| E/P ratio | 0.06 | 0.06 | 0.00 | 2.03 | ** | 0.06 | 0.06 | 0.00 | 1.54 | |
| ESG score | 54.44 | 55.38 | 0.94 | -2.59 | *** | 52.26 | 58.01 | 5.75 | -4.61 | *** |
| *Panel B: Investor attention* | | | | | | | | | | |
| SEC EDGAR downloads (k) | 7.66 | 8.70 | 1.04 | -4.62 | *** | 6.74 | 9.99 | 3.25 | -4.19 | *** |
| Google SVI | 4.37 | 4.98 | 0.62 | -4.72 | *** | 3.73 | 5.84 | 2.11 | -4.70 | *** |
| News articles | 24.02 | 27.22 | 3.20 | -3.96 | *** | 20.65 | 31.81 | 11.16 | -4.00 | *** |
| Analyst reports | 17.84 | 21.61 | 3.76 | -3.79 | *** | 15.13 | 25.64 | 10.51 | -3.06 | *** |
| *Panel C: 10-K sentiment* | | | | | | | | | | |
| Sentiment in 10-K MD&A section | 0.22 | 0.21 | -0.01 | 2.21 | ** | 0.21 | 0.21 | 0.00 | -0.11 | |

**Note:** The table displays equal-weighted averages for the fundamental stock and investor attention measures for stocks, separately by whether they are included or not included in portfolio recommendations generated by human financial advisors (column 1) and LLMs (column 2). The table also reports $t$ statistics of two-sample t-tests, for which significance levels are indicated by asterisks ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table 9: Contributions to explained variation in portfolio weights

| | Human | LLM | | | | |
|---|---|---|---|---|---|---|
| DV: Stock weight | (1) Baseline | (2) Baseline | (3) $T_1$ (theory) | (4) $T_2$ (quant.) | (5) $T_3$ (qual.) | (6) $T_4$ (all) |
| **Fundamentals** | **0.62** | **0.77** | **0.82** | **0.89** | **0.58** | **0.84** |
| Market beta | 0.03 | 0.09 | 0.10 | 0.06 | 0.09 | 0.05 |
| Book-to-market ratio | 0.07 | 0.10 | 0.12 | 0.09 | 0.07 | 0.08 |
| Market cap | 0.41 | 0.14 | 0.10 | 0.06 | 0.13 | 0.09 |
| Momentum | 0.03 | 0.02 | 0.02 | 0.04 | 0.01 | 0.04 |
| Earnings-to-price ratio | 0.04 | 0.05 | 0.05 | 0.03 | 0.04 | 0.04 |
| ESG score | 0.04 | 0.37 | 0.44 | 0.60 | 0.24 | 0.56 |
| **Investor attention** | **0.34** | **0.11** | **0.08** | **0.05** | **0.10** | **0.07** |
| SEC EDGAR downloads | 0.34 | 0.11 | 0.08 | 0.05 | 0.10 | 0.07 |
| **Sentiment** | **0.04** | **0.12** | **0.10** | **0.06** | **0.31** | **0.08** |
| Sentiment in 10-K MD&A section | 0.04 | 0.12 | 0.10 | 0.06 | 0.31 | 0.08 |
| R2 | 0.02 | 0.21 | 0.14 | 0.30 | 0.10 | 0.21 |
| Obs. | 12,456 | 1,008 | 996 | 1,008 | 984 | 996 |

**Note:** The table reports the standardized dominance statistics of independent variables of regressions of each stock's portfolio weight on that stock's fundamentals, investor attention (as measured by the number of download requests registered in the SEC EDGAR database), and sentiment in the stock's 10-K management discussion and analysis (MD&A) section. Alternative attention measures are omitted due to collinearity (see Table A12 in the Appendix). Missing values arise from missing portfolio recommendations and portfolio recommendations with 100% fixed income allocation.

Table 10: Investor characteristics and portfolio risk

| | Human | LLM | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| DV: Risky share | Baseline | Baseline | $T_1$(theory) | $T_2$(quant.) | $T_3$(qual.) | $T_4$(all) |
| $\mathbb{1}$(Risk tolerance = high) | 0.155*** | 0.461*** | 0.490*** | 0.264*** | 0.412*** | 0.302*** |
| | (0.026) | (0.046) | (0.039) | (0.057) | (0.052) | (0.046) |
| $\mathbb{1}$(Horizon = 6 months) | 0.024** | 0.055 | 0.007 | -0.033 | 0.067** | 0.008 |
| | (0.010) | (0.047) | (0.032) | (0.054) | (0.024) | (0.028) |
| $\mathbb{1}$(Horizon = 12 months) | 0.038*** | 0.043 | 0.068* | -0.005 | 0.128*** | -0.020 |
| | (0.011) | (0.036) | (0.030) | (0.052) | (0.031) | (0.034) |
| Constant | 0.745*** | 0.427*** | 0.425*** | 0.664*** | 0.494*** | 0.698*** |
| | (0.029) | (0.045) | (0.053) | (0.074) | (0.054) | (0.069) |
| Obs. | 1,068 | 84 | 84 | 84 | 83 | 84 |
| Adj. R2 | 0.111 | 0.762 | 0.792 | 0.452 | 0.710 | 0.510 |
| Exp. condition | Baseline | Baseline | T1 (theory) | T1 (quant.) | T3 (qual.) | T4 (all) |
| Advisor | Human | LLM | LLM | LLM | LLM | LLM |
| *Covariates* | | | | | | |
| Sustainability preference | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Note:** The table reports regression coefficients from OLS regressions with the risky share as the dependent variable and investor characteristics as the independent variables. $\mathbb{1}$(High risk tolerance) equals 1 if the investor profile indicates high risk tolerance. The omitted category for the investment horizon dummies is 1 month. $\mathbb{1}$(Sustainability preference) equals 1 if the investor profile indicates a preference for sustainable investing, and 0 otherwise. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$). Missing values arise from missing portfolio recommendations.

Table 11: Investor characteristics and portfolio-level ESG score

| | Human | LLM | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| DV: Avg. ESG score | Baseline | Baseline | $T_1$(theory) | $T_2$(quant.) | $T_3$(qual.) | $T_4$(all) |
| $\mathbb{1}$(Sustainability preference = yes) | 0.067 | 2.524* | 6.710** | 7.582*** | 5.619*** | 6.915* |
| | (0.499) | (1.223) | (2.021) | (1.671) | (1.230) | (2.957) |
| Constant | 56.202*** | 65.101*** | 62.568*** | 67.313*** | 58.778*** | 63.583*** |
| | (0.878) | (1.879) | (3.661) | (0.744) | (2.148) | (1.549) |
| Obs. | 1,038 | 84 | 83 | 84 | 82 | 83 |
| Adj. R2 | 0.008 | 0.270 | 0.300 | 0.412 | 0.375 | 0.184 |
| *Covariates* | | | | | | |
| Risk tolerance | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Investment horizon | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Note:** The table reports regression coefficients from OLS regressions with the weighted portfolio-level Thomson Reuters ESG score as the dependent variable. The variable $\mathbb{1}$(Sustainability preference) equals 1 if the investor profile indicates a preference for sustainable investing, and 0 otherwise. All other investor profile characteristics (dummy variable indicating high risk tolerance, dummy variables indicating 6-month and 12-month investment horizons) are included as control variables, but omitted from the table. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$). Missing values arise from missing portfolio recommendations and portfolio recommendations with 100% fixed income allocation.

Table 12: Treatment impact on diversification

| | (1) No. securities | (2) No. securities | (3) HHI | (4) HHI | (5) HHI (equity) | (6) HHI (equity) |
|---|---|---|---|---|---|---|
| $\mathbb{1}(LLM)$ | -1.280*** | | 0.021 | | 0.001 | |
| | (0.459) | | (0.023) | | (0.020) | |
| $\mathbb{1}(T_1)$ | | -0.333 | | 0.010 | | 0.003 |
| | | (0.218) | | (0.008) | | (0.011) |
| $\mathbb{1}(T_2)$ | | -0.310 | | -0.048*** | | -0.020 |
| | | (0.304) | | (0.012) | | (0.015) |
| $\mathbb{1}(T_3)$ | | 0.257 | | -0.019 | | 0.005 |
| | | (0.504) | | (0.019) | | (0.042) |
| $\mathbb{1}(T_4)$ | | -0.131 | | -0.042** | | -0.014 |
| | | (0.285) | | (0.013) | | (0.015) |
| Constant | 7.739*** | 5.726*** | 0.230*** | 0.265*** | 0.245*** | 0.232*** |
| | (0.431) | (0.408) | (0.020) | (0.025) | (0.021) | (0.012) |
| Obs. | 1,152 | 419 | 1,152 | 419 | 1,122 | 416 |
| Adj. R2 | 0.010 | 0.443 | 0.037 | 0.474 | 0.009 | 0.401 |
| Profile FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Advisor FE | | ✓ | | ✓ | | ✓ |
| Exp. condition | Baseline | All | Baseline | All | Baseline | All |
| Advisor | Human + LLM | LLM | Human + LLM | LLM | Human + LLM | LLM |

**Note:** The table reports regression coefficients for OLS regressions of diversification measures on a dummy variable indicating that a recommendation is generated by an LLM (versus human financial advisor, $\mathbb{1}(LLM)$), and dummies indicating the four experimental conditions ($\mathbb{1}(T_1)$). Columns 1, 4, and 7 only include recommendations in the baseline condition; columns 2, 3, 5, 6, 8, and 9 only include LLM-generated recommendations. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

Table 13: Treatment impact on portfolio risk

| | (1) Risky share | (2) Risky share | (3) Vola (%) | (4) Vola (%) | (5) IVOL (%) | (6) IVOL (%) | (7) FF6 $\beta_M$ | (8) FF6 $\beta_M$ |
|---|---|---|---|---|---|---|---|---|
| $\mathbb{1}(LLM)$ | -0.138*** | | -0.896*** | | -0.564*** | | -0.112*** | |
| | (0.022) | | (0.138) | | (0.116) | | (0.017) | |
| $\mathbb{1}(T_1)$ | | -0.001 | | 0.119*** | | 0.126** | | 0.000 |
| | | (0.006) | | (0.030) | | (0.043) | | (0.003) |
| $\mathbb{1}(T_2)$ | | 0.121*** | | 0.551*** | | 0.255*** | | 0.104*** |
| | | (0.014) | | (0.106) | | (0.058) | | (0.019) |
| $\mathbb{1}(T_3)$ | | 0.077*** | | 0.543*** | | 0.323** | | 0.080*** |
| | | (0.021) | | (0.133) | | (0.115) | | (0.020) |
| $\mathbb{1}(T_4)$ | | 0.142*** | | 0.663*** | | 0.290** | | 0.120*** |
| | | (0.027) | | (0.132) | | (0.101) | | (0.022) |
| Constant | 0.927*** | 0.876*** | 4.889*** | 4.623*** | 2.859*** | 2.457*** | 0.806*** | 0.823*** |
| | (0.018) | (0.031) | (0.126) | (0.315) | (0.113) | (0.187) | (0.018) | (0.045) |
| Obs. | 1,152 | 419 | 1,152 | 419 | 1,152 | 419 | 1,152 | 419 |
| Adj. R2 | 0.160 | 0.694 | 0.155 | 0.723 | 0.087 | 0.480 | 0.163 | 0.778 |
| Profile FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Advisor FE | | ✓ | | ✓ | | ✓ | | ✓ |
| Exp. condition | Baseline | All | Baseline | All | Baseline | All | Baseline | All |
| Advisor | Human + LLM | LLM | Human + LLM | LLM | Human + LLM | LLM | Human + LLM | LLM |

**Note:** The table reports regression coefficients for OLS regressions of risk measures on a dummy variable indicating that a recommendation is generated by an LLM (versus human financial advisor, $\mathbb{1}(LLM)$), and dummies indicating the four experimental conditions ($\mathbb{1}(T_1)$). Columns 1, 4, and 7 only include recommendations in the baseline condition; columns 2, 3, 5, 6, 8, and 9 only include LLM-generated recommendations. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table 14: Treatment impact on performance: In-sample

| | (1) Excess return (%) | (2) Excess return (%) | (3) Sharpe ratio | (4) Sharpe ratio | (5) FF6 $\alpha$ (%) | (6) FF6 $\alpha$ (%) | (7) Annual turnover | (8) Annual turnover |
|---|---|---|---|---|---|---|---|---|
| $\mathbb{1}(LLM)$ | -0.154*** | | -0.070*** | | -0.101*** | | -0.079*** | |
| | (0.025) | | (0.025) | | (0.017) | | (0.016) | |
| $\mathbb{1}(T_1)$ | | -0.010 | | -0.027* | | 0.001 | | 0.017 |
| | | (0.011) | | (0.013) | | (0.011) | | (0.011) |
| $\mathbb{1}(T_2)$ | | 0.097*** | | 0.069*** | | -0.006 | | 0.019 |
| | | (0.021) | | (0.013) | | (0.016) | | (0.014) |
| $\mathbb{1}(T_3)$ | | 0.079** | | 0.030* | | 0.030 | | 0.038** |
| | | (0.022) | | (0.014) | | (0.021) | | (0.015) |
| $\mathbb{1}(T_4)$ | | 0.121*** | | 0.072** | | 0.015 | | 0.025 |
| | | (0.030) | | (0.021) | | (0.019) | | (0.014) |
| Constant | 0.667*** | 0.581*** | 0.473*** | 0.404*** | -0.084*** | -0.240*** | 0.517*** | 0.427*** |
| | (0.020) | (0.037) | (0.017) | (0.018) | (0.013) | (0.018) | (0.014) | (0.028) |
| Obs. | 1,152 | 419 | 1,152 | 419 | 1,152 | 419 | 1,152 | 419 |
| Adj. R2 | 0.173 | 0.746 | 0.077 | 0.480 | 0.100 | 0.420 | 0.126 | 0.677 |
| Profile FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Advisor FE | | ✓ | | ✓ | | ✓ | | ✓ |
| Exp. condition | Baseline | All | Baseline | All | Baseline | All | Baseline | All |
| Advisor | Human + LLM | LLM | Human + LLM | LLM | Human + LLM | LLM | Human + LLM | LLM |

**Note:** The table reports regression coefficients for OLS regressions of performance measures on a dummy variable indicating that a recommendation is generated by an LLM (versus human financial advisor, $\mathbb{1}(LLM)$), and dummies indicating the four experimental conditions ($\mathbb{1}(T_1)$). Columns 1, 4, and 7 only include recommendations in the baseline condition; columns 2, 3, 5, 6, 8, and 9 only include LLM-generated recommendations. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

Table 15: Mean-variance efficiency of risky portfolio shares, by ESG preferences

**Panel A: Mean monthly return (%)**

| Treatment condition | ESG profiles | Non-ESG profiles | Δ | t-stat. | | z score | |
|---|---|---|---|---|---|---|---|
| Human | 0.796 | 0.795 | 0.001 | 0.173 | | 0.279 | |
| LLM (baseline) | 0.700 | 0.768 | -0.069 | -2.752 | *** | -2.733 | *** |
| LLM (T1, theory) | 0.681 | 0.770 | -0.089 | -3.091 | *** | -3.060 | *** |
| LLM (T2, quant.) | 0.697 | 0.776 | -0.080 | -3.713 | *** | -3.686 | *** |
| LLM (T3, qual.) | 0.713 | 0.783 | -0.070 | -2.539 | ** | -2.533 | ** |
| LLM (T4, all) | 0.707 | 0.779 | -0.072 | -2.952 | *** | -3.315 | *** |

**Panel B: Mean monthly volatility (%)**

| Treatment condition | ESG profiles | Non-ESG profiles | Δ | t-stat. | | z score | |
|---|---|---|---|---|---|---|---|
| Human | 5.018 | 5.010 | 0.008 | 0.141 | | 0.267 | |
| LLM (baseline) | 4.522 | 4.587 | -0.065 | -0.444 | | -0.219 | |
| LLM (T1, theory) | 4.681 | 4.790 | -0.108 | -0.608 | | -0.610 | |
| LLM (T2, quant.) | 4.472 | 4.808 | -0.336 | -2.170 | ** | -0.930 | |
| LLM (T3, qual.) | 4.848 | 4.896 | -0.049 | -0.230 | | -1.002 | |
| LLM (T4, all) | 4.518 | 4.851 | -0.334 | -2.165 | ** | -1.585 | |

**Panel C: Euclidean distance from frontier**

| Treatment condition | ESG profiles | Non-ESG profiles | Δ | t-stat. | | z score | |
|---|---|---|---|---|---|---|---|
| Human | 0.495 | 0.521 | -0.026 | 0.914 | | 5.079 | *** |
| LLM (baseline) | 0.414 | 0.402 | 0.012 | -0.402 | | -0.796 | |
| LLM (T1, theory) | 0.480 | 0.458 | 0.022 | -0.384 | | -1.621 | |
| LLM (T2, quant.) | 0.412 | 0.436 | -0.024 | 0.739 | | 0.081 | |
| LLM (T3, qual.) | 0.546 | 0.476 | 0.070 | -0.637 | | -0.603 | |
| LLM (T4, all) | 0.420 | 0.421 | -0.001 | 0.038 | | 0.984 | |

**Note:** Panel A displays mean monthly portfolio returns for all portfolios with sustainability preference and without. Panel B displays the mean monthly portfolio volatility for all portfolios with sustainability preference and without. Panel C reports the mean Euclidean distances between the portfolio risk and return profile and the respective efficient frontier, given the investor's sustainability preference. Results are reported for all treatment groups separately including baseline and all treatments and baseline together. The table also reports $t$ statistics of two-sample t-tests and Mann-Whitney U-test statistics, for which significance levels are indicated by asterisks ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table 16: Domain-specific information and mean-variance efficiency

| | ESG profiles | | | | | Non-ESG profiles | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Eucl. distance $T_i$ | Eucl. distance baseline | $\Delta$ | z score | | Eucl. distance $T_i$ | Eucl. distance baseline | $\Delta$ | z score |
| T1 (theory) | 0,480 | 0,414 | 0,066 | -1,691 | * | 0,458 | 0,402 | 0,056 | -0,956 |
| T2 (quant.) | 0,412 | 0,414 | -0,003 | 0,000 | | 0,436 | 0,402 | 0,034 | -0,953 |
| T3 (qual.) | 0,546 | 0,414 | 0,132 | -0,711 | | 0,476 | 0,402 | 0,074 | -1,211 |
| T4 (all) | 0,420 | 0,414 | 0,005 | 0,206 | | 0,421 | 0,402 | 0,019 | -1,548 |

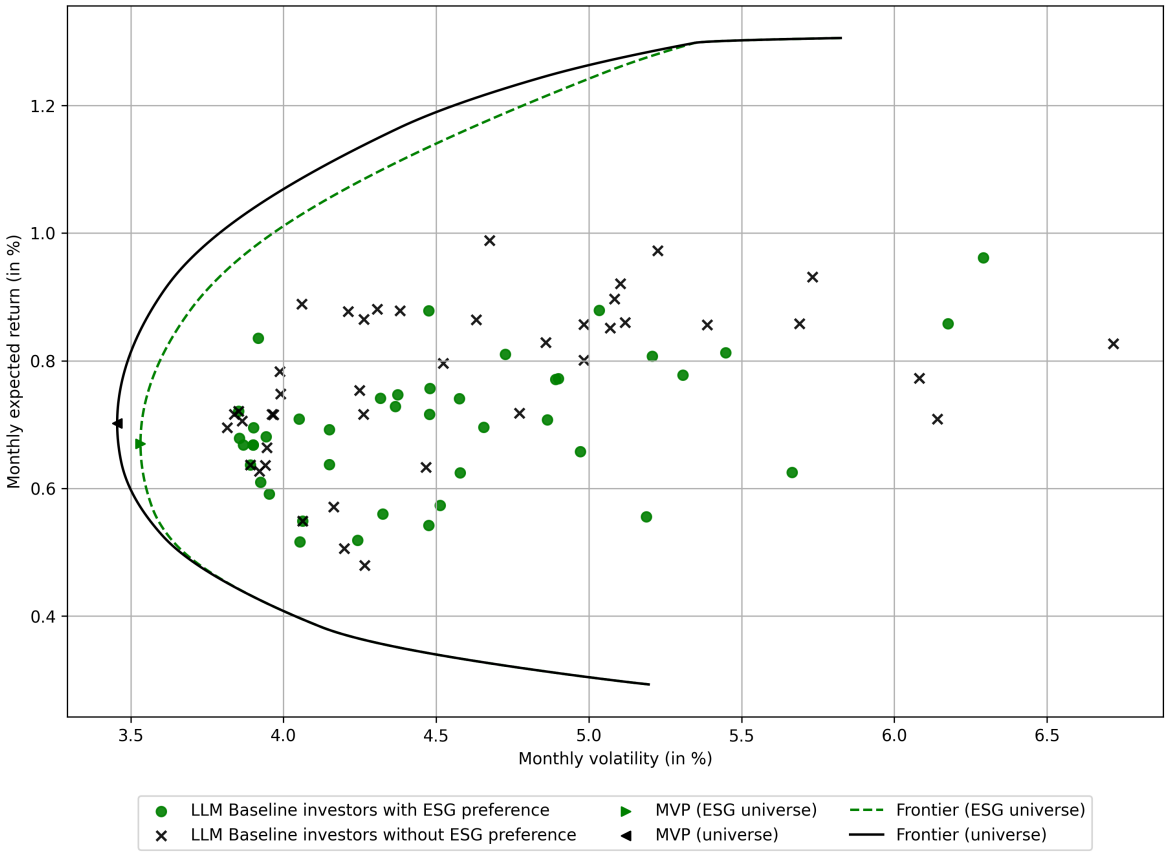
**Note:** Significance levels are reported for Mann-Whitney U-tests ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table 17: Treatment impact on performance: Out-of-sample

| | (1) Mean daily return (%) | (2) Mean daily return (%) | (3) Daily volatility (%) | (4) Daily volatility (%) |
|---|---|---|---|---|
| $\mathbb{1}(LLM)$ | -0.012*** | | -0.193*** | |
| | (0.004) | | (0.023) | |
| $\mathbb{1}(T_1)$ | | 0.005* | | 0.020 |
| | | (0.002) | | (0.011) |
| $\mathbb{1}(T_2)$ | | -0.026** | | 0.125*** |
| | | (0.007) | | (0.024) |
| $\mathbb{1}(T_3)$ | | 0.016** | | 0.120*** |
| | | (0.004) | | (0.026) |
| $\mathbb{1}(T_4)$ | | -0.019*** | | 0.150*** |
| | | (0.004) | | (0.029) |
| Constant | 0.025*** | -0.001 | 1.027*** | 0.933*** |
| | (0.005) | (0.008) | (0.023) | (0.047) |
| Obs. | 1,152 | 419 | 1,152 | 419 |
| Adj. R2 | 0.016 | 0.185 | 0.123 | 0.637 |
| Profile FE | ✓ | ✓ | ✓ | ✓ |
| Advisor FE | | ✓ | | ✓ |
| Exp. condition | Baseline | All | Baseline | All |
| Advisor | Human + LLM | LLM | Human + LLM | LLM |

**Note:** The table reports regression coefficients for OLS regressions of performance measures on a dummy variable indicating that a recommendation is generated by an LLM (versus human financial advisor, $\mathbb{1}(LLM)$), and dummies indicating the four experimental conditions ($\mathbb{1}(T_1)$). Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

Figure 1: Mean-variance frontiers (LLM recommendations, baseline)




**Note:** The figure displays risk-return profiles of portfolios recommended to investors by LLMs in the baseline condition. Portfolios recommended to investors with (without) sustainability preference are displayed as green circles (black crosses). The mean variance frontier for the ESG universe is based on the universe of stocks with ESG scores exceeding 40 (i.e., 10 stocks, see Table 2) The minimum-variance portfolios (MVP) for each frontier depict the portfolios with the lowest historical volatility.

# Online Appendix:
# Making GenAI Smarter:
# Evidence From A Portfolio Allocation Experiment

*Lars Hornuf, David Streich, Niklas Töllich*

## Appendix A: Supplementary tables and figures

Table A1: Investor profiles

| Profile | Risk tolerance | Sustainability preference | Investment horizon |
|---|---|---|---|
| 1 | High | Yes | 1 month |
| 2 | High | Yes | 6 months |
| 3 | High | Yes | 12 months |
| 4 | High | No | 1 month |
| 5 | High | No | 6 months |
| 6 | High | No | 12 months |
| 7 | Low | Yes | 1 month |
| 8 | Low | Yes | 6 months |
| 9 | Low | Yes | 12 months |
| 10 | Low | No | 1 month |
| 11 | Low | No | 6 months |
| 12 | Low | No | 12 months |

Table A2: Questions to determine model cutoff dates

| Date | Question | Answer |
|---|---|---|
| January 2023 | How many people died when two helicopters collided near the Sea World theme park in Gold Coast, Queensland, Australia, on January 2, 2023? | Four people were killed in the accident |
| January 2023 | On January 8, 2023, during the 2023 Brazilian Congress attack, three government buildings in Brasília were breached by protesters. Name at least two of them | National Congress, Supreme Federal Court, and the Palácio do Planalto |
| January 2023 | What was the name of the New Zealand Prime Minister, that resigned in late January 2023? | Jacinda Ardern |
| February 2023 | Who was elected President of Cyprus on February 12, 2023? | Nikos Christodoulides |
| February 2023 | Who succeeded Susan Wojcicki as CEO of YouTube in February 2023? | Neal Mohan |
| February 2023 | How many people were killed in an airstrike on Damascus by the Israeli Air Force in February 2023? | 15 people were killed in the attack |
| March 2023 | On March 6, 2023, in Bolan, Balochistan province, Pakistan, a terrorist attack targeting a van traveling from Sibi to Quetta occured. Of which occupation were the passangers of the van and how many people were killed? | The targeted van was carrying police officers and at least nine of them died |
| March 2023 | Which US Bank collapsed after a bank run in march 2023, marking the largest U.S. bank failure since the 2008 financial crisis? | Silicon Valley Bank |
| March 2023 | Who was voted president in the People's Republic of China on March 10, 2023? | Xi Jinping |
| April 2023 | Which country joined NATO as the 31st member on April 4, 2023? | Finland |
| April 2023 | Who was the Prime Minister of Japan targeted in an assassination attempt involving a pipe bomb, and where did it happen? | Fumio Kishida in Wakayama (Kansai Region) |
| April 2023 | Which homeware giant filed for bankruptcy on April 23, 2023? | Bed Bath & Beyond |
| May 2023 | Where was the coronation of King Charles III held and what was the date?" | King Charles III's coronation took place on May 6, 2023, at Westminster Abbey in London |
| May 2023 | Which company was fined $1.3 billion by the Irish Data Protection Commission in late May 2023 for violating General Data Protection Regulation (GDPR) protections? | Meta |
| May 2023 | How many people were killed in the suicide car bombing at a checkpoint in North Waziristan District, Khyber Pakhtunkhwa, Pakistan, on May 24, 2023? | Four people were killed in the suicide bombing |
| June 2023 | In which country did the train collision on June 2, 2023, occur, resulting in approximately 288 deaths and making it one of the deadliest railway accidents in recent history? | Odisha India |
| June 2023 | On what date did the third inauguration of Recep Tayyip Erdogan as President of Turkey take place at the Presidential Complex in Ankara? | June 3, 2023 |
| June 2023 | On June 23, 2023, which private military group seized and occupied the cities of Rostov-on-Don and Voronezh, capturing the Southern Military District headquarters in Rostov-on-Don during clashes across both oblasts? | Wagner Group |
| July 2023 | On July 12, 2023, which organization approved a billion bailout deal for Pakistan to help avert potential debt defaults and regain economic stability? | International Monetary Fund, |
| July 2023 | What happened to the President of Niger, Mohamed Bazoum, on July 26, 2023? | Niger's President Mohamed Bazoum was detained by members of his presidential guard in a coup d'état. |

| Date | Question | Answer |
|---|---|---|
| July 2023 | What significant event took place in Kolok on July 29, 2023, resulting in at least 12 deaths and 118 injuries? | Firework warehouse exploded |
| August 2023 | On August 1, 2023, which major credit rating agency downgraded the United States' bond credit rating from AAA to AA+, and what reason did they cite for this downgrade? | Fitch |
| August 2023 | How did the leader of the Wagner Group, Yevgeny Prigozhin, die? | Plane crash in Tver Oblast, Russia |
| August 2023 | What happened to President Ali Bongo of Gabon on August 30, 2023? | Gabon's President Ali Bongo Ondimba was ousted in a military coup |
| September 2023 | In early September 2023, which country implemented a ban on the use of iPhones for government officials, leading to a sharp decline in Apple Inc.'s market value? | China |
| September 2023 | What was the magnitude of the Al Haouz, Marocco, earthquake that occurred on September 8, 2023? | The magnitude was 6.8 |
| September 2023 | On September 23, 2023, how many people were killed in the truck bombing in Somalia, and in which city did it occur? | Truck bombing occurred in Beledweyne, Somalia, resulting in at least 30 fatalities |
| October 2023 | On October 1, 2023, where did a suicide bombing occur in Ankara, Turkey, injuring two police officers, and what happened to the second attacker? | The suicide bombing occurred near the entrance of the Ministry of Interior Affairs in Ankara, Turkey. The second assailant was killed in a shootout with police. |
| October 2023 | On October 3, 2023, who became the first Speaker of the United States House of Representatives to be ousted, and what was the primary reason for this unprecedented removal? | Kevin McCarthy |
| October 2023 | On which exact date in October 2023 did the surprise attack on Israel by Hamas-led militant groups occur, which included a rocket barrage and a breach of the GazaIsrael barrier? | October 7, 2023 |
| November 2023 | How strong was the earthquake that struck Karnali Province, Nepal, on November 17, 2023, and what were its effects? | The magnitude was 5.6 |
| November 2023 | On November 19, 2023, who was elected president of Argentina and who did he defeat in the race? | Javier Milei won and Sergio Massa lost in the run-off |
| November 2023 | On November 29, 2023, which property and retail giant declared insolvency after attempts to secure fresh funding failed, becoming the biggest casualty of Europe's property crash? | Signa Holding |
| December 2023 | On December 2, 2023, how many people were killed and injured in a knife and hammer attack near the Eiffel Tower in Paris, and what happened to the suspect? | The attack in Paris resulted in one fatality and two injuries. |
| December 2023 | In which Chinese city did the collision on subway occur on December 14, 2023, resulting in at least 515 injuries but no fatalities? | Beijing |
| December 2023 | On December 21, 2023, how many people were killed and injured in the mass shooting at Charles University in Prague, and what was the perpetrator's connection to the university? | 15 people lost their lives, including 14 individuals (students and staff) at the university, the perpetrator's father. The gunman was a student. |
| January 2024 | In January 2024, how strong was the earthquake that struck the Noto Peninsula in Ishikawa Prefecture, Japan, and what were its effects? | The magnitude was 7.6 |
| January 2024 | On January 3, 2024, how many people were killed and injured in the double bombing in Kerman, Iran, during a ceremony marking the fourth anniversary of Qasem Soleimani's assassination? | 89 killed, 284 injured |

| Date | Question | Answer |
|---|---|---|
| January 2024 | On January 13, 2024, who was elected president of Taiwan, and which political party does the new president represent? | Lai Ching-te, Democratic Progressive Party |
| February 2024 | On February 3, 2024, how many people were killed by the wildfires in Chile, and which two cities were most affected by the fires? | 51 fatalities and the most affected cities were Viña del Mar and Quilpué |
| February 2024 | On February 7, 2024, how many people were killed in the bombings outside electoral offices in Balochistan, Pakistan, and which terrorist group claimed responsibility for the attacks? | At least 30 fatalities and the IS claimed responsibility for the attack |
| February 2024 | On February 11, 2024, who won the Finnish presidential election in the second-round runoff, and which canidate was defeated? | Alexander Stubb won the runoff against Pekka Haavisto |
| March 2024 | On March 7, 2024, which country officially joined NATO? | Sweden |
| March 2024 | On March 26, 2024, what caused the collapse of the Francis Scott Key Bridge in Baltimore, Maryland? | Container ship Dali hits the bridge |
| March 2024 | On March 30, 2024, how many people were killed and injured in the car bomb explosion in the market place of Azaz, Syria? | 8 killed, 30+ injured |
| April 2024 | On April 7, 2024, how many people died in the sinking of the makeshift ferry Zico off the coast of northern Mozambique? | Zico sank off the northern coast of Mozambique, resulting in the deaths of over 100 of the approximately 130 passengers on board. |
| April 2024 | On April 8, 2024, which budget retailer filed for Chapter 11 bankruptcy protection and announced plans to close all of its stores in the U.S.? | 99 Cents Only |
| April 2024 | On April 13, 2024, how many people were killed and injured in the stabbing attack at the Westfield Bondi Junction shopping center in Sydney, and what happened to the attacker? | Six fatalities and ten injuries. The attacker, 40-year-old Joel Cauchi, was shot dead |
| May 2024 | Which performer was disqualified from the Eurovision Song Contest in 2024? | Joost Klein |
| May 2024 | On May 15, 2024, where was Slovak Prime Minister Robert Fico shot and critically injured, and who was arrested at the scene of the attempted assassination? | Slovak Prime Minister Robert Fico was shot and critically injured in Handlová. The assailant, identified as 71-year-old poet Juraj Cintula. |
| May 2024 | On May 19, 2024, which Iranian government officials were killed in the helicopter crash near the village of Uzi in East Azerbaijan, Iran? | President Ebrahim Raisi, Foreign Minister Hossein Amir-Abdollahian, Governor-General Malek Rahmati, representative Mohammad Ali Ale-Hashem |
| June 2024 | On June 11, 2024, what caused the crash of the Dornier 228 aircraft in Malawi, and who was among the ten people confirmed dead? | Bad weather caused the crash and among the deceased was Vice-President Saulos Chilima |
| June 2024 | On June 14, 2024, who was elected as President of South Africa? | Cyril Ramaphosa |
| June 2024 | On June 21, 2024, how many people were killed and wounded in the shooting at the Mad Butcher supermarket in Fordyce, Arkansas? | Four people are killed and ten others are wounded in the shooting |

Table A3: Model cut-off date estimation

| Model | Reported cut-off date | Jan 23 | Feb 23 | Mar 23 | Apr 23 | May 23 | Jun 23 | Jul 23 | Aug 23 | Sep 23 | Oct 23 | Nov 23 | Dec 23 | Jan 24 | Feb 24 | Mar 24 | Apr 24 | May 24 | Jun 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | October 2023 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| GPT-4o-mini | October 2023 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | | | | | |
| GPT-4-turbo | December 2023 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | |
| Llama-3.1-8B-Instruct | December 2023 | ✓ | ✓ | ✓ | | ✓ | | | | | | | | | | | | | |
| Mistral-7B-Instruct-v0.3 | Not reported | ✓ | | | | | | ✓ | | | | | | | | | | | |
| Phi-3-small-128k-instruct | October 2023 | ✓ | | | | ✓ | | | | | | | | | | | | | |
| Qwen2.5-7B-Instruct | Not reported | ✓ | | ✓ | | | | ✓ | | | ✓ | | | | | | | | |

**Note:** The table reports whether at least one of three falsifiable questions in a given month was answered correctly by a model. The questions are listed in Table **??**.

Table A4: Textbook summary

Stock valuation forms the basis of understanding an asset's intrinsic worth. Rooted in the Law of One Price, the valuation process centers on estimating future cash flows, such as dividends and potential sale proceeds, and discounting them to the present using the equity cost of capital. This method, known as the dividend-discount model, applies across different investment horizons, extending even to an infinite period where dividends are presumed to continue indefinitely. Changes in corporate decisions, like increased investments for growth or shifts in management, can influence stock prices as they alter investor expectations about future cash flows. This model provides investors with an analytical framework for understanding the factors driving stock prices, emphasizing the importance of forecast accuracy and corporate financial policies in influencing stock valuation.

Capital markets and risk pricing introduce the interplay between risk and return, differentiating between diversifiable (idiosyncratic) and non-diversifiable (systematic) risks. Only systematic risk, driven by broad economic factors, warrants a risk premium since it affects all assets and cannot be mitigated through diversification. Historical returns reveal that stocks generally offer higher long-term returns than safer assets like bonds or Treasury bills, although with greater volatility. Diversification, achieved by spreading investments across unrelated assets, reduces idiosyncratic risk and overall portfolio volatility. Over shorter time frames, returns are more variable, stressing the need for investment strategies aligned with the investor's horizon and risk profile. Different return metrics, like arithmetic and compound averages, assist in understanding both expected and realized returns over time. This exploration of risk-return tradeoffs, particularly in large portfolios, provides a framework for evaluating investment performance and cost of capital.

Optimal portfolio choice builds on these foundations with a focus on mean-variance optimization, where investors seek to maximize returns within their risk tolerance. By allocating assets with low correlation, investors can reduce portfolio volatility, creating a more stable return profile than individual stocks allow. The Capital Asset Pricing Model (CAPM) further refines this by linking expected returns to an asset's market beta, demonstrating that higher expected returns align with higher market-related risk. While unsystematic risk diminishes as more assets are added to a portfolio, systematic risk persists, underscoring the limits of diversification. The efficient frontier concept emerges from this framework, representing portfolios with the optimal risk-return balance. The inclusion of additional assets can enhance the efficient frontier, offering investors improved risk-return combinations.

Table A5: Definition of quantitative metrics

| Metric | Definition |
|---|---|
| Market beta | A measure of a stock's sensitivity to overall market movements. |
| Book-to-market ratio | The ratio of a company's book value of equity to its market value of equity. |
| Market cap | The total market value of a company's outstanding shares. |
| Momentum | The cumulative stock return over the previous twelve months. |
| Earnings-to-price ratio | The ratio of a company's net income to its market value of equity. |
| ESG score | The ESG score measures the company's self-reported performance in the environmental, social and corporate governance pillars. A higher score (on a scale from 0 to 100) indicates stronger adherence to ESG principles. |

Table A6: 10-K filing and reporting dates

| Company | Stock ticker | 10-K filing date | 10-K reporting date |
|---|---|---|---|
| Air Lease Corp | AL | 15-Feb-24 | 31-Dec-23 |
| Alkermes Plc | ALKS | 21-Feb-24 | 31-Dec-23 |
| Berkshire Hathaway Inc | BRK.B | 26-Feb-24 | 31-Dec-23 |
| Cincinnati Financial Corp | CINF | 26-Feb-24 | 31-Dec-23 |
| Eastman Chemical Co | EMN | 25-Feb-24 | 31-Dec-23 |
| Evertec Inc | EVTC | 29-Feb-24 | 31-Dec-23 |
| Kimberly-Clark Corp | KMB | 08-Feb-24 | 31-Dec-23 |
| Lockheed Martin Corp | LMT | 23-Jan-24 | 31-Dec-23 |
| PepsiCo Inc | PEP | 09-Feb-24 | 30-Dec-23 |
| S&T Bancorp Inc | STBA | 27-Feb-24 | 31-Dec-23 |
| St Joe Co | JOE | 21-Feb-24 | 31-Dec-23 |
| Sturm Ruger & Company Inc | RGR | 21-Feb-24 | 31-Dec-23 |

Table A7: Summary of 10-K MD&A sections

| Company | Summary of latest 10-K's MD&A section |
|---|---|
| Evertec Inc | EVERTEC, a significant transaction-processing enterprise in Latin America, Puerto Rico, and the Caribbean, experienced a productive year in 2023. The company provides comprehensive merchant acquiring, payment services, and business solutions, serving 26 countries from 20 offices. EVERTEC operates the ATH network, a leading PIN debit network in Latin America, processing over six billion transactions annually. Its diverse offerings include core banking, cash processing, technology outsourcing, and fraud monitoring across multiple regions.<br>In 2023, EVERTEC expanded its operations through the acquisitions of paySmart and Sinqia, enhancing its services in Brazil. These acquisitions are part of its strategy to diversify and increase its market presence, expected to drive synergies and augment the company's growth trajectory. The company's revenue model is predominantly recurring, supported by multi-year contracts, contributing to strong operating margins and moderate capital expenditure requirements.<br>EVERTEC's relationship with Popular remains substantial, generating around 35% of its revenues. Adjustments to their agreements in 2022, including extended terms and revised revenue-sharing provisions, have fortified this partnership, although EVERTEC is no longer considered a subsidiary of Popular.<br>The transition from cash to electronic payments continues to benefit the industry, with EVERTEC positioned to capitalize on the growing electronic payment adoption in its markets. The company's broad service range across the transaction-processing value chain allows it to offer unique, integrated solutions, providing competitive advantages in customer acquisition and retention.<br>Financially, EVERTEC reported a 12% increase in revenues in 2023, totaling $694.7 million, driven by growth in all segments, particularly in payment services and merchant acquiring in Puerto Rico and the Caribbean. However, operating income decreased by 13% due to higher costs associated with acquisitions and expansion efforts. The company's robust business model and strategic acquisitions support its positive outlook, despite economic uncertainties affecting consumer confidence and spending patterns in its operating regions. |
| Alkermes Plc | In 2023, the company significantly improved its financial performance with a net income of $519.2 million, a marked increase from a net loss of $33.2 million in 2022. This growth was driven by increases in product sales, manufacturing, and royalty revenues, especially notable from the arbitration resolution favoring the long-acting INVEGA products. The company also benefits from a portfolio of competitively positioned products, including VIVITROL, ARISTADA, ARISTADA INITIO, and LYBALVI, which are expected to continue generating substantial revenues.<br>The company strategically enhanced its focus by separating its oncology business into Mural Oncology plc, allowing shareholders to directly participate in Mural's potential success. It also agreed to sell the Athlone Facility, aiming to simplify operations and leverage subcontracting arrangements through 2025. This decision aligns with broader efforts to streamline operations and concentrate on high-performing areas.<br>Product sales were positively impacted by increased units sold and favorable price adjustments, contributing to a growth in net product sales. However, competitive pressures in addiction treatment and mental health markets persist, challenging future sales of key products like VIVITROL and LYBALVI.<br>Financially, the company has managed to improve its liquidity and capital resources, ending the year with significant increases in cash and investments. It continues to maintain a robust portfolio of U.S. and international government and agency debt securities to manage investment risks effectively.<br>Despite these improvements, there are ongoing risks from potential generic competition and litigation over intellectual property, which could affect future revenue streams from key products. The company also faces typical industry challenges like regulatory changes, market dynamics, and economic conditions that could impact financial performance.<br>Overall, the company's strategic divestitures, focus on core product lines, and effective management of financial resources position it for potential growth, but it must navigate significant competitive and regulatory challenges to sustain its success. |

| Company | Summary of latest 10-K's MD&A section |
|---|---|
| Air Lease Corp | Air Lease Corporation, an industry leader in aircraft leasing founded by Steven F. Udvar-Házy, focuses on acquiring fuel-efficient, new technology commercial jet aircraft from manufacturers like Airbus and Boeing, and leasing them globally. In 2023, the company expanded its fleet by purchasing 71 new aircraft and selling 27, ending the year with 463 owned aircraft. Its managed fleet included 78 aircraft, down from 85 in 2022. The net book value of the fleet increased by 6.9% to $26.2 billion. With an average fleet age of 4.6 years and a lease term of 7.0 years, Air Lease boasts a lease utilization rate of 99.9% and a customer base of 119 airlines across 62 countries.<br>Financially, Air Lease committed to purchasing 334 aircraft by 2028, valued at $21.7 billion, with plans to finance these through cash, operational cash flows, aircraft sales, and debt, primarily unsecured. The company ended 2023 with a strong liquidity position of $6.8 billion and a total debt of $19.4 billion, predominantly at a fixed rate. Revenues rose by 15.9% to $2.7 billion, driven by fleet growth and increased lease and sales activity. Net income for 2023 was notable at $572.9 million, rebounding from a net loss in 2022 largely due to a significant write-off from their Russian fleet operations.<br>Strategically, Air Lease focuses on maintaining a young, efficient fleet, well-positioned to benefit from global air travel growth and demands for newer, more efficient aircraft amid ongoing OEM delivery delays and environmental sustainability pressures. The company anticipates increased demand for leasing due to OEM supply chain challenges, which, along with rising interest rates, might further tighten capital access, potentially boosting lease rates. |
| Berkshire Hathaway Inc | Insurance Operations: Berkshire's insurance underwriting was highly profitable in 2023, generating $5.4 billion in after-tax earnings, a significant increase from the previous years. This success is largely due to fewer catastrophic events and improved underwriting at GEICO, which also benefited from premium rate increases and lower claims frequencies. Investment income from insurance increased markedly due to higher short-term interest rates, enhancing earnings from short-term investments.<br>Railroad (BNSF): BNSF's earnings slightly declined due to a mix of lower freight volumes and higher non-fuel operating costs, despite a reduction in fuel costs. The railroad's revenue per car/unit saw a slight increase, which was not enough to offset the lower volumes.<br>Utilities and Energy (BHE): Earnings in this sector decreased, particularly due to lower contributions from U.S. regulated utilities, affected by increased wildfire loss estimates. However, gains in other energy businesses like tax equity investments and natural gas pipeline businesses offset some declines.<br>Manufacturing, Service, and Retailing: Sectors like industrial products saw growth due to demands in infrastructure, while service revenues were bolstered by acquisitions like Alleghany. However, some manufacturing and service sectors faced weakening demand.<br>Investment and Derivative Gains/Losses: There were significant fluctuations in earnings from investments and derivatives, reflecting the volatile nature of equity markets. Despite these fluctuations, the investments continue to be a crucial component of Berkshire's revenue strategy.<br>Corporate Governance and Structure: Berkshire operates with a decentralized management structure, where significant capital allocation and investment decisions are made by the senior management team without centralized business functions.<br>Overall, Berkshire Hathaway's diverse portfolio allowed it to manage risks and capitalize on opportunities across its operating sectors, despite some challenges posed by economic conditions and market volatility. The financial strategies and management decisions highlighted in this report underscore the company's robust approach to navigating complex market dynamics. |

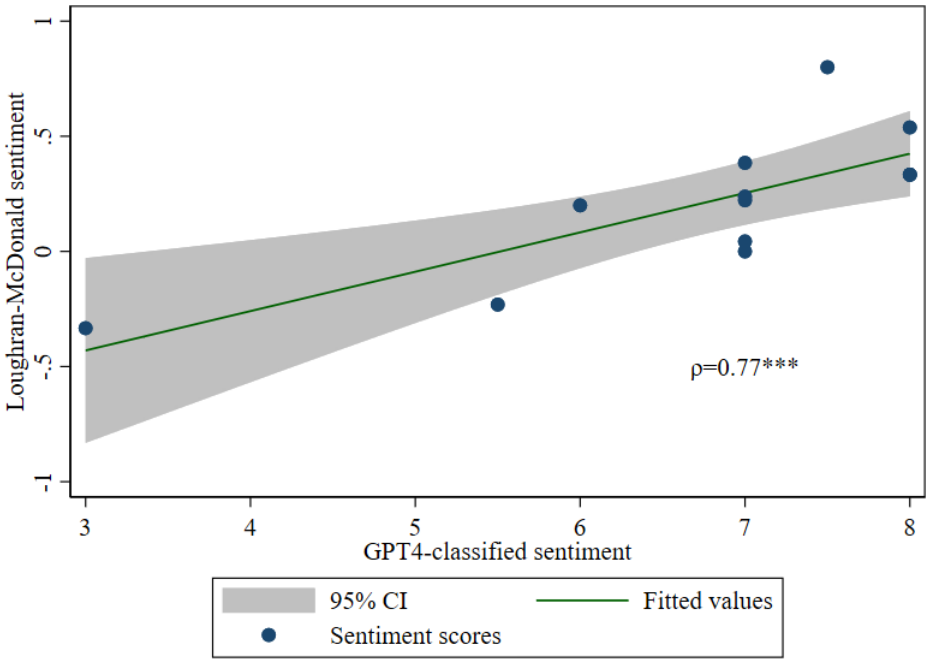
| Company | Summary of latest 10-K's MD&A section |
|---|---|
| Lockheed Martin Corp | The company discussed in the 10-K annual report is a global security and aerospace firm involved in advanced technology systems, products, and services, focusing primarily on defense, space, intelligence, and information technology, including cybersecurity. Most of its business, about 73%, comes from U.S. government contracts, with significant portions also coming from international customers and a small percentage from U.S. commercial sectors. The company operates across four main segments: Aeronautics, Missiles and Fire Control (MFC), Rotary and Mission Systems (RMS), and Space, each providing distinct products and services ranging from aircraft and missile systems to radar and satellite technologies.<br>For the fiscal year 2023, the company reported net sales of $67.6 billion, a slight increase over the previous year, primarily driven by its Space segment due to ramped-up activities in missile defense and fleet ballistic missile programs. The company is involved in several growth areas, including hypersonics and classified programs, expected to transition from development to production soon. It continually seeks to expand through strategic acquisitions, partnerships, and internal innovations, particularly in technologies that integrate advanced networking and operational technologies into defense platforms.<br>The company faces challenges from the U.S. budget environment, with a significant portion of its revenue dependent on government defense spending, which is subject to political and fiscal changes. A continuing resolution and the Fiscal Responsibility Act may impact funding levels. Moreover, geopolitical tensions, like Russia's activities in Ukraine, have increased demand for its defense products, though supply chain disruptions and inflation have affected costs and financial performance.<br>Overall, the company's strategy focuses on leveraging its technology leadership to meet current and future customer needs, optimizing its business portfolio through selective acquisitions and divestitures, and maintaining a robust pipeline of innovations to secure its position in the defense and aerospace sectors. |
| Eastman Chemical Co | Eastman Chemical Company's annual report for the year ending December 31, 2023, provides a comprehensive overview of its financial and operational performance. The company, which prepares its consolidated financial statements in accordance with U.S. GAAP, reported critical accounting estimates requiring significant management judgment, particularly in areas like impairment of assets, environmental costs, and postretirement benefits. Management disclosed a total goodwill of $3.6 billion, with an impairment test confirming this value exceeds the carrying amounts of the reporting units tested.<br>Revenue for 2023 was noted at $9.21 billion, a decrease from the previous year's $10.58 billion, primarily due to reduced sales volumes across various segments and customer destocking. The operating income, excluding non-core and unusual items, was reported at $1.097 billion. The report also highlighted a decrease in gross profit, down by 4% from 2022, due to lower sales volume and higher manufacturing costs, albeit partially offset by reduced raw material and energy costs.<br>Eastman emphasized its proactive approach to liquidity and financial management, with significant mentions of managing environmental remediation costs, which ranged from $252 million to $497 million in undiscounted remediation costs. These reflect liabilities expected to be paid over approximately 30 years.<br>The company also maintains defined benefit pension and other postretirement benefit plans. The cost and obligations related to these plans are significantly influenced by assumptions related to discount rates and expected return on plan assets. The year-end 2023 figures assumed weighted average discount rates of 5.22% for U.S. plans and 3.83% for non-U.S. plans, with expected returns on plan assets at 7.50% and 4.74%, respectively.<br>Further, Eastman discussed its income tax positions, noting a provision for income taxes at $191 million with an effective tax rate of 18% for 2023. The company's strategic maneuvers included the sale of its Texas City operations and ongoing investments in its methanolysis plastic-to-plastic molecular recycling manufacturing facilities, indicating a strong focus on sustainability and innovation-driven growth models.<br>Overall, Eastman's financial discourse for 2023 reflects a complex balance of managing operational costs, navigating market fluctuations, and investing in technology and sustainability to foster long-term growth. |

| Company | Summary of latest 10-K's MD&A section |
|---|---|
| St Joe Co | The 2023 10-K report from St. Joe reveals a focus on real estate development, asset management, and operations within Northwest Florida. The company plans to utilize its existing assets across residential, hospitality, and commercial ventures, actively seeking better uses through development activities and partnerships. A significant part of St. Joe's strategy includes forming joint ventures and limited partnerships, enhancing growth and diversifying business exposure.<br>Financially, St. Joe reported substantial increases in key metrics: revenue rose by 54.3% to $389.2 million, operating income increased by 47.7% to $90.7 million, and net income attributable to the company grew by 9.6% to $77.7 million. These figures reflect a robust operational performance despite challenges from macroeconomic factors like inflation, interest rate hikes, and supply chain disruptions. The company benefits from the growth and appeal of Northwest Florida, which continues to attract new residents and investment due to its high quality of life and natural beauty.<br>St. Joe operates predominantly in three segments: residential, hospitality, and commercial, with a notable focus on developing properties close to the Gulf of Mexico. The residential segment, which plans and develops diverse residential communities, reported increased revenue and profit from homesite sales. The hospitality segment, boosted by new amenities and properties, showed significant revenue growth, although operating costs linked to new openings slightly lowered the profit margins.<br>The commercial segment continues to grow with new developments, focusing on retail, office, and mixed-use projects that complement residential communities. Despite the diverse revenue streams, the company notes potential delays and operational impacts due to external economic pressures, although it remains well-positioned with its strategic land holdings and capital resources.<br>St. Joe's financial condition appears solid, with plans to fund future developments through cash proceeds from completed projects and financing arrangements. The company's strategy is set to leverage its substantial land entitlements and favorable location to generate long-term shareholder value, focusing on meeting market demand across its key business segments. |
| S&T Bancorp Inc | S&T Bancorp, Inc. provided a comprehensive analysis of its financial condition for the fiscal years ending 2023, 2022, and 2021. The company's focus was on key financial metrics, such as earnings, asset quality, and operational efficiency, while also highlighting future outlooks and potential risks.<br>In 2023, S&T Bancorp reported a net income of $144.8 million, marking a 6.83% increase from 2022, driven by higher net interest income due to increased interest rates. The net interest margin improved, with interest and dividend income rising significantly, offset by a sharp increase in interest expenses. The bank managed an effective tax rate of 19.0%, benefiting from Low Income Housing Tax Credits.<br>Credit quality was a focal point, with a noted increase in the provision for credit losses due to heightened charge-offs. The allowance for credit losses incorporated both historical data and future economic forecasts, reflecting management's cautious stance amidst uncertain economic conditions.<br>Operational metrics indicated solid management of resources, evidenced by a decrease in the efficiency ratio, signifying improved profitability. Noninterest income remained relatively stable, with minor fluctuations in various revenue streams such as mortgage banking and service charges.<br>S&T Bancorp continued to emphasize strategic growth, focusing on expanding its deposit base and enhancing core profitability through various financial services in Pennsylvania and Ohio. The balance sheet showed strength with a total capital ratio of 15.27% and robust liquidity positions, ensuring resilience against potential financial stresses.<br>The bank also provided detailed forward-looking statements, warning of potential risks related to market volatility, interest rate changes, and economic downturns which could affect future performance. This cautious outlook underscores the bank's proactive approach to navigating anticipated financial challenges while capitalizing on growth opportunities in its regional markets. |

| Company | Summary of latest 10-K's MD&A section |
|---|---|
| Sturm Ruger & Company Inc | Sturm, Ruger & Company, Inc. specializes in the design, manufacture, and sale of firearms, primarily serving the domestic U.S. market, with about 99% of its sales derived from firearms and 6% from exports. The company's products are predominantly sold through a limited number of independent wholesale distributors to the commercial sporting market. It also produces steel alloy investment castings and metal injection molding parts, contributing less than 1% to its total sales. The company observes seasonal trends in orders, with stronger demand in the first quarter and weaker in the third.<br>In 2023, the company reported a 7% decline in product sell-through from distributors to retailers compared to 2022, a steeper decline than the 4% drop in adjusted National Instant Criminal Background Check System (NICS) checks, suggesting competitive pressures from rivals offering more aggressive sales promotions. Total adjusted NICS checks stood at 15,848,000, down from 16,425,000 in 2022. These checks are adjusted by the National Shooting Sports Foundation to exclude non-purchase related checks.<br>The financial results for 2023 reflected challenges, with net sales of firearms dropping to $540.7 million from $593.3 million in 2022, an 8.9% decrease. Gross profit also fell significantly from $180.1 million to $133.6 million, impacted by increased promotional costs and inflationary pressures on materials and services. This resulted in a gross margin decrease from 30.2% to 24.6%. Despite these headwinds, new product sales, which include items launched within the past two years like the MAX-9 pistol and Marlin lever-action rifles, constituted 23% of the firearm sales.<br>The company's operating income declined sharply to $52.1 million, down from $103.5 million in 2022, primarily due to reduced sales volumes and increased operational costs, although partially offset by higher pricing strategies. Selling, general, and administrative expenses also increased, driven by higher professional service costs and marketing activities. Notably, the company utilizes both GAAP and non-GAAP financial measures, including EBITDA, to provide clearer insights into its financial health and operational performance, with EBITDA for 2023 recorded at $75,947,000.<br>Overall, Sturm, Ruger & Company is navigating a challenging market environment with fluctuating demand and competitive pressures, while continuing to innovate and optimize its product offerings and manufacturing efficiency. |
| PepsiCo Inc | In the 2023 annual report, PepsiCo describes a challenging year marked by global disruptions such as supply chain issues, inflation, and geopolitical conflicts, yet it continued to advance its strategic transformation initiative, pep+. The company focused on sustainability, aiming to become net-zero in emissions by 2040 and net water positive by 2030. Significant investments were made to promote regenerative agriculture and develop sustainable packaging solutions.<br>PepsiCo's financial performance showed a 6% increase in net revenue to $91.471 billion, while operating profit grew by 4% to $11.986 billion, despite a slight dip in operating margin. The company attributes these results to effective net pricing and productivity savings, although offset by higher commodity and operational costs. PepsiCo continued its efforts in innovation and market expansion, particularly in digital and e-commerce platforms, to adapt to the rapidly changing retail landscape.<br>The report also outlines risks related to commodity price fluctuations, climate change regulations, and the impacts of the ongoing conflict in Ukraine, which has notably affected its operations and financials in the region. PepsiCo's approach to managing these risks includes fixed-price contracts, pricing agreements, and hedging strategies.<br>Internationally, the company faces challenges from economic instability and currency volatility, particularly in emerging markets. Despite these challenges, PepsiCo is committed to its strategic priorities of growth, resilience, and sustainability, aiming to drive long-term shareholder value and societal benefits. |

| Company | Summary of latest 10-K's MD&A section |
|---|---|
| Kimberly-Clark Corp | The Management's Discussion and Analysis (MD&A) from Kimberly-Clark's 2023 annual report outlines the company's financial performance, highlighting a modest revenue increase of 1% to $20.4 billion, attributed to organic sales growth of 5%. The company has expanded its market presence in over 175 countries, focusing on renowned brands like Kleenex and Huggies across three business segments: Personal Care, Consumer Tissue, and K-C Professional.<br>In 2023, the firm successfully sold its Brazilian tissue and K-C Professional businesses, netting a pre-tax gain of $44 million. It also increased its stake in Thinx, acquiring full ownership by year-end. These transactions influenced the financials, alongside non-GAAP adjustments like intangible asset impairments and pension settlements.<br>Operationally, North America showed strong performance with organic sales growth in both consumer products and professional segments, driven by strategic price increases and product mix enhancements. However, global operations were tempered by foreign exchange impacts and the divestiture of certain businesses. Despite these challenges, the company achieved a net income of $1.76 billion, although down from $1.93 billion in 2022, and continued its 51-year tradition of dividend growth.<br>Kimberly-Clark's forward-looking statements emphasize sustained investment in innovation and market expansion, particularly in personal care products, supported by digital and e-commerce capabilities. The company also remains vigilant about various macroeconomic challenges, including the ongoing impacts of COVID-19, which are expected to continue affecting global operations and supply chain dynamics.<br>The financial outlook is cautiously optimistic, focusing on strategic growth initiatives and maintaining rigorous cost management to bolster profitability in a competitive and uncertain global market environment. |
| Cincinnati Financial Corp | Cincinnati Financial Corporation's Management's Discussion and Analysis provides insight into the company's financial condition and operational results, emphasizing the importance of long-term strategic decisions despite ongoing economic and industry challenges. The corporation's primary performance metric, the Value Creation Ratio (VCR), significantly exceeded its target in 2023, reaching 19.5% due to strong contributions from net income and investment portfolio valuations.<br>The company, one of the top 25 property casualty insurers in the U.S., experienced a 10% growth in net written premiums in 2023, reflecting the success of its strategic initiatives and its ability to outpace industry growth rates. The financial highlights include an underwriting profit, driven by reduced losses and expenses compared to revenues, and substantial investment income, primarily from bond interest and stock dividends.<br>Cincinnati Financial has maintained a strong commitment to shareholder returns, consistently increasing its dividend over 63 consecutive years. Its financial stability is underlined by a solid increase in total assets and shareholders' equity in 2023, along with a decrease in debt levels. Moreover, the company's strategic focus on maintaining strong agency relationships and premium growth initiatives is expected to sustain its above-average industry performance, particularly in property casualty insurance.<br>Despite potential risks like economic downturns or significant legal expenses, Cincinnati Financial is positioned for future growth based on a robust balance sheet, proactive investment strategies, and effective underwriting practices. This foundation enables the company to navigate uncertainties while striving to meet or exceed its performance targets, thereby continuing to create substantial value for its stakeholders. |

Figure A1: Correlation of 10-K sentiment measures




**Note:** The figure plots sentiment measures for the summaries of the management discussion & analysis sections of the 12 stocks' 2023 10-K filings. The Loughran-McDonald sentiment scores are computed as the difference between positive and negative words divided by the sum of positive and negative words (*tone*). As a comparison, we have OpenAI's GPT-4 classify the sentiment in the summaries on a scale from 0 (extremely negative) to 10 (extremely positive).

Table A8: Univariate determinants of stock recommendations, by treatment condition

| | (1) Baseline | | | (2) T1 (theory) | | | (3) T2 (quant.) | | | (4) T3 (qual.) | | | (5) T4 (all) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Stocks not recommended | Stocks recommended | t stat. | Stocks not recommended | Stocks recommended | t stat. | Stocks not recommended | Stocks recommended | t stat. | Stocks not recommended | Stocks recommended | t stat. | Stocks not recommended | Stocks recommended | t stat. |
| *Panel A: Fundamentals* | | | | | | | | | | | | | | | |
| Market beta | 0.84 | 0.79 | 1.67 * | 0.83 | 0.80 | 1.33 | 0.87 | 0.75 | 4.42 *** | 0.81 | 0.82 | -0.60 | 0.85 | 0.78 | 2.36 ** |
| Book-to-market ratio | 0.48 | 0.39 | 3.78 *** | 0.48 | 0.39 | 3.37 *** | 0.50 | 0.37 | 5.01 *** | 0.46 | 0.42 | 1.67 * | 0.48 | 0.40 | 3.16 *** |
| Market cap (B USD) | 84.70 | 160.68 | -4.56 *** | 99.74 | 145.97 | -2.75 *** | 115.96 | 127.05 | -0.66 | 103.82 | 136.51 | -1.95 * | 113.07 | 129.12 | -0.96 |
| Momentum | 0.15 | 0.16 | -0.42 | 0.15 | 0.15 | 0.03 | 0.13 | 0.19 | -4.53 *** | 0.15 | 0.16 | -0.33 | 0.13 | 0.18 | -4.15 *** |
| E/P ratio | 0.06 | 0.06 | 1.54 | 0.06 | 0.06 | 1.54 | 0.06 | 0.06 | -0.99 | 0.06 | 0.06 | 0.08 | 0.06 | 0.06 | -0.68 |
| ESG score | 52.26 | 58.01 | -4.61 *** | 52.78 | 57.61 | -3.85 *** | 48.06 | 63.04 | -12.80 *** | 52.99 | 56.82 | -3.05 *** | 49.52 | 60.49 | -9.06 *** |
| *Panel B: Investor attention* | | | | | | | | | | | | | | | |
| SEC EDGAR downloads (k) | 6.74 | 9.99 | -4.19 *** | 7.39 | 9.35 | -2.51 ** | 8.19 | 8.42 | -0.30 | 7.60 | 8.92 | -1.69 * | 8.06 | 8.54 | -0.62 |
| Google SVI | 3.73 | 5.84 | -4.70 *** | 4.16 | 5.42 | -2.79 *** | 3.49 | 6.19 | -6.06 *** | 4.24 | 5.19 | -2.11 ** | 3.76 | 5.72 | -4.36 *** |
| News articles | 20.65 | 31.81 | -4.00 *** | 23.15 | 29.32 | -2.20 ** | 23.10 | 29.35 | -2.23 ** | 23.44 | 28.29 | -1.73 * | 23.51 | 28.48 | -1.78 * |
| Analyst reports | 15.13 | 25.64 | -3.06 *** | 17.36 | 23.43 | -1.76 * | 22.91 | 17.00 | 1.71 * | 18.34 | 21.79 | -1.00 | 21.87 | 18.47 | 0.99 |
| *Panel C: 10-K sentiment* | | | | | | | | | | | | | | | |
| Sentiment in MD&A section | 0.21 | 0.21 | -0.11 | 0.22 | 0.20 | 0.61 | 0.23 | 0.18 | 2.76 *** | 0.19 | 0.23 | -2.58 ** | 0.22 | 0.20 | 1.32 |

**Note:** The table displays equal-weighted averages for the fundamental stock and investor attention measures for stocks, separately by treatment conditions. The table also reports $t$ statistics of two-sample t-tests, for which significance levels are indicated by asterisks ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

Table A9: Univariate determinants of stock recommendations (placebo)

| | (1) Baseline 2 | | | (2) Placebo 1 | | | (3) Placebo 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Stocks not recommended | Stocks recommended | t stat. | Stocks not recommended | Stocks recommended | t stat. | Stocks not recommended | Stocks recommended | t stat. |
| *Panel A: Fundamentals* | | | | | | | | | |
| Market beta | 0.84 | 0.78 | 2.22 ** | 0.82 | 0.81 | 0.64 | 0.84 | 0.79 | 1.88 * |
| Book-to-market ratio | 0.48 | 0.40 | 3.41 *** | 0.47 | 0.41 | 2.22 ** | 0.49 | 0.38 | 4.50 *** |
| Market cap | 94.57 | 149.84 | -3.30 *** | 82.85 | 159.68 | -4.62 *** | 81.53 | 166.95 | -5.13 *** |
| Momentum | 0.15 | 0.16 | -1.21 | 0.15 | 0.16 | -0.61 | 0.15 | 0.16 | -0.28 |
| E/P ratio | 0.06 | 0.06 | 0.92 | 0.06 | 0.06 | -0.57 | 0.06 | 0.06 | 1.53 |
| ESC score | 51.96 | 58.33 | -5.12 *** | 52.27 | 57.79 | -4.43 *** | 51.26 | 59.37 | -6.56 *** |
| *Panel B: Investor attention* | | | | | | | | | |
| SEC EDGAR downloads | 7.17 | 9.52 | -3.02 *** | 6.64 | 9.98 | -4.32 *** | 6.62 | 10.24 | -4.67 *** |
| Google SVI | 3.80 | 5.77 | -4.37 *** | 3.68 | 5.82 | -4.78 *** | 3.41 | 6.30 | -6.49 *** |
| News articles | 21.69 | 30.67 | -3.21 *** | 19.89 | 32.16 | -4.41 *** | 19.33 | 33.72 | -5.18 *** |
| Analyst reports | 16.98 | 23.62 | -1.93 * | 14.56 | 25.82 | -3.28 *** | 14.87 | 26.30 | -3.32 *** |
| *Panel C: 10-K sentiment* | | | | | | | | | |
| Sentiment in MD&A section | 0.21 | 0.21 | 0.24 | 0.21 | 0.21 | -0.27 | 0.23 | 0.19 | 2.03 ** |

**Note:** The table displays equal-weighted averages for the fundamental stock and investor attention measures for stocks, separately by whether they are included or not included in portfolio recommendations generated by LLMs in the various placebo conditions. The table also reports $t$ statistics of two-sample t-tests, for which significance levels are indicated by asterisks ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table A10: Contributions to explained variation in portfolio weights (non-ESG investors)

| | Human | LLM | | | | |
|---|---|---|---|---|---|---|
| DV: Stock weight | (1) Baseline | (2) Baseline | (3) $T_1$ (theory) | (4) $T_2$ (quant.) | (5) $T_3$ (qual.) | (6) $T_4$ (all) |
| **Fundamentals** | **0.62** | **0.79** | **0.80** | **0.86** | **0.67** | **0.79** |
| Market beta | 0.03 | 0.21 | 0.26 | 0.06 | 0.20 | 0.07 |
| Book-to-market ratio | 0.07 | 0.09 | 0.10 | 0.05 | 0.05 | 0.05 |
| Market cap | 0.41 | 0.13 | 0.14 | 0.08 | 0.10 | 0.10 |
| Momentum | 0.03 | 0.01 | 0.02 | 0.11 | 0.20 | 0.08 |
| Earnings-to-price ratio | 0.05 | 0.04 | 0.05 | 0.04 | 0.09 | 0.06 |
| ESG score | 0.04 | 0.31 | 0.24 | 0.52 | 0.04 | 0.43 |
| **Investor attention** | **0.34** | **0.11** | **0.11** | **0.07** | **0.09** | **0.08** |
| SEC EDGAR downloads | 0.34 | 0.11 | 0.11 | 0.07 | 0.09 | 0.08 |
| **Sentiment** | **0.04** | **0.10** | **0.08** | **0.07** | **0.24** | **0.13** |
| Sentiment in 10-K MD&A section | 0.04 | 0.10 | 0.08 | 0.07 | 0.24 | 0.13 |
| R2 | 0.02 | 0.15 | 0.06 | 0.18 | 0.07 | 0.13 |
| Obs. | 6,228 | 504 | 492 | 504 | 480 | 492 |

**Note:** The table reports the standardized dominance statistics of independent variables of regressions of each stock's portfolio weight on that stock's fundamentals, investor attention (as measured by the number of download requests registered in the SEC EDGAR database), and sentiment in the stock's 10-K management discussion and analysis (MD&A) section. The sample only includes portfolios recommended to investors with no sustainability preference. Alternative attention measures are omitted due to collinearity (see Table A12 in the Appendix). Missing values arise from missing portfolio recommendations and portfolio recommendations with 100% fixed income allocation.

Table A11: Contributions to explained variation in portfolio weights (placebo)

| DV: Stock weight | (1) Baseline | (2) Baseline 2 | (3) $P_1$ | (4) $P_2$ |
|---|---|---|---|---|
| **Fundamentals** | **0.77** | **0.76** | **0.73** | **0.78** |
| Market beta | 0.09 | 0.10 | 0.07 | 0.06 |
| Book-to-market ratio | 0.10 | 0.09 | 0.09 | 0.12 |
| Market cap | 0.14 | 0.13 | 0.18 | 0.17 |
| Momentum | 0.02 | 0.02 | 0.03 | 0.03 |
| Earnings-to-price ratio | 0.05 | 0.04 | 0.03 | 0.04 |
| ESG score | 0.37 | 0.38 | 0.32 | 0.36 |
| **Investor attention** | **0.11** | **0.11** | **0.15** | **0.15** |
| SEC EDGAR downloads | 0.11 | 0.11 | 0.15 | 0.15 |
| **Sentiment** | **0.12** | **0.13** | **0.12** | **0.08** |
| Sentiment in 10-K MD&A section | 0.12 | 0.13 | 0.12 | 0.08 |
| R2 | 0.21 | 0.20 | 0.19 | 0.23 |
| Obs. | 1,008 | 996 | 1,008 | 984 |

**Note:** The table reports the standardized dominance statistics of independent variables of regressions of each stock's portfolio weight on that stock's fundamentals, investor attention (as measured by the number of download requests registered in the SEC EDGAR database), and sentiment in the stock's 10-K management discussion and analysis (MD&A) section. Alternative attention measures are omitted due to collinearity (see Table **??** in the Appendix). Missing values arise from missing portfolio recommendations and portfolio recommendations with 100% fixed income allocation.

Table A12: Security selection regressions: Variance inflation factors

| | (1) $\mathbb{1}(w_i > 0)$ | (2) $\mathbb{1}(w_i > 0)$ | (3) $\mathbb{1}(w_i > 0)$ |
|---|---|---|---|
| *Fundamentals* | | | |
| Market beta | 3.3 | 1.7 | |
| Book-to-market ratio | 20.0 | 4.0 | |
| Market cap | 387.3 | 1.2 | |
| Momentum | 3.3 | 1.3 | |
| E/P ratio | 6.5 | 3.0 | |
| ESG score | 17.9 | 1.5 | |
| *Investor attention* | | | |
| SEC EDGAR downloads | 1,762.7 | | 84.1 |
| Google SVI | 131.4 | | 10.2 |
| News articles | 289.0 | | 22.2 |
| Analyst reports | 1,060.4 | | 76.2 |

Table A13: Investor characteristics and portfolio risk (placebo)

| DV: Risky share | (1) Baseline 2 | (2) Placebo 1 (qual.) | (3) Placebo 2 (quant.) |
|---|---|---|---|
| 𝟙(Risk tolerance = high) | 0.455*** | 0.446*** | 0.448*** |
| | (0.027) | (0.028) | (0.042) |
| 𝟙(Horizon = 6 months) | 0.038 | 0.027 | 0.041 |
| | (0.034) | (0.034) | (0.051) |
| 𝟙(Horizon = 12 months) | 0.000 | 0.078** | 0.068 |
| | (0.034) | (0.034) | (0.051) |
| Constant | 0.456*** | 0.465*** | 0.411*** |
| | (0.031) | (0.031) | (0.047) |
| Obs. | 84 | 84 | 84 |
| Adj. R2 | 0.767 | 0.756 | 0.579 |
| *Covariates* | | | |
| Sustainability preference | ✓ | ✓ | ✓ |

**Note:** The table reports regression coefficients from OLS regressions with the risky share as the dependent variable and investor characteristics as the independent variables. 𝟙(High risk tolerance) equals 1 if the investor profile indicates high risk tolerance. The omitted category for the investment horizon dummies is 1 month. 𝟙(Sustainability preference) equals 1 if the investor profile indicates a preference for sustainable investing, and 0 otherwise. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$). Missing values arise from missing portfolio recommendations.

Table A14: Investor characteristics and portfolio-level ESG score (placebo)

| DV: Avg. ESG score | (1) Baseline 2 | (2) Placebo 1 (qual.) | (3) Placebo 2 (quant.) |
|---|---|---|---|
| $\mathbb{1}$(Sustainability preference = yes) | 6.089*** | 3.802** | 4.505** |
| | (1.568) | (1.531) | (2.155) |
| Constant | 62.534*** | 63.968*** | 66.304*** |
| | (1.797) | (1.711) | (2.533) |
| Obs. | 83 | 84 | 82 |
| Adj. R2 | 0.387 | 0.324 | 0.181 |
| *Covariates* | | | |
| Risk tolerance | ✓ | ✓ | ✓ |
| Investment horizon | ✓ | ✓ | ✓ |

**Note:** The table reports regression coefficients from OLS regressions with the weighted portfolio-level Thomson Reuters ESG score as the dependent variable. The variable $\mathbb{1}$(Sustainability preference) equals 1 if the investor profile indicates a preference for sustainable investing, and 0 otherwise. All other investor profile characteristics (dummy variable indicating high risk tolerance, dummy variables indicating 6-month and 12-month investment horizons) are included as control variables, but omitted from the table. Standard errors are robust to clustering on the advisor level and reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$). Missing values arise from missing portfolio recommendations and portfolio recommendations with 100% fixed income allocation.

Figure A2: Diversification measures, by condition

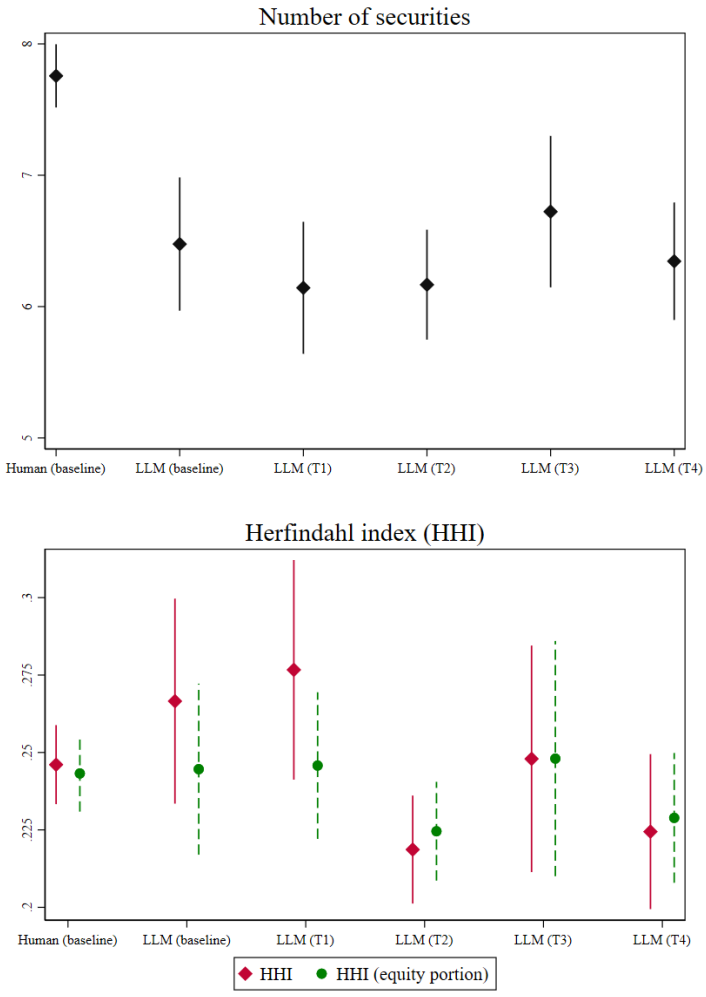


Figure A3: Portfolio risk measures, by condition

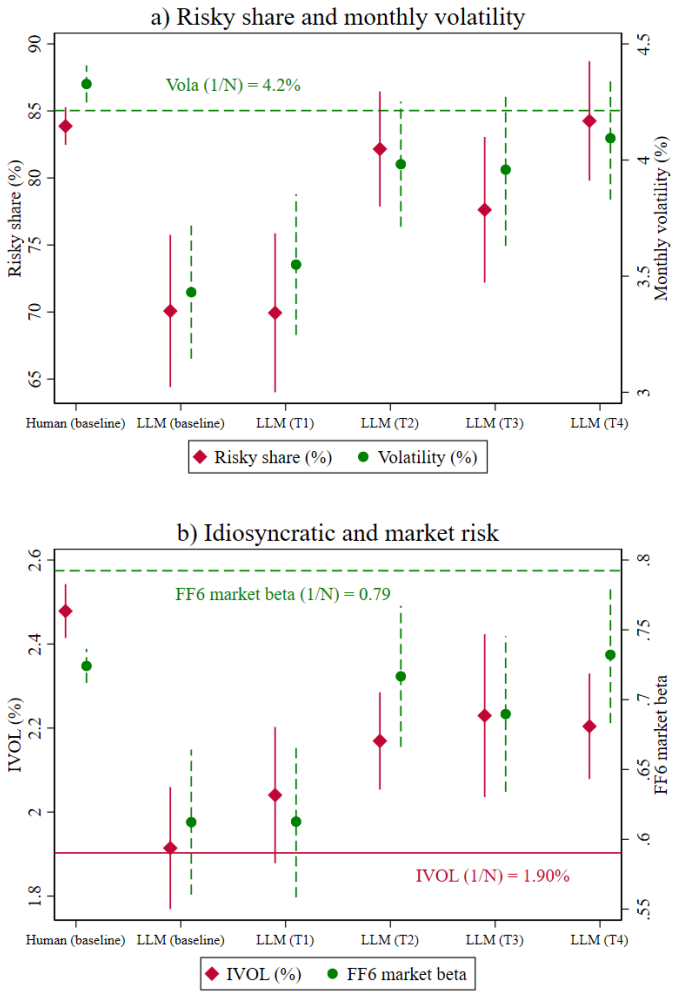


Figure A4: Risky share, by condition and risk tolerance

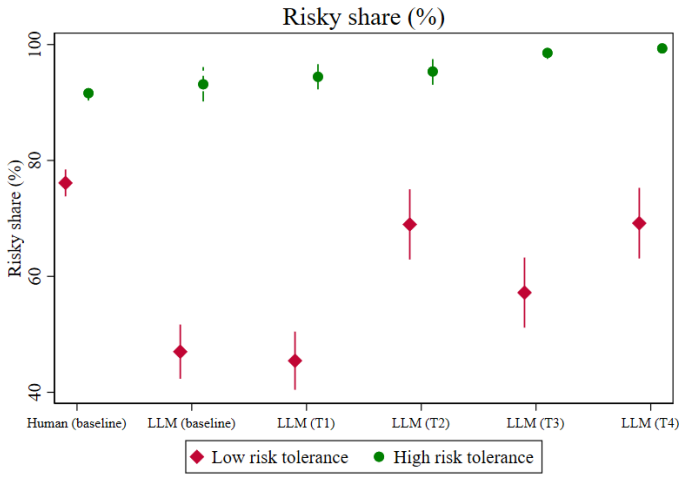
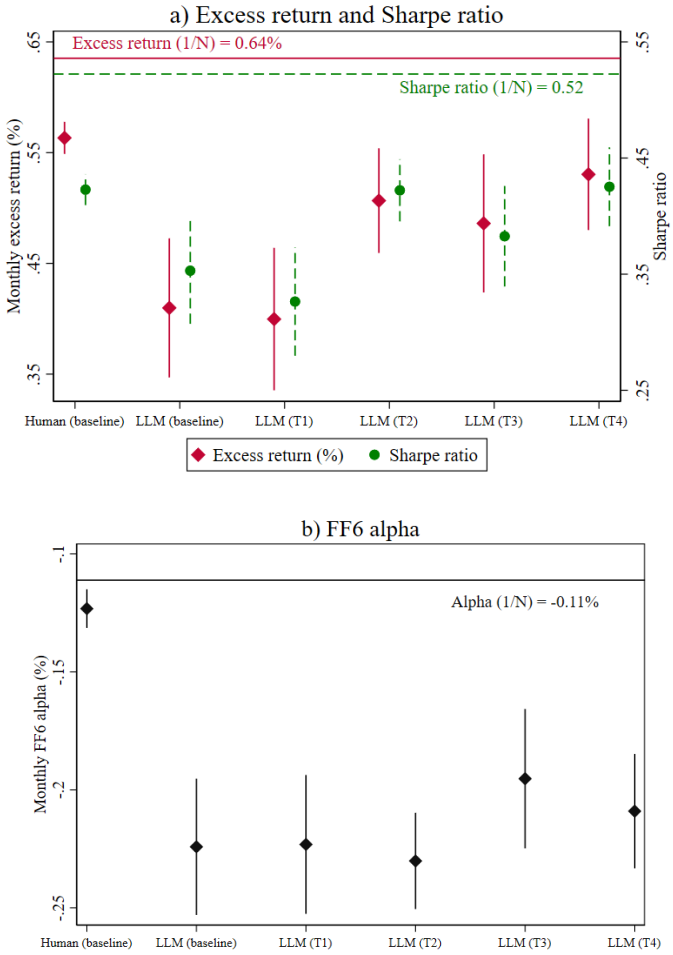


Figure A5: Performance measures, by condition



## Appendix B: Survey data preparation

This section details the process we use to pre-process the survey data to ensure high-quality responses in our analyses. We proceed as follows:

1. We record a total number of **105** completed survey responses.
2. We exclude **16** respondents who stated they were not in fact financial advisors.[34] The remaining **89** respondents make up our full sample.
3. As a robustness test, we exclude respondents who fail to pass our attention checks:
   (a) First, we exclude **9** respondents who, when asked for the risk tolerance level of the last displayed scenario, answered "medium" (risk tolerance was either "high" or "low," but never "medium").
   (b) Second, we exclude a further **11** respondents who finished the entire survey in less than 10 minutes. Assuming it takes 4 minutes to (i) carefully read the instructions and consent statement, (ii) answer the 2 screening questions, (iii) answer the manipulation check question and provide a rationale for the decision-making, and (iv) answer the 5 questions aimed at eliciting the respondents' experience and financial literacy, this leaves the respondent 30 seconds per scenario, which is not sufficient to provide a sensible recommendation.
   (c) Third, we exclude **1** further respondent whose response to the question "Please briefly describe the rationale behind your recommendations." suggested the responses were generated by an AI-based tool.[35]
   (d) We will use the remaining **68** attentive respondents as an alternative benchmark in robustness tests.

Table B1 reports summary statistics for the full sample of respondents (panel A) and the limited sample of respondents who passed all attention checks (panel B). On average, respondents were 35 years old, more likely male than female (30% female), and mostly in full-time employment. It took an average of 27 minutes (30 minutes in the more restrictive specification) to complete the survey. Respondents were experienced financial advisors in most cases, with 50% of both samples having at least 5 years of experience in the financial advice industry and having served an average of 244 clients over their careers (293 in the more restrictive specification). Finally, while there is some variation in financial literacy, 70% of all respondents answered all three financial literacy questions correctly. As a comparison,

[34] Three of the 16 respondents stated they were not (or have never been) financial advisors the first time they attempted to complete the survey, then re-entered the survey and stated they were in fact (or have been) financial advisors.

[35] The response was "Rationale is a term that refers to the reasoning or underlying logic behind a decision, action, or belief. It explains why something is done or chosen, providing the justification or motivation. A rationale typically includes factors that were considered, goals to be achieved, and any relevant evidence or arguments that support the decision. For example, in a business setting, a company might present a rationale for launching a new product. This rationale would include market research, anticipated customer demand, competitive advantages, and the expected benefits to the company's growth. In academic and research contexts, the rationale helps outline why a particular study or experiment is undertaken, clarifying the purpose and significance of the research. A strong rationale is clear, logical, and backed by evidence, making it easier for others to understand and potentially support the reasoning behind the action."

only 43% of the US respondents of the 2019 Survey of Consumer Finance correctly answered all "Big Three" financial literacy questions (Lusardi and Mitchell, 2023, p. 140), which are arguably slightly easier to answer than the set of questions we use.[36] Univariate tests reveal that respondents who have passed all attention checks self-report a higher number of clients (293 vs. 84, $p < 0.05$), longer duration (1837 vs. 972 seconds, $p < 0.01$), and are less likely to be employed full-time (82% vs. 100%, $p < 0.01$).

Table B1: Summary statistics: Survey respondents

| | N | Mean | SD | p5 | Median | p95 |
|---|---|---|---|---|---|---|
| *Panel A: Full sample (N=89)* | | | | | | |
| Age | 89 | 34.5 | 11.6 | 19 | 33 | 54 |
| 𝟙(female) | 89 | 0.3 | 0.5 | 0 | 0 | 1 |
| 𝟙(student) | 80 | 0.5 | 0.5 | 0 | 0 | 1 |
| 𝟙(full-time employment) | 87 | 0.9 | 0.3 | 0 | 1 | 1 |
| Duration (in seconds) | 89 | 1632.5 | 903.1 | 479 | 1,504 | 3,377 |
| No. clients (self-reported) | 89 | 243.7 | 629.9 | 3 | 48 | 1,000 |
| 𝟙(≥5y experience) | 89 | 0.5 | 0.5 | 0 | 0 | 1 |
| 𝟙(all FL questions correct) | 89 | 0.7 | 0.5 | 0 | 1 | 1 |
| *Panel B: Passed attention checks (N=68)* | | | | | | |
| Age | 68 | 35.5 | 12.1 | 19 | 34 | 54 |
| 𝟙(female) | 68 | 0.3 | 0.5 | 0 | 0 | 1 |
| 𝟙(student) | 60 | 0.5 | 0.5 | 0 | 0 | 1 |
| 𝟙(full-time employment) | 66 | 0.8 | 0.4 | 0 | 1 | 1 |
| Duration (in seconds) | 68 | 1836.5 | 860.8 | 843 | 1,696 | 3,744 |
| No. clients (self-reported) | 68 | 292.9 | 712.4 | 3 | 49 | 1,000 |
| 𝟙(≥5y experience) | 68 | 0.5 | 0.5 | 0 | 0 | 1 |
| 𝟙(all FL questions correct) | 68 | 0.7 | 0.5 | 0 | 1 | 1 |

**Note:** Number of clients and experience in the financial advice industry are self-reported. The financial literacy (FL) questions are (correct answers underlined): 1) "Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you thing you would have in the account if you left the money to grow?" [More than $102 · Exactly $102 · Less than $102 · Don't know · Prefer not to say] 2) "If interest rates rise, what will typically happen to bond prices?" [They will rise · They will fall · They will stay the same · There is no relationship between bond prices and the interest rate · Don't know · Prefer not to say] 3) "Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?" [More than today · Exactly the same · Less than today · Don't know ·Prefer not to say] (cf. Lusardi and Mitchell, 2011).

[36] While two of our questions are taken from the Big Three, we ask for the relationship between interest rates and bond prices rather than the risk profiles of individual stocks versus stock mutual funds.

## Appendix C: Investor attention measures

This section details the construction of our investor attention measures. We use four investor attention measures. First, we proxy for investor interest by the number of downloads of the respective stocks' documents filed with the SEC. Second, we use three specifications for the Google SVI. Third, we use company-related news coverage. Fourth, we use analyst reports. As Table C1 shows, the measures are highly correlated. Because SEC download requests arguably capture investor attention (versus e.g., consumer attention) best, we use average download volumes as our preferred attention measure.

**Attention measure 1: Downloads of SEC filings**

The SEC publishes daily log files in its EDGAR database, which list all downloads of SEC-filed documents on that day. We collect all data relating to the 12 stocks we use in our sample for the year 2023 and construct a daily attention measure. As Figure C1 shows, download requests are highly seasonal around quarterly filings. We use the average daily download requests as a cross-sectional measure of investor attention on the 12 stocks. The variation in the number of download requests is large, ranging from less than 2,000 for St. Joe to almost 50,000 for Berkshire Hathaway.

Figure C1: SEC download requests (Berkshire Hathaway, 2023)

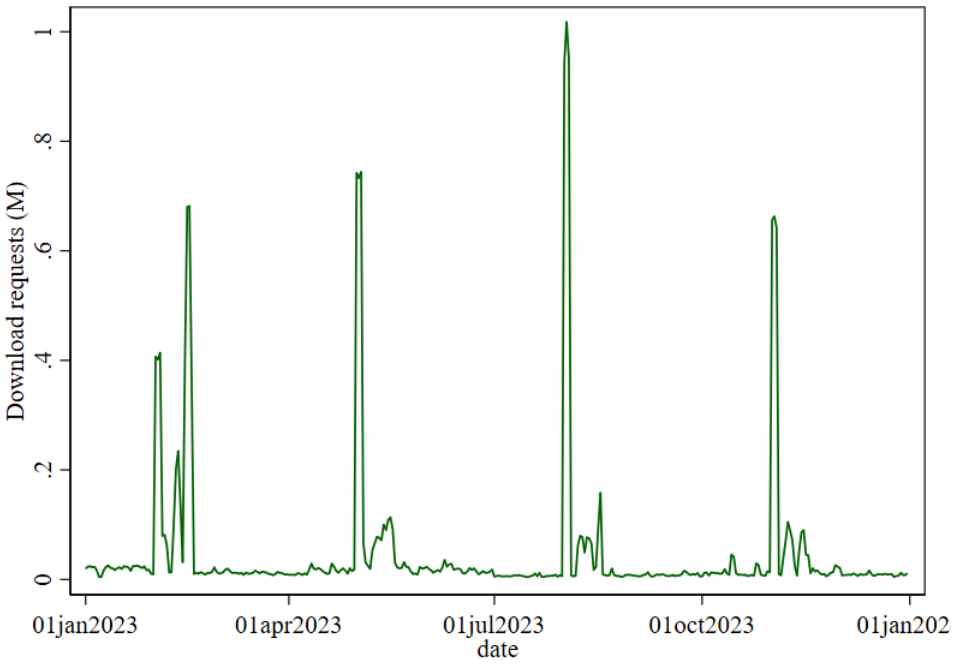
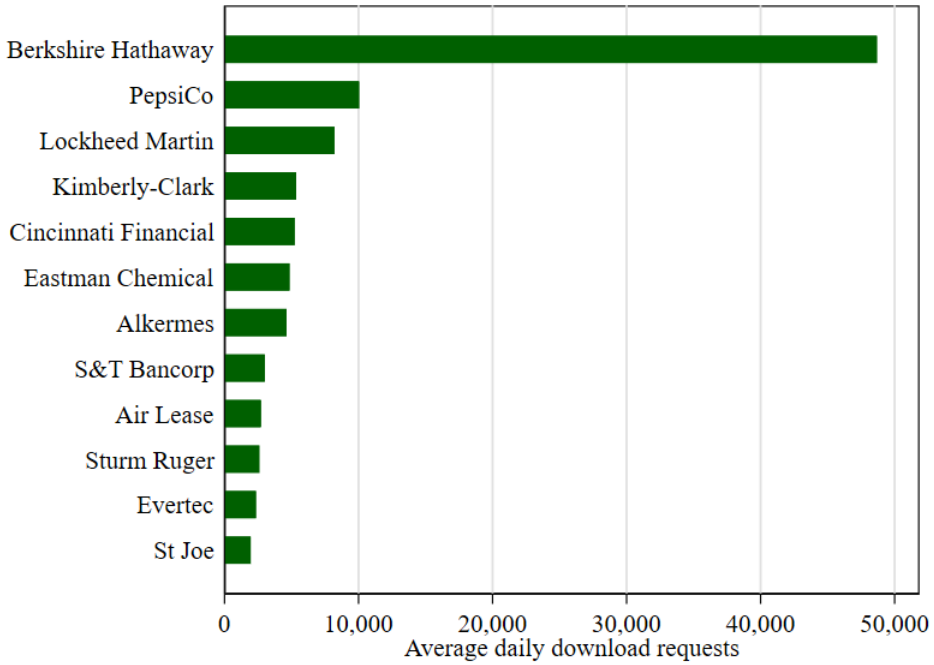


Figure C2: Average daily SEC download requests (2023)



**Attention measure 2: Google SVI**

We use three ways to obtain Google SVIs for the 12 stocks in our sample. First, we use company names as keywords. Second, to filter out noise from non-investors (e.g., consumers) searching for the underlying companies, we use stock tickers and the exchange-linked keyword suggested by Google. Third, we follow deHaan et al. (2024) in further reducing noise from non-investor-related searches by adding " stock" to each keyword. Because the ticker for St. Joe Co. is ambiguous ("JOE") and would have been highly overstated compared to the other two approaches, we omit this company for approach 3.

Figure C3 displays the weekly SVI time series for the past 5 years separately for the three approaches. While there are some structural differences between the three approaches, some pattern are robust: The companies with most Google searches are — perhaps not surprisingly — Pepsi Co., Berkshire Hathaway, and Lockheed Martin. Intuitively, PepsiCo is the most searched company when company names considered, while Berkshire Hathaway and Lockheed Martin register higher search volumes when the stock ticker (or the stock ticker and the word "stock") is considered as a keyword. This suggests that PepsiCo likely attracts more non-investor-related search traffic than the other two companies.

In the next step, we compute average (weekly) Google SVI scores for 2023 and aggregate over the three approaches by using the median SVI score for each company. As Figure C4 shows, the ranking of companies does not change materially when we use the mean instead of the median, but the SVI is less affected by outliers.[37]

[37] Specifically, Air Lease Corp. only displays high search volumes when the keyword "AL stock" is used. This likely captures search volume not related to the company or its stock.

Figure C3: Google SVI (time series)

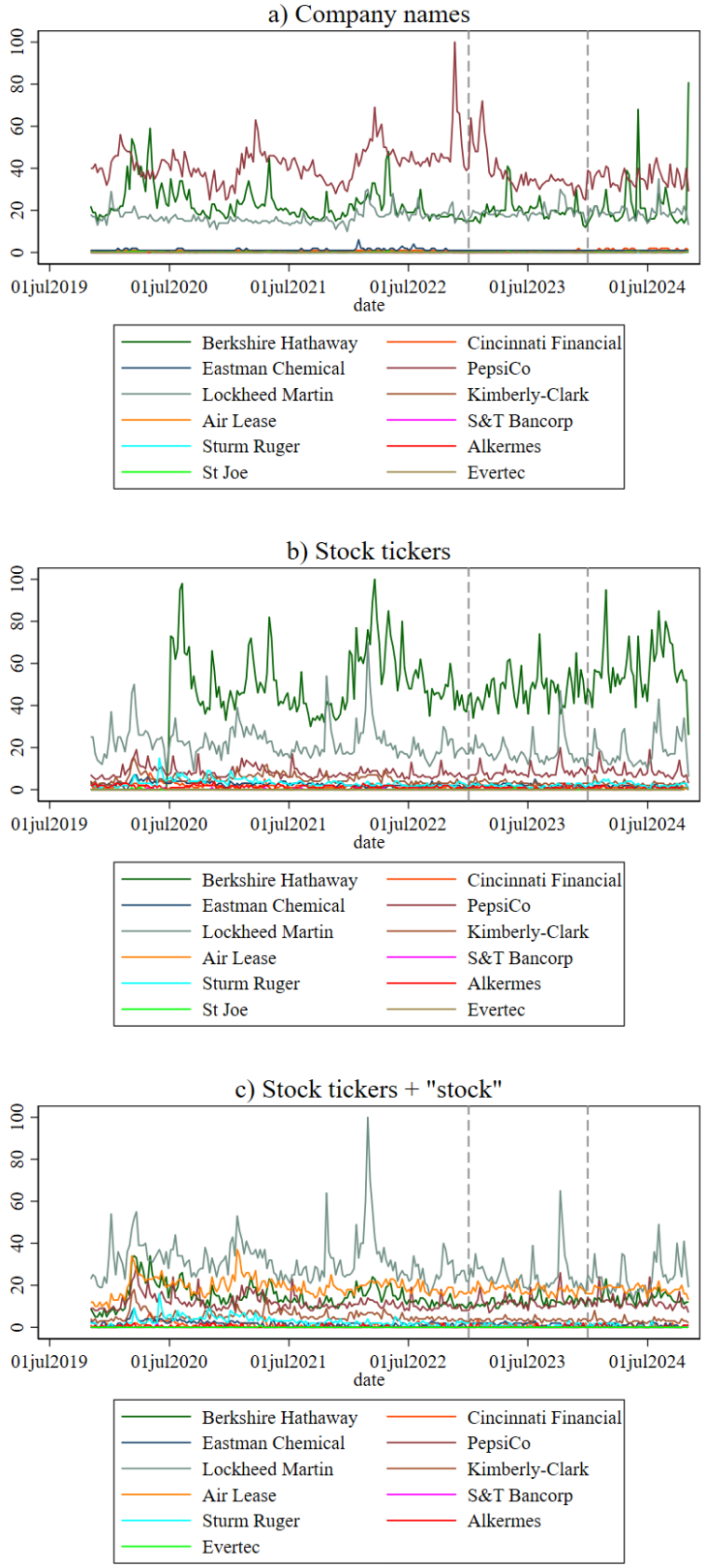
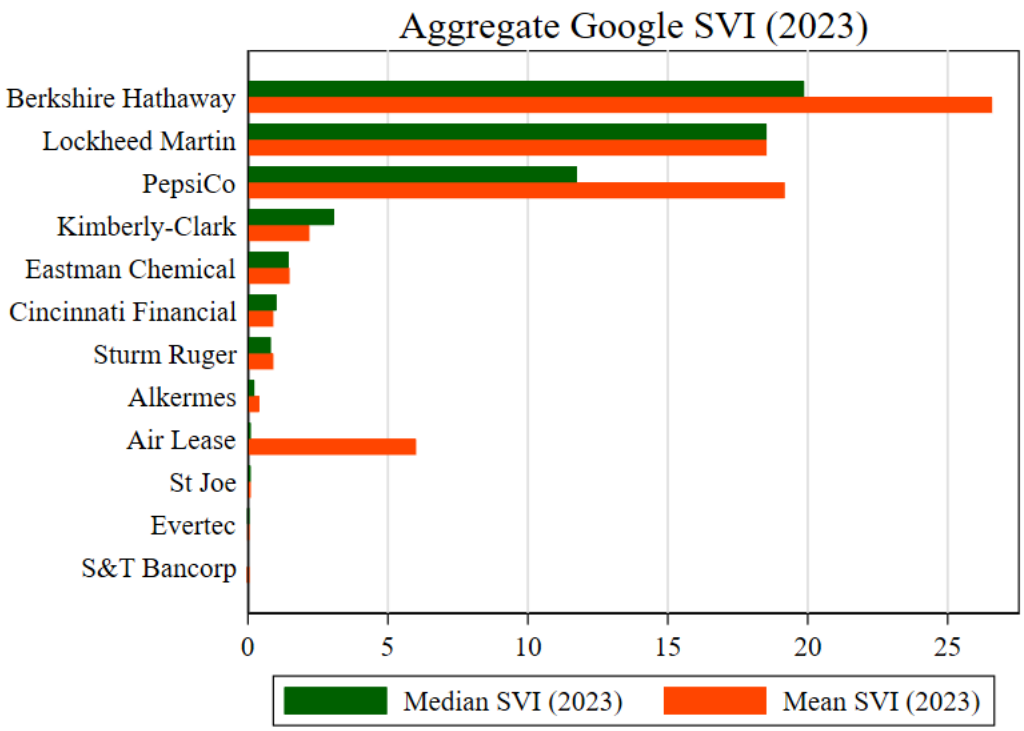


Figure C4: Google SVI (aggregated)



**Attention measure 3: News coverage**

We obtain an alternative investor attention measure by identifying company-related news articles in the *Business Source Complete* database. Specifically, we look for articles including the companies' names (not including legal suffixes) published in popular financial media outlets (*Bloomberg*, *Economist*, *Kiplinger*, *Wall Street Journal*) in 2023. We use the total number of news articles as a measure of investor attention (Da et al., 2011).

**Attention measure 4: Analyst coverage**

Analyst coverage and reporting has been shown to increase investor recognition of a stock (Mola et al., 2013; Claussen et al., 2020). Thus, we use the number of analyst reports published in 2023 for each of the 12 stocks. Our data comes from Thomson Reuters Eikon.

Figure C5 displays the attention measures constructed from news articles and analyst reports. In line with Google SVI, by far the companies generating the most attention are Berkshire Hathaway, Lockheed Martin, and PepsiCo.

Figure C5: News articles and analyst reports

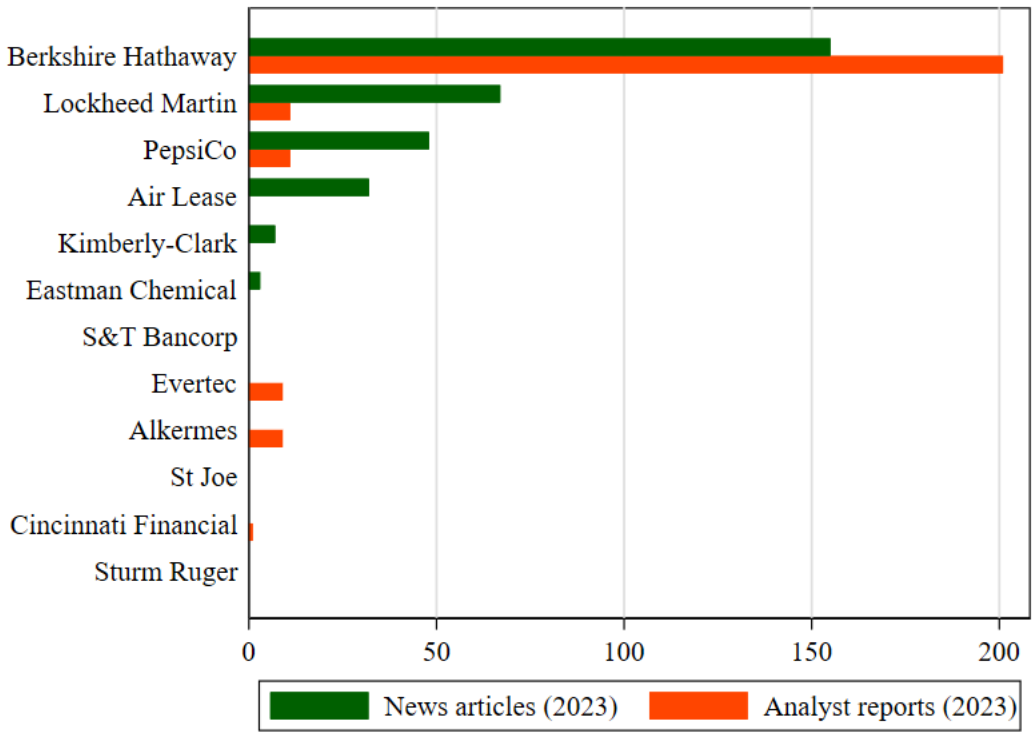


Table C1: Pairwise correlations: Attention measures

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) Median Google SVI 2023 | 1 | | | | |
| (2) Mean Google SVI 2023 | 0.959*** | 1 | | | |
| (3) News articles 2023 | 0.895*** | 0.935*** | 1 | | |
| (4) Analyst reports 2023 | 0.674* | 0.713** | 0.894*** | 1 | |
| (5) Average daily SEC download requests 2023 | 0.746** | 0.782** | 0.925*** | 0.988*** | 1 |

**Note:** The table displays pairwise correlation for the five attention measures detailed in Online Appendix C. Stars indicate statistical significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

## Appendix D: Robustness and Placebo treatments

To ensure that our treatment effects are caused by the actual information content in the treatments rather than by the mere provision of information, we conduct two placebo tests. First, we provide unrelated qualitative information in the form of a summary of the technology acceptance model (TAM), a seminal information systems theory (Davis, 1989). We use the following summary of the TAM:

> **TAM summary:**
> The theoretical framework of the technology acceptance model revolves around the constructs of "Perceived Usefulness" and "Perceived Ease of Use," which are posited as fundamental determinants of user acceptance of information technology. The framework builds on the premise that users are likely to adopt a technology if they perceive it as useful in enhancing their job performance, and if they believe it does not require substantial effort to use.
> Perceived Usefulness is defined as the degree to which a person believes that using a particular system would enhance their job performance. This belief is rooted in the organizational context where performance improvements are typically rewarded, thus motivating individuals to adopt technologies they perceive as beneficial. The concept suggests that users assess the potential of a technology based on its ability to provide tangible benefits in performing their job duties.
> Perceived Ease of Use, on the other hand, refers to the degree to which a person believes that using a particular system would be free of effort. This notion stems from the understanding that users are more likely to embrace a technology that is not only beneficial but also easy to use. It underscores the idea that the effort required to learn and use a new system is a significant factor in the decision to adopt it. If a system is perceived as easy to use, it reduces the cognitive load and potential frustration, making it more appealing for adoption.
> The interplay between these two constructs forms the core of the theoretical model proposed in the paper, suggesting that both perceived usefulness and ease of use are crucial predictors of user acceptance. The model hypothesizes that while both factors independently influence acceptance, perceived usefulness is a stronger predictor than perceived ease of use. Moreover, perceived ease of use might also indirectly affect user acceptance by influencing perceived usefulness.
> Empirical validation of the model involves the development and testing of measurement scales for both constructs. These scales are designed to assess the extent to which users perceive a system as useful and easy to use, with the results demonstrating significant correlations with actual system usage. The reliability and validity of these scales are rigorously tested through empirical studies involving users and various applications.
> In conclusion, the technology acceptance model offers a robust model for understanding the dynamics of user acceptance of information technology. It highlights the critical role of user perceptions in the acceptance process and establishes a foundation for further research to explore the nuances of how these perceptions influence the broader adoption of new technologies.

Second, we provide unrelated quantitative information in the form of average annual precipitation data for the 50 US state capitals for 2023.[38]

We collect the data for the baseline specification, as well as the two placebo specifications on November 21, 2024. We re-collect the baseline data (i) to investigate the robustness of the baseline recommendations over time, and (ii) to ensure that changes in the recommended portfolio weights in the placebo conditions are not caused by methodological changes to the LLMs between the first (November 4, 2024) and second (November 21, 2024) data collections.[39]

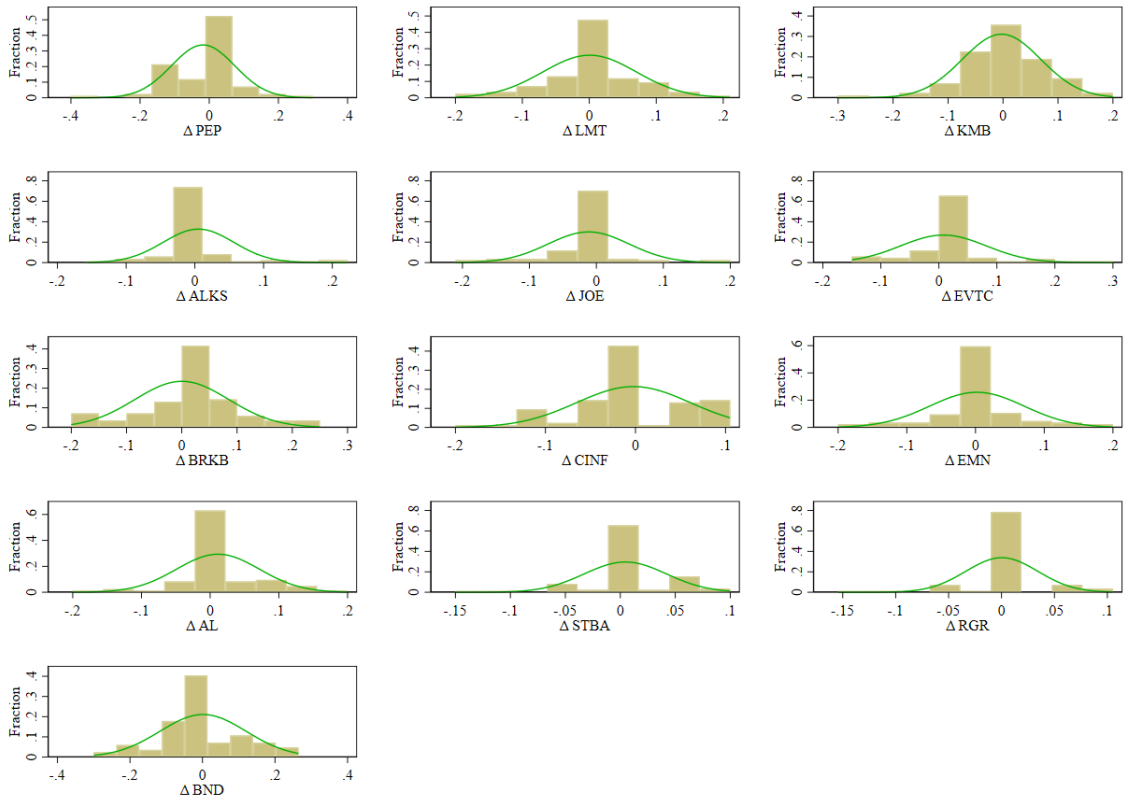
First, we investigate the robustness of recommendations in the baseline by comparing portfolio weights for each security and in each scenario for the two data collection dates. While we ensure that all models' temperatures are set to zero, suggesting that models provide the same answer to identical questions, other parameters in the LLMs may have changed, causing the recommendations to change in repeated requests. Figure D1 displays the aggregate distribution of changes in security weights between collection dates. The charts suggest that while there is some variation in portfolio weights for all securities, there is no difference on average and in the majority of cases. Table D1 compares changes in security weights by models. While the average change is zero by definition, standard deviations and percentiles suggest that recommendations by OpenAI models are more stable than recommendations by the other (and smaller) models.

Second, we investigate whether adding irrelevant additional information to the queries affects the models' portfolio suggestions.

[38] We use precipitation data from https://open-meteo.com/.

[39] The placebo specifications were suggested to us after the pre-registration, which is why it is not contained in the pre-registered experimental design.

Figure D1: Robustness of security weights (baseline)




**Note:** The figure displays for each of the 12 scenarios the distribution of differences in security weights for the respective security from the first data collection to the second data collection. Positive (negative) values suggest that the security's average allocation has increased (decreased) from the first to the second collection date. The green line depicts a normal distribution representation of the distribution.

Table D1: Robustness of security weights, by model

| Model | N | Mean | SD | p25 | p50 | p75 |
|---|---|---|---|---|---|---|
| GPT-4-Turbo | 156 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 |
| GPT-4o | 156 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 |
| GPT-4o-mini | 156 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 |
| Llama-3.1-8B-Instruct | 156 | 0.00 | 0.07 | -0.03 | 0.00 | 0.01 |
| Mistral-7B-Instruct-v0.3 | 156 | 0.00 | 0.08 | -0.03 | 0.00 | 0.02 |
| Phi-small-128k-instruct | 156 | 0.00 | 0.07 | -0.03 | 0.00 | 0.02 |
| Qwen2.5-7B-Instruct | 156 | 0.00 | 0.09 | -0.01 | 0.00 | 0.05 |

**Note:** The table displays descriptive statistics for differences in portfolio weights between the two collection dates, separately by model. The data is based on 156 observations (12 scenarios $\times$ 13 securities).

Table D2: Placebo effects

| | Diversification | | | Risk | | | | Performance | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) No. securities | (2) HHI | (3) HHI (equity) | (4) Risky share | (5) Vola (%) | (6) IVOL (%) | (7) FF6 $\beta_M$ | (8) Excess return (%) | (9) Sharpe ratio | (10) FF6 $\alpha$ (%) |
| $\mathbb{1}(P_1)$ | 0.190 | -0.010 | 0.009 | 0.033 | 0.134 | 0.001* | 0.027 | 0.047** | 0.027* | 0.034** |
| | (0.399) | (0.015) | (0.013) | (0.019) | (0.086) | (0.001) | (0.014) | (0.013) | (0.012) | (0.011) |
| $\mathbb{1}(P_2)$ | -0.167 | 0.028 | 0.024* | -0.008 | -0.016 | 0.000 | -0.007 | 0.006 | -0.026 | 0.018 |
| | (0.145) | (0.016) | (0.011) | (0.022) | (0.097) | (0.001) | (0.019) | (0.015) | (0.023) | (0.015) |
| Constant | 6.373*** | 0.174*** | 0.162*** | 0.921*** | 4.850*** | 0.026*** | 0.847*** | 0.604*** | 0.435*** | -0.213*** |
| | (0.359) | (0.027) | (0.015) | (0.051) | (0.345) | (0.002) | (0.049) | (0.052) | (0.040) | (0.035) |
| Obs. | 252 | 252 | 249 | 252 | 252 | 252 | 252 | 252 | 252 | 252 |
| Adj. R2 | 0.542 | 0.590 | 0.473 | 0.764 | 0.792 | 0.703 | 0.799 | 0.806 | 0.580 | 0.661 |
| Profile FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Advisor FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Exp. condition | All | All | All | All | All | All | All | All | All | All |
| Advisor | LLM | LLM | LLM | LLM | LLM | LLM | LLM | LLM | LLM | LLM |

**Note:** The table reports the interaction coefficients of logit regressions of a dummy variable indicating that the respective security is included in the portfolio on (binary) fundamental stock metrics, a binary investor attention proxy, a binary sentiment measure of a company's latest 10-K filing's MD&A section, as well as interaction terms between binary placebo indicators and those measures. The sample for the regressions is restricted to observations from the baseline condition (7 models $\times$ 12 scenarios $\times$ 12 stocks = 1,008 observations) and the respective placebo condition (1,008 observations). The table only reports the coefficients for the interaction terms for improved readability. Standard errors are reported in parentheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$) and are robust to clustering on the LLM level.